Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Domain-Held-Out Validation and Temperature Calibration: Stabilizing Rejection Performance of CLIP Adapters for Open-Set Domain Generalization

Hongyi Yang
Guizhou Minzu University
Jiegang Liu
liujiegang2025@163.com
Guizhou Minzu University
Cheng Fang
Guizhou Minzu University
Jin Zhang
Guizhou Minzu University
Shuang Li
Guizhou Minzu University
Ke Jang
Guizhou Minzu University
Taoze Liu
Guizhou Minzu University
Zhanghong Wang
Guizhou Minzu University

Research Article

Keywords: Open-set recognition, domain generalization, CLIP, adapter tuning, temperature scaling, FPR@95

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8631400/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Domain-Held-Out Validation and Temperature Calibration: Stabilizing Rejection Performance of CLIP Adapters for Open-Set Domain Generalization

Hongyi Yang[1], Jiegang Liu[1,2*], Cheng Fang[1], Jin Zhang[1], Shuang Li[1], Ke Jang[1,2], Taoze Liu[1,2], Zhanghong Wang[1,2]

[1]Guizhou Minzu University, Guiyang, 550025, China.
[2]Engineering Research Center for Green and Low-carbon Technology for Plastic Application in Guizhou Province, Guiyang, 550025, China.

*Corresponding author(s). E-mail(s): liujiegang2025@163.com;

Abstract

Open-set recognition (OSR) systems must correctly classify known classes while rejecting unknown-class inputs. In practical image-processing pipelines, OSR is frequently deployed under domain shift (e.g., photo vs. sketch), where the *operating point* of a rejection module can become unstable: the scale and shape of logit/score distributions drift across domains, so a threshold tuned on one domain may yield a higher false-alarm rate or missed rejections on another. Motivated by this reliability issue, we study CLIP-based open-set domain generalization (OSDG) on the PACS benchmark and use a lightweight CLIP adapter trained on source-domain labeled known classes to improve closed-set discrimination. We then propose **DHV-DS** (*Domain-Held-out Validation with Decoupled Scoring*), a minimal calibration pipeline that explicitly accounts for domain shift: (1) *Domain-held-out validation* holds out an entire source domain as a validation domain that is never used for training; (2) *Decoupled scoring with automatic selection* keeps adapter logits for classification, while allowing the rejection score to be computed from zero-shot logits, adapter logits, or their mixture, selected automatically on the held-out domain; (3) *Joint temperature and threshold calibration* tunes temperature and chooses a rejection threshold on the held-out domain to better transfer the desired operating point. On PACS (macro-averaged over four target domains, 10 unknown-class splits, and two random seeds), DHV-DS reduces FPR@95 from 0.1148 to 0.1004 and improves AUROC from 0.9606 to 0.9677, achieving the lowest mean FPR@95 on all four target domains. Paired

comparisons show consistent gains across unknown partitions and seeds. DHV-DS requires no extra data and adds negligible overhead, making it a practical default procedure for stabilizing cross-domain OSR rejection in CLIP-adapter-based systems.

Keywords: Open-set recognition, domain generalization, CLIP, adapter tuning, temperature scaling, FPR@95

1 Introduction

Open-set recognition (OSR) requires a model to not only classify samples from a set of known classes but also detect and reject inputs from unknown classes at test time [1, 2]. Understanding the failure modes of OSR under realistic deployment conditions is essential for building reliable open-world vision systems. Unlike closed-set classification, OSR performance depends not only on class separability, but also critically on the *confidence scale* and the *shape* of the model's score distribution. When unknown samples overlap with known samples in score space, a system can exhibit high known-class accuracy while still producing a large number of false rejections (false positives) or missed rejections, leading to operational risks.

In practice, OSR is often coupled with *domain shift*: training and deployment domains differ in style, texture, and shape statistics (e.g., photo vs. cartoon vs. sketch), causing representation drift and global shifts in similarity/logit distributions. In such settings, a fixed rejection threshold chosen on the training domain tends to be non-transferable, and metrics such as AUROC and FPR@95 can vary significantly across target domains and across unknown-class splits [3–5]. This phenomenon can be interpreted as *cross-domain threshold failure*: even if known-class classification remains strong, the rejection module may become unreliable because temperature/threshold parameters are misaligned with the shifted logit scale.

Vision–language models such as CLIP provide strong zero-shot generalization and serve as a natural starting point for cross-domain recognition [6]. However, CLIP logits are typically obtained from scaled cosine similarity, and the effective scale can drift with domain style, which directly affects standard OSR scores such as maximum softmax probability (MSP). A common strategy to improve known-class recognition is to add a lightweight adapter (e.g., CLIP-Adapter) trained with labeled known classes while freezing the CLIP backbone [7]. While adapters often boost closed-set performance, they can also sharpen logits and raise confidence, potentially amplifying scale drift under domain shift and making rejection thresholds even more brittle. This can lead to the counterintuitive outcome of "better classification but worse or less stable rejection".

To address this tension, we propose **DHV-DS** (Domain-Held-out Validation with Decoupled Scoring), a simple calibration pipeline designed for CLIP-based open-set domain generalization (OSDG). The core idea is to explicitly incorporate domain shift into calibration by holding out an entire source domain as a validation domain, and to decouple the logits used for classification and rejection. Specifically, classification

always uses adapter logits to retain discrimination gains, while rejection is computed from a *separately chosen* score source: zero-shot logits, adapter logits, or their mixture. On the held-out domain, we automatically select the most robust score source and jointly tune temperature and the rejection threshold to obtain a more transferable operating point.

Our contributions are: (i) **System-oriented calibration under domain shift**: we propose *domain-held-out validation* as a simple model-selection and operating-point calibration strategy for OSR when deployment domains differ from training domains [8]. (ii) **Decoupled rejection module**: we introduce *decoupled scoring* with automatic selection (zero-shot / adapter / mixture) so that classification benefits from the adapter while rejection uses the most stable score source on a held-out domain. (iii) **Reliability-focused evaluation**: on PACS OSDG, we demonstrate consistent reductions in false positives at a fixed recall operating point (FPR@95) and improved AUROC across target domains, unknown splits, and random seeds, and provide a lightweight recipe that can be integrated into CLIP-adapter-based recognition pipelines.

2 Related work

Recent studies on open-set and out-of-distribution (OOD) recognition emphasize that reliable deployment depends not only on classification accuracy but also on the stability of operating points under distribution shift. Temperature scaling and other post-hoc calibration techniques are widely used to align confidence with empirical risk, but they typically assume that calibration data are drawn from the same distribution as training data. Under domain shift, this assumption breaks, leading to threshold drift and unstable false-positive rates. Meanwhile, most domain generalization methods (e.g., adversarial alignment, invariant risk minimization, and distributionally robust optimization) primarily target closed-set accuracy and do not explicitly calibrate the rejection operating point. In contrast, we focus on the reliability of the rejection module and use domain-held-out validation as a practical surrogate for unseen target conditions.

Open-set recognition (OSR) studies recognition when test data may include classes that are not observed during training. Early work formalized OSR by explicitly accounting for unknowns and evaluating the joint behavior of correct known-class recognition and unknown rejection [1]. Deep OSR extends these ideas to modern neural networks [2]. Representative directions include reducing "agnostophobia" to improve unknown rejection [5], reconstruction-based OSR via class-conditioned auto-encoders [9], and broader surveys that organize OSR settings and evaluation protocols [10].

Out-of-distribution (OOD) detection is closely related to OSR when unknown samples are treated as out-of-distribution. Common baselines include maximum softmax probability [3] and perturbation/temperature-based scoring such as ODIN [4]. Subsequent work proposed unified confidence scoring frameworks [11], energy-based scores [12], and data-driven improvements using outlier exposure [13]. Recent methods also explore alternative logit representations, such as virtual-logit matching [14].

Calibration aims to align predicted confidences with empirical correctness. Temperature scaling is a simple and widely used post-hoc method for modern neural networks [15], building on classical ideas such as Platt scaling [16]. Multi-class calibration methods such as Dirichlet calibration [17] and Bayesian binning into quantiles (BBQ) [18] can further improve probability calibration. In contrast to calibrating only probabilities, our work calibrates both temperature and a rejection threshold to stabilize a safety-relevant operating point under domain shift.

Domain generalization (DG) learns from multiple source domains and evaluates on an unseen target domain. PACS is a standard benchmark for this setting [8]. Recent analyses emphasize that the choice of evaluation protocol and model-selection strategy can substantially affect reported DG performance and conclusions [19].

A long-standing line of work seeks domain-invariant representations via adversarial alignment, e.g., domain-adversarial training (DANN) [20]. Complementary objectives enforce invariance or robustness at the risk level, including invariant risk minimization (IRM) [21], distributionally robust optimization for group shifts (GroupDRO) [22], and risk extrapolation (REx) [23]. Data-augmentation and gradient-based strategies such as cross-gradient training offer another practical route to improving domain robustness [24].

More recent DG methods include feature-statistics perturbation (MixStyle) [25] and optimization-based approaches such as SWAD, which improves generalization by seeking flatter minima [26]. Gradient matching and variance regularization methods, including Fish and Fishr, further encourage stable updates across domains [27, 28]. In addition, open-domain DG incorporates meta-learning to better handle distribution shifts in open-world settings [29].

Our problem is also related to open-set and universal domain adaptation, where unknown categories appear together with domain shift [30–32].

Finally, we build on vision–language transfer. CLIP learns transferable visual representations from natural language supervision and has become a strong foundation for recognition under distribution shift [6].

Parameter-efficient adaptation is commonly achieved via lightweight add-on modules. CLIP-Adapter improves transfer while keeping most CLIP parameters frozen [7], and Tip-Adapter provides a training-free alternative for efficient few-shot adaptation [33].

Prompt-based adaptation offers another effective mechanism for steering pretrained models. Representative approaches include learning prompts (CoOp) [34], conditional prompt learning (CoCoOp) [35], and visual prompt tuning [36]. These strategies are closely related to parameter-efficient transfer ideas in NLP [37] and low-rank adaptation (LoRA) [38].

3 Method

Implementation and overhead.

Our calibration is a lightweight grid search on the held-out source domain over temperature and the rejection threshold (selected to satisfy the target TPR constraint). Score-source selection is performed by comparing validation risk across candidate

scores. The extra cost is limited to several forward passes on the held-out domain and introduces no additional training, making the procedure easy to integrate into existing CLIP-adapter pipelines with negligible memory and time overhead.

We describe DHV-DS as a three-stage pipeline (*train–calibrate–deploy*) and explain why it improves rejection stability under domain shift.

3.1 Problem formulation

System view and metrics.

From a system perspective, open-set recognition can be viewed as a two-stage pipeline: a feature extraction and classification stage, followed by a rejection stage that controls the operating point. In safety-critical or large-scale deployments, thresholds are often chosen to satisfy a minimum true positive rate (e.g., TPR=95%) while minimizing false alarms. We therefore use FPR@95 as the primary metric, and additionally report cross-domain variance and worst-case performance to quantify operating-point stability under domain shift.

Let $\mathcal{D}_s = \{\mathcal{D}^1, \ldots, \mathcal{D}^m\}$ denote source domains and $\mathcal{D}_t$ be a target domain. Let $\mathcal{C}_k$ and $\mathcal{C}_u$ be the sets of known and unknown classes with $\mathcal{C}_k \cap \mathcal{C}_u = \emptyset$. In OSDG, training accesses only labeled samples from known classes in the source domains. We further split the source domains into a training domain set and a held-out validation domain $\mathcal{D}_{\text{val}}$; the adapter is trained on $\mathcal{D}_{\text{tr}}$ only, and calibration is performed on $\mathcal{D}_{\text{val}}$.

At test time, samples may belong to either $\mathcal{C}_k$ or $\mathcal{C}_u$. We report: Known Acc (classification accuracy on known samples), OSR Acc (overall accuracy where "unknown" predictions on unknown samples are counted correct), and two standard detection metrics AUROC and FPR@95 [3]. FPR@95 measures the false-positive rate (known samples incorrectly rejected) when the true-positive rate (unknown correctly rejected) is 95%, which is particularly relevant for safety-critical operating points.

3.2 CLIP zero-shot classification with prompt ensembling

CLIP embeds images and texts into a shared space. Given an input image x, the image encoder produces a normalized feature $\mathbf{f}(x) \in \mathbb{R}^d$. For each known class $c \in \mathcal{C}_k$, we use a set of prompt templates $\mathcal{P}$ to create natural-language prompts and obtain text embeddings. Prompt ensembling builds a class prototype by averaging template embeddings:

$$\mathbf{t}_c \;=\; \frac{\sum_{p\in\mathcal{P}} \operatorname{norm}(\mathbf{g}(p(c)))}{\left\|\sum_{p\in\mathcal{P}} \operatorname{norm}(\mathbf{g}(p(c)))\right\|_2}, \tag{1}$$

where $\mathbf{g}(\cdot)$ is the CLIP text encoder and $\operatorname{norm}(\cdot)$ denotes ℓ_2 normalization. Stacking all class prototypes gives $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_{|\mathcal{C}_k|}] \in \mathbb{R}^{d\times|\mathcal{C}_k|}$. CLIP zero-shot logits are computed as scaled cosine similarities:

$$\mathbf{z}_{\text{zs}}(x) \;=\; s\,\mathbf{f}(x)^\top \mathbf{T}, \tag{2}$$

where s is CLIP's learned logit scale.

3.3 Lightweight adapter

To improve known-class discrimination without updating the frozen CLIP backbone, we attach a lightweight MLP adapter $A(\cdot)$ to the image feature and use a residual form:

$$\mathbf{f}_{\text{ab}}(x) \;=\; \text{norm}\big(\mathbf{f}(x) + \gamma\, A(\mathbf{f}(x))\big), \tag{3}$$

where $\gamma \in [0, \gamma_{\max}]$ controls the adaptation strength. The adapter is trained on $\mathcal{D}_{\text{tr}}$ using cross-entropy on known classes with an additional cosine regularization term to constrain feature drift:

$$\mathcal{L} \;=\; \text{CE}\big(y,\, \text{softmax}(\mathbf{z}_{\text{ab}}(x))\big) \;+\; \lambda_{\cos}\big(1 - \cos(\mathbf{f}_{\text{ab}}(x), \mathbf{f}(x))\big), \tag{4}$$

where $\mathbf{z}_{\text{ab}}(x) = s\,\mathbf{f}_{\text{ab}}(x)^{\top}\mathbf{T}$ and $\lambda_{\cos}$ is a regularization weight.

3.4 Decoupled scoring for rejection

Adapters tend to sharpen logits and increase MSP, which can amplify domain-induced scale drift and destabilize fixed thresholds. We therefore decouple *classification logits* and *rejection logits*. Classification always uses adapter logits $\mathbf{z}_{\text{ab}}$ to retain discrimination gains. For rejection, we allow the score source to be chosen from:

$$\mathbf{z}_{\text{score}}(x) \in \{\mathbf{z}_{\text{zs}}(x),\, \mathbf{z}_{\text{ab}}(x),\, \mathbf{z}_{\text{mix}}(x)\}, \tag{5}$$

$$\mathbf{z}_{\text{mix}}(x) = (1-\lambda)\,\mathbf{z}_{\text{zs}}(x) + \lambda\,\mathbf{z}_{\text{ab}}(x), \tag{6}$$

where $\lambda \in [0, 1]$ is a mixture weight. We use MSP as the rejection score:

$$S(x; T) \;=\; \max_{k} \text{softmax}\left(\frac{\mathbf{z}_{\text{score}}(x)}{T}\right)_{k}, \tag{7}$$

where $T > 0$ is a temperature parameter tuned on $\mathcal{D}_{\text{val}}$ (equivalently rescaling the effective logit scale).

A sample is rejected as unknown if $S(x; T) < \tau$. Otherwise, it is treated as known and classified by $\arg\max_k \mathbf{z}_{\text{ab}}(x)_k$.

3.5 Domain-held-out validation and joint calibration (DHV-DS)

DHV-DS holds out an entire source domain $\mathcal{D}_{\text{val}}$ to simulate deployment shift and uses it to select: (i) the rejection score source (zero-shot / adapter / mixture), (ii) the mixture weight λ (if applicable), (iii) the temperature T, and (iv) the rejection threshold τ.

We perform grid search over $(\text{score_source}, \lambda, T)$ on $\mathcal{D}_{\text{val}}$ and select the configuration that minimizes FPR@95; ties are broken by higher AUROC. Then, under the selected configuration, we choose τ such that TPR on unknown samples in $\mathcal{D}_{\text{val}}$ reaches 95% (thereby defining FPR@95). Algorithm 1 summarizes the procedure.

Algorithm 1 DHV-DS calibration on a held-out domain $\mathcal{D}_{\text{val}}$.

Require: trained adapter; temperature candidates $\mathcal{T}$; mixture candidates Λ; score sources $\mathcal{S} = \{\text{zs}, \text{ab}, \text{mix}\}$.

Ensure: selected $(s^\star, \lambda^\star, T^\star, \tau)$.

1: **for all** $(s, \lambda, T) \in \mathcal{S} \times \Lambda \times \mathcal{T}$ **do**
2: compute $S(x; T)$ for all $x \in \mathcal{D}_{\text{val}}$ using score source s (Eq. 7).
3: compute AUROC and FPR@95 on $\mathcal{D}_{\text{val}}$.
4: **end for**
5: select $(s^\star, \lambda^\star, T^\star)$ that minimizes FPR@95 (tie-break by AUROC).
6: under $(s^\star, \lambda^\star, T^\star)$, set threshold τ so that unknown TPR on $\mathcal{D}_{\text{val}}$ is 95%.

3.6 Pipeline illustration

Fig. 1 illustrates the overall train–calibrate–deploy procedure and the decoupling between classification and rejection.

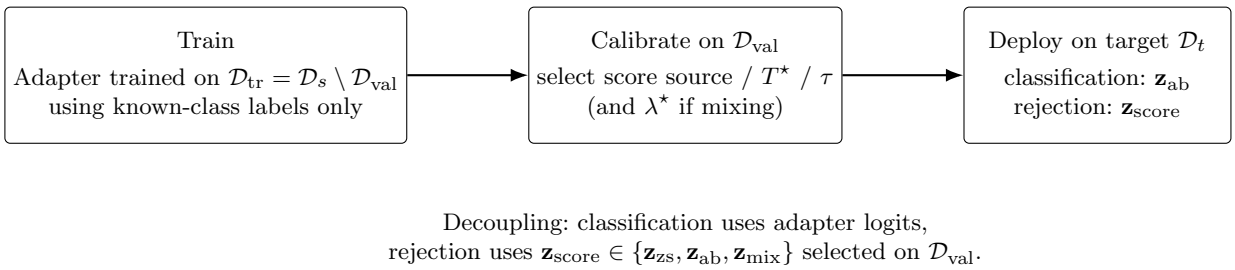


Fig. 1 DHV-DS: domain-held-out validation, decoupled scoring, and joint temperature/threshold calibration.

4 Experimental setup

4.1 Dataset and protocol

We evaluate on the PACS benchmark [8], which contains four domains (`photo`, `art_painting`, `cartoon`, `sketch`) and seven object categories. We follow the leave-one-domain-out protocol: each run selects one domain as the target $\mathcal{D}_t$ and uses the remaining three as source domains. For the open-set setting, we choose two categories (out of seven) as unknown; these unknown categories are removed from source-domain training data and appear only during validation and testing as unknown inputs.

We implement all experiments in PyTorch and optionally use automatic mixed precision (AMP) for efficiency.

Our main results average over 4 target domains $\times$ 10 unknown splits $\times$ 2 random seeds, yielding 80 runs. We use an inner domain holdout: within the three source domains, we hold out one entire domain as $\mathcal{D}_{\text{val}}$ for calibration and train on the remaining two domains $\mathcal{D}_{\text{tr}}$ (no leakage). Fig. 2 illustrates the protocol.

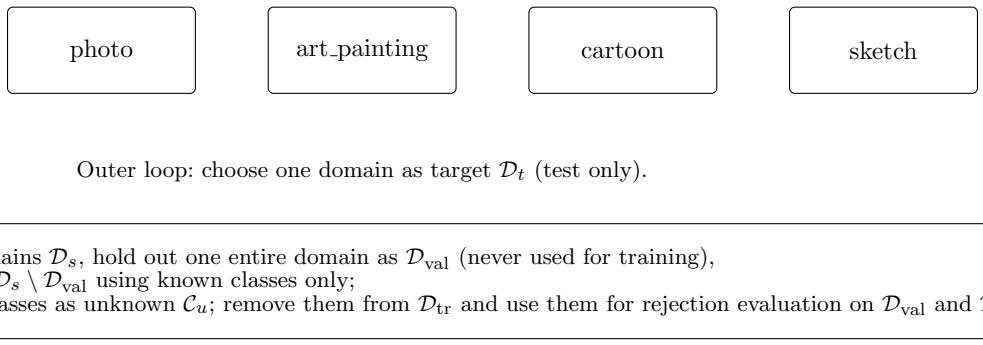


Fig. 2 PACS OSDG protocol: outer leave-one-domain-out with an inner domain-held-out validation split.

4.2 Metrics

Let $\mathcal{K}$ and $\mathcal{U}$ denote the known and unknown test sets. Known Acc is the classification accuracy on $\mathcal{K}$. OSR Acc counts a prediction as correct if a known sample is accepted and classified correctly, or if an unknown sample is rejected. AUROC is computed from the rejection score $S(x;T)$ treating unknown as the positive class. FPR@95 is the false-positive rate on known samples at the operating point where unknown TPR is 95%.

4.3 Baselines

Scope of comparison. Our goal is system-oriented calibration within a CLIP-adapter pipeline: we keep the recognition backbone and adapter training fixed, and evaluate how different *score sources and calibration policies* transfer an operating point under domain shift. Accordingly, we compare plug-in variants that differ only in the rejection score and calibration procedure.

We compare: **(i) Adapter baseline**: both classification and rejection use adapter logits; rejection score is MSP with a fixed default temperature. **(ii) Zero-shot + calibration**: use zero-shot logits for rejection; tune temperature and threshold on $\mathcal{D}_{\text{val}}$ with the same selection criterion (minimize FPR@95). **(iii) Ours (DHV-DS)**: classification uses adapter logits; rejection uses an automatically selected score source (zs / ab / mix) with joint temperature and threshold calibration on $\mathcal{D}_{\text{val}}$.

4.4 Implementation details

We use CLIP ViT-B/32 with OpenAI pre-trained weights [6]. The CLIP backbone is frozen. The adapter is a lightweight MLP (hidden ratio 0.0625) trained with AdamW [39] on $\mathcal{D}_{\text{tr}}$ known-class samples. We use automatic mixed precision for efficiency. Temperature candidates and mixture candidates are searched on $\mathcal{D}_{\text{val}}$; the selection criterion prioritizes lower FPR@95 and uses AUROC as tie-break.

5 Results and analysis

5.1 Macro-averaged results

Table 1 reports macro-averaged results over all runs. Compared with the adapter baseline (classification and rejection both tied to adapter logits), DHV-DS improves open-set detection: FPR@95 decreases from 0.1148 to 0.1004 and AUROC increases from 0.9606 to 0.9677. Known Acc remains unchanged (0.9625), while OSR Acc changes slightly, indicating that the gain mainly comes from improved rejection reliability under domain shift rather than incidental changes in closed-set accuracy. Compared with zero-shot + calibration, DHV-DS further reduces FPR@95 ($0.1075 \rightarrow 0.1004$) while maintaining comparable Known Acc, suggesting that adapter discrimination and robust calibration can be combined.

Table 1 Macro-averaged results on PACS (4 target domains, 10 unknown splits, 2 seeds). Lower is better for FPR@95.

Method	Known Acc	OSR Acc	AUROC	FPR@95↓
Adapter (MSP)	0.9625	0.8941	0.9606	0.1148
Zero-shot + calibration (FPR@95)	0.9629	0.8860	0.9651	0.1075
Ours (DHV-DS: AutoMix + Temp, FPR@95 calib.)	0.9625	0.8917	0.9677	0.1004

5.2 Distributional view of rejection stability

Average metrics can hide long-tail risk. Fig. 3 shows the distribution of FPR@95 over all runs. DHV-DS shifts the distribution toward lower false rejections and reduces the frequency of high-FPR outliers, consistent with our hypothesis that domain-held-out calibration and decoupled score selection mitigate cross-domain logit scale drift.

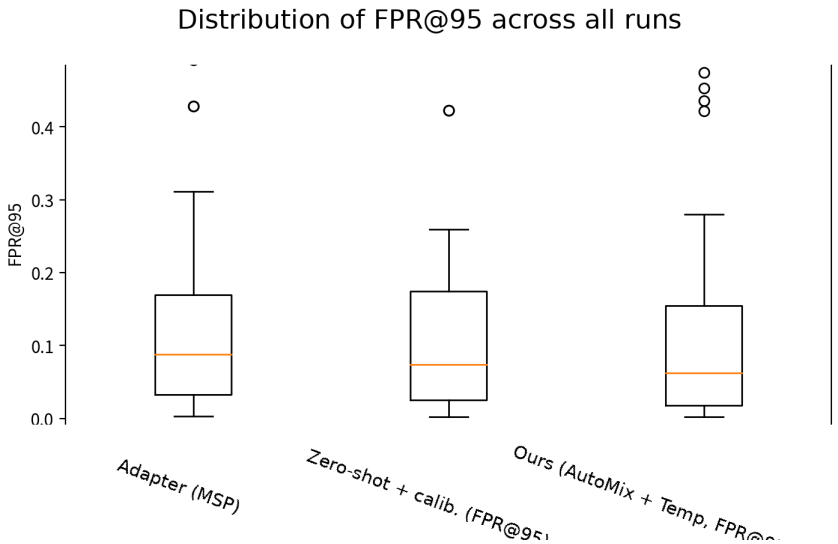


Fig. 3 Distribution of FPR@95 across all runs (box plots). DHV-DS reduces median FPR@95 and suppresses high-FPR outliers.

5.3 Per-domain results and hard-domain analysis

Table 2 reports per-target-domain mean±std results (each domain has $n = 20$ runs). The `sketch` domain is the most challenging due to stronger style shift, leading to larger overlap between known and unknown scores. DHV-DS achieves the lowest mean FPR@95 on all four target domains, including the hard `sketch` domain (0.2568 $\rightarrow$ 0.2152), while improving AUROC (0.9085 $\rightarrow$ 0.9313) on `sketch`.

Table 2 Per-target-domain results on PACS (mean±std, $n = 20$ per domain).

Target domain	Method	Known Acc	OSR Acc	AUROC	FPR@95↓
art_painting	Adapter (MSP)	0.9736±0.0051	0.9017±0.0385	0.9592±0.0229	0.1212±0.0635
	Zero-shot + calib.	0.9731±0.0054	0.8983±0.0404	0.9600±0.0257	0.1209±0.0684
	Ours (DHV-DS)	0.9736±0.0051	0.9046±0.0426	0.9622±0.0250	0.1136±0.0707
cartoon	Adapter (MSP)	0.9859±0.0062	0.9311±0.0264	0.9780±0.0126	0.0700±0.0362
	Zero-shot + calib.	0.9863±0.0064	0.9298±0.0336	0.9788±0.0159	0.0696±0.0448
	Ours (DHV-DS)	0.9859±0.0062	0.9377±0.0275	0.9809±0.0149	0.0623±0.0443
photo	Adapter (MSP)	0.9983±0.0016	0.9509±0.0400	0.9967±0.0021	0.0112±0.0115
	Zero-shot + calib.	0.9987±0.0013	0.8985±0.1824	0.9948±0.0035	0.0158±0.0151
	Ours (DHV-DS)	0.9983±0.0016	0.9024±0.1837	0.9966±0.0024	0.0106±0.0143
sketch	Adapter (MSP)	0.8922±0.0527	0.7926±0.0578	0.9085±0.0540	0.2568±0.1192
	Zero-shot + calib.	0.8935±0.0550	0.8174±0.0543	0.9268±0.0628	0.2236±0.1420
	Ours (DHV-DS)	0.8922±0.0527	0.8222±0.0477	0.9313±0.0523	0.2152±0.1360

Fig. 4 further visualizes FPR@95 distributions on the hard `sketch` domain. The adapter baseline exhibits a heavier high-FPR tail, while DHV-DS concentrates more mass in lower-FPR regions, suggesting improved robustness under stronger domain shift.

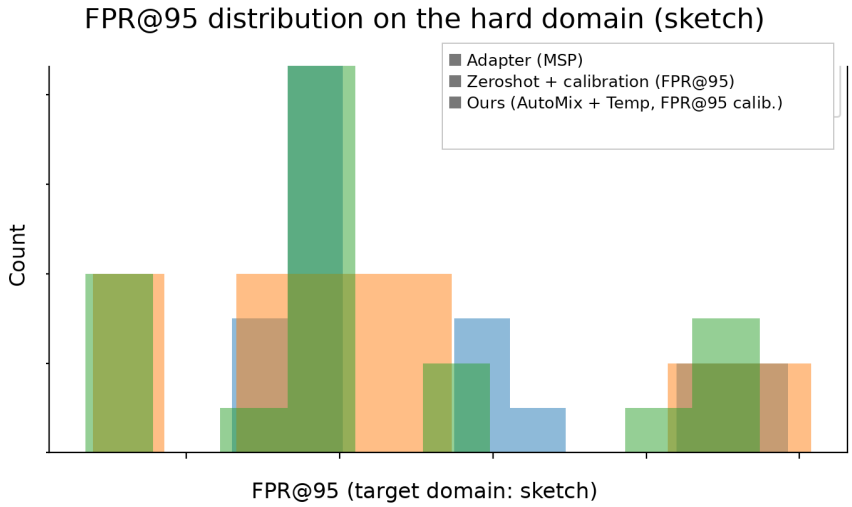


Fig. 4 FPR@95 distribution on the hard `sketch` domain. DHV-DS reduces the high-FPR tail compared to baselines.

5.4 Optimal configuration statistics

To better understand the behavior of automatic score-source selection, we count how often each score source is chosen across all runs. As shown in Fig. 5, adapter logits are chosen most frequently, but mixture and zero-shot are also selected in a non-trivial fraction of runs. This supports our design motivation: the most stable rejection score source depends on the target domain and unknown split, and a flexible, decoupled selection is beneficial.

Table 3 Best-performing configuration per target domain (selected by minimizing FPR@95 on the held-out validation domain).

Target domain	Best method	FPR@95↓	AUROC↑	Known Acc↑	OSR Acc↑
art_painting	Ours (AutoMix + Temp, FPR@95 calibration)	0.1136	0.9622	0.9736	0.9046
cartoon	Ours (AutoMix + Temp, FPR@95 calibration)	0.0623	0.9809	0.9859	0.9377
photo	Ours (AutoMix + Temp, FPR@95 calibration)	0.0106	0.9966	0.9983	0.9024
sketch	Ours (AutoMix + Temp, FPR@95 calibration)	0.2152	0.9313	0.8922	0.8222

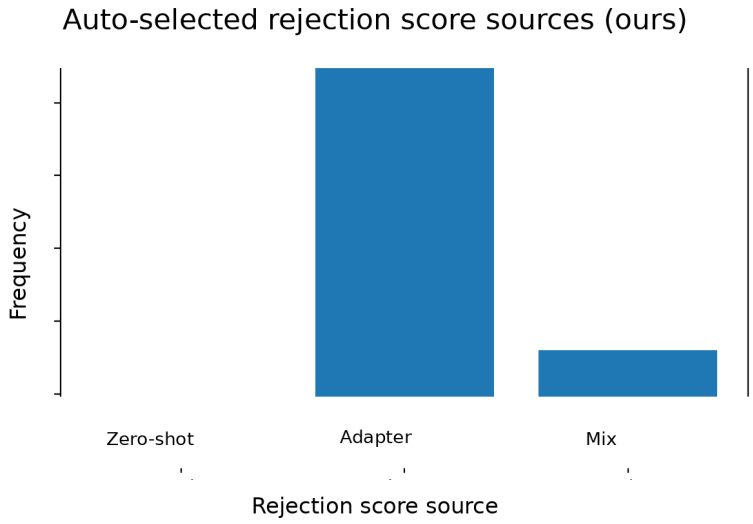


Fig. 5 Frequency of the automatically selected rejection score source across 80 runs.

5.5 Paired gains

We perform strict paired comparisons under identical settings (target domain, unknown split, and seed). We then compute differences relative to the adapter baseline. Table 4 shows that DHV-DS yields a consistent reduction of FPR@95 and an increase of AUROC under paired evaluation.

Table 4 Paired improvements over the adapter baseline (80 paired runs). Negative ΔFPR@95 is better.

Compared to baseline	Pairs	ΔOSR Acc	ΔAUROC	ΔFPR@95↓
Ours (DHV-DS)	80	-0.0024 ± 0.0993	0.0071 ± 0.0141	-0.0144 ± 0.0275
Zero-shot + calib.	80	-0.0081 ± 0.0991	0.0045 ± 0.0143	-0.0073 ± 0.0290

Fig. 6 provides a visual paired comparison of FPR@95 between DHV-DS and the adapter baseline. Most points fall below the diagonal, indicating lower FPR@95 for DHV-DS in the majority of settings.

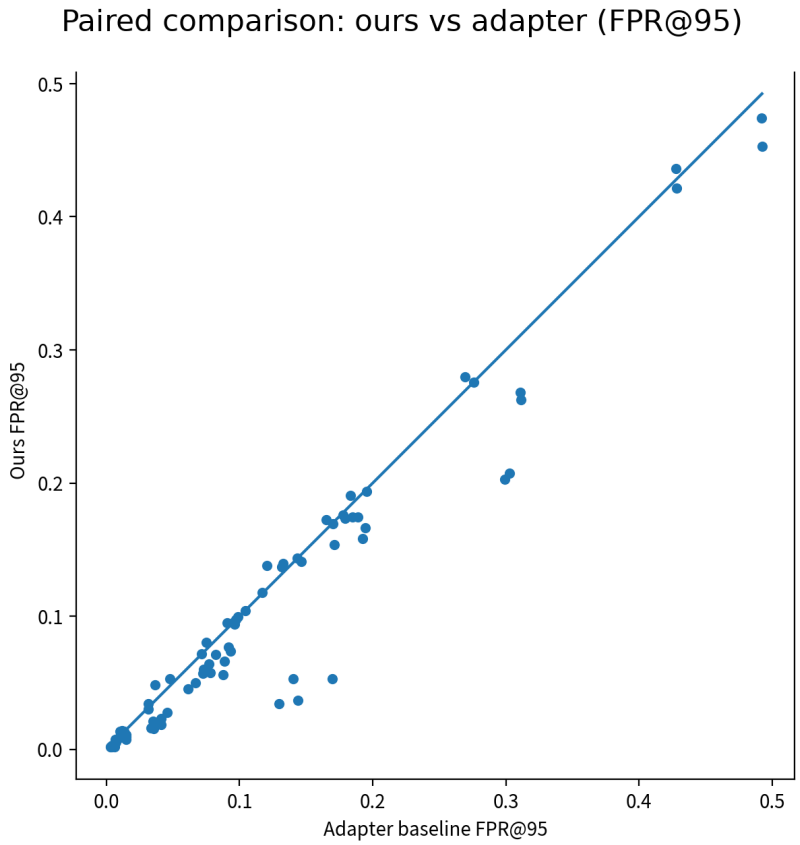


Fig. 6 Paired comparison of FPR@95: ours vs. adapter baseline. Points below the diagonal indicate improvement by DHV-DS.

6 Discussion and limitations

6.1 Deployment notes and optional configurations

For practical deployment, we recommend domain-held-out calibration whenever multiple source domains are available, since it provides a conservative proxy for unseen target behavior and reduces operating-point drift. The approach can be less effective when (i) the held-out domain is not representative of the target shift, or (ii) the number of source domains is very limited. In such cases, conservative thresholding and monitoring confidence/score drift are advisable to maintain stable rejection behavior.

DHV-DS is intended as a practical default for stabilizing rejection under domain shift. Below we summarize optional configurations that practitioners may adjust depending on the deployment setting and available data:

- **Validation split choice:** we recommend holding out an entire source domain as $\mathcal{D}_{\text{val}}$ to reflect cross-domain deployment. If domain-level holdout is infeasible, a

grouped split that preserves domain-specific style (rather than i.i.d. random splits) is a better approximation for threshold transfer.

- **Score source policy:** for rejection, one may either (a) fix the score source to {zero-shot, adapter, mix} for simplicity, or (b) use automatic selection (DHV-DS) to adapt the score source to the most stable option on $\mathcal{D}_{\text{val}}$.
- **Temperature and threshold:** jointly tuning T and τ on $\mathcal{D}_{\text{val}}$ improves operating-point transfer when score scales drift. In deployments with strict latency constraints, one may fix T and tune only τ as a lightweight fallback.
- **Prompt construction:** for zero-shot and mixed scores, prompt ensembling (multiple templates) can reduce variance; single-template prompts are cheaper but may be more sensitive to wording.

6.2 Failure modes and limitations

The method is designed for multi-source settings where holding out an entire source domain provides a conservative proxy for the target shift. It can be less effective when (i) the held-out domain is not representative of the target deployment shift, or (ii) the number of source domains is very limited. In these cases, practitioners may adopt more conservative thresholds and monitor score drift to maintain stable rejection behavior.

This study focuses on a single benchmark and a specific family of vision–language models and adapters. Extending the analysis to larger-scale domain generalization benchmarks and broader OOD scoring functions is an important direction for future work. Another promising avenue is to combine domain-held-out calibration with online monitoring or periodic recalibration to improve long-term stability in evolving environments.

DHV-DS prioritizes a safety-relevant operating point (TPR=95% on unknowns) and therefore may slightly trade off OSR Acc in some settings. In practice, users can shift the operating point by adjusting τ to match their risk tolerance.

6.3 Future work

We see three practical directions for follow-up studies: (1) evaluating the proposed calibration protocol on larger-scale domain generalization benchmarks and more diverse target shifts; (2) extending the set of rejection scores and calibration objectives to cover broader OOD scoring families; and (3) enabling label-free operating-point calibration under domain shift. Our current procedure sets the threshold to achieve a target unknown TPR on a held-out domain, which assumes access to labeled unknown samples for calibration. When such labels are unavailable, one may estimate the operating point using proxy out-of-distribution data, synthetic unknowns, or unsupervised tail modeling. Developing reliable label-free calibration and pairing it with online monitoring or periodic recalibration remains an important direction for robust long-term deployment.

7 Conclusion

We presented DHV-DS, a lightweight calibration pipeline for CLIP-based open-set domain generalization. By holding out a full domain for validation, decoupling classification and rejection logits, and jointly calibrating temperature and threshold with automatic score-source selection, DHV-DS stabilizes rejection behavior under domain shift. On PACS, DHV-DS consistently lowers FPR@95 and improves AUROC over both an adapter baseline and a zero-shot calibration baseline, with negligible additional overhead. We hope DHV-DS serves as a practical default calibration strategy for reliable open-world deployment.

Acknowledgements. The authors acknowledge Guizhou Minzu University for providing institutional resources and support during this study (Science and Technology Achievement Transformation Program; Project Leader: Jiegang Liu).

Declarations

Funding. This research received no external funding.

Conflict of interest. The authors declare that they have no conflict of interest.

Use of generative AI. During the preparation of this work, the author(s) used *ChatGPT (OpenAI)* for language polishing and to assist with code drafting/debugging for the experiments. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the content.

References

[1] Scheirer, W.J., Rezende Rocha, A., Sapkota, A., Boult, T.E.: Toward open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(7), 1757–1772 (2013) https://doi.org/10.1109/TPAMI.2012.256

[2] Bendale, A., Boult, T.E.: Towards open set deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016). https://arxiv.org/abs/1511.06233

[3] Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (ICLR) (2017). https://arxiv.org/abs/1610.02136

[4] Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks (ODIN). In: International Conference on Learning Representations (ICLR) (2018). https://arxiv.org/abs/1706.02690

[5] Dhamija, A.R., Günther, M., Boult, T.E.: Reducing network agnostophobia. In: Advances in Neural Information Processing Systems (NeurIPS) (2018). https://arxiv.org/abs/1811.04110

[6] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML) (2021). https://arxiv.org/abs/2103.00020

[7] Gao, P., Geng, S., Zhang, R., et al.: CLIP-Adapter: Better Vision-Language Models with Feature Adapters (2021). https://arxiv.org/abs/2110.04544

[8] Li, D., Yang, Y., Song, Y.-Z., Hospedales, T.: Deeper, broader and artier domain generalization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017). https://arxiv.org/abs/1710.03077

[9] Oza, P., Patel, V.M.: C2AE: Class conditioned auto-encoder for open-set recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019). https://arxiv.org/abs/1904.07674

[10] Geng, C., Huang, S.-J., Chen, S.: Recent Advances in Open Set Recognition: A Survey (2018). https://arxiv.org/abs/1811.08581

[11] Lee, K., Lee, H., Lee, K., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Advances in Neural Information Processing Systems (NeurIPS) (2018). https://arxiv.org/abs/1807.03888

[12] Liu, W., Wen, Y., Yu, Z., *et al.*: Energy-based out-of-distribution detection. In: Advances in Neural Information Processing Systems (NeurIPS) (2020). https://arxiv.org/abs/2010.03759

[13] Hendrycks, D., Mazeika, M., Dietterich, T.: Outlier Exposure with Confidence Control for Out-of-Distribution Detection (2018). https://arxiv.org/abs/1812.04606

[14] Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022). https://arxiv.org/abs/2203.10807

[15] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML) (2017). https://arxiv.org/abs/1706.04599

[16] Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in Large Margin Classifiers, (1999). https://www.microsoft.com/en-us/research/publication/probabilistic-outputs-for-support-vector-machines-and-comparisons-to-regularized-likelihood-methods/

[17] Kull, M., Silva Filho, T., Flach, P.: Beyond Temperature Scaling: Obtaining Well-Calibrated Multi-Class Probabilities with Dirichlet Calibration (2019). https://arxiv.org/abs/1910.12656

[18] Naeini, M.P., Cooper, G.F., Hauskrecht, M.: Obtaining well calibrated probabilities using Bayesian binning into quantiles. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2015). https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9667

[19] Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization. In: International Conference on Learning Representations (ICLR) (2021). https://arxiv.org/abs/2007.01434

[20] Ganin, Y., Ustinova, E., Ajakan, H., *et al.*: Domain-adversarial training of neural networks. Journal of Machine Learning Research **17**(59), 1–35 (2016) arXiv:1505.07818

[21] Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant Risk Minimization (2019). https://arxiv.org/abs/1907.02893

[22] Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization (2019). https://arxiv.org/abs/1911.08731

[23] Krueger, D., Caballero, E., Jacobsen, J.-H., et al.: Out-of-Distribution Generalization via Risk Extrapolation (REx) (2020). https://arxiv.org/abs/2003.00688

[24] Shankar, S., Piratla, V., Chakrabarti, S., *et al.*: Generalizing across domains via cross-gradient training. In: International Conference on Learning Representations (ICLR) (2018). https://arxiv.org/abs/1804.10745

[25] Zhou, K., Yang, J., Qiao, Y., Xiang, T.: Domain generalization with MixStyle. In: International Conference on Learning Representations (ICLR) (2021). https://arxiv.org/abs/2104.02008

[26] Cha, J., Chun, S., Lee, K., *et al.*: SWAD: Domain generalization by seeking flat minima. In: Advances in Neural Information Processing Systems (NeurIPS) (2021). https://arxiv.org/abs/2102.08604

[27] Shi, Y., Seely, J., Torr, P., *et al.*: Fish: A method for gradient matching for domain generalization. In: International Conference on Learning Representations (ICLR) (2022). https://arxiv.org/abs/2104.09937

[28] Rame, A., Dancette, C., Cord, M., *et al.*: Fishr: Invariant gradient variances for out-of-distribution generalization. In: Proceedings of the 39th International Conference on Machine Learning (ICML) (2022). https://arxiv.org/abs/2109.0

2934

[29] Shu, Y., Xu, J., Liao, S., *et al.*: Open domain generalization with domain-augmented meta-learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021). https://arxiv.org/abs/2104.03620

[30] Panareda Busto, P., Gall, J.: Open set domain adaptation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017). https://arxiv.org/abs/1709.09475

[31] Saito, K., Yamamoto, S., Ushiku, Y., Harada, T.: Open set domain adaptation by backpropagation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018). https://arxiv.org/abs/1807.07506

[32] You, K., Long, M., Wang, J., Jordan, M.I.: Universal domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019). https://arxiv.org/abs/1812.04803

[33] Zhang, R., Fang, R., Gao, P., *et al.*: Training-free adaption of CLIP for few-shot classification (Tip-Adapter). In: Proceedings of the European Conference on Computer Vision (ECCV) (2022). https://arxiv.org/abs/2207.09519

[34] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to Prompt for Vision-Language Models (2021). https://arxiv.org/abs/2109.01134

[35] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022). https://arxiv.org/abs/2203.05557

[36] Jia, M., Tang, L., Chen, B.-C., *et al.*: Visual prompt tuning. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022). https://arxiv.org/abs/2203.12119

[37] Houlsby, N., Giurgiu, A., Jastrzebski, S., *et al.*: Parameter-efficient transfer learning for NLP. In: Proceedings of the 36th International Conference on Machine Learning (ICML) (2019). https://arxiv.org/abs/1902.00751

[38] Hu, E.J., Shen, Y., Wallis, P., et al.: LoRA: Low-Rank Adaptation of Large Language Models (2021). https://arxiv.org/abs/2106.09685

[39] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019). https://arxiv.org/abs/1711.05101