Stabilizing Few-Shot Object Detection with Language-Conditioned Probabilistic Prototypes

Jiaying Wu[1] **Yuchun Zhao**[1] **Jia Luo**[2,3,4] **Lei Shi**[5,6,*]
Dongmei Wei[1] **Feifei Kou**[7] **Pengfei Zhang**[8] **Zhaoman Zhong**[1]

[1]School of Computer Engineering, Jiangsu Ocean University, China
[2]Interdisciplinary Faculty of Science and Engineering, Shimane University, Japan
[3]College of Economics and Management, Beijing University of Technology, China
[4]Chongqing Research Institute, Beijing University of Technology, China
[5]State Key Laboratory of Media Convergence and Communication, Communication University of China, China
[6]Key Laboratory of Public Opinion Governance and Computational Communication, Sichuan Police College, China
[7]School of Computer Science (National Pilot Software Engineering School),
Beijing University of Posts and Telecommunications, China
[8]School of Computer Science and Engineering, Anhui University of Science and Technology, China
jiaying@jou.edu.cn 2025220957@jou.edu.cn jialuo@riko.shimane-u.ac.jp leiky_shi@cuc.edu.cn
weidm0807@jou.edu.cn koufeifei000@bupt.edu.cn zpf.bupt@bupt.cn zhaomanzhong@jou.edu.cn
[*]Corresponding author: Lei Shi (leiky_shi@cuc.edu.cn)

Abstract

Few-shot object detection (FSOD) aims to adapt detectors to novel categories using only a few annotated instances. Prototype-based methods address this task by extracting class representations from support examples, making prototype quality critical when visual evidence is scarce. However, visual query transfer lacks an explicit semantic reference for novel categories, while deterministic prototype matching does not account for differences in support reliability. To address these limitations, we propose Semantic Fine-Grained Prototype Distillation (SFPD), comprising Language-Conditioned Feature Query (LC-FQ) and Uncertainty-Aware Prototype Distillation (UA-PD). LC-FQ combines projected class-name embeddings from a frozen CLIP text encoder with learnable visual queries to guide class-relevant feature extraction, while UA-PD predicts feature-dependent variances to weight support fusion and modulate prototype assignment. On PASCAL VOC, SFPD exceeds the published FPD mean in 11 of 12 evaluated multi-run settings, achieving an average gain of 1.4 points in novel-class AP_{50}; results on MS-COCO also match or improve upon FPD. Five-seed evaluations show low performance dispersion, and the displayed learning curves indicate stronger early adaptation. These findings suggest that semantic guidance and support reliability modeling provide complementary information for prototype construction under limited supervision. With class embeddings computed offline, SFPD incorporates this information without requiring online text encoding, offering a practical approach to extending object detection to categories with scarce annotations.

1 Introduction

Object detection is a fundamental computer vision task that localizes and recognizes objects in images. Modern detectors typically learn from large datasets with bounding-box annotations, but extending them to additional categories requires collecting and annotating new examples. This requirement becomes difficult to meet when objects are rare or annotation requires specialist knowledge, motivating detection methods that can adapt with limited supervision.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Few-shot object detection (FSOD) addresses this challenge by transferring knowledge from data-rich base categories to novel categories with only a few annotated instances. Existing approaches commonly use episodic meta-learning or two-stage transfer learning [16, 39, 30, 27]. We follow the latter setting, where a pretrained detector is fine-tuned using a small support set. Because this set covers only a narrow range of object appearances, poses, and backgrounds, adaptation must extract class information that generalizes beyond the observed instances.

Prototype-based methods make support evidence explicit by constructing class representations that guide detection. Recent studies improve these representations through visual feature refinement, sampling-invariant metric learning, and conditional prototype construction [32, 25, 20, 14]. Fine-grained prototype distillation (FPD) uses learnable class queries to extract representative support features and assigns the resulting prototypes to query-image features [31]. These advances strengthen visual prototype construction, but leave two issues insufficiently addressed under sparse supervision: the semantic reference used to guide class queries and the reliability of the features contributing to a prototype.

First, transferred visual queries provide a starting point for novel categories without explicitly encoding the meaning of their class names. With one or a few examples, query adaptation can follow appearance cues specific to the sampled supports. Second, support features differ in reliability: occlusion, background clutter, and imprecise localization can contaminate the class representation [13, 20, 21]. Deterministic matching does not explicitly model this variation in uncertainty. Consequently, selecting class-relevant evidence and controlling its contribution remain complementary challenges for prototype-based adaptation.

Addressing these challenges matters because each support instance has substantial influence when annotations are scarce. A prototype dominated by incidental appearance or unreliable features can affect both feature assignment and subsequent detection. Additional visual supervision may reduce these effects, but requiring more annotations weakens the motivation for FSOD. Class-level semantics can instead supply information beyond the sampled images, while feature-dependent reliability estimates can regulate the use of the available evidence. Combining these sources offers a way to improve prototype construction without requiring online text encoding for every query image.

We therefore propose Semantic Fine-Grained Prototype Distillation (SFPD), which augments FPD with two complementary components. Language-Conditioned Feature Query (LC-FQ) projects class-name embeddings from a frozen CLIP text encoder [28] into the detector query space and combines them with learnable visual queries. Uncertainty-Aware Prototype Distillation (UA-PD) predicts Gaussian variances for fine-grained support prototypes, using them to weight same-class support fusion and modulate assignment to query-image features. SFPD retains the subsequent detection pipeline and caches text embeddings offline. Experiments on PASCAL VOC and MS-COCO assess accuracy, component contributions, five-seed consistency, and early adaptation.

Our contributions are threefold:

- We introduce LC-FQ, which combines offline CLIP class-name embeddings with learnable visual queries to guide prototype extraction for novel categories.
- We develop UA-PD, which estimates feature-dependent prototype variance and uses it in support fusion and query-image assignment. Together, the two components integrate semantic guidance with reliability-aware visual adaptation.
- We evaluate SFPD on PASCAL VOC and MS-COCO through comparisons, ablations, five-seed evaluations, and learning-curve analyses. On VOC, SFPD exceeds the published FPD mean in 11 of 12 multi-run settings with an average gain of 1.4 points. SFPD's inference cost is comparable to FPD's.

2 Related Work

FSOD and Prototype-Centered Transfer. FSOD is commonly addressed by meta-learning, which learns support–query interactions episodically [16, 39, 8], or by transfer learning, which pretrains on base classes and fine-tunes on a small balanced support set [30, 27]. We follow the latter setting. Prototype- and metric-centered methods represent a novel class with support-derived features and match query RoIs to that representation [33, 29, 11]. Related prototype research also models object parts in few-shot semantic segmentation [24]. Within FSOD, recent work improves this

representation through feature refinement [25], sampling-invariant metric learning [20], background suppression [21], conditional prototypes [14], and fine-grained prototype distillation (FPD) [31]. SFPD builds directly on this efficient route: whereas these methods strengthen the visual prototype, UA-PD additionally estimates which scarce support regions are reliable enough to shape it.

Language- and VLM-Based FSOD. CLIP [28] and open-vocabulary detectors such as F-VLM [18] and Grounding DINO [23] establish language as transferable category knowledge. Detection prompt learning further adapts language representations to open-vocabulary detection [5]. In FSOD, language knowledge is used through visual–textual alignment with VLMs [26] and multimodal knowledge transfer using generated or auxiliary text [4]. Related few-shot classification work selects semantically related base-class priors by MAP [34] or combines class-level semantics with caption-derived instance cues for probabilistic prototype refinement [36]. These two classification approaches operate on image-level labels. In contrast, LC-FQ uses frozen, offline CLIP class embeddings to guide detector queries; no online text encoding is required in the detection path.

Uncertainty for Reliable Adaptation. Bayesian models, ensembles, and heteroscedastic regression use uncertainty for calibration and robust prediction under imperfect observations [17, 19, 2]; in generalized FSOD, uncertainty has also been used to reduce forgetting [10]. Noise-robust few-shot classification further suppresses anomalous supports and augments features generatively [35]; attention prompting reduces background distraction in ViT classifiers [37]. In contrast, UA-PD predicts prototype uncertainty and uses it before prediction to downweight unreliable support regions in detector prototype assignment. SFPD therefore combines a language prior for cold-start query initialization with reliability-aware prototype construction without online text encoding; deployment measurements are reported in Appendix A.

3 Methodology

3.1 Few-shot Object Detection Setting

We follow two-stage transfer learning. The category space is divided into disjoint base and novel sets, $\mathcal{C}_{\text{base}} \cap \mathcal{C}_{\text{novel}} = \emptyset$. A detector is first trained on abundant base-class annotations and then fine-tuned on a balanced support set containing K annotated instances per class. Evaluation measures detection on novel classes. A support image supplies the class evidence used to construct prototypes; a query image is an image on which detections are predicted. These roles are distinct from the learnable class queries used inside prototype distillation.

3.2 Revisiting Prototype-Centered FSOD

In FPD, a class is represented by several fine-grained prototypes rather than a single average of support RoI features. Learnable class queries attend to support feature maps to extract these prototypes. Prototype Assignment (PA) then uses them to condition query-image features before proposal generation, while Non-Linear Fusion (NLF) combines query and support RoI features before detection. SFPD extends prototype distillation and assignment within this pipeline.

Query adaptation without explicit semantics. The baseline learns class queries from visual supervision and transfers base-class query representations to novel classes. Such transfer supplies a visual starting point, but does not explicitly encode the meaning of a novel class name. When support is scarce, the selected examples can strongly influence which features the queries extract. A language-derived class reference can complement this visual initialization.

Support evidence with unequal reliability. Fine-grained attention selects support features, but deterministic matching does not explicitly represent their uncertainty. Occluded objects and background-contaminated features can therefore contribute to prototype construction and assignment without a separate reliability estimate. SFPD predicts feature-dependent variance to regulate these contributions. We study semantic guidance and uncertainty weighting as complementary design choices; the experiments do not isolate a causal interaction between them.

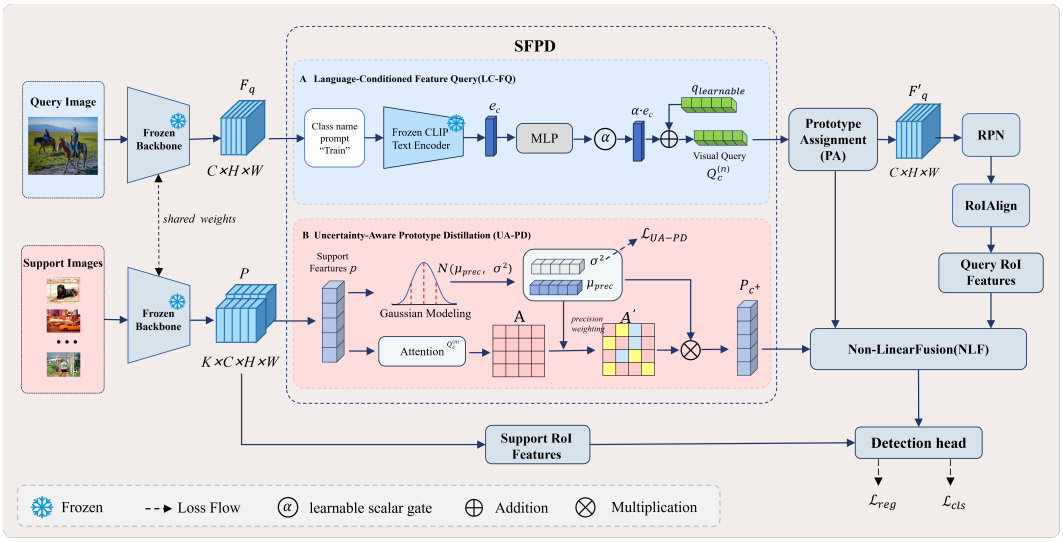


Figure 1: **Overview of the proposed SFPD framework.** Query and support images are processed by a shared frozen backbone. Support features are used to generate uncertainty-aware class prototypes through prototype distillation and to obtain support RoI features. Offline CLIP text embeddings guide LC-FQ, and UA-PD reweights support evidence when forming class prototypes. The subsequent PA/NLF detection path follows FPD.

3.3 Semantic Fine-Grained Prototype Distillation (SFPD)

Design rationale. The two baseline failure modes yield two requirements: a class-specific semantic reference for query adaptation and reliability-weighted support aggregation. LC-FQ implements the former with a frozen language prior, and UA-PD implements the latter with heteroscedastic uncertainty. Ablations assess the contribution of each component, while seed and convergence analyses characterize the combined model.

Overall framework. Figure 1 presents the complete feature flow. Query and support images pass through a frozen backbone with shared parameters. Support features p, together with offline CLIP class embeddings, enter Prototype Distillation: LC-FQ produces semantic class queries and UA-PD estimates feature uncertainty to obtain reliability-aware class prototypes. PA uses these class prototypes to condition the query feature map before proposal generation and RoIAlign. The query and support RoI features are then combined by the baseline Non-Linear Fusion (NLF) path and sent to the detection head. The three dashed loss flows make the supervision explicit: $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{reg}}$ optimize detection, while $\mathcal{L}_{\text{UA-PD}}$ regularizes the uncertainty estimate during prototype distillation.

Fine-tuning procedure. During novel-class adaptation, we (1) draw the balanced K-shot support set and extract support features with the shared frozen backbone; (2) encode each class name once with frozen CLIP and use the resulting offline embedding to construct LC-FQ class queries; (3) use UA-PD to estimate support uncertainty and form reliability-aware class prototypes for PA; (4) extract query and support RoI features, fuse them through the baseline NLF path, and predict detections; and (5) optimize the network with Eq. (5). Thus, SFPD contains exactly two proposed components: language provides a precomputed semantic prior through LC-FQ, whereas uncertainty determines how much each support feature contributes through UA-PD.

3.3.1 Language-Conditioned Feature Query (LC-FQ)

LC-FQ supplies a class-specific semantic reference during adaptation by incorporating prior language information. Specifically, we first project pretrained text embeddings e_c from a frozen CLIP text encoder [28] into the detector query dimension via a two-layer (Linear-ReLU-Linear) MLP. The language-conditioned query $\tilde{Q}_c$ is generated by reshaping the MLP output for each class. To balance the prior semantic information with the learned visual representations, we introduce a learnable

residual scale $\alpha \in \mathbb{R}$ (initialized to 1.0). The resulting query Q_c is then constructed as follows:

$$Q_c = q_{\text{learnable},c} + \alpha \cdot \tilde{Q}_c, \tag{1}$$

where $q_{\text{learnable},c}$ denotes the class-specific visual query embeddings. For base classes, these embeddings are randomly initialized and optimized during base training. For novel classes, we initialize them by copying compatible base-class query embeddings, selected according to their cross-attention similarity to the support features. The projected language residual is then added as in Eq. (1). All main experiments use this visual-query transfer followed by language-residual addition; the copied visual embeddings remain learnable during adaptation.

3.3.2 Uncertainty-Aware Prototype Distillation (UA-PD)

Let $P_i^{(c)} \in \mathbb{R}^{M\times d}$ denote the M fine-grained prototypes distilled by the FPD attention path from support instance i of class c. UA-PD augments each prototype with a channel-wise heteroscedastic variance [17, 2], predicted by a lightweight head h:

$$\Sigma_i^{(c)} = \text{softplus}\Big(h(P_i^{(c)})\Big) + \sigma^2_{\min}, \tag{2}$$

where $\Sigma_i^{(c)}$ has the same shape as $P_i^{(c)}$ and $\sigma^2_{\min}$ ensures numerical stability. During support initialization, multiple same-class supports are fused element-wise by inverse-variance weighting, so uncertain prototype channels contribute less; their fused variance is the reciprocal of the summed precision. During training, the same uncertainty estimate directly regulates the instance-level prototypes passed to Prototype Assignment (PA).

Specifically, let $X \in \mathbb{R}^{N\times d}$ be the flattened query feature map and stack the available prototypes as keys K and values V. For prototype r with variance vector σ_r^2, define its scalar precision $\rho_r = (d^{-1}\sum_j \sigma_{r,j}^{-2})^\gamma$ and log-variance penalty $u_r = d^{-1}\sum_j \log \sigma_{r,j}^2$. UA-PD modifies the PA logits and residual output as

$$A'_{nr} = \frac{\rho_r (W_q x_n)^\top (W_k k_r)/\sqrt{d_k} - \beta u_r}{\tau}, \qquad X' = X + \eta\, \text{softmax}(A')V, \tag{3}$$

where d_k is the key dimension, τ is the temperature, γ controls precision scaling, β controls the uncertainty penalty, and η is the residual scale inherited from FPD. Thus, variance controls support fusion and modulates query assignment, while X' proceeds through the unchanged detection path.

Finally, a KL-type regularizer penalizes deviations from a finite prior variance, discouraging both vanishing and unbounded variance:

$$\mathcal{L}_{\text{UA-PD}} = \frac{1}{2|\mathcal{S}|}\sum_{s\in\mathcal{S}}\left(\frac{s}{\sigma^2_{\text{prior}}} - \log\frac{s}{\sigma^2_{\text{prior}}} - 1\right), \tag{4}$$

where $\mathcal{S}$ contains all entries of $\{\Sigma_i^{(c)}\}$ and σ^2_{prior} is the prior variance.

3.3.3 Optimization

The overall training objective consists of the standard Faster R-CNN classification loss $\mathcal{L}_{\text{cls}}$, bounding-box regression loss $\mathcal{L}_{\text{reg}}$, and the proposed uncertainty-aware regularizer $\mathcal{L}_{\text{UA-PD}}$:

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}}}_{\mathcal{L}_{\text{det}}} + \mathcal{L}_{\text{UA-PD}}. \tag{5}$$

Here, $\mathcal{L}_{\text{cls}}$ encourages each fused RoI feature to be assigned to its ground-truth class (or background), and $\mathcal{L}_{\text{reg}}$ refines its predicted box coordinates. $\mathcal{L}_{\text{UA-PD}}$ regularizes the support-feature variance used during prototype construction. Accordingly, the first two losses are applied at the detection head, whereas the last loss is applied to the UA-PD branch, as visualized in Figure 1.

4 Experiments

We test whether SFPD makes prototype-based FSOD more reliable when visual supervision is scarce. VOC/COCO comparisons assess low-shot accuracy; ablations isolate the two core mechanisms; and seed, learning-curve, and failure analyses assess reproducibility and early optimization. Controlled comparisons retain the backbone and downstream detection pipeline.

4.1 Experimental Setup

Protocol. We evaluate on PASCAL VOC 2007+2012 [6, 7] (15 base / 5 novel, three standard splits, $K \in \{1, 2, 3, 5, 10\}$) and MS-COCO 2017 [22] (60/20, $K \in \{10, 30\}$). We use Faster R-CNN with ResNet-101, standard two-stage transfer, and frozen CLIP (ViT-B/32). FPD [31], which shares the downstream detection pipeline, is our primary prototype-based baseline: it provides a reference for evaluating semantic query guidance and probabilistic support weighting under the same downstream detector configuration.

Experimental workflow. For each split and shot setting, we pretrain on base classes, fine-tune on the balanced novel-class support set, and report novel-class AP_{50} (nAP_{50}) on PASCAL VOC and novel-class AP averaged over IoU thresholds from 0.50 to 0.95 in steps of 0.05 (nAP) on MS-COCO. Controlled comparisons fix the backbone, transfer schedule, support protocol, and downstream detection configuration. We repeat SFPD with five seeds (42, 123, 456, 789, 2024), report mean and standard deviation, trace learning curves, and conduct cumulative and leave-one-out ablations.

Implementation details. We follow the two-stage optimization and data protocol of FPD for the controlled ablations, freezing the shared ResNet-101 backbone during novel-class adaptation. Frozen CLIP ViT-B/32 class embeddings are computed once and cached, avoiding online text encoding at inference. LC-FQ copies compatible base-class visual queries and adds a projected language residual with learnable scale α, initialized to 1. Removing LC-FQ retains visual-query transfer but omits the language residual; the Naive CLIP control instead uses raw CLIP initialization. All controlled comparisons retain the same PA/NLF configuration and transfer schedule. Experiments use two NVIDIA RTX 3090 GPUs and take approximately two wall-clock hours per run, equivalent to four GPU-hours if both GPUs remain occupied. An aggregate compute total across all experiments is unavailable.

4.2 Comparison with Existing Methods

Tables 1–2 first show that reliability does not come at the cost of accuracy. The largest single-run VOC gain occurs at $K = 1$ (+3.3 nAP_{50} on Split 2; 33.1 versus 29.8), in an extreme low-shot setting. Across 12 multi-run VOC settings, SFPD exceeds the published FPD mean in 11 and gains 1.4 points on average; the sole exception is -0.3 on Split 1 at 5-shot. On COCO, SFPD matches or improves FPD in every reported setting, including +0.4/+0.9 at 10/30-shot on average. The following analyses test whether this competitive performance is also obtained reliably.

4.3 Ablation Study

Contribution of each component. Table 3 evaluates the two components with simplified substitutes, cumulative integration, and leave-one-out removal. The simplified controls remove parts of the proposed design: raw CLIP initialization removes learned residual calibration, while the uncertainty-head-only control retains variance-driven precision scaling in Eq. (3) but removes the explicit log-variance penalty and KL regularizer. Thus, predicted variance still influences prototype assignment through the precision factor ρ_r.

LC-FQ: contribution of semantic guidance. Naive CLIP (raw embeddings, no residual) yields a 1.0-point gain (Panel A); the calibrated pipeline (MLP projection + learnable gate α) contributes a 2.2-point gain (Panel B), and removal costs 2.1 points (Panel C). Its leave-one-out effect decays monotonically with shot count ($-2.1 \rightarrow -1.3 \rightarrow -0.9$, Section 4.4). This pattern is consistent with semantic guidance being most useful with scarce support: additional visual evidence can reduce dependence on the external semantic prior.

UA-PD: contribution of uncertainty weighting. The variance-head control with precision scaling adds 0.7 points (Panel A); adding the full UA-PD component to LC-FQ yields a further 1.1 points (Panel B), and removal costs 1.1 points (Panel C). These comparisons support using variance in prototype assignment, although they do not isolate the individual effects of the attention penalty and regularizer. Its effect also decays monotonically ($-1.1 \rightarrow -0.8 \rightarrow -0.6$), consistent with greater benefit from reliability weighting when support is scarce.

Table 1: Performance comparison on PASCAL VOC across three novel splits (nAP_{50}, %). The largest gains over the matched FPD baseline occur in the most data-starved settings, illustrating its benefit under scarce support. Red: highest per column; blue: second-highest. Tied values share the same rank. Multi-run baseline results and single-run Meta-DETR results are from FPD [31]. SFPD results are ours.

Method / shot	Novel Set 1					Novel Set 2					Novel Set 3				
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
Single-run results:															
FSRW [16]	14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9
Meta R-CNN [39]	19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	45.4	14.3	18.2	27.5	41.2	48.1
TFA w/ cos [30]	39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8
MPSR [33]	41.7	42.5	51.4	55.2	61.8	24.4	29.3	39.2	39.9	47.8	35.6	41.8	42.3	48.0	49.7
Retentive [9]	42.4	45.8	45.9	53.7	56.1	21.7	27.8	35.2	37.0	40.3	30.2	37.6	43.0	49.7	50.1
FSCE [29]	44.2	43.8	51.4	61.9	63.4	27.3	29.5	43.5	44.2	50.2	37.2	41.9	47.5	54.6	58.5
Meta FR-CNN [11]	43.0	54.5	60.6	66.1	65.4	27.7	35.5	46.1	47.8	51.4	40.6	46.4	53.4	59.9	58.6
Meta-DETR [40]	40.6	51.4	58.0	59.2	63.6	37.0	36.6	43.7	49.1	54.6	41.6	45.9	52.7	58.9	60.6
FCT [12]	49.9	57.1	57.9	63.2	67.1	27.6	34.5	43.7	49.2	51.2	39.5	54.7	52.3	57.0	58.7
VFA [13]	57.7	64.6	64.7	67.2	67.4	41.4	46.2	51.1	51.8	51.6	48.9	54.8	56.6	59.0	58.9
SCDC [3]	50.3	60.5	63.7	68.2	69.8	33.6	42.8	47.7	50.8	54.5	44.2	55.2	57.3	62.2	62.4
FPD [31]	48.1	62.2	64.0	67.6	68.4	29.8	43.2	47.7	52.0	53.9	44.9	53.8	58.1	61.6	62.9
SFPD (Ours)	49.4	63.3	65.5	66.7	68.5	33.1	44.1	49.3	52.1	57.1	44.9	54.6	59.0	62.9	64.7
Average results over multiple runs:															
FSDetView [38]	24.2	35.3	42.2	49.1	57.4	21.6	24.6	31.9	37.0	45.7	21.2	30.0	37.2	43.8	49.6
DCNet [15]	33.9	37.4	43.7	51.1	59.6	23.2	24.8	30.6	36.7	46.6	32.3	34.9	39.7	42.6	50.7
Meta-DETR [40]	35.1	49.0	53.2	57.4	62.0	27.9	32.3	38.4	43.2	51.8	34.9	41.8	47.1	54.1	58.2
DeFRCN [27]	40.2	53.6	58.2	63.6	66.5	29.5	39.7	43.4	48.1	52.8	35.0	38.3	52.9	57.7	60.8
FCT [12]	38.5	49.6	53.5	59.8	64.3	25.9	34.2	40.1	44.9	47.4	34.7	43.9	49.3	53.1	56.3
VFA [13]	47.4	54.4	58.5	64.5	66.5	33.7	38.2	43.5	48.3	52.4	43.8	48.9	53.3	58.1	60.0
FPD [31]	41.5	52.8	58.4	64.9	67.1	28.2	38.7	43.8	50.3	53.6	34.9	48.6	54.0	58.4	61.5
SFPD (Ours)	43.2	54.5	59.8	64.6	67.4	32.0	39.9	45.2	50.8	56.2	36.1	49.5	55.1	59.6	63.0

Table 2: MS-COCO novel-class performance (nAP, %). SFPD matches or exceeds the matched FPD baseline in all four reported settings (single-run and multi-run averages at 10/30-shot), providing an additional accuracy comparison beyond VOC. Red: highest per column; blue: second-highest. Previously published multi-run results for comparison methods are reproduced from [31]; SFPD results are ours.

			Single-run results		**Average results**	
Method	Venue (Year)	Framework	10	30	10	30
TFA w/ cos [30]	ICML (2020)	FR-CNN	10.0	13.7	9.1	12.1
Retentive [9]	CVPR (2021)	FR-CNN	10.5	13.8	—	—
FSCE [29]	CVPR (2021)	FR-CNN	11.9	16.4	—	—
FADI [1]	NeurIPS (2021)	FR-CNN	12.2	16.1	—	—
DeFRCN [27]	ICCV (2021)	FR-CNN	18.5	22.6	16.8	21.2
FCT [12]	CVPR (2022)	Transformer	17.1	21.4	15.3	20.2
FSRW [16]	ICCV (2019)	YOLOv2	5.6	9.1	—	—
Meta R-CNN [39]	ICCV (2019)	FR-CNN	8.7	12.4	—	—
FSDetView [38]	ECCV (2020)	FR-CNN	12.5	14.7	10.7	15.9
Meta FR-CNN [11]	AAAI (2022)	FR-CNN	12.7	16.6	—	—
Meta-DETR [40]	TPAMI (2023)	Deformable DETR	—	—	19.0	22.2
DCNet [15]	CVPR (2021)	FR-CNN	—	—	12.8	18.6
VFA [13]	AAAI (2023)	FR-CNN	16.2	18.9	15.9	18.4
SCDC [3]	ICASSP (2026)	FR-CNN	15.5	18.4	—	—
FPD [31]	AAAI (2024)	FR-CNN	16.5	20.1	15.9	19.3
SFPD (Ours)	—	FR-CNN	16.5	20.3	16.3	20.2
Δ over FPD			0.0	**+0.2**	**+0.4**	**+0.9**

4.4 Adaptation Consistency and Convergence

The preceding comparisons show competitive accuracy, but mean AP alone cannot establish reliability. We therefore examine seed-to-seed dispersion, cold-start optimization, and whether each module matters less as support grows. These analyses test the two failure sources in Section 3.2.

Competitive accuracy with low seed dispersion. Table 4 compares the published FPD mean with SFPD mean $\pm$ standard deviation over five seeds. SFPD exceeds FPD in 11 of 12 settings and gains 1.4 points on average; the sole exception is Split 1 at 5-shot (-0.3). Its standard deviations

Table 3: **Core-mechanism ablation on VOC Split 2 1-shot (nAP_{50} %).** (A) Simplified baselines; (B) cumulative integration; (C) leave-one-out removal.

(A) Simplified Baselines (vs. FPD)			
Config	nAP_{50}	Δ	**Implication**
FPD (baseline)	29.8	—	visual query transfer + deterministic matching
FPD + Naive CLIP init	30.8	+1.0	raw CLIP embeddings, no residual
FPD + Uncertainty head only	30.5	+0.7	precision scaling retained; no penalty or KL reg
(B) Cumulative Core Design			
Stage	nAP_{50}	Δ	**Module Added**
s0	29.8	—	Baseline (FPD)
s1	32.0	+2.2	+ LC-FQ (residual-scaled CLIP, α learnable)
s2	33.1	+1.1	+ UA-PD (variance head + attention penalty + KL reg)
(C) Leave-One-Out (from SFPD = 33.1)			
Config	nAP_{50}	Δ	**Module Removed**
SFPD	**33.1**	—	LC-FQ + UA-PD
w/o LC-FQ	31.0	-2.1	LC-FQ removed
w/o UA-PD	32.0	-1.1	UA-PD removed

Table 4: **Competitive accuracy with low seed dispersion (nAP_{50}, mean $\pm$ std over 5 runs).** SFPD exceeds the published FPD mean in 11 of 12 settings (average gain: 1.4) with standard deviations of 0.15–0.69. The exception is Split 1 at 5-shot (-0.3). FPD values from Table 1 are accuracy references; no paired variance comparison is claimed.

		FPD (mean)				**SFPD (Ours) (mean $\pm$ std)**			
Dataset	**Split**	**1-shot**	**3-shot**	**5-shot**	**10-shot**	**1-shot**	**3-shot**	**5-shot**	**10-shot**
VOC	Split1	41.5	58.4	64.9	67.1	**43.2** $\pm$ 0.44	**59.8** $\pm$ 0.29	**64.6** $\pm$ 0.38	**67.4** $\pm$ 0.23
	Split2	28.2	43.8	50.3	53.6	**32.0** $\pm$ 0.48	**45.2** $\pm$ 0.54	**50.8** $\pm$ 0.18	**56.2** $\pm$ 0.15
	Split3	34.9	54.0	58.4	61.5	**36.1** $\pm$ 0.69	**55.1** $\pm$ 0.22	**59.6** $\pm$ 0.35	**63.0** $\pm$ 0.24

remain between 0.15 and 0.69. Thus, SFPD retains competitive accuracy while producing tightly grouped results not attributable to one favorable seed. Since per-seed FPD results are unavailable, we claim low SFPD dispersion, not lower variance than FPD. The tight 1-shot results are consistent with LC-FQ anchoring initialization and UA-PD limiting unreliable support evidence.

Cold-start optimization. Figure 2 traces nAP_{50} over evaluation rounds. SFPD starts from a higher accuracy at the first recorded evaluation and improves earlier than the baseline in the displayed runs. These curves are consistent with improved early adaptation, but do not separately attribute the change to either component or establish lower optimization variance. The separation is strongest at 1-shot in the displayed runs.

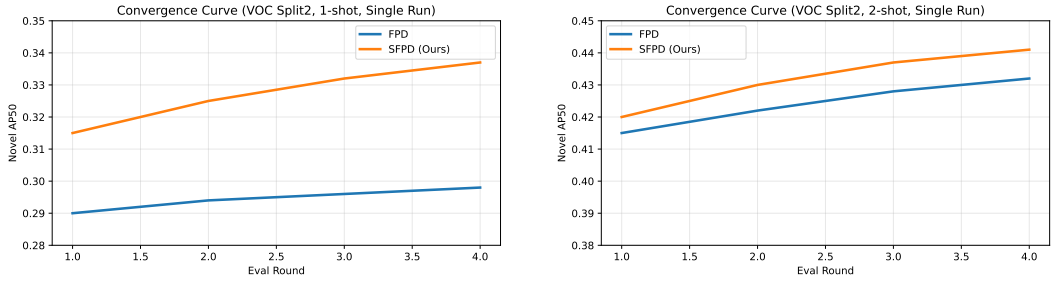
Dependence on support size. The reported leave-one-out effects decrease across 1/3/5-shot: removing LC-FQ costs 2.1/1.3/0.9 points, while removing UA-PD costs 1.1/0.8/0.6 points. This pattern is consistent with both components being most useful when visual evidence is scarce. It does not uniquely identify the source of the improvement.

4.5 Qualitative Analysis

Definition of Missing Rate. Missing Rate (MR) measures the proportion of annotated novel-class objects that the detector fails to recover. Let TP be the number of ground-truth objects matched to a retained prediction of the correct class, and FN the number left unmatched. Then

$$\mathrm{MR} = \frac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}} \times 100\% = (1 - \mathrm{Recall}) \times 100\%. \tag{6}$$

A valid match requires the correct category and sufficient bounding-box intersection over union (IoU); confidence filtering and one-to-one matching also affect the counts. Thus, a missed object may reflect a localization error, an incorrect class, or a prediction removed by filtering. Lower MR indicates fewer misses at the chosen operating point. For example, an MR of 58.1% means approximately 58 of every 100 annotated objects are missed. MR does not measure false-positive detections and therefore complements AP rather than replacing it.



(a) FPD vs. SFPD (1-shot)

(b) FPD vs. SFPD (2-shot)

Figure 2: **Cold-start convergence on VOC Split 2.** SFPD has higher accuracy at the first recorded evaluation and improves earlier in the displayed runs; the separation is strongest at 1-shot.

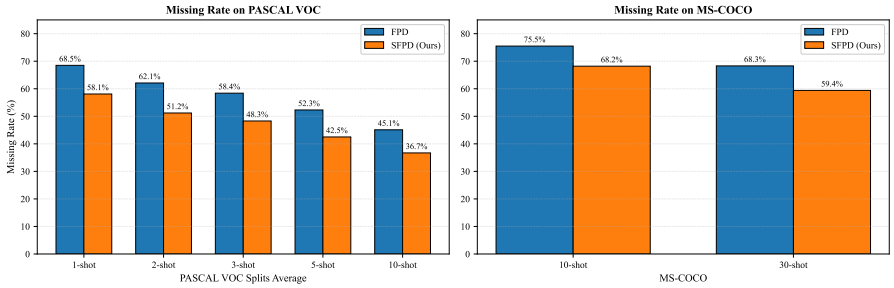
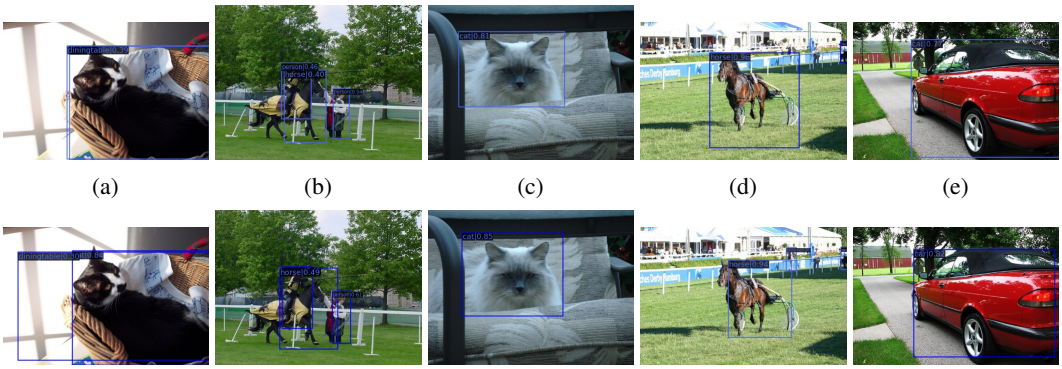


Figure 3: **Missing rate analysis on PASCAL VOC and MS-COCO.** SFPD reduces the missing rate under all shot settings, achieving a drop of 10.4 percentage points on 1-shot VOC. All results are obtained with unified evaluation hyperparameters: confidence threshold = 0.05, NMS IoU = 0.3, matching IoU = 0.5.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 4: **Qualitative comparison under 1-shot novel-class detection. Top:** the baseline misses occluded or cluttered objects and gives low-confidence predictions. **Bottom:** SFPD recovers these detections with higher confidence. Each column shows the same test image. Best viewed in color.

Interpretation of the reported values. Figure 3 reports archived aggregate percentages evaluated under a fixed, unified evaluation protocol. The VOC values are averages across splits; the 1-shot value decreases from 68.5% to 58.1%, a reduction of 10.4 percentage points. All methods use confidence filtering at 0.05, followed by non-maximum suppression (NMS) with an IoU threshold of 0.3. A valid ground-truth-prediction match for MR calculation requires an IoU threshold of 0.5. The full evaluation pipeline and configuration are released in our code repository. These MR values provide quantitative comparisons under the unified evaluation protocol. Figure 4 provides selected paired detections that illustrate recovered objects in the displayed examples.

5 Discussion

Complementary adaptation signals. LC-FQ guides feature extraction with class semantics; UA-PD regulates support fusion and prototype assignment. Removing LC-FQ costs 2.1/1.3/0.9 points at 1/3/5-shot, suggesting greater benefit with scarce visual evidence. On VOC Split 2 1-shot, SFPD reduces NLL from 0.2436 to 0.2049 and Brier score from 0.0613 to 0.0505 (Appendix B). These combined-model results do not establish causal interactions between components or explain individual recovered detections.

Interpreting reliability. Predicted variance captures feature-dependent uncertainty, not calibrated occlusion or labeling-error probabilities. It regulates fusion and assignment, but normalized attention does not guarantee that higher variance lowers every assignment weight. Ambiguous category names and domain shifts may weaken both semantic and visual representations.

Deployment cost. Cached text embeddings avoid online CLIP encoding; query projection and variance estimation still incur computation. SFPD adds 0.3M parameters, with equal FLOPs at the reported precision and similar throughput to FPD (Appendix A). Embeddings are reusable for a fixed vocabulary; prototypes depend on the support set. Latency may vary across hardware and support configurations.

6 Limitations

Our evaluation is limited to prototype-based few-shot object detection on PASCAL VOC and MS-COCO, and the conclusions may not directly generalize to other detection architectures or substantial domain shifts. LC-FQ relies on semantic representations produced by a frozen CLIP text encoder, whose category knowledge may be less reliable in specialized domains. Although SFPD shows low dispersion across five seeds, per-seed results for the published FPD baseline are unavailable, preventing a paired comparison of adaptation variance.

7 Conclusion

We identify semantically unanchored query initialization and deterministic prototype matching as two sources of instability in prototype-based FSOD. SFPD addresses these issues through LC-FQ, which anchors queries with precomputed class-name embeddings from a frozen text encoder, and UA-PD, which uses uncertainty estimates to regulate prototype fusion and assignment. Together, these components improve the adaptation signal while preserving the downstream detection pipeline. Experiments on PASCAL VOC and MS-COCO show improvements in most reported settings, including a single-run 3.3-point nAP_{50} gain on VOC Split 2 in the 1-shot setting, together with low dispersion across SFPD runs and stronger early adaptation in the displayed curves. The multi-run accuracy results and low seed dispersion indicate that the reported benefits are not limited to a single favorable SFPD run. More broadly, this design offers a practical route for adapting detectors to rare or newly emerging categories when annotation and inference-time resources are limited.

Acknowledgments and Disclosure of Funding

This research is supported by a grant from the Shimane University Diversity Promotion Office under the "Project Creation–Type Support Program for Fostering Women Researchers as Leaders at Shimane University," JSPS KAKENHI Grant Number JP26K25556, the Beijing Natural Science Foundation (Grant Nos. 9242003 and L2603037), the National Natural Science Foundation of China (Grant Nos. 72674024 and 62677010), the Key Laboratory of Public Opinion Governance and Computational Communication (Grant No. YQKFYB202501), the Fundamental Research Funds for the Central Universities (Grant No. CUC26TD13), the Foundation of Yunnan Key Laboratory of Service Computing (Grant No. YNSC26108), the Youth Project of the Lianyungang Basic Research Program (Grant No. JCYJ2523), and the Postgraduate Research and Practice Innovation Program of Jiangsu Province (Grant Nos. 26CXJH6471 and 26CXJH6480). The authors declare no competing interests.

References

[1] Yuhang Cao, Jiaqi Wang, Ying Jin, Tong Wu, Kai Chen, Ziwei Liu, and Dahua Lin. Few-shot object detection via Association and DIscrimination. *Advances in Neural Information Processing Systems*, 34:16570–16581, 2021.

[2] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee. Gaussian YOLOv3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 502–511, 2019.

[3] Hao Ding, Wei Zhuo, Zhiyue Tang, and Linlin Shen. Support-conditioned dynamic convolution for few-shot object detection. In *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4171–4175. IEEE, 2026.

[4] Yaoyang Du, Fang Liu, Licheng Jiao, Shuo Li, Zehua Hao, Pengfang Li, Jiahao Wang, Hao Wang, and Xu Liu. Text generation and multi-modal knowledge transfer for few-shot object detection. *Pattern Recognition*, 161:111283, 2025.

[5] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022.

[6] 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(2):303–338, September 2009. ISSN 1573-1405.

[7] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. `https://www.robots.ox.ac.uk/~vgg/projects/pascal/VOC/voc2012/workshop/index.html`, 2012.

[8] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with Attention-RPN and Multi-Relation Detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2020.

[9] Zhibo Fan, Yuchen Ma, Zeming Li, and Jian Sun. Generalized few-shot object detection without forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4527–4536, 2021.

[10] Karim Guirguis, George Eskandar, Mingyang Wang, Matthias Kayser, Eduardo Monari, Bin Yang, and Jürgen Beyerer. Uncertainty-based forgetting mitigation for generalized few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2586–2595, 2024.

[11] Guangxing Han, Shiyuan Huang, Jiawei Ma, Yicheng He, and Shih-Fu Chang. Meta Faster R-CNN: Towards accurate few-shot object detection with attentive feature alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 780–789, 2022.

[12] Guangxing Han, Jiawei Ma, Shiyuan Huang, Long Chen, and Shih-Fu Chang. Few-shot object detection with Fully Cross-Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5321–5330, 2022.

[13] Jiaming Han, Yuqiang Ren, Jian Ding, Ke Yan, and Gui-Song Xia. Few-shot object detection via Variational Feature Aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 755–763, 2023.

[14] Zhenwei He, Xinye Liao, and Xin Feng. Conditional prototype learning for few-shot object detection. In *The 36th British Machine Vision Conference, Sheffield, UK*, pages 1–14, 2025.

[15] Hanzhe Hu, Shuai Bai, Aoxue Li, Jinshi Cui, and Liwei Wang. Dense relation distillation with context-aware aggregation for few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10185–10194, 2021.

[16] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8420–8429, 2019.

[17] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 2017.

[18] Weicheng Kuo, Yin Cui, Xiuye Gu, A. J. Piergiovanni, and Anelia Angelova. F-VLM: Open-vocabulary object detection upon frozen vision and language models. In *International Conference on Learning Representations*, 2023.

[19] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

[20] Jiaxu Leng, Taiyue Chen, Xinbo Gao, Mengjingcheng Mo, Yongtao Yu, and Yan Zhang. Sampling-invariant fully metric learning for few-shot object detection. *Neurocomputing*, 511: 54–66, 2022.

[21] Ning Li, Mingliang Wang, Gaochao Yang, Bo Li, Baohua Yuan, Shoukun Xu, and Jun Qi. Background suppression and comprehensive prototype pyramid distillation for few-shot object detection. *Robotics and Autonomous Systems*, 187:104938, 2025.

[22] 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 *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer International Publishing, 2014. ISBN 978-3-319-10602-1.

[23] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024.

[24] Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He. Part-aware prototype network for few-shot semantic segmentation. In *European Conference on Computer Vision*, pages 142–158. Springer, 2020.

[25] Xiaonan Lu, Wenhui Diao, Yongqiang Mao, Junxi Li, Peijin Wang, Xian Sun, and Kun Fu. Breaking immutable: Information-coupled prototype elaboration for few-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1844–1852, 2023.

[26] Anish Madan, Neehar Peri, Shu Kong, and Deva Ramanan. Revisiting few-shot object detection with vision-language models. *Advances in Neural Information Processing Systems*, 37: 19547–19560, 2024.

[27] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. DeFRCN: Decoupled Faster R-CNN for few-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8681–8690, 2021.

[28] 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*, pages 8748–8763. PMLR, 2021.

[29] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. FSCE: Few-shot object detection via contrastive proposal encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7352–7362, 2021.

[30] Xin Wang, Thomas Huang, Joseph Gonzalez, Trevor Darrell, and Fisher Yu. Frustratingly simple few-shot object detection. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9919–9928. PMLR, 2020. URL `https://proceedings.mlr.press/v119/wang20j.html`.

[31] Zichen Wang, Bo Yang, Haonan Yue, and Zhenghao Ma. Fine-grained prototypes distillation for few-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5859–5866, 2024.

[32] Aming Wu, Yahong Han, Linchao Zhu, and Yi Yang. Universal-prototype enhancing for few-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9567–9576, 2021.

[33] Jiaxi Wu, Songtao Liu, Di Huang, and Yunhong Wang. Multi-scale positive sample refinement for few-shot object detection. In *European Conference on Computer Vision*, pages 456–472. Springer, 2020.

[34] Jiaying Wu and Jinglu Hu. Improved prior selection using semantics in maximum a posteriori for few-shot learning. *Knowledge-Based Systems*, 237:107688, February 2022. ISSN 0950-7051.

[35] Jiaying Wu, Jingyu Chen, Jia Luo, Wenqian Yu, Jinglu Hu, and Hui Li. Denoised generative fusion networks for noise-robust few-shot image classification. *Complex & Intelligent Systems*, 12(2):58, February 2026. ISSN 2198-6053.

[36] Jiaying Wu, Can Gao, Jinglu Hu, Hui Li, Xiaofeng Cao, and Jingcai Guo. PMCE: Probabilistic multi-granularity semantics with caption-guided enhancement for few-shot learning. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI-2026, pages 1848–1856. International Joint Conferences on Artificial Intelligence Organization, September 2026.

[37] Jiaying Wu, Zheng Huang, Jia Luo, Wenqian Yu, Jinglu Hu, and Zhaoman Zhong. Revisiting few-shot image classification: Harnessing attention prompt to diminish background noise. *Expert Systems with Applications*, 331:133252, December 2026. ISSN 0957-4174.

[38] Yang Xiao and Renaud Marlet. Few-shot object detection and viewpoint estimation for objects in the wild. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 192–210, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58520-4.

[39] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta R-CNN: Towards general solver for instance-level low-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9577–9586, 2019.

[40] Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, Shijian Lu, and Eric P. Xing. Meta-DETR: Image-level few-shot detection with inter-class correlation exploitation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12832–12843, 2023.

Appendix: Supplementary Material

This supplementary material reports model complexity and inference efficiency, confidence calibration across multiple shot settings, UA-PD uncertainty visualizations, and a t-SNE analysis of LC-FQ query embeddings. It concludes with broader societal impacts.

A Model Complexity and Inference Efficiency

Table 5: **Model complexity and inference efficiency.** SFPD adds 0.3M parameters. FLOPs are equal at the reported precision (180.2G), and throughput is similar (17.8 versus 17.9 FPS).

Model	Params (M)	FLOPs (G)	Time (ms/img)	FPS
Baseline (FPD)	41.3	180.2	56.0	17.9
SFPD (Ours)	41.6	180.2	56.2	17.8

B Confidence Calibration Results

Expected Calibration Error (ECE) measures the difference between predicted confidence and actual accuracy. SFPD consistently improves calibration across all shot settings compared to the baseline.

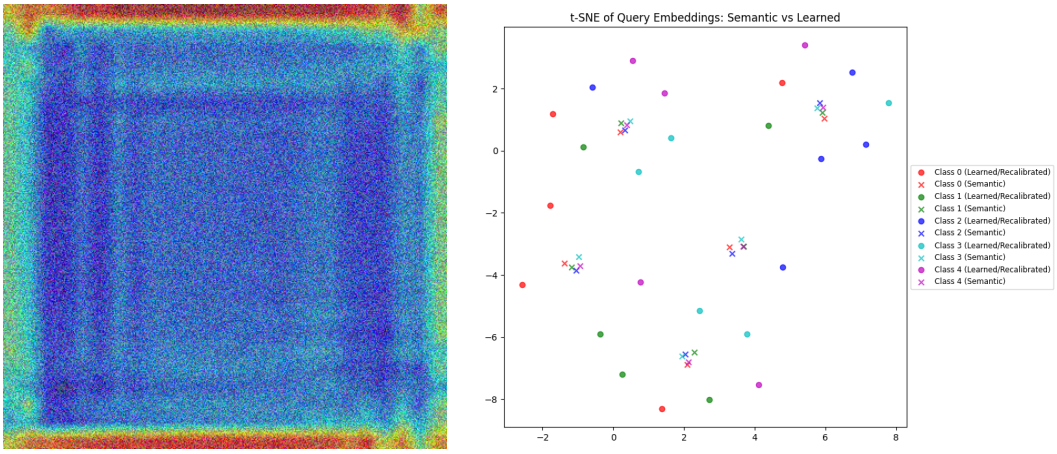
Table 6: Confidence calibration metrics on VOC Split 2 (lower is better).

Shot	**Method**	**ECE**	**NLL**	**Brier**
1-shot	FPD	0.0608	0.2436	0.0613
1-shot	Only-UA-PD	0.0612	0.2170	0.0555
1-shot	SFPD (Ours)	0.0585	0.2049	0.0505
3-shot	FPD	0.0748	0.3721	0.1111
3-shot	Only-UA-PD	0.0615	0.3647	0.1087
3-shot	SFPD (Ours)	0.0578	0.3495	0.1018
5-shot	FPD	0.0878	0.3175	0.0910
5-shot	Only-UA-PD	0.0652	0.2854	0.0825
5-shot	SFPD (Ours)	0.0560	0.2510	0.0715

We note that UA-PD alone achieves comparable ECE at 1-shot while markedly improving NLL ($0.2436 \rightarrow 0.2170$) and Brier ($0.0613 \rightarrow 0.0555$), suggesting that probabilistic modeling sharpens predictive distributions despite similar ECE values. The LC-FQ+UA-PD model consistently reduces all three metrics across all shot settings.

C Additional Qualitative Visualizations

In this paper, we provide supplementary visualizations for UA-PD uncertainty estimation and LC-FQ query embeddings, complementing the quantitative analyses in the main text.



(a) UA-PD uncertainty heatmap. Warmer colors indicate higher uncertainty at visually ambiguous object boundaries and in cluttered regions.

(b) LC-FQ query embeddings: t-SNE comparison of semantic embeddings (crosses) and learned/recalibrated embeddings (circles). Colors denote class identity.

Figure 5: **Supplementary visualizations of UA-PD and LC-FQ.** (a) The variance head predicts per-region uncertainty, with higher values at visually ambiguous object boundaries and in cluttered regions. (b) The LC-FQ panel compares semantic and learned/recalibrated query embeddings using t-SNE, following the plot legend.

D Broader Impacts

Positive impacts. SFPD reduces the need for large-scale manual annotation, lowering the barrier to deploying object detectors in resource-constrained domains such as wildlife conservation, medical imaging, and disaster response. The similar measured inference cost may be useful where data and compute are limited; edge-device performance has not been evaluated.

Negative impacts and mitigation. As with any detection system, SFPD could be misused for unauthorized surveillance or mass tracking. The frozen CLIP encoder may introduce biases inherited from web-scale training data, which can affect under-represented categories or regions. Training and repeated evaluations consume GPU resources; an aggregate compute total cannot be verified from the retained records. We encourage practitioners to audit model fairness on target populations, evaluate domain-specific biases, and consider domain-adapted visual representations when deploying in sensitive applications.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately state the two core contributions (LC-FQ and UA-PD) and the empirical evaluation scope on PASCAL VOC and MS-COCO.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses the evaluation scope, possible limitations of frozen CLIP representations in specialized domains, and the absence of per-seed FPD results for a paired variance comparison.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: Our paper is empirical and does not present formal theorems requiring mathematical proofs.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide details on dataset splits, benchmark protocols, and core-module ablations in Sections 4.1 and 4.3.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The current source implementation is publicly available at `https://github.com/zyc010228/SFPD`. The repository provides the model components and training and evaluation entry points.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training details, including the ResNet-101 backbone, frozen CLIP encoder, and K-shot sampling, are specified in Section 4.1.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the mean and standard deviation over five seeds in Table 4 to ensure statistical robustness. Table 4 is specifically dedicated to multi-seed stability analysis ($mean \pm std$), while Tables 1 and 3 report nAP_{50} values and gains (Δ).

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We report two NVIDIA RTX 3090 GPUs and approximately two wall-clock hours per run (approximately four GPU-hours if both GPUs are fully occupied). The aggregate compute across all experiments is not verified, so complete resource accounting remains unavailable.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics and confirm our work adheres to all guidelines, including the acknowledgement of potential biases in multimodal models.

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the positive impact of data-efficient detection and the potential risks of unauthorized surveillance in Appendix D.

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This work does not release high-risk generative models or newly scraped datasets.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We correctly cite the original owners of PASCAL VOC, MS-COCO, CLIP, and the FPD baseline.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: We do not introduce any new datasets or benchmark assets in this paper.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No human subjects or crowdsourcing were involved in this research.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This study does not involve human participants.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [No]

Justification: No LLM is used as a component of the submitted method.