Open-World AI Image Detection with Abstention

Omar Ghabayen
ECE
oghabaye@andrew.cmu.edu

Gorkem Yar
ECE
gyar@andrew.cmu.edu

Abdullah Almansour
ECE
aalmanso@andrew.cmu.edu

Zongchi Xie
ECE
zongchix@andrew.cmu.edu

Youlin Qu
ECE
ynq@andrew.cmu.edu

Abstract

Binary AI-image detectors often perform well on curated benchmarks but fail under distribution shift. We study open-world AI image detection through a reliability-centered framework that augments a CLIP-based detector with calibration and abstention. We evaluate three architectures (CLIP baseline, DIRE-inspired residual fusion, and SGF-Net) across four datasets and three decision modes (forced, threshold, and conformal). Our results show that a lightweight residual branch consistently improves robustness across domains, while selective prediction significantly increases accuracy on answered samples. We further demonstrate that conformal abstention provides a principled reliability and coverage tradeoff with statistical guarantees. Finally, we report a negative result as multi-branch forensic model (SGF-Net) overfits to training artifacts and fails under unseen-generator shift. These findings highlight that in open-world detection, reliability depends not only on model architecture but also on calibrated uncertainty and abstention.

1 Introduction

The generative landscape is inherently non-stationary. New text-to-image models are introduced continuously, and detectors trained on past data often fail to generalize under this open-world shift. For example, the Visual Counter Turing Test (VCT^2) [7] evaluates detectors on modern generators (e.g., Stable Diffusion (SD3/SD3.5) [4], DALL·E 3 [10], Midjourney 6 [8]) and shows that many methods degrade to near-chance performance under unified evaluation settings [7]. Beyond generator shift, real-world transformations such as JPEG compression and reposting can destroy fragile forensic signals, while adversarial benchmarks such as RAID [3] demonstrate that even state-of-the-art detectors can be systematically deceived.

These limitations highlight a fundamental issue as conventional detectors are typically optimized for forced classification. More specifically, the model must always predict real or fake, even when the input lies outside its training distribution. In open-world settings, such forced decisions can lead to confidently incorrect predictions, which are often more harmful than uncertainty.

To address this, we adopt a selective prediction framework that allows the model to abstain on uncertain inputs instead of committing to potentially incorrect decisions. However, introducing abstention raises a critical challenge, as if not properly regulated, the model may overuse abstention and avoid difficult tasks altogether, reducing practical utility. Therefore, abstention must be coupled with mechanisms that balance reliability and coverage.

In this work, we incorporate calibrated confidence and principled abstention through temperature scaling and split conformal prediction, enabling the model to make predictions with controlled risk while maintaining meaningful coverage. Rather than optimizing only for raw accuracy, our goal is to improve the trustworthiness of AI image detection systems under distribution shift, ensuring that the model is accurate when it chooses to answer and appropriately uncertain otherwise.

18-786 Course Project .

Our goal is to build a classification system that determines whether an image is AI-generated or real when the generator is unknown, and safely abstains when uncertain. More specifically, our detector predicts AI or Real only when confidence exceeds a threshold and abstains otherwise. We evaluate accuracy and calibration on the images it chooses to label, reflecting real-world settings where low-confidence guesses can be harmful.

2 Related Work

AI-image detection under distribution shift. Recent work shows that detector performance can degrade sharply when test data differs from training data in generator family, image pipeline, or post-processing [9, 3]. Community Forensics emphasizes scale and diversity in detector training, arguing that broad generator coverage improves generalization relative to small-generator settings [11]. However, benchmark studies such as VCT^2 highlight that many detectors remain brittle on newer text-to-image systems and unified evaluation protocols [7]. This motivates open-world evaluation, in which the training and test distributions are intentionally mismatched.

Robustness and adversarial stress testing. Beyond natural shift, adversarial transfer further challenges detector reliability. RAID provides large-scale transferable attacks that can substantially reduce detection performance across models, exposing gaps between standard benchmark accuracy and real robustness [3]. These findings suggest that forced binary decisions are often unsafe in deployment settings where manipulations, reposting, and compression are common.

Forensic features and selective prediction. On the modeling side, Contrastive Language-Image Pretraining (CLIP) style representations provide strong generic visual features for lightweight downstream detectors [12], while DIRE proposes diffusion-reconstruction residual signals that can improve cross-diffusion forensic behavior [13]. On the uncertainty side, temperature scaling is a simple and effective post-hoc calibration method for reducing overconfidence [6]. Conformal prediction adds a complementary set-valued uncertainty layer with finite-sample coverage guarantees under exchangeability assumptions [2]. In selective classification, abstention is used to trade coverage for lower error on answered samples, typically summarized with risk–coverage curves and Area Under the Risk Curves (AURC) [5].

Positioning of this work. Our work combines these threads in a single open-world pipeline: a CLIP-based binary detector (with optional DIRE-style residual fusion), post-hoc calibration, and abstention via thresholding and split conformal prediction. The focus is not only on raw detection accuracy but also on reliability under shifts, measured by calibration quality and selective-performance metrics on several benchmarks.

3 Method

Datasets. Our goal is to build an AI-image detector that remains reliable under distribution shift rather than only performing well on a single curated benchmark. To this end, we train on a mixed-domain dataset and evaluate on both in-distribution and out-of-distribution test sets.

Our primary training source is CommunityForensics-Small [1], which contains 278K generated images collected from 4,803 generator models, and paired with 278K real images in multiple shards. In our experiments, we use a 4,000-sample training shard together with a 4,000-sample `aiart` training shard, distributed uniformely per batch. The `aiart` set contains stylized paintings and AI-generated art, and is included to expose the detector to a modality missing from the original photorealistic training distribution. Example samples from `aiart` are shown in Figure below. This mixed-domain setting is important because after analyzing our earlier experiments, it showed that a detector trained only on Community Forensics failed badly on art-style imagery.

In our evaluation setting, we use Community Forensics and `aiart` as in-distribution test sets after applying the new mixed-domain training protocol. RAID [3] is used as an unseen-generator out-of-distribution benchmark consisting of adversarially challenging fake images. CIFAKE is used as a modality-shift benchmark. It contains low-resolution CIFAR-style real and fake images, upsampled to the CLIP input size, and tests whether features learned on higher-resolution photos and paintings transfer to a substantially different image regime.

Importantly, all evaluation sets are disjoint from both training and calibration data. We reserve a held-out 10% split from the training union for model selection, temperature scaling, and conformal quantile estimation.

Figure 1: Example samples from the `aiart` dataset.

Model Family Overview. All detectors in our system share the same CLIP-compatible 224×224 input pipeline and produce a binary prediction head with two logits: class 0 for real and class 1 for fake. Abstention is implemented only at inference time, after calibration, using either a fixed confidence threshold or split conformal prediction. This design keeps the underlying classification problem strictly binary while allowing the system to decline uncertain predictions when needed. In our work, we studied three architectures: (1) a frozen-CLIP baseline, (2) a CLIP with a lightweight DIRE-inspired residual fusion model, and (3) SGF-Net, a multi-branch forensic fusion model with learned per-image gating. Figure 2 presents these three architectures while each is discussed below.

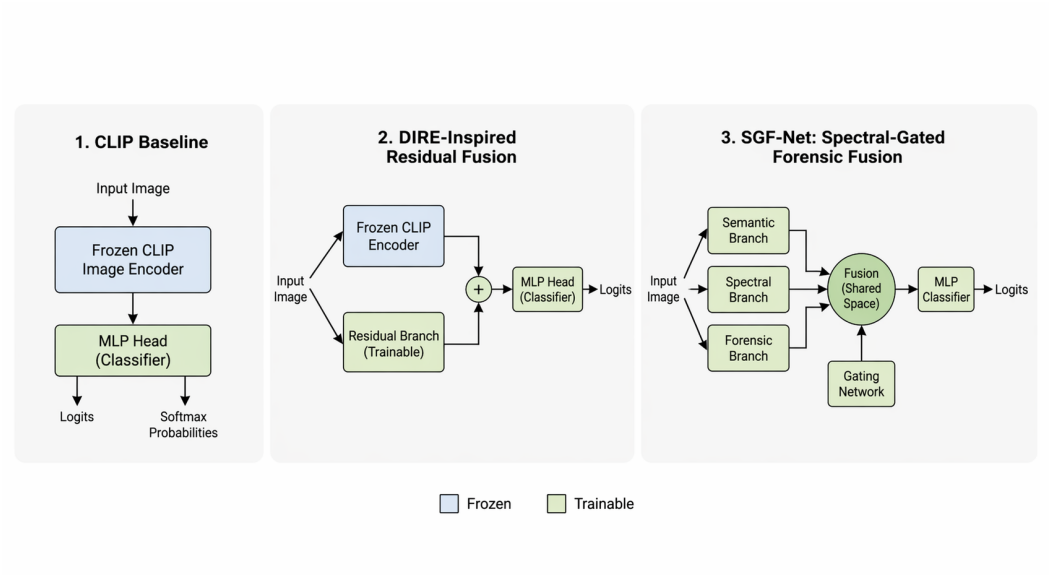

Figure 2: General overview of the three architectures we experimented with in our work.

CLIP Baseline Detector. Our baseline detector uses a pretrained OpenCLIP ViT-B/32 image encoder as a frozen feature extractor. For an input image $x \in \mathbb{R}^{3\times224\times224}$, the encoder produces a semantic embedding, $z_{\text{clip}} = f_{\text{CLIP}}(x) \in \mathbb{R}^{512}$. The CLIP backbone remains frozen throughout training. On top of z_{clip}, we train a lightweight a Multi-Layer Perceptron (MLP) classifier ($512 \to 512 \to 256 \to 2$), with ReLU activations and dropout ($p = 0.1$) between hidden layers. In this baseline, only the MLP head is trainable. Further, the model outputs the two logits, the corresponding softmax probabilities, and the feature representation used for downstream domain-adversarial training.

DIRE-Inspired Residual Fusion Detector. While CLIP provides strong semantic features, it may miss local high-frequency artifacts that help separate real and generated images. To complement the semantic branch, we introduce a lightweight residual branch inspired by DIRE [13]. Given an input image x, we first compute a smoothed approximation $\hat{x}$ using 3×3 average pooling form a residual map $R = |x - \hat{x}|$. This residual highlights local structure, edges, and fine-grained high-frequency content. Unlike the full DIRE method, which relies on diffusion reconstruction, our branch uses this low-cost proxy to retain some of the same inductive bias without incurring the computational cost of a diffusion pass.

The residual map is encoded by a small Convolution Neural Network (CNN) followed by adaptive average pooling and a linear layer to produce $z_{\text{res}} \in \mathbb{R}^{128}$. We concatenate the semantic and residual features, $[z_{\text{clip}}; z_{\text{res}}] \in \mathbb{R}^{640}$ and feed the fused vector into a two-layer MLP classifier. The CLIP encoder remains frozen while only the residual encoder and fusion head are trainable.

SGF-Net: Spectral-Gated Forensic Fusion. Our third model, SGF-Net, is a forensic fusion architecture that combines three complementary streams: a semantic CLIP branch, a spectral branch, and a pixel-forensic branch. Rather than concatenating all features with fixed weights, SGF-Net uses a learned per-image gating network to decide how much each branch should contribute.

The first branch is the same frozen CLIP encoder used in the baseline. The second branch operates on log-magnitude Fast Fourier Transform (FFT) spectrum to capture frequency-domain structure. The third branch uses Neighboring Pixel Relation (NPR) maps and fixed Steganalysis Rich Model (SRM) style high pass filters to emphasize pixel-level forensic cues. Each branch is projected into a shared fusion space, and a small gating MLP computes softmax weights over the three branches from cheap spectral and pixel summary statistics. The final fused representation is a convex combination of the projected branch embeddings. Further, to improve downstream abstention behavior, SGF-Net is trained with a compound objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{conf}} \mathcal{L}_{\text{conf}} \tag{1}$$

where $\mathcal{L}_{\text{conf}}$ encourages high confidence on correct predictions and lower confidence on incorrect ones. In our experiments, we assigned the value of 0.3 to λ_{conf}.

Training Pipeline. Our current training pipeline has three main components. First, Tier-1 bias-removal augmentation. Preliminary analysis showed that CommunityForensics-Small contains a JPEG-history shortcut. More specifically, real images often carry web JPEG artifacts, while many generated images do not. Therefore, a detector can exploit this shortcut instead of learning more general authenticity cues. To reduce this effect, we apply JPEG recompression, random resized crop, and mild color jitter during training. These augmentations disrupt compression-history and framing shortcuts while preserving task-relevant signal. Second, evenly mixed-domain training. Tier-1 augmentation removes spurious cues, but does not teach the model to recognize unseen modalities. We therefore train on a mixture of 50% of the data from Community Forensics and the rest from `aiart`. This makes `aiart` an in-distribution evaluation set rather than a pure Out-Of-Distribution (OOD) benchmark, but substantially improves robustness to stylistic variation. CIFAKE is then used as the new modality-shift frontier. Third, Domain-adversarial training (DAT). As an optional regularizer, we attach a domain-classification head to the detector features through a gradient reversal layer. The objective encourages feature representations that remain predictive for the real-vs-fake task while becoming less informative about the source domain (Community Forensics vs. `aiart`). We linearly ramp the gradient-reversal strength during training to avoid destabilizing early optimization.

Calibration and Abstention. Abstention in our system is not learned as a separate class. Instead, it is an inference-time mechanism applied to the calibrated binary logits. We first perform temperature scaling [6] on a held-out calibration split as shown in Equation 2.

$$T^* = \arg\min_T \sum_i -\log \text{softmax}(z_i/T)[y_i] \tag{2}$$

This improves confidence calibration without changing argmax predictions. We then support three evaluation modes. First, Forced, which always predicting the argmax class. Second, Threshold, which predicts only if $\max(\text{softmax}(z/T)) \geq \tau$, with $\tau = 0.9$. Finally, Conformal, where we use split conformal prediction [2] to produce a prediction set; abstain when the set contains both classes.

For conformal abstention, we compute nonconformity scores on the calibration split by first obtaining s_i as shown in Equation 3. After that, we estimate the $(1 - \alpha)$ quantile $\hat{q}$, and then output the

prediction set as demonstrated in Equation 4. For binary detection, $C(x) = \{0\}$ means predict real, $C(x) = \{1\}$ means predict fake, and $C(x) = \{0, 1\}$ means abstain since the model seems not sure what is the correct label. This gives a selective prediction mechanism with a coverage guarantee under the standard exchangeability assumption.

$$s_i = 1 - \text{softmax}(z_i/T)[y_i] \tag{3}$$

$$C(x) = \{c : \text{softmax}(z(x)/T)[c] \geq 1 - \hat{q}\} \tag{4}$$

Evaluation Metrics. The main success criterion of our work is that the system remains trustworthy under open-world shifts. The goal is to be accurate when the system gives an answer, and to abstain rather than make confident mistakes when the evidence is weak or degraded. Therefore, we evaluate every model under all three decision modes using several complementary metrics. Forced and selective accuracy measure classification quality with and without abstention. Coverage and abstention rate quantify how often the system commits to a prediction. Moreover, several metrics are to be used in order to evaluate our work. Area Under the Receiver Operating Characteristic curve (AUROC) and True Positive Rate (TPR) at 1% False Positive Rate (FPR) measure ranking quality in low-false-positive regimes. To address the limitations of accuracy alone, we also report class-conditional precision, recall, F1, TPR, True Negative Rate (TNR), FPR, and False Negative Rate (FNR) on the answered subset, together with reliability diagrams and Expected Calibration Error (ECE). Finally, we use risk coverage curves and AURC to summarize the reliability and coverage tradeoff for each abstention policy.

4 Results

Experimental Setup. We evaluate six setups across four datasets and three decision modes. The setups are the baseline CLIP detector, the DIRE-inspired fusion model, and SGF-Net, each trained with and without domain-adversarial training (DAT). The four evaluation datasets are Community Forensics, RAID, `aiart`, and CIFAKE. Every setup is evaluated in forced, threshold, and conformal modes in a single pass so that the reported numbers are directly comparable across abstention policies. Figure 3 visualizes conformal selective accuracy across all 24 run–dataset cells. The strongest overall row is `dire_mix_aug`, while the most striking failure mode is the SGF collapse on RAID.

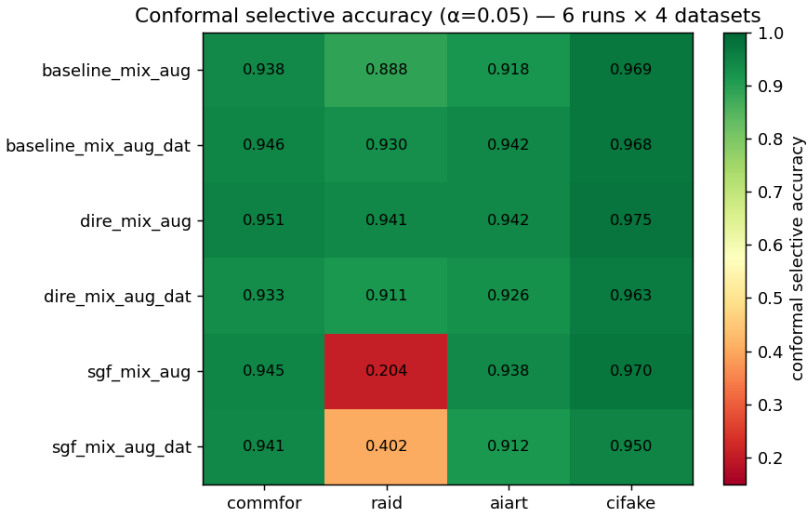

Figure 3: Conformal selective accuracy ($\alpha = 0.05$) across six runs and four evaluation datasets. DIRE is the most uniformly strong model; SGF exhibits a sharp failure on RAID.

Main Quantitative Results. Table 1 summarizes the main results where several trends are clear. First, all non-SGF models achieve strong in-distribution and mixed-domain performance. On Com-

Table 1: Main results across six runs, four datasets, and three abstention modes. Forced accuracy (`f_acc`) reports no-abstention performance. Threshold accuracy (`t_acc`) uses $\tau = 0.9$. Conformal accuracy (`c_acc`) uses $\alpha = 0.05$.

Run	Dataset	AUROC	f_acc	t_acc	t_abst	c_acc	c_abst
baseline_mix_aug	commfor	0.966	0.909	0.967	0.258	0.938	0.090
baseline_mix_aug	raid	–	0.844	0.937	0.342	0.888	0.131
baseline_mix_aug	aiart	0.933	0.868	0.966	0.554	0.918	0.191
baseline_mix_aug	cifake	0.975	0.918	0.994	0.565	0.969	0.177
baseline_mix_aug_dat	commfor	0.967	0.907	0.975	0.318	0.946	0.126
baseline_mix_aug_dat	raid	–	0.883	0.969	0.352	0.930	0.148
baseline_mix_aug_dat	aiart	0.943	0.882	0.976	0.623	0.942	0.224
baseline_mix_aug_dat	cifake	0.970	0.915	0.992	0.611	0.968	0.211
dire_mix_aug	commfor	0.969	0.903	0.978	0.314	0.951	0.149
dire_mix_aug	raid	–	0.880	0.973	0.385	0.941	0.185
dire_mix_aug	aiart	0.936	0.863	0.971	0.622	0.942	0.275
dire_mix_aug	cifake	0.971	0.909	0.996	0.638	0.975	0.256
dire_mix_aug_dat	commfor	0.965	0.907	0.969	0.251	0.933	0.077
dire_mix_aug_dat	raid	–	0.868	0.962	0.354	0.911	0.126
dire_mix_aug_dat	aiart	0.940	0.866	0.972	0.612	0.926	0.186
dire_mix_aug_dat	cifake	0.971	0.909	0.995	0.618	0.963	0.171
sgf_mix_aug	commfor	0.952	0.893	0.975	0.417	0.945	0.171
sgf_mix_aug	raid	–	0.250	0.149	0.407	0.204	0.196
sgf_mix_aug	aiart	0.944	0.876	0.985	0.677	0.938	0.260
sgf_mix_aug	cifake	0.969	0.905	0.994	0.639	0.970	0.239
sgf_mix_aug_dat	commfor	0.931	0.861	0.966	0.520	0.941	0.303
sgf_mix_aug_dat	raid	–	0.430	0.389	0.721	0.402	0.459
sgf_mix_aug_dat	aiart	0.869	0.792	0.982	0.830	0.912	0.532
sgf_mix_aug_dat	cifake	0.909	0.833	0.973	0.814	0.950	0.511

munity Forensics, forced accuracy lies between 0.903 and 0.909 for the baseline and DIRE variants, with AUROC between 0.965 and 0.969. On `aiart`, which had previously been a failure case, all baseline and DIRE variants now achieve AUROC between 0.933 and 0.943, confirming that Tier-1 augmentation and evenly mixed-domain training successfully addressed the earlier modality-shift collapse.

Second, abstention consistently improves selective accuracy. For example, `dire_mix_aug` improves from 0.903 forced accuracy to 0.978 threshold selective accuracy on Community Forensics, from 0.880 to 0.973 on RAID, from 0.863 to 0.971 on `aiart`, and from 0.909 to 0.996 on CIFAKE. The conformal rule is less aggressive than the fixed threshold and therefore typically sits between forced and threshold accuracy while providing a coverage guarantee.

Third, the best overall model is `dire_mix_aug`. It achieves the strongest or near-strongest conformal selective accuracy across all four datasets: 0.951 on Community Forensics, 0.941 on RAID, 0.942 on `aiart`, and 0.975 on CIFAKE. This suggests that a lightweight residual branch adds useful cross-domain structure beyond the CLIP semantic embedding.

Metrics Beyond Accuracy. Accuracy alone is insufficient for understanding detector behavior, especially once abstention is introduced. We therefore report class-conditional confusion rates, precision, recall, and F1 on the answered subset. Table 2 shows a representative example for `baseline_mix_aug` on Community Forensics.

In forced mode, the model achieves 0.909 accuracy, 0.911 precision, 0.903 recall, and 0.907 F1. Under threshold abstention, coverage drops to 0.742 but selective accuracy rises to 0.967, with precision 0.972 and recall 0.963. Under conformal abstention, coverage is higher at 0.910, and the

model still achieves 0.938 accuracy, 0.944 precision, 0.931 recall, and 0.938 F1. This confirms that the selective gains are not an artifact of a single metric. As precision, recall, and F1 all improve when the model is allowed to abstain on uncertain inputs.

Table 2: Representative precision, recall, and F1 breakdown for `baseline_mix_aug` on Community Forensics; class 1 is fake.

Mode	Coverage	Acc	Precision	Recall	F1	TNR
Forced	1.000	0.909	0.911	0.903	0.907	0.916
Threshold	0.742	0.967	0.972	0.963	0.967	0.971
Conformal	0.910	0.938	0.944	0.931	0.938	0.946

Calibration and Abstention Behavior. Temperature scaling materially improves confidence calibration without changing argmax predictions. Figure 4 shows the reliability diagram for `baseline_mix_aug` on Community Forensics before and after temperature scaling. ECE drops from 0.042 to 0.016, and the post-scaling bars track the diagonal much more closely.

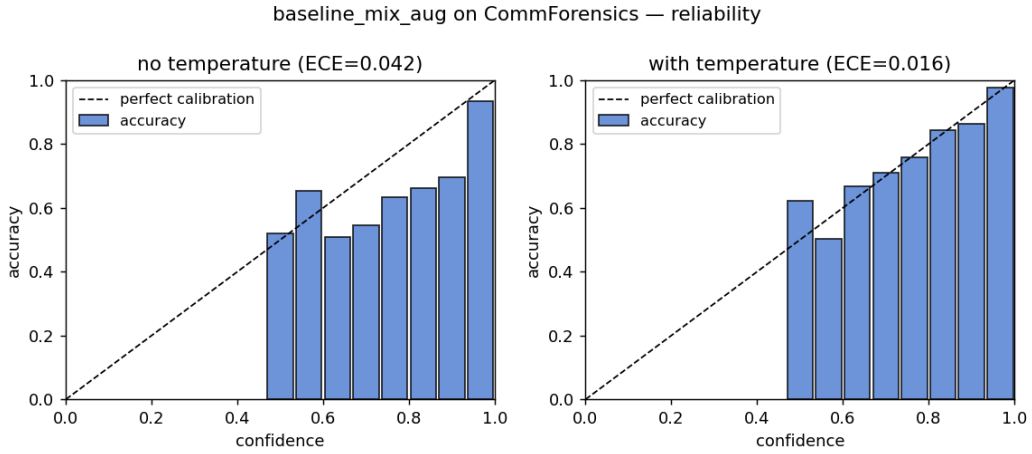

Figure 4: Reliability diagram for `baseline_mix_aug` on Community Forensics before and after temperature scaling. Temperature scaling reduces ECE from 0.042 to 0.016 without changing argmax accuracy.

Figure 5 compares forced, threshold, and conformal selective accuracy for `dire_mix_aug`. As expected, threshold abstention is the most aggressive and achieves the highest selective accuracy, while conformal abstention offers a middle point between forced prediction and fixed-threshold rejection. The key advantage of conformal abstention is not that it always maximizes accuracy, but that it provides a principled coverage guarantee rather than relying on a hand-tuned threshold.

Figure 6 shows the risk–coverage curves for `dire_mix_aug`. On `aiart` and CIFAKE, forced prediction starts with visibly higher risk, which falls rapidly as low-confidence samples are removed. On Community Forensics, the curve is flatter because the model is already relatively confident and accurate for in-distribution samples. These curves support the main claim of our work. Which states that abstention is most valuable exactly where the distribution is harder.

Comparison to Prior Work. Our models sit at three different points in the prior-work design space. The baseline is mostly comparable to the frozen-CLIP binary detectors demonstrated in UnivFD [9] and Community Forensics [11]. The DIRE-inspired model is closest to reconstruction or residual-augmented methods such as DIRE [13]. SGF-Net is our learned multi-branch fusion model, combining semantic, spectral, and pixel-forensic features with input-dependent gating.

Quantitatively, our best DIRE-inspired model lies in the same range as published CLIP-based detectors on CommunityForensics-style benchmarks. In particular, `dire_mix_aug` achieves 0.969 AUROC and 0.978 threshold selective accuracy on Community Forensics, while also performing strongly

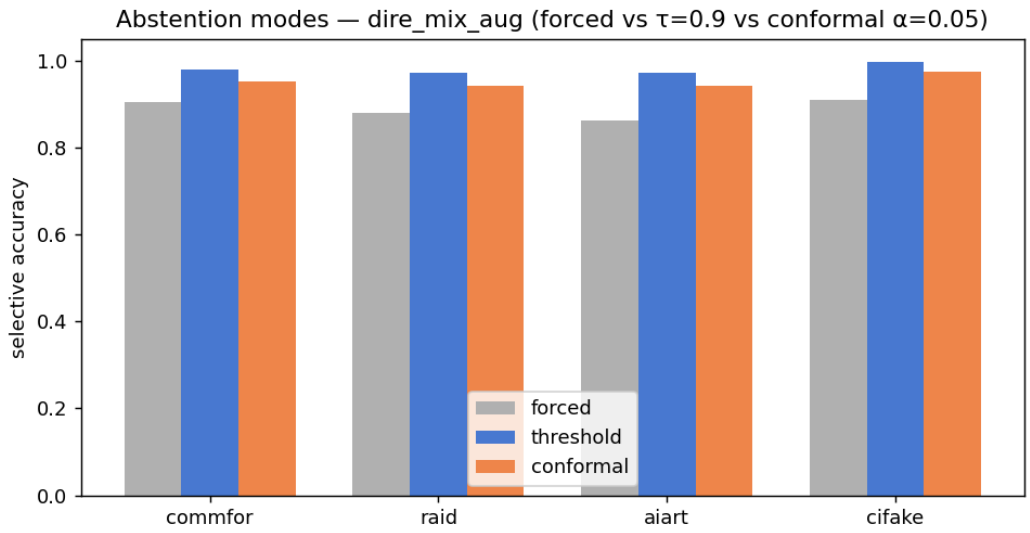

Figure 5: Selective accuracy for `dire_mix_aug` under forced, threshold ($\tau = 0.9$), and conformal ($\alpha = 0.05$) prediction. Threshold is more aggressive; conformal trades some selective accuracy for a coverage guarantee.

on RAID and CIFAKE. More importantly, our work adds a dimension that most prior detection papers do not report: calibrated selective prediction with conformal coverage guarantees. Rather than only reporting forced accuracy, we explicitly characterize what happens when the detector is allowed to abstain and measure the resulting reliability and coverage tradeoff. At the same time, this is not a strict fair comparison. Since we did not retrain external baselines such as UnivFD under our exact Tier-1, mixed-domain and conformal evaluation protocol. Therefore, we interpret prior-work comparison as contextual positioning rather than a direct leader board claim.

Key Findings. Our updated experiments from our milestone report, support five main conclusions. First, the earlier modality-shift failure is resolved. Previously, all models performed poorly on `aiart`. While after adding Tier-1 shortcut removal and mixed-domain training, every model clears 0.87 AUROC on `aiart`. Second, `dire_mix_aug` is the strongest overall single model. It is the most uniformly strong architecture across Community Forensics, RAID, `aiart`, and CIFAKE, suggesting that a lightweight residual branch is a better robustness investment than a more complex from-scratch forensic fusion network. Third, abstention improves reliability in a controlled way rather than acting as a crutch. Forced accuracy remains strong for all non-failed runs, while threshold and conformal abstention improve selective accuracy further. Conformal abstention is especially appealing because it regulates abstention through coverage guarantees rather than an arbitrary rejection threshold. Fourth, the SGF results reveal an important negative finding: handcrafted or from-scratch forensic branches can overfit to training-distribution artifacts and fail badly under unseen-generator shift. This is most visible on RAID, where SGF collapses while CLIP-based models remain strong. Finally, domain-adversarial training is architecture-dependent. It helps the generic baseline, but unnecessary or slightly harmful for DIRE and it further destabilizes SGF-Net.

5 Ablation Study

We performed ablations along three axes: architecture, domain-adversarial training (DAT), abstention and calibration setups.

Effect of Domain-Adversarial Training. To isolate the effect of DAT, we pair each `mix_aug` run with its corresponding `mix_aug_dat` run while holding Tier-1 augmentation and mixed-domain training fixed. We use conformal selective accuracy as the primary comparison metric because it is more stable across datasets than a single forced operating point. Figure 7 shows the marginal effect

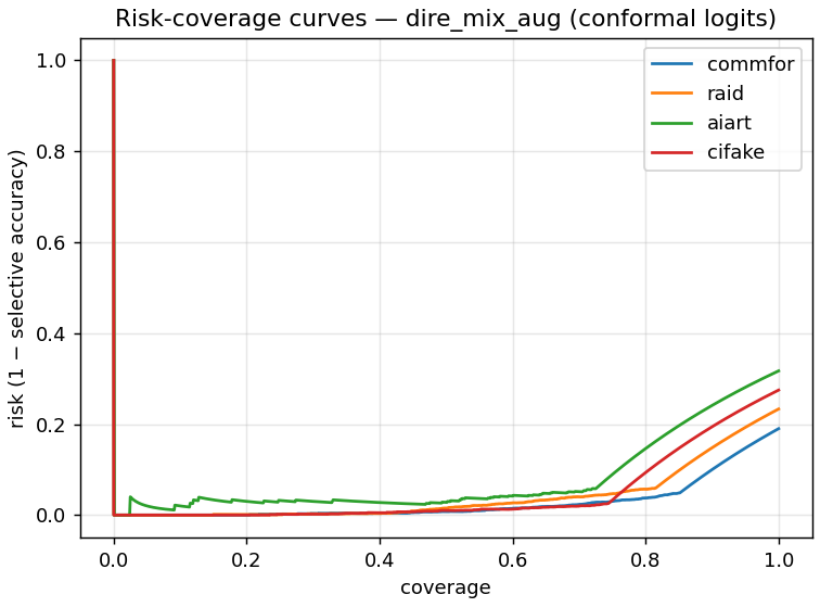

Figure 6: Risk–coverage curves for `dire_mix_aug`. Abstention yields the largest gains on harder distributions such as `aiart` and CIFAKE, while the in-distribution Community Forensics curve is already relatively flat.

of DAT on each architecture. DAT is clearly beneficial for the baseline CLIP detector, as it improves conformal accuracy by +0.008 on Community Forensics, +0.042 on RAID, and +0.024 on `aiart`, while leaving CIFAKE essentially unchanged (-0.001). This suggests that feature-level domain invariance is helpful when the model relies mainly on a generic semantic representation.

In contrast, DAT is neutral to negative for the DIRE-inspired model: -0.018 on Community Forensics, -0.030 on RAID, -0.016 on `aiart`, and -0.012 on CIFAKE. A likely explanation is that the residual branch already injects cross-domain structure, so additional domain-adversarial pressure removes useful task signal rather than improving invariance.

For SGF-Net, DAT is actively harmful. Although the RAID number increases from 0.204 to 0.402, the model remains far below every other architecture and still fails catastrophically. Moreover, SGF with DAT shows runaway abstention and degraded performance on the other datasets. We therefore interpret the apparent RAID gain as partial recovery from a collapsed run, not as evidence that DAT is beneficial for SGF.

Architecture Ablation. The architecture comparison itself is a central ablation. The baseline tests whether frozen CLIP semantics alone are sufficient. The DIRE-inspired model tests whether adding a lightweight residual branch improves robustness. SGF-Net tests whether richer spectral and pixel-forensic signals, combined by learned gating, outperform simpler fusion.

As demonstrated by the experiments, the lightweight DIRE-style fusion is the best overall tradeoff. It outperforms the baseline on the hardest datasets and avoids the catastrophic OOD failures of SGF. This results suggests that adding a small, targeted forensic signal to a strong pretrained semantic backbone is more robust than building a larger from-scratch forensic architecture around handcrafted branches.

Calibration Ablation. Temperature scaling produces a substantial improvement in confidence calibration. On Community Forensics, ECE for `baseline_mix_aug` falls from 0.042 to 0.016 after scaling, a 2.6$\times$ reduction, while argmax accuracy remains unchanged. This confirms that calibration is not merely cosmetic and it materially improves the quality of confidence estimates used by both threshold and conformal abstention.

The reliability diagrams in Figure 4 show that post-scaling confidence bins align much more closely with empirical accuracy. Since both threshold-based and conformal abstention operate on the

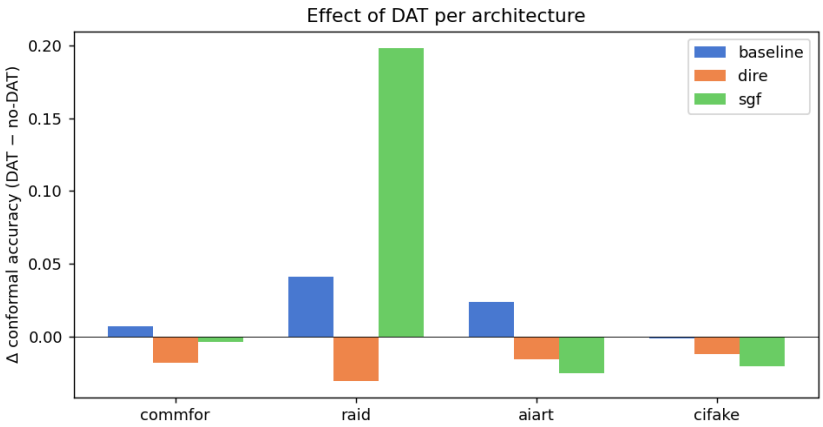

Figure 7: Effect of DAT on conformal selective accuracy. DAT helps the baseline, neutral-to-negative for DIRE, and harmful for SGF-Net.

calibrated confidence distribution, this calibration step is a necessary component of the full selective prediction pipeline.

Abstention Policy Ablation. We compare three decision rules: forced prediction, fixed-threshold abstention, and conformal abstention. This ablation addresses a key concern raised in the feedback of our milestone report as whether abstention simply lets the model avoid difficult cases. The results show that abstention is additive rather than substituting for a weak classifier. Forced accuracy remains strong across all successful runs, typically around 0.86–0.91 on the balanced datasets and 0.84–0.88 on RAID for the baseline and DIRE variants. Threshold abstention then improves selective accuracy further, often into the 0.97–0.99 range, at the cost of substantial abstention. Conformal abstention provides a less aggressive but more principled alternative, usually landing between forced and threshold performance.

For example, on Community Forensics, `dire_mix_aug` moves from 0.903 forced accuracy to 0.978 threshold selective accuracy and 0.951 conformal selective accuracy. On RAID, the same model moves from 0.880 forced accuracy to 0.973 threshold and 0.941 conformal. This demonstrates the intended tradeoff since threshold maximizes answered-sample accuracy, while conformal yields slightly lower selective accuracy but regulates abstention through coverage guarantees.

Negative Result: SGF Collapse Under OOD Shift. The strongest negative result in our work is the SGF failure on RAID. Despite competitive in-distribution or near-distribution performance on Community Forensics, `aiart`, and CIFAKE, `sgf_mix_aug` drops to 0.250 forced accuracy and 0.204 conformal selective accuracy on RAID. The DAT variant remains broken, achieving only 0.430 forced and 0.402 conformal accuracy despite much higher abstention.

The training-loss comparison in Figure 8 provides supporting evidence that SGF is harder to optimize than the baseline even before considering OOD generalization. More importantly, the evaluation pattern suggests that SGF's from-scratch forensic branches overfit to training-distribution artifacts. Unlike CLIP, these branches do not have pretrained semantic priors to fall back on when confronted with an unseen generator distribution.

This negative result is useful because it refines the design hypothesis of the report; not all additional forensic structure helps. In our experiments, pretrained semantics plus a lightweight residual cue generalize better than a larger from-scratch spectral and pixel fusion model.

Ablation Summary. Taken together, the ablations support three conclusions. First, the best-performing system is not the most complex one. The DIRE-inspired residual fusion model is more robust than both the plain baseline and SGF-Net. Second, calibration and abstention are essential parts of the system rather than optional post-processing, because they substantially improve reliability on hard distributions. Third, robustness interventions are architecture-dependent. As mentioned earlier, DAT helps the generic baseline but it is unnecessary for DIRE and harmful for SGF-Net.

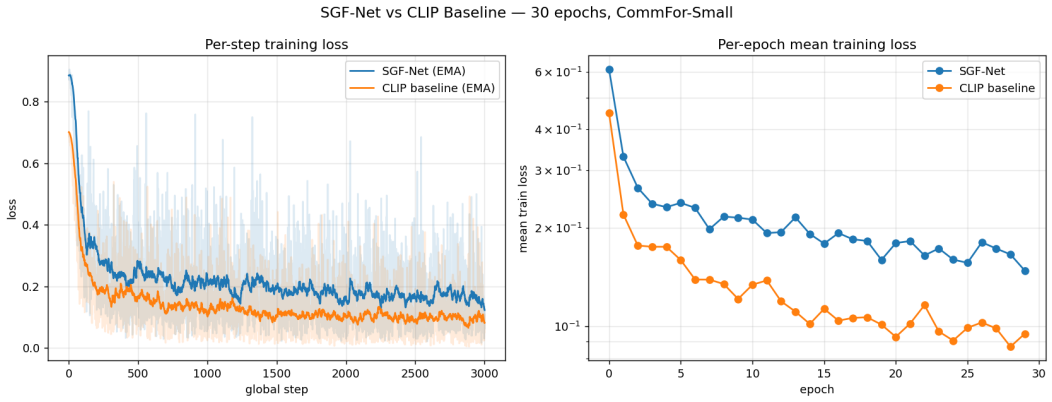

Figure 8: Training-loss comparison between SGF-Net and the CLIP baseline on CommunityForensics-Small. SGF remains harder to optimize and converges more slowly, which is consistent with its later instability under OOD evaluation.

6 Discussion & Conclusion

Discussion. Our final results show that strong AI-image detection performance is achievable with a relatively simple frozen-CLIP backbone, but that raw classification accuracy alone is not an adequate measure of trustworthiness under shift. Across the six-run matrix in Figure 3, the strongest overall model is the DIRE-inspired residual fusion detector, `dire_mix_aug`, which achieves the most uniform performance across Community Forensics, RAID, `aiart`, and CIFAKE. This indicates that augmenting pretrained semantic features with a lightweight residual forensic signal is a better robustness investment than either the plain semantic baseline or a larger from-scratch forensic fusion model.

A central finding of this work is that abstention improves reliability in a controlled and measurable way. Under both fixed-threshold and conformal decision rules, selective accuracy increases substantially over forced prediction, especially on harder distributions such as RAID, `aiart`, and CIFAKE. At the same time, this gain comes with a real coverage cost. The most aggressive threshold rule attains the highest answered-sample accuracy but abstains on a larger fraction of inputs, while conformal abstention provides a more conservative middle ground with a formal coverage guarantee. This confirms that abstention is useful not because it hides model weakness, but because it converts low-confidence errors into explicit uncertainty.

The calibration results further support this interpretation. Temperature scaling substantially improves confidence quality without changing argmax predictions, reducing ECE on the baseline from 0.042 to 0.016. Since both threshold-based and conformal abstention operate on calibrated confidence, this step is not cosmetic post-processing; it is a necessary part of making abstention behave sensibly.

Our ablation results also show that robustness interventions are architecture-dependent. Domain-adversarial training (DAT) improves the generic CLIP baseline, especially on cross-domain evaluation, but is neutral-to-negative for the DIRE-inspired model and actively harmful for SGF-Net. This suggests that when a model already contains a useful cross-domain residual branch, additional domain-invariance pressure may remove task-relevant structure rather than helping. More importantly, the SGF collapse on RAID reveals a meaningful negative resul. As from-scratch forensic branches can overfit to training-distribution artifacts and fail catastrophically under unseen-generator shift. Whereas pretrained semantic features provide a much stronger fallback.

Relative to prior work, our results place the baseline and DIRE-inspired models in the same general performance band as published CLIP-based and residual-augmented detectors on CommunityForensics-style benchmarks. The key difference is that our work does not stop at forced accuracy. Instead, we explicitly evaluate selective prediction under both threshold and split-conformal abstention, and report accuracy, calibration, coverage, and risk–coverage behavior together. To our knowledge, this selective-prediction perspective with conformal coverage guarantees is still largely

absent from the AI-image detection literature, and it constitutes the main practical contribution of this project.

Conclusion. Overall, the project demonstrates that open-world AI-image detection benefits from combining three ingredients: a strong pretrained semantic backbone, lightweight forensic augmentation, and calibrated abstention. Among the tested models, the DIRE-inspired residual fusion architecture provides the best overall robustness and accuracy tradeoff. More broadly, the results support a simple conclusion, in open-world detection, the goal should not be to force a prediction on every input, but to make accurate predictions when the model is confident and to abstain in a principled way when it is not.

References

[1] Andrew Owens Research Lab. OwensLab/CommunityForensics-Small dataset card. `https://huggingface.co/datasets/OwensLab/CommunityForensics-Small`, 2025. Accessed 2026-02-03.

[2] Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021.

[3] Hicham Eddoubi, Jonas Ricker, Federico Cocchi, Lorenzo Baraldi, Angelo Sotgiu, Maura Pintor, Marcella Cornia, Asja Fischer, Rita Cucchiara, and Battista Biggio. Raid: A dataset for testing the adversarial robustness of ai-generated image detectors, 2025.

[4] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[5] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks, 2017.

[6] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. arXiv:1706.04599.

[7] Nasrin Imanpour, Abhilekh Borah, Shashwat Bajpai, Subhankar Ghosh, Sainath Reddy Sankepally, Hasnat Md Abdullah, Nishoak Kosaraju, Shreyas Dixit, Ashhar Aziz, Shwetangshu Biswas, Vinija Jain, Aman Chadha, Song Wang, Amit Sheth, and Amitava Das. The visual counter turing test (vct^2): A benchmark for evaluating ai-generated image detection and the visual ai index (vai), 2024. Last revised 2025.

[8] Midjourney, Inc. Midjourney. `https://www.midjourney.com`, 2023.

[9] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24480–24489, 2023.

[10] OpenAI. Dall·e 3: Improving image generation with better captions. `https://openai.com/research/dall-e-3`, 2023.

[11] Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 8245–8257, June 2025. arXiv:2411.04125.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. arXiv:2103.00020.

[13] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. arXiv:2303.09295.