ASNET: ENERGY-CONDITIONED RELIABILITY FOR ROBUST AUDIO DEEPFAKE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio deepfake detectors that are accurate in-domain can fail under unseen generators, codec/channel shifts, and adversarial perturbations while remaining confident, making confidence-based rejection unreliable. Our key insight is that a detector's own confidence can serve two reliability roles: allocating robustness pressure across training samples and deciding which test inputs to reject. If this confidence drifts under codec or channel changes, however, it can misdirect training. We therefore introduce ASNet, which stabilizes a bounded, class-symmetric, detector-derived energy score before it controls training. Spectral and prosodic views provide complementary evidence; energy anchoring and OOD consistency stabilize the score's scale and limit its drift; Energy-Conditioned Robustness Modulation then sets per-sample robustness weights, and the same score is reused unchanged for validation-fixed abstention. Under single-source training and adaptation-free evaluation on four external corpora, ASNet raises external Robust AUC from 0.59 to 0.65 over a compute-matched baseline and yields lower source-domain ECE before post-hoc recalibration. The reused score rejects 85% of PGD-10 errors but only 48% under a matched-budget energy-aware attack, exposing the shared score as an attack surface.

1 INTRODUCTION

Modern speech synthesis produces highly realistic speech (van den Oord et al., 2016; Shen et al., 2017; Kong et al., 2020; 2021; Popov et al., 2021), raising risks of impersonation, fraud, and audio-driven misinformation (Yi et al., 2023; Li et al., 2025). For forensic and biometric use, high in-domain accuracy is insufficient: predictions must remain reliable under unseen generators, codecs, channels, and waveform perturbations, particularly when confidence also determines whether a prediction is accepted or rejected. Prior work typically addresses cross-domain generalization (Wang et al., 2019; Channing et al., 2024; Ren et al., 2024), adversarial robustness (Wu et al., 2024a;b), and calibration or selective prediction (Guo et al., 2017; Geifman & El-Yaniv, 2019) with separate objectives. Our key insight is that a detector's own confidence can serve two reliability roles at once: allocating robustness pressure across training samples and deciding which test inputs to reject. This dual use couples the two roles: a confidence scale that drifts under codec or channel transformations can weaken the very allocation it controls. We therefore study a shared **reliability interface**, in which one detector-derived score is stabilized before it controls robustness training and is then reused unchanged at deployment.

We introduce ASNet, organized around a bounded, class-symmetric energy score $E(x)$ and a **Construct–Stabilize–Use** design. **Construct:** spectral and prosodic evidence, which respond differently to codec distortion and bounded perturbations (Liu et al., 2023; Zeng et al., 2023), is combined through learned cross-view fusion and margin-capped Cross-Modal Robustness Alignment (CMRA) to discourage excessive redundancy. **Stabilize:** energy anchoring stabilizes the score's operating scale, while OOD consistency limits codec/channel-induced drift before the score controls robustness training. **Use:** Energy-Conditioned Robustness Modulation (ECRM) converts the stabilized score into sample-wise robustness weights, while the same $E(x)$ is reused unchanged for validation-fixed abstention. Our contribution is not a new uncertainty ranking or generic reweighting rule, but a **stabilize-before-control** dependency: one detector-derived score is stabilized before it allocates robustness training and then reused unchanged for abstention. We test this dependency with matched stabilization–allocation controls and an adaptive attacker that directly targets the deployed score; alternative train–deploy pairings remain outside the present claim.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We evaluate whether the resulting reliability profile can instead be explained by additional compute, model capacity, uniform robustness weighting, or simultaneous optimization. All protocol-matched models train only on LibriSeVoc and are evaluated without adaptation on the designated test splits of ASVspoof21, WaveFake, FakeAVCeleb, and CodecFake; no external corpus is used for training, validation, checkpoint or abstention-threshold selection, or corpus-specific retuning. Compute-matched baselines, targeted counterfactuals, held-out conditioning-signal comparisons, per-corpus robustness evaluation, and abstention-aware adaptive attacks test alternative explanations and expose the limits of the shared reliability interface.

Contributions. We introduce a **stabilize-before-control reliability design** in which a detector-derived signal is made stable before it controls robust training and is then reused unchanged at deployment. ASNet realizes this design through three stages:

1. **Construct—complementary evidence:** spectral and prosodic views are combined through learned cross-view fusion and CMRA; the resulting detector prediction defines the shared class-symmetric energy score.
2. **Stabilize—reliability before control:** energy anchoring stabilizes the score's operating scale, while OOD consistency limits its drift under codec and channel transformations before the score controls robustness training.
3. **Use—shared train–deploy reliability:** ECRM converts the stabilized score into stop-gradient, unit-mean sample weights for robustness regularization, while the same score is reused unchanged for validation-fixed abstention.

Compute-matched, component, and schedule controls test alternative explanations for the complete pipeline; cross-corpus, held-out-signal, and adaptive-attack evaluations characterize transfer and failure modes.

2 Related Work and Threat Model

Audio deepfake detection. Early countermeasures combined MFCC or CQCC features with GMM or i-vector classifiers (Wu et al., 2015; Tak et al., 2020). The ASVspoof challenges (Wang et al., 2020a; Delgado et al., 2021) motivated neural detectors including LCNN (Lavrentyeva et al., 2019), RawNet2 (Tak et al., 2021), and AASIST (weon Jung et al., 2021), while HuBERT (Hsu et al., 2021) and WavLM (Chen et al., 2022) strengthened learned representations. These approaches improve detector representations; ASNet addresses a different question: how one detector-derived score can be stabilized before controlling robustness allocation and then reused unchanged for abstention. Because the score is computed from the detector output, the formulation does not require a particular front-end; however, this work evaluates only the encoders in Sec. 3 and does not claim front-end portability.

Generalization under distribution shift. Prior work uses multi-domain training (Huang et al., 2025; Channing et al., 2024), codec and channel augmentation (Liu et al., 2023; Zeng et al., 2023), contrastive learning (Wu et al., 2024b), augmentation-based domain generalization (Ren et al., 2024), and domain-mix training in RawNetLite (Pierno et al., 2025). ASNet instead uses codec and channel transformations to stabilize the energy score used by the reliability interface. Augmentation-matched baselines in Tab. 2 control for the additional augmentation exposure this introduces.

Adversarial robustness in audio. Speech systems remain vulnerable to gradient-based and transfer perturbations (Carlini & Wagner, 2018; Qin et al., 2019; Schönherr et al., 2018). Prior defenses use adversarial training, feature regularization, and modified objectives (Wu et al., 2024a;b; Hussain et al., 2021; Zhang et al., 2019). Instance-reweighted methods such as MART (Wang et al., 2020b) and GAIRAT (Zhang et al., 2021) allocate robustness pressure using true-class confidence or attack-derived geometry. ASNet does not claim instance reweighting itself as novel or superiority to these weighting families; it instead stabilizes a label-free, class-symmetric detector score before using it for robustness allocation and reuses the same score unchanged for inference-time abstention. Tab. 3 uses uniform weighting as a matched counterfactual to isolate signal-conditioned allocation within ASNet.

Calibration, OOD detection, and selective prediction. OOD detection (Liang et al., 2020; Lee et al., 2018), calibration (Guo et al., 2017), energy scoring (Liu et al., 2021), and selective classification (Geifman & El-Yaniv, 2019) provide uncertainty signals for trained models. ASNet instead reuses one detector-derived score for training-time robustness allocation and inference-time absten-

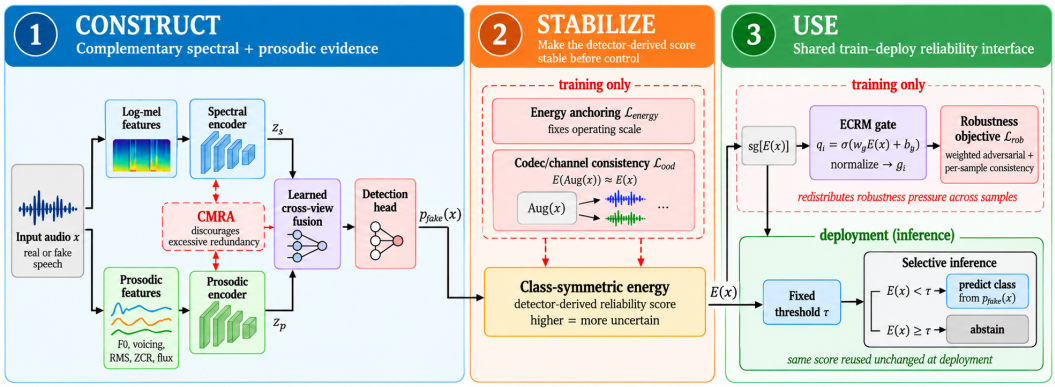


Figure 1: **ASNet architecture and shared reliability interface.** Spectral and prosodic evidence is fused under training-only CMRA. The resulting class-symmetric energy score $E(x)$ is stabilized by energy anchoring and OOD consistency, then used by ECRM to allocate robustness pressure. At deployment, the same $E(x)$ is reused unchanged with a validation-fixed threshold for prediction or abstention.

tion. ECE is reported before post-hoc recalibration for every method; matched retraining in App. E.8 compares energy with margin, entropy, Mahalanobis distance, and a learned confidence head.

2.1 Threat Model

Deployed system. Fig. 1 separates training from deployment. ECRM is used only during training; deployment retains the encoders, fusion module, detection head, and deterministic energy score. For an input x', the system computes $p_{\text{fake}}(x')$ and $E(x')$ and either returns a binary prediction or abstains using a validation-fixed threshold τ (Eq. 10).

Adversary classes. We consider digital inference-time adversaries that modify test audio without changing training data, labels, or model parameters. Depending on the evaluation, perturbation attacks use white-box model access or black-box transfer, and abstention-aware attacks additionally know the deterministic energy-based rejection rule. A **domain-shift adversary** changes the synthesis pipeline, vocoder, codec, digital post-processing chain, or transmission encoding, inducing conditions absent from training (Liu et al., 2023; Zeng et al., 2023). A **perturbation adversary** applies untargeted ℓ_∞-bounded waveform perturbations (Carlini & Wagner, 2018; Qin et al., 2019).

Attack budgets and scope. Training uses waveform-level PGD (Madry et al., 2019; Zhang et al., 2019) with $\epsilon_{\text{train}} = 2 \times 10^{-3}$. The primary comparative metric uses PGD-10 at $\epsilon = 10^{-3}$, while additional matched classification stress tests use APGD-CE (Croce & Hein, 2020), CW (Carlini & Wagner, 2017), and BPDA-APGD (Athalye et al., 2018) (App. C.8). These finite-budget attacks support only empirical comparisons, not attack-convergence claims. A separate 100-iteration, multi-trajectory adaptive evaluation targets both classification and the deployed energy-based rejection rule and characterizes selective-system failure rather than baseline classifier ranking (App. D.2). Training-time poisoning, backdoors, model stealing, physical playback, and unbounded adaptive attacks are outside scope.

3 Method

Problem formulation and design. Given an input waveform x with label $y \in \{0, 1\}$, where $y = 1$ denotes fake speech, the detector produces a spoof probability $p_{\text{fake}}(x)$. Our goal is not to introduce a separate confidence model, but to derive one score from this detector and use it in two roles: allocating robustness regularization during training and deciding when to abstain at inference. This motivates stabilizing the score before it is used as a training-time controller. ASNet implements this dependency through three stages. **Construct:** spectral and prosodic encoders, learned cross-view fusion, and CMRA form complementary evidence for the detector. **Stabilize:** energy anchoring and OOD consistency stabilize the detector-derived energy score under codec and channel

transformations before it is used as a training-time controller. **Use:** ECRM reads the stabilized score to redistribute robustness regularization across samples, while the same score is reused unchanged for validation-fixed abstention at inference. Notation and acronyms are summarized in Apps. B.1 and A.

3.1 CONSTRUCT: SPECTRAL–PROSODIC EVIDENCE

Dual evidence streams. A detector that relies on a single spectral representation may depend heavily on localized artifacts that change under codec distortion or bounded waveform perturbations, whereas slowly varying prosodic trajectories can be affected differently. ASNet therefore represents each waveform through spectral and prosodic views. The spectral encoder f_s operates on log-mel features. The prosodic encoder f_p receives five frame-level sequences: fundamental frequency (f_0), voicing, RMS energy, zero-crossing rate, and spectral flux, temporally aligned and stacked channel-wise to form $\mathbf{X}_{\text{pro}} \in \mathbb{R}^{5\times L}$:

$$\mathbf{z}_s = f_s(\mathbf{X}_{\text{mel}}) \in \mathbb{R}^d, \qquad \mathbf{z}_p = f_p(\mathbf{X}_{\text{pro}}) \in \mathbb{R}^d. \tag{1}$$

The prosodic stream is not treated as an independently stronger detector; its role is to provide a second evidence source that complements the spectral representation after fusion. Feature extraction and encoder details are given in Apps. B.2–B.3.

Cross-view fusion. The two embeddings are combined by a learned gate,

$$\mathbf{z}^* = \text{LayerNorm}(\text{MLP}([\mathbf{z}_s; \mathbf{z}_p]) \odot \sigma(W_f[\mathbf{z}_s; \mathbf{z}_p] + \mathbf{b}_f)), \tag{2}$$

where $[\mathbf{z}_s; \mathbf{z}_p]$ denotes concatenation and $\odot$ element-wise multiplication. This fusion gate operates on representations and is distinct from the ECRM gate introduced below, which weights training losses. The fused representation $\mathbf{z}^*$ is passed to the binary detection head and to training-only auxiliary heads (App. B.4).

Cross-Modal Robustness Alignment. Learned fusion alone does not prevent the two streams from encoding highly similar representations. We therefore use Cross-Modal Robustness Alignment (CMRA), a margin-capped cosine penalty,

$$\mathcal{L}_{\text{cmra}} = \mathbb{E}_x\big[\max\big(0, \cos(\mathbf{z}_s, \mathbf{z}_p) - m\big)\big]. \tag{3}$$

The penalty activates only when cross-view similarity exceeds the margin m, so CMRA is designed to discourage excessive redundancy without forcing the embeddings to be artificially dissimilar. The CMRA margin is selected once on clean validation data, while cross-domain controls for structured cross-view interaction are reported in App. E.5.

3.2 STABILIZE: THE DETECTOR-DERIVED ENERGY SCORE

Class-symmetric energy score. The detection head produces $p_{\text{fake}}(x) = \sigma(f_\theta(\mathbf{z}^*))$, from which we define

$$E(x) = -\log(\max\{p_{\text{fake}}(x),\, 1 - p_{\text{fake}}(x)\}). \tag{4}$$

The score is class-symmetric and bounded, $E(x) \in [0, \log 2]$: it is low for confident predictions of either class and maximal at the decision boundary. For the binary sigmoid head, $E(x) = \text{softplus}(-|f_\theta(\mathbf{z}^*)|) = -\log \text{MSP}$; hence energy, binary predictive entropy, and logit margin induce equivalent uncertainty rankings on a fixed detector (up to confidence direction). We therefore do not claim a distinct post-hoc ranking. The methodological contribution is the score's controlled use: stabilize this bounded quantity before it determines sample-wise robustness weights, then reuse the identical score for abstention. Unlike log-sum-exp energy for OOD detection (Liu et al., 2021), this adds no separate inference model.

Energy anchoring and OOD consistency. Before $E(x)$ controls robustness regularization, we stabilize it with

$$\mathcal{L}_{\text{energy}} = \mathbb{E}_{x\sim\mathcal{D}_{\text{train}}}\left[(E(x) - \tau_{\text{id}})^2\right], \tag{5}$$

$$\mathcal{L}_{\text{ood}} = \mathbb{E}_{x'\sim\text{Aug}(x)}\left[(E(x') - E(x))^2\right], \tag{6}$$

where τ_{id} is estimated from correctly classified validation samples and $\text{Aug}(\cdot)$ applies codec and channel transformations to LibriSeVoc training audio (App. B.5); no external evaluation corpus is

used. $\mathcal{L}_{\text{energy}}$ softly anchors the energy/confidence scale while remaining jointly optimized with the detection objectives; $\mathcal{L}_{\text{ood}}$ limits transformation-induced score drift.

Why stabilize before control. $\mathcal{L}_{\text{ood}}$ directly bounds the transformation sensitivity of the raw ECRM gate. For $q(E) = \sigma(w_g E + b_g)$, since $|\sigma'(u)| \le 1/4$,

$$\mathbb{E}\Big[\big(q(E(x')) - q(E(x))\big)^2\Big] \;\le\; \frac{w_g^2}{16}\,\mathcal{L}_{\text{ood}}, \qquad x' \sim \text{Aug}(x). \tag{7}$$

Reducing $\mathcal{L}_{\text{ood}}$ limits raw-gate sensitivity through Eq. 7; App. B.6 extends this bound to the normalized ECRM weights g_i (Eq. 15), up to a finite gate-dependent constant.

3.3 Use: Energy-Conditioned Robustness and Selective Inference

ECRM turns the stabilized detector-derived energy into a sample-wise robustness controller. For a mini-batch of size B,

$$q_i = \sigma(w_g\,\text{sg}[E(x_i)] + b_g)\,, \qquad g_i = \frac{q_i}{\frac{1}{B}\sum_{j=1}^{B} q_j}, \tag{8}$$

where $\text{sg}[\cdot]$ denotes stop-gradient. Thus, $\frac{1}{B}\sum_i g_i = 1$: ECRM redistributes, rather than globally changes, robustness pressure. Stop-gradient blocks gradients from g_i through $E(x_i)$, while the weighted $\mathcal{L}_{\text{ood}}^{\text{gate}}$ term remains differentiable through its own loss path. The resulting robustness objective is

$$\mathcal{L}_{\text{rob}} = \frac{1}{B}\sum_{i=1}^{B} g_i\big(\lambda_{\text{adv}}\mathcal{L}_{\text{adv}}(x_i) + \lambda_{\text{ood}}\mathcal{L}_{\text{ood}}^{\text{gate}}(x_i)\big)\,, \tag{9}$$

where $\mathcal{L}_{\text{adv}}$ penalizes clean–adversarial logit drift and $\mathcal{L}_{\text{ood}}^{\text{gate}}(x) = (E(\text{Aug}(x)) - E(x))^2$ is the per-sample consistency term. The control $g_i \equiv 1$ therefore removes energy-conditioned allocation while preserving the same average robustness weight. Full objective and optimization details are given in App. B.6.

Staged schedule. The security-specific objectives are activated as

$$\mathcal{L}_{\text{energy}} \;\longrightarrow\; \mathcal{L}_{\text{ood}} \;\longrightarrow\; \mathcal{L}_{\text{rob}},$$

at 20%, 40%, and 60% of training. Anchoring first stabilizes the energy scale, OOD consistency then limits codec/channel-induced score drift, and only afterward does ECRM use that score to allocate robustness pressure. The complete training objective and algorithm are provided in Apps. B.6–B.7.

Selective inference. At deployment, ECRM and all training-only objectives are removed. The same energy score determines whether the detector predicts or abstains:

$$E(x) < \tau \Rightarrow \hat{y} = \mathbb{1}\{p_{\text{fake}}(x) \ge 0.5\}, \qquad E(x) \ge \tau \Rightarrow \text{abstain}. \tag{10}$$

The threshold τ is selected once on clean LibriSeVoc validation and held fixed at evaluation. Thus, the score stabilized and consumed during training is the same quantity reused for selective inference.

4 Experimental Setup

Protocol and datasets. All gradient-bearing mini-batches in the protocol-matched experiments contain only samples from the LibriSeVoc training split (Sun et al., 2023). For ASNet, LibriSeVoc validation determines τ_{id}, τ, m, and checkpoint selection; its waveforms never enter gradient-bearing training. The $20/40/60\%$ schedule and unit method-specific coefficients are fixed before test evaluation, and the sensitivity studies in Apps. B.9–B.10 are post-hoc checks, not tuning procedures. Evaluation is performed without adaptation on the held-out LibriSeVoc test set and four external datasets: ASVspoof21 LA (Delgado et al., 2021), WaveFake (Frank & Schönherr, 2021), FakeAVCeleb (audio) (Khalid et al., 2022), and CodecFake (Xie et al., 2024). We use the designated evaluation/test split of each external dataset; for CodecFake, this is the complete official test release (C1–C7 and A1–A3). No external dataset contributes to training, validation, checkpoint selection, abstention-threshold selection, adaptation, or corpus-specific retuning. Dataset statistics, preprocessing, and evaluation details are given in Apps. C.1–C.3.

Metrics. We report detection AUC (↑) and EER (↓), calibration error ECE before post-hoc recalibration (↓), adversarial Robust AUC, and OOD-AUROC. Corpus-level Robust AUC is the detection AUC under waveform-level PGD-10 at $\epsilon = 10^{-3}$; External Robust AUC is its unweighted mean over the four external corpora. Mechanism ablations instead report LibriSeVoc Robust AUC, and we keep these source-domain and external quantities distinct. OOD-AUROC is a directional clean–shift rank-separation diagnostic, with the shifted condition as the positive class. Values above or below 0.5 indicate opposite score directions; neither direction nor a larger value is interpreted as better. Because OOD-AUROC is rank-based rather than a paired score-drift measure, it is not used to assess stabilization or unseen-OOD detection. Stabilization is evaluated separately by the paired stage diagnostics in App. B.8. For selective prediction, each evaluated model selects its abstention threshold using only its own clean LibriSeVoc validation scores to target 20% abstention, after which that threshold is frozen for test evaluation. For ASNet, this procedure gives $\tau = 0.22$, which is then used unchanged across all test corpora and attacks. No test corpus is used for threshold selection or retuning. Retained-set discrimination is reported as a supporting diagnostic, while error catch characterizes ASNet's fixed rejection rule. Selective and attack-specific metrics are defined in App. C.7. Main comparisons use three seeds and are interpreted descriptively rather than as formal significance tests.

Baselines and fairness controls. We compare with LCNN (Lavrentyeva et al., 2019), RawNet2 (Tak et al., 2021), AASIST (weon Jung et al., 2021), RawNetLite (Pierno et al., 2025), and Ren et al. (Ren et al., 2024). Default-protocol baselines use the same LibriSeVoc-only data, preprocessing, crop length, optimizer family, training budget, and evaluation protocol; configurations and native reimplementation checks are reported in Apps. C.5–C.6. To control for the additional codec and channel exposure introduced by OOD consistency, RawNetLite and Ren et al. are also retrained with the same security augmentations (`SecAug`), while Ren et al. is additionally evaluated with adversarial training. Because ASNet requires approximately $1.4\times$ the training compute, we also compare against Ren et al.$^{\texttt{SecAug+Adv}}$ at the same cumulative A100 GPU-hour budget, changing only its training horizon and corresponding cosine-decay schedule (Tab. 1).

Mechanism and robustness controls. Tab. 3 provides matched Construct, Stabilize, and Use controls, including uniform instead of signal-conditioned robustness allocation; Tab. 22 transfers these robustness controls across corpora. To test energy against alternative signal-conditioned allocations, App. E.8 retrains the same framework with logit margin, entropy, Mahalanobis distance, and a learned confidence head while holding the remaining design choices fixed. Beyond the primary PGD-10 comparison, we report matched classification-focused APGD-CE and CW stress tests, plus black-box transfer. Separately, a stronger abstention-aware evaluation jointly targets classification error and energy suppression at $\epsilon = 2\times10^{-3}$. The former support finite-budget classifier comparisons; the latter stress-tests the deployed selective rule rather than baseline ranking (Apps. C.8–D.2).

5 Results

5.1 Compute-Controlled Cross-Corpus Reliability

Table 1: **Compute-matched comparison.** Compute is normalized to the default Ren et al.$^{\texttt{SecAug+Adv}}$ run. External AUC and External Robust AUC are unweighted means over ASVspoof21, WaveFake, FakeAVCeleb, and CodecFake. ECE and OOD-AUROC are evaluated on LibriSeVoc; OOD-AUROC is reported descriptively and is not used for ranking. Compute-matched rows report mean±SD over the same three seeds.

Method	Compute	External AUC↑	External Robust AUC↑	LibriSeVoc ECE↓	LibriSeVoc OOD-AUROC
Ren et al.$^{\texttt{SecAug+Adv}}$	1.0×	0.836	0.580	0.160	0.680
Ren et al.$^{\texttt{SecAug+Adv}}$ (compute-matched)	1.4×	0.838±0.004	0.591±0.007	0.148±0.005	0.693±0.006
ASNet (ours)	1.4×	**0.881±0.003**	**0.650±0.020**	**0.101±0.004**	0.740±0.030

The compute-matched comparison separates additional training budget from the proposed reliability interface. Extending Ren et al.$^{\texttt{SecAug+Adv}}$ from $1.0\times$ to $1.4\times$ compute changes External Robust AUC from 0.580 to 0.591, whereas at the same $1.4\times$ cumulative A100 GPU-hour budget ASNet reaches 0.650 External Robust AUC, 0.881 External AUC, and 0.101 LibriSeVoc ECE (Tab. 1). Relative to the compute-matched control, these differences are $+0.059$, $+0.043$, and -0.047, respectively. Although energy anchoring directly regularizes confidence magnitude, the matched Uniform control

retains stabilization without reproducing Full ASNet's ECE (App. E.2); we therefore treat ECE as a full-system training outcome measured before post-hoc recalibration for all methods.

Table 2: **Single-source cross-corpus evaluation.** All models train only on LibriSeVoc and are evaluated without adaptation. Corpus entries report AUC/EER. External Robust AUC is the unweighted PGD-10 mean over ASVspoof21, WaveFake, FakeAVCeleb, and CodecFake. ECE and OOD-AUROC are source-domain LibriSeVoc diagnostics; OOD-AUROC denotes clean-vs.-mp3-64 kbps separation as defined in Sec. 4. Reported values are three-seed means, with mean±SD shown for Robust AUC and OOD-AUROC. Bold indicates the best result only for metrics with a preferred direction.

Model	LibriSeVoc	ASVspoof21	WaveFake	FakeAVCeleb	CodecFake	LibriSeVoc ECE↓	External Robust AUC↑	LibriSeVoc OOD-AUROC
LCNN (Lavrentyeva et al., 2019)	0.946/0.087	0.742/0.287	0.804/0.236	0.731/0.291	0.552/0.433	0.26	0.40±0.01	0.50±0.01
RawNet2 (Tak et al., 2021)	0.981/0.043	0.763/0.273	0.901/0.173	0.791/0.242	0.602/0.392	0.22	0.45±0.01	0.62±0.01
AASIST (weon Jung et al., 2021)	0.986/0.035	0.781/0.252	0.912/0.162	0.812/0.224	0.621/0.371	0.20	0.48±0.01	0.59±0.01
RawNetLite (Pierno et al., 2025)	0.993/0.028	0.821/0.203	0.893/0.188	0.833/0.191	0.613/0.381	0.19	0.50±0.01	0.61±0.01
Ren et al. (Ren et al., 2024)	**0.999**/0.005	0.842/0.184	0.931/0.154	0.861/0.170	0.664/0.341	0.16	0.53±0.01	0.66±0.01
Matched security-augmentation and adversarial-training baselines								
RawNetLite $^{\texttt{SecAug}}$ (Pierno et al., 2025)	0.994/0.025	0.829/0.198	0.901/0.178	0.841/0.183	0.621/0.374	0.18	0.52±0.01	0.63±0.01
Ren et al.$^{\texttt{SecAug}}$ (Ren et al., 2024)	0.997/0.011	0.851/0.178	0.938/0.148	0.869/0.163	0.672/0.333	0.15	0.55±0.01	0.67±0.01
Ren et al.$^{\texttt{SecAug+Adv}}$ (Ren et al., 2024)	0.998/0.007	0.855/0.176	0.940/0.146	0.872/0.161	0.676/0.329	0.16	0.58±0.01	0.68±0.01
ASNet (ours)	**0.999/0.002**	**0.892/0.158**	**0.971/0.124**	**0.922/0.142**	**0.739/0.292**	**0.10**	**0.65±0.02**	0.74±0.03

Tab. 2 shows that the reliability improvements do not come at the expense of clean detection performance. ASNet retains near-ceiling performance on the held-out LibriSeVoc test set (AUC 0.999, EER 0.002) and obtains the highest reported AUC and lowest EER among the protocol-matched methods on each of the four external corpora. The largest remaining degradation occurs on CodecFake (AUC 0.739, EER 0.292), making it the most difficult external dataset under the present single-source evaluation. The robustness mechanisms also retain their relative effect beyond the source corpus. Under PGD-10, Full ASNet obtains Robust AUC values of 0.69, 0.71, 0.64, and 0.56 on ASVspoof21, WaveFake, FakeAVCeleb, and CodecFake, respectively, for an external mean of 0.650. Although these values are lower than the LibriSeVoc Robust AUC of 0.90, Full ASNet exceeds the No-$\mathcal{L}_{\text{rob}}$ and Uniform controls on every evaluated external corpus, with mean differences of $+0.068$ and $+0.040$, respectively (Tab. 22). This pattern indicates that the robustness block and energy-conditioned allocation retain measurable benefits under cross-corpus transfer, while the absolute values also expose the remaining robustness gap under distribution shift. Cross-domain removals of fusion/CMRA, energy stabilization, and adversarial/ECRM training are reported in App. E.5, and results under multi-domain training, which lies outside the controlled protocol, appear in App. D.5.

5.2 Score Reuse for Abstention and Digital Attack Robustness

Each model selects its abstention threshold only on clean LibriSeVoc validation and freezes it before test evaluation. Retained-set discrimination is near ceiling for ASNet and matched Ren controls (App. D.1). Within the ASNet framework, separately trained score-replacement variants isolate the rejection signal: at 78.9–80.5% realized LibriSeVoc coverage, energy catches 80.1% of PGD-10 errors versus 65.1–73.9% for margin, entropy, Mahalanobis distance, and a learned confidence head (App. E.8; reruns trained separately from the final model). The final ASNet threshold, selected once on clean LibriSeVoc validation, is reused without retuning and rejects 81% of clean LibriSeVoc errors, 66% of CodecFake errors, and 85% under PGD-10 (Fig. 3), showing useful but incomplete transfer.

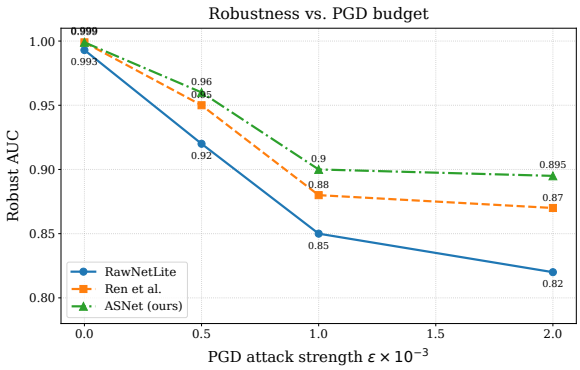


Figure 2: **PGD robustness on LibriSeVoc.** Robust AUC under fixed 10-step PGD for ASNet, the compute-matched Ren et al.$^{\texttt{SecAug+Adv}}$ control, and RawNetLite; $\epsilon = 0$ denotes clean AUC.

Classifier robustness under finite-budget attacks. Fig. 2 gives the fixed 10-step PGD sweep. At $\epsilon = 10^{-3}$, ASNet reaches Robust AUC 0.90, versus 0.88 for the compute-matched Ren et

al.$^{\texttt{SecAug+Adv}}$ control and 0.85 for RawNetLite; at 2×10^{-3}, the values are 0.895, 0.87, and 0.82. Matched APGD-CE, CW, and BPDA-APGD results are reported in Tab. 15.

Robustness of the rejection rule. Robustness of the classifier does not imply robustness of the rejection mechanism once the adversary targets the deployed energy score. Under energy-aware PGD, error catch falls to 48%. A stronger joint APGD-CE and energy-suppression attack at $\epsilon = 2 \times 10^{-3}$ reduces catch to 35% and increases UASR to 31.2% (Tab. 14). Because this evaluation also changes the perturbation budget, search length, and number of trajectories, the reduction reflects the stronger joint adaptive evaluation rather than energy suppression alone. Detector-derived energy is therefore a useful rejection signal but also an attack surface: the fixed-threshold rule degrades substantially under informed adaptive attack. These results support empirical error rejection under the evaluated conditions, not an adaptive-robustness guarantee.

5.3 Ablation Study

Tab. 3 reports matched controls under a common LibriSeVoc protocol. Construct controls test capacity and bare dual-view alternatives; Stabilize removes energy stabilization; Use replaces energy-conditioned allocation with uniform weighting; and the final control uses simultaneous rather than staged optimization. These are the controls used for mechanism attribution. The cost-oriented Lite/Mid variants in App. E.7 jointly change mechanism set and model cost, so cross-table numerical orderings are not interpreted as component effects.

Table 3: **Mechanism ablation on LibriSeVoc.** The first three groups provide matched controls for Construct, Stabilize, and Use; the final group tests staged optimization. All rows use the same optimizer, training budget, data split, and evaluation protocol. Values are three-seed means; the completed stabilization–allocation 2×2 is reported in App. E.2. "Energy stab." denotes $\mathcal{L}_{\text{energy}} + \mathcal{L}_{\text{ood}}$. ✓ = active; ✗ = removed or replaced; — = not applicable. †The bare dual-view control excludes CMRA.

Variant	Design elements				AUC↑	LibriSeVoc Robust AUC↑	OOD-AUROC	ECE↓
	Dual views	Energy stab.	ECRM	Staged schedule				
Construct								
Capacity-matched spectral-only	✗	✓	✓	✓	0.975	0.77	0.61	0.21
Bare dual-view detector	✓†	✗	✗	—	0.953	0.71	0.55	0.24
Stabilize								
w/o energy stabilization	✓	✗	✓	—	0.986	0.82	0.64	0.20
Use								
Uniform robustness, $g_i \equiv 1$	✓	✓	✗	✓	0.983	0.83	0.66	0.19
Staged schedule								
All-at-once training	✓	✓	✓	✗	0.994	0.85	0.69	0.15
Full ASNet	✓	✓	✓	✓	**0.999**	**0.90**	0.74	**0.10**

Stabilize–Use dependency. A matched 2×2 control crosses stabilization with conditioned versus uniform allocation (App. E.2). Conditioning improves Robust AUC by 0.07 with stabilization and 0.04 without it; the resulting descriptive interaction is $+0.03$. Thus, stabilization strengthens the observed benefit of conditioning without being required for that benefit. The instantaneous $1.4\times$ gate span is a per-step quantity; ECRM acts throughout the final 40% of training and weights both $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{ood}}^{\text{gate}}$.

Construct. The capacity-matched spectral-only and bare dual-view controls reach Robust AUC 0.77 and 0.71, respectively, versus 0.90 for Full ASNet. They rule out parameter count and a bare second view as sufficient explanations, but do not attribute the gain to any Construct subcomponent in isolation. Grouped cross-domain controls and a per-vocoder breakdown are reported in Apps. E.5 and D.3.

Staged schedule. All-at-once optimization retains high clean AUC (0.994) but reaches Robust AUC/ECE $0.85/0.15$, versus $0.90/0.10$ for the proposed schedule. Because objectives remain active after introduction, changing their sequence also changes cumulative exposure. We therefore interpret this as evidence for the complete staged schedule, not an order-only intervention. OOD-AUROC is reported only as a shift-separation diagnostic and is not used to establish schedule superiority or score invariance. Full schedule results and the joint clean–robustness comparison are reported in App. E.3 and Fig. 6 (App. E.6), respectively.

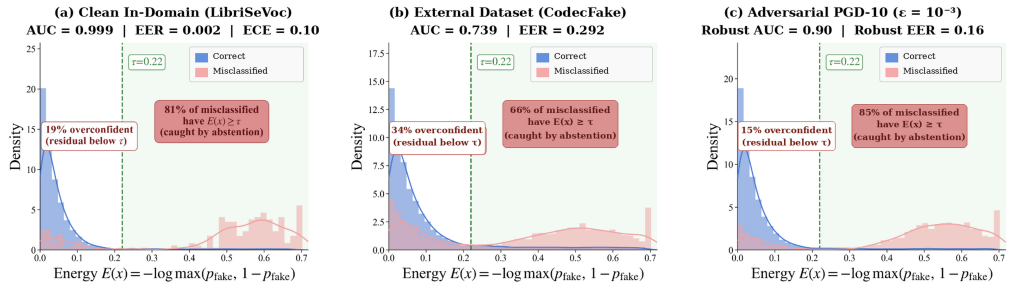


Figure 3: **Inference-time reuse of the shared energy score.** Distributions of detector-derived energy $E(x)$ for correct and misclassified predictions under (a) clean LibriSeVoc, (b) the external CodecFake test set, and (c) LibriSeVoc under PGD-10. The threshold $\tau = 0.22$ is selected once on clean LibriSeVoc validation and applied unchanged. It rejects 81%, 66%, and 85% of errors, respectively. Residual errors below the threshold expose the remaining limitation of confidence-derived abstention.

Inference-time reuse. Fig. 3 completes the shared-interface analysis at deployment. Across clean LibriSeVoc, CodecFake, and adversarial LibriSeVoc PGD-10, errors shift toward higher energy under the same validation-fixed threshold, supporting score reuse without condition-specific retuning while exposing incomplete separation, particularly on CodecFake.

Alternative conditioning controls. Matched held-out variants replace energy with margin, entropy, Mahalanobis, or a learned confidence head under the same framework. Although energy, margin, and entropy are post-hoc rank-equivalent, ECRM uses score magnitude, so retraining yields different normalized weights. Energy improves Robust AUC by 0.02 over the strongest non-energy alternative on both LibriSeVoc and ASVspoof21, with positive paired-seed differences throughout. Thus, energy is strongest among the tested rules at the complete-variant level; we do not attribute this advantage to the learned gate alone.

6 Discussion and Conclusion

Under the evaluated single-source protocol, the results support ASNet's central design claim: stabilize a detector-derived score before it controls robustness allocation, then reuse that score unchanged for abstention. At matched compute, ASNet raises External Robust AUC from 0.59 to 0.65 and lowers source-domain ECE from 0.15 to 0.10 before post-hoc recalibration; additional training budget alone does not reproduce this gain. The completed 2×2 shows a larger conditioning benefit after stabilization ($+0.07$ vs. $+0.04$), a descriptive end-to-end stabilization–allocation interaction not attributed to any single gradient path. Matched held-out variants favor energy over four alternative conditioning signals in Robust AUC, ECE, and error catch, while the validation-fixed score transfers across clean, shifted, and adversarial conditions without retuning. Under the strongest abstention-aware evaluation, error catch falls to 35% and UASR reaches 31.2%, exposing the confidence-derived score as an adaptive attack surface. Accordingly, the reused score provides useful but incomplete error rejection under the evaluated digital shifts, not an adaptive-robustness guarantee.

Limitations and future work. Scope. Main claims use single-source LibriSeVoc training, the encoders in Sec. 3, and the digital inference-time threat model; physical playback/replay, telephony, poisoning, backdoors, model stealing, self-supervised front-ends, and broader multilingual settings remain outside scope. Classifier robustness is evaluated under finite attack budgets rather than as a convergence guarantee. **Evidence.** Absolute external robustness remains limited, most notably on CodecFake (Robust AUC 0.56), and 15–34% of errors remain below the fixed abstention threshold. Selective comparisons use matched score-replacement variants within the ASNet framework at validation-targeted coverage; separately trained selective architectures and alternative train–deploy signal pairings are not evaluated. The tested weighting rules are not exhaustive, ECE is reported before post-hoc recalibration, schedule comparisons jointly vary order and exposure, and three-seed differences are interpreted descriptively. **Cost.** ASNet requires approximately $1.4\times$ the training compute of the default Ren et al. control; runtime and parameter counts are reported in App. D.6. Future work should test broader weighting rules, front-ends, and reliability signals less coupled to detector confidence.

Reproducibility Statement

The single-source protocol, datasets, baselines, and attack budgets are specified in Sec. 4 and Sec. 2.1. Feature extraction and encoder architectures are given in Apps. B.2–B.3; the complete objective, gate formulation, and staged training algorithm in Apps. B.6–B.7; dataset splits and preprocessing in Apps. C.1–C.2; optimizer settings and loss coefficients in App. C.4; the hyperparameter audit in App. B.10; baseline configurations and reimplementation checks in Apps. C.5–C.6; metric definitions in App. C.7; and attack protocols in Apps. C.8 and D.2. All main comparisons use seeds $\{0, 1, 2\}$. Anonymized code is available at `https://github.com/aiai-9/ASNet`.

Ethics Statement

This work studies defensive detection of synthetic and adversarially perturbed speech. No new human-subject data were collected; experiments use existing research datasets. Adversarial evaluations are included to characterize detector and abstention failure modes under bounded digital perturbations rather than to facilitate impersonation or evasion of deployed systems. We report these failure modes and limitations to support more reliable evaluation of audio-forensic systems.

Statement on AI Use

Generative AI tools were used to assist with manuscript organization, language editing, readability, and critical feedback on the presentation of the methodology and experiments. All experimental design decisions, implementations, reported numerical results, mathematical statements, citations, and scientific claims were reviewed and verified by the authors. The authors take full responsibility for the content of the submission.

References

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples, 2018. URL https://arxiv.org/abs/1802.00420.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks, 2017. URL https://arxiv.org/abs/1608.04644.

Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text, 2018. URL https://arxiv.org/abs/1801.01944.

Georgia Channing, Juil Sock, Ronald Clark, Philip Torr, and Christian Schroeder de Witt. Toward robust real-world audio deepfake detection: Closing the explainability gap, 2024. URL https://arxiv.org/abs/2410.07436.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, October 2022. ISSN 1941-0484. doi: 10.1109/jstsp.2022.3188113. URL http://dx.doi.org/10.1109/JSTSP.2022.3188113.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks, 2020. URL https://arxiv.org/abs/2003.01690.

Héctor Delgado, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, Xuechen Liu, Andreas Nautsch, Jose Patino, Md Sahidullah, Massimiliano Todisco, Xin Wang, and Junichi Yamagishi. Asvspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan, 2021. URL https://arxiv.org/abs/2109.00535.

Joel Frank and Lea Schönherr. Wavefake: A data set to facilitate audio deepfake detection, 2021. URL https://arxiv.org/abs/2111.02813.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option, 2019. URL https://arxiv.org/abs/1901.09192.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *ArXiv*, abs/1706.04599, 2017. URL https://api.semanticscholar.org/CorpusID:28671436.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021. URL https://arxiv.org/abs/2106.07447.

Wen Huang, Yanmei Gu, Zhiming Wang, Huijia Zhu, and Yanmin Qian. Generalizable audio deepfake detection via latent space refinement and augmentation, 2025. URL https://arxiv.org/abs/2501.14240.

Shehzeen Samarah Hussain, Paarth Neekhara, Shlomo Dubnov, Julian McAuley, and Farinaz Koushanfar. Waveguard: Understanding and mitigating audio adversarial examples. In *USENIX Security Symposium*, 2021. URL https://api.semanticscholar.org/CorpusID:232134992.

Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset, 2022. URL https://arxiv.org/abs/2108.05080.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *ArXiv*, abs/2010.05646, 2020. URL https://api.semanticscholar.org/CorpusID:222291664.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis, 2021. URL https://arxiv.org/abs/2009.09761.

Galina Lavrentyeva, Sergey Novoselov, Andzhukaev Tseren, Marina Volkova, Artem Gorlanov, and Alexandr Kozlov. Stc antispoofing systems for the asvspoof2019 challenge, 2019. URL https://arxiv.org/abs/1904.05576.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks, 2018. URL https://arxiv.org/abs/1807.03888.

Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. A survey on speech deepfake detection. *ACM Computing Surveys*, 57(7):1–38, 2025.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks, 2020. URL https://arxiv.org/abs/1706.02690.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection, 2021. URL https://arxiv.org/abs/2010.03759.

Xuechen Liu, Xin Wang, Md Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas Evans, Andreas Nautsch, and Kong Aik Lee. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2507–2522, 2023. ISSN 2329-9304. doi: 10.1109/taslp.2023.3285283. URL http://dx.doi.org/10.1109/TASLP.2023.3285283.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019. URL https://arxiv.org/abs/1706.06083.

Andrea Di Pierno, Luca Guarnera, Dario Allegra, and Sebastiano Battiato. End-to-end audio deepfake detection from raw waveforms: a rawnet-based approach with cross-dataset evaluation, 2025. URL https://arxiv.org/abs/2504.20923.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech, 2021. URL https://arxiv.org/abs/2105.06337.

Yao Qin, Nicholas Carlini, Ian Goodfellow, Garrison Cottrell, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition, 2019. URL https://arxiv.org/abs/1903.10346.

Hainan Ren, Li Lin, Chun-Hao Liu, Xin Wang, and Shu Hu. Improving generalization for ai-synthesized voice detection, 2024. URL https://arxiv.org/abs/2412.19279.

Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding, 2018. URL https://arxiv.org/abs/1808.05665.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Z. Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, 2017. URL https://api.semanticscholar.org/CorpusID:206742911.

Chengzhe Sun, Shan Jia, Shuwei Hou, and Siwei Lyu. Ai-synthesized voice detection using neural vocoder artifacts, 2023. URL https://arxiv.org/abs/2304.13085.

Hemlata Tak, Jose Patino, Andreas Nautsch, Nicholas Evans, and Massimiliano Todisco. An explainability study of the constant q cepstral coefficient spoofing countermeasure for automatic speaker verification, 2020. URL https://arxiv.org/abs/2004.06422.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2, 2021. URL https://arxiv.org/abs/2011.01108.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio, 2016. URL https://arxiv.org/abs/1609.03499.

Hongji Wang, Heinrich Dinkel, Shuai Wang, Yanmin Qian, and Kai Yu. Cross-Domain Replay Spoofing Attack Detection Using Domain Adversarial Training. In *Interspeech 2019*, pp. 2938–2942, 2019. doi: 10.21437/Interspeech.2019-2120.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Hector Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sebastien Le Maguer, Markus Becker, Fergus Henderson, Rob Clark, Yu Zhang, Quan Wang, Ye Jia, Kai Onuma, Koji Mushika, Takashi Kaneda, Yuan Jiang, Li-Juan Liu, Yi-Chiao Wu, Wen-Chin Huang, Tomoki Toda, Kou Tanaka, Hirokazu Kameoka, Ingmar Steiner, Driss Matrouf, Jean-Francois Bonastre, Avashna Govender, Srikanth Ronanki, Jing-Xuan Zhang, and Zhen-Hua Ling. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech, 2020a. URL https://arxiv.org/abs/1911.01601.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020b. URL https://api.semanticscholar.org/CorpusID:211548864.

Jee weon Jung, Hee-Soo Heo, Hemlata Tak, Hye jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks, 2021. URL https://arxiv.org/abs/2110.01200.

Haibin Wu, Xu Li, Andy T. Liu, Zhiyong Wu, Helen Meng, and Hung yi Lee. Improving the adversarial robustness for speaker verification by self-supervised learning, 2024a. URL https://arxiv.org/abs/2106.00273.

Haolin Wu, Jing Chen, Ruiying Du, Cong Wu, Kun He, Xingcan Shang, Hao Ren, and Guowen Xu. Clad: Robust audio deepfake detection against manipulation attacks with contrastive learning, 2024b. URL https://arxiv.org/abs/2404.15854.

Zhizheng Wu, Tomi H. Kinnunen, Nicholas W. D. Evans, Junichi Yamagishi, Cemal Hanilçi, Md. Sahidullah, and Aleksandr Sizov. Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Interspeech*, 2015. URL https://api.semanticscholar.org/CorpusID:11206540.

Yuankun Xie, Yi Lu, Ruibo Fu, Zhengqi Wen, Zhiyong Wang, Jianhua Tao, Xin Qi, Xiaopeng Wang, Yukun Liu, Haonan Cheng, Long Ye, and Yi Sun. The codecfake dataset and countermeasures for the universally detection of deepfake audio, 2024. URL https://arxiv.org/abs/2405.04880.

Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. Audio deepfake detection: A survey. *arXiv preprint arXiv:2308.14970*, 2023.

Chang Zeng, Xin Wang, Xiaoxiao Miao, Erica Cooper, and Junichi Yamagishi. Improving generalization ability of countermeasures for new mismatch scenario by combining multiple advanced regularization terms, 2023. URL https://arxiv.org/abs/2305.10940.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy, 2019. URL https://arxiv.org/abs/1901.08573.

Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training, 2021. URL https://arxiv.org/abs/2010.01736.

A ACRONYMS

This section summarizes the acronyms and shorthand notation used throughout the main paper and appendix.

Table 4: List of frequently used acronyms and shorthand terms.

Acronym / shorthand	Meaning
ASNet	Audio Security Network (this paper)
CMRA	Cross-Modal Robustness Alignment
ECRM	Energy-Conditioned Robustness Modulation
`SecAug`	Security Augmentation
OOD	Out-of-Distribution
w/o	without
AUC	Area Under the ROC Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
OOD-AUROC	Clean-vs.-mp3-64 kbps separation AUROC using the evaluated uncertainty signal
EER	Equal Error Rate
ECE	Expected Calibration Error
Robust AUC	AUC under adversarial perturbation evaluation
Robust EER	EER under adversarial perturbation evaluation
PGD	Projected Gradient Descent
CW	Carlini–Wagner attack
APGD-CE	Auto-PGD with cross-entropy loss
LibriSeVoc	LibriSeVoc dataset
ASVspoof21	ASVspoof 2021 Logical Access evaluation corpus
WaveFake	WaveFake corpus
FakeAVCeleb	FakeAVCeleb audio corpus
CodecFake	External deepfake-audio corpus with neural-codec and audio-language-model conditions
FakeOrReal	FakeOrReal dataset
LCNN	Light Convolutional Neural Network
RawNet2	RawNet2 detector
AASIST	Audio Anti-Spoofing using Integrated Spectro-Temporal graph attention
RawNetLite	Lightweight RawNet-style detector
WavLM	WavLM self-supervised speech encoder
XLS-R	Cross-Lingual Speech Representation model
MFCC	Mel-Frequency Cepstral Coefficients
CQCC	Constant-Q Cepstral Coefficients
GMM	Gaussian Mixture Model
STFT	Short-Time Fourier Transform
TCN	Temporal Convolutional Network
SE	Squeeze–Excitation
RMS	Root-Mean-Square energy
ZCR	Zero-Crossing Rate
RAPT	Robust Algorithm for Pitch Tracking

B ADDITIONAL METHOD DETAILS

This section provides the implementation details underlying Sec. 3. We follow the same Construct–Stabilize–Use dependency: spectral and prosodic representations construct the evidence used by the detector; energy anchoring and OOD consistency stabilize the detector-derived energy score; and ECRM uses that score to allocate robustness regularization. We specify feature extraction, encoder architectures, auxiliary objectives, energy stabilization, adversarial training, the ECRM gate, and the complete staged training procedure.

B.1 NOTATION

Tab. 5 summarizes the notation used in Sec. 3 and throughout this appendix.

B.2 FEATURE EXTRACTION

Given a waveform $x \in \mathbb{R}^T$ sampled at 16 kHz, ASNet constructs the spectral and prosodic inputs used by the two evidence streams:

$$\mathbf{X}_{\text{mel}} = \text{LogMel}(\text{STFT}(x)) \in \mathbb{R}^{F \times L}, \qquad \mathbf{X}_{\text{pro}} = [f_0, v, \text{RMS}, \text{ZCR}, \text{Flux}] \in \mathbb{R}^{5 \times L}.$$

Table 5: **Notation used in Sec. 3.**

Symbol	Meaning
$x \in \mathbb{R}^T$	Input waveform of length T
$y \in \{0, 1\}$	Binary label, where 1 denotes fake speech
$\mathbf{X}_{\text{mel}}$	Log-mel spectral feature map
$\mathbf{X}_{\text{pro}}$	Prosodic feature map containing f_0, voicing, RMS, ZCR, and flux
f_s, f_p	Spectral and prosodic encoders
$\mathbf{z}_s, \mathbf{z}_p \in \mathbb{R}^d$	Spectral and prosodic embeddings
$\mathbf{z}^*$	Fused spectral–prosodic representation
$W_f, \mathbf{b}_f$	Cross-view fusion-gate parameters
$f_\theta(\cdot)$	Binary detection logit head
$p_{\text{fake}}(x)$	Predicted probability that x is fake
$E(x)$	Class-symmetric energy score
τ_{id}	In-distribution energy anchor estimated from validation data
τ	Inference-time abstention threshold
$\text{Aug}(\cdot)$	Codec and channel transformation operator used for OOD consistency
q_i, g_i	Raw and unit-mean-normalized ECRM weights for sample i, used only during training
$\text{sg}[\cdot]$	Stop-gradient operator applied to the ECRM gate input
w_g, b_g	Learned ECRM gate parameters
x_{adv}	PGD-perturbed waveform
$\epsilon_{\text{train}}, \epsilon_{\text{eval}}$	Training and evaluation PGD budgets
$\mathcal{L}_{\text{det}}$	Binary detection loss
$\mathcal{L}_{\text{dom}}$	Domain/vocoder classification loss
$\mathcal{L}_{\text{con}}$	Supervised contrastive loss
$\mathcal{L}_{\text{cmra}}$	Cross-Modal Robustness Alignment loss
$\mathcal{L}_{\text{energy}}$	Energy anchoring loss
$\mathcal{L}_{\text{ood}}$	OOD consistency loss
$\mathcal{L}_{\text{ood}}^{\text{gate}}$	Per-sample OOD-consistency term inside ECRM
$\mathcal{L}_{\text{adv}}$	Adversarial logit-consistency loss
$\mathcal{L}_{\text{rob}}$	Energy-conditioned robustness loss
$\mathcal{L}_{\text{base}}$	$\mathcal{L}_{\text{det}} + \lambda_{\text{dom}}\mathcal{L}_{\text{dom}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}$
$\mathcal{L}_{\text{total}}$	Complete stage-wise training objective
$\mathbb{1}_{\text{eng}}, \mathbb{1}_{\text{ood}}, \mathbb{1}_{\text{adv}}$	Stage indicators for energy, OOD-consistency, and robustness phases

We use $F = 80$ mel bins with a 25 ms analysis window and 10 ms hop. The prosodic representation contains fundamental frequency f_0, a voiced/unvoiced indicator v, root-mean-square energy, zero-crossing rate, and spectral flux. Fundamental frequency is estimated using RAPT. Missing values in unvoiced frames are linearly interpolated and masked so that these regions do not introduce spurious positive pairs for the contrastive head. All features are mean–variance normalized per utterance.

B.3 Encoder Architectures

Spectral encoder. The spectral stream combines a convolutional front-end, frequency attention, a four-block dilated temporal convolutional network, and attentive pooling:

$$\mathbf{z}_s = \text{AttnPool}\big(\text{TCN}_{[1,2,4,8]}(\text{FreqAttn}(\text{Conv2D}(\mathbf{X}_{\text{mel}})))\big)\,.$$

The TCN uses kernel size 3 and dilations $[1, 2, 4, 8]$. Frequency attention uses a channel-wise squeeze over the frequency axis followed by a sigmoid gate.

Prosodic encoder. The prosodic stream uses a one-dimensional convolutional front-end followed by a six-block dilated TCN, squeeze–excitation, and attentive pooling:

$$\mathbf{z}_p = \text{AttnPool}\big(\text{SE}\big(\text{TCN}_{[1,2,4,8,16,32]}(\text{Conv1D}(\mathbf{X}_{\text{pro}}))\big)\big)\,.$$

Its dilations are $[1, 2, 4, 8, 16, 32]$, giving the prosodic stream a wider temporal receptive field than the spectral stream. Both encoders output $d = 256$ dimensional representations and use ReLU activations, layer normalization, and dropout $p = 0.1$. The resulting embeddings are combined by the learned fusion mechanism in Eq. 2.

B.4 Auxiliary Training Heads

During training, the fused representation $\mathbf{z}^*$ feeds two auxiliary heads included in $\mathcal{L}_{\text{base}}$. The domain head predicts the vocoder or synthesis condition using cross-entropy loss $\mathcal{L}_{\text{dom}}$. The supervised contrastive head maps $\mathbf{z}_i^*$ to $\mathbf{u}_i = g_{\text{con}}(\mathbf{z}_i^*) \in \mathbb{R}^{128}$ using a two-layer MLP. With temperature $\tau_c = 0.07$,

$$\mathcal{L}_{\text{con}} = \sum_i \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\text{sim}(\mathbf{u}_i, \mathbf{u}_p)/\tau_c)}{\sum_{a \in A(i)} \exp(\text{sim}(\mathbf{u}_i, \mathbf{u}_a)/\tau_c)}.$$

Here $P(i)$ contains examples sharing the same speaker–utterance identity but differing in vocoder or channel, and $A(i)$ contains the remaining batch examples. This objective acts across samples and synthesis conditions, whereas CMRA in Eq. 3 acts within each sample to regulate redundancy between the spectral and prosodic representations.

B.5 Energy Stabilization Details

Energy anchor. At the beginning of the energy-stabilization phase, τ_{id} is set to the median energy of correctly classified LibriSeVoc validation samples and is refreshed every ten epochs thereafter. Only the resulting scalar anchor enters $\mathcal{L}_{\text{energy}}$; validation waveforms are not included in gradient-bearing training mini-batches. The anchor is therefore estimated only after the initial base-training phase rather than from the randomly initialized detector.

OOD-consistency augmentations. The transformation distribution $\text{Aug}(\cdot)$ in Eq. 6 contains mp3 compression at $\{64, 96, 128, 192\}$ kbps, 7 kHz low-pass filtering, and convolution with measured room impulse responses. These transformations are applied only to LibriSeVoc training samples; external evaluation datasets are not used by the energy-anchoring or OOD-consistency objectives.

The two stabilization objectives play distinct roles. $\mathcal{L}_{\text{energy}}$ anchors the absolute in-distribution scale of $E(x)$, whereas $\mathcal{L}_{\text{ood}}$ constrains changes in that score under the specified codec and channel transformations. Their individual effects are evaluated in App. E.1.

B.6 Adversarial Perturbation and ECRM Details

Training-time PGD. Adversarial examples are generated with untargeted waveform-level ℓ_∞ PGD-2 using $\epsilon_{\text{train}} = 2 \times 10^{-3}$ and $\alpha = \epsilon/4$:

$$x^{(k+1)} = \Pi_{B_\epsilon(x)} \left(x^{(k)} + \alpha \, \text{sign}\Big(\nabla_{x^{(k)}} \mathcal{L}_{\text{det}}(x^{(k)}) \Big) \right).$$

The main test-time evaluation instead uses PGD-10 with $\epsilon_{\text{eval}} = 10^{-3}$ and step size 2.5×10^{-4}. Both attacks perturb waveforms before feature extraction; standard PGD backpropagates through the differentiable path while treating RAPT, voicing, and ZCR as non-differentiable. BPDA-APGD separately stress-tests these operators using backward surrogates (App. C.8).

Adversarial logit consistency. For the final adversarial example x_{adv}, the robustness objective penalizes the change in the detector logit:

$$\mathcal{L}_{\text{adv}}(x) = \| f_\theta(\mathbf{z}^*(x)) - f_\theta(\mathbf{z}^*(x_{\text{adv}})) \|_2^2 .$$

This term is used only during training and does not introduce an additional inference branch.

ECRM gate optimization. The gate parameters are initialized as $(w_g, b_g) = (1.0, 0.0)$ and optimized jointly with the detector. For each mini-batch, $q_i = \sigma(w_g \, \text{sg}[E(x_i)] + b_g)$ is normalized according to Eq. 8, yielding $\frac{1}{B} \sum_i g_i = 1$. Thus, the gate redistributes a fixed mean robustness weight rather than changing its batch-average magnitude. Stop-gradient blocks gradients from the gate weights through E, while $\mathcal{L}_{\text{ood}}^{\text{gate}}$ remains differentiable through its own loss path.

Writing $R_i = \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}(x_i) + \lambda_{\text{ood}} \mathcal{L}_{\text{ood}}^{\text{gate}}(x_i)$ and $S = \sum_j q_j$, we have

$$\mathcal{L}_{\text{rob}} = \frac{\sum_i q_i R_i}{S}, \qquad \frac{\partial \mathcal{L}_{\text{rob}}}{\partial q_i} = \frac{R_i - \mathcal{L}_{\text{rob}}}{S}. \tag{11}$$

If the q_i were independently adjustable, direct minimization would therefore favor relatively larger weights on samples with below-average R_i; the objective alone does not guarantee hard-example

emphasis. In ECRM, however, all sample weights are coupled through the shared two-parameter map $q_i = \sigma(w_g E_i + b_g)$, for which

$$\frac{\partial \mathcal{L}_{\mathrm{rob}}}{\partial w_g} = \frac{1}{S} \sum_i q_i(1 - q_i) E_i (R_i - \mathcal{L}_{\mathrm{rob}}). \tag{12}$$

The update direction therefore depends on the joint batch relationship between energy and the centered robustness terms and is not sign-constrained. Accordingly, we treat the sign and magnitude of the final gate as an empirical property of the trained system rather than a guarantee of the objective.

All reported ECRM-enabled results use this normalized gate formulation; no unnormalized gate is used in the reported main or appendix experiments. For the Uniform control, the normalized gate is replaced by $g_i \equiv 1$ while $\mathcal{L}_{\mathrm{adv}}$, $\mathcal{L}_{\mathrm{ood}}^{\mathrm{gate}}$, the stage-level $\mathcal{L}_{\mathrm{ood}}$, optimizer, schedule, and training budget remain unchanged.

Across three default-seed runs, the optimized parameters are $w_g = 1.12 \pm 0.06$ and $b_g = -0.08 \pm 0.04$. At the mean endpoint and over $E \in [0, \log 2]$, the raw gate spans approximately $q(E) \in [0.48, 0.67]$, corresponding to at most about a $1.4\times$ relative weight ratio. Unit-mean normalization preserves this pairwise ratio because $g_i/g_j = q_i/q_j$. The positive reported endpoint therefore implements a moderate monotone allocation in which higher-energy samples receive greater robustness weight. We interpret this as an observed property of the trained models, not as a guaranteed consequence of minimizing $\mathcal{L}_{\mathrm{rob}}$ or as evidence that the positive orientation is learned independently of initialization.

The $1.4\times$ span characterizes instantaneous relative weighting, not the difference between models obtained after training. We therefore interpret the Full–Uniform comparison only as the end-to-end effect of replacing uniform allocation with the complete energy-conditioned ECRM mechanism under otherwise matched training. It does not isolate the contribution of optimizing w_g, b_g from the energy-conditioned parameterization, and the instantaneous $1.4\times$ span is not interpreted as a direct effect size. The matched stabilization–allocation control is reported in App. E.2.

The ECRM parameters w_g, b_g are distinct from the fusion parameters $W_f, \mathbf{b}_f$ in Eq. 2: the former determine training-loss weights, whereas the latter combine the spectral and prosodic representations. Initializing $w_g \in \{0.5, 1.0, 2.0\}$ while keeping all other settings fixed changes final Robust AUC by at most 0.02. We use this experiment only as an end-to-end initialization-sensitivity check; because it does not establish convergence of the gate parameters to a common endpoint, we do not claim that the final slope is initialization-independent. Perturbing τ_{id} by $\pm 20\%$ changes ECE and Robust AUC by at most 0.01 each. Broader coefficient sensitivity is reported in App. B.10.

Normalized-gate stability. The raw-gate bound in Eq. 7 also extends to the normalized weights used by ECRM. Since $E(x) \in [0, \log 2]$, define

$$q_{\min} = \sigma(\min\{b_g,\, b_g + w_g \log 2\}) > 0. \tag{13}$$

For paired clean/transformed mini-batches, with $g_i = q_i/\bar{q}$ and $g_i' = q_i'/\bar{q}'$, we have

$$|g_i' - g_i| \le \frac{|q_i' - q_i|}{q_{\min}} + \frac{1}{B q_{\min}^2} \sum_{j=1}^{B} |q_j' - q_j|. \tag{14}$$

Using $|\sigma'(u)| \le 1/4$ and Jensen's inequality gives

$$\frac{1}{B} \sum_{i=1}^{B} (g_i' - g_i)^2 \le \frac{w_g^2}{8} \left(q_{\min}^{-2} + q_{\min}^{-4}\right) \frac{1}{B} \sum_{i=1}^{B} (E_i' - E_i)^2. \tag{15}$$

Thus, $\mathcal{L}_{\mathrm{ood}}$ also bounds the mean-squared transformation sensitivity of the normalized robustness weights up to a finite gate-dependent constant.

Complete training objective. The full objective used during staged training is

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{cmra}} \mathcal{L}_{\mathrm{cmra}} + \mathbb{1}_{\mathrm{eng}} \lambda_{\mathrm{eng}} \mathcal{L}_{\mathrm{energy}} + \mathbb{1}_{\mathrm{ood}} \lambda_{\mathrm{ood}} \mathcal{L}_{\mathrm{ood}} + \mathbb{1}_{\mathrm{adv}} \mathcal{L}_{\mathrm{rob}}, \tag{16}$$

where $\mathcal{L}_{\mathrm{base}} = \mathcal{L}_{\mathrm{det}} + \lambda_{\mathrm{dom}} \mathcal{L}_{\mathrm{dom}} + \lambda_{\mathrm{con}} \mathcal{L}_{\mathrm{con}}$. The base objectives and CMRA remain active throughout training, while the stage indicators activate energy anchoring, OOD consistency, and robustness modulation sequentially.

Once the robustness stage is active, the stage-level $\mathcal{L}_{\text{ood}}$ continues to apply uniform consistency, while $\mathcal{L}_{\text{ood}}^{\text{gate}}$ provides its per-sample counterpart inside $\mathcal{L}_{\text{rob}}$. The coefficient λ_{ood} is shared between these terms. The ECRM parameters w_g, b_g are distinct from the fusion parameters $W_f, \mathbf{b}_f$, and the gate is removed at inference.

B.7 Staged Training Algorithm

The security-specific objectives are activated at fixed fractional boundaries of the total training budget. For the default 80-epoch schedule, energy anchoring begins at epoch 16, OOD consistency at epoch 32, and ECRM-based robustness training at epoch 48. The base objectives and CMRA remain active throughout. Algorithm 1 gives the complete procedure.

Algorithm 1 ASNet staged training.

Require: Dataset $\mathcal{D} = \{(x_i, y_i)\}$; epochs E; batch size B; PGD budget ϵ_{train}; boundaries $(e_{\text{eng}}, e_{\text{ood}}, e_{\text{adv}}) = (0.2E, 0.4E, 0.6E)$.

Initialize detector parameters θ, ECRM parameters w_g, b_g, and AdamW optimizer.
for epoch $e = 1, \ldots, E$ **do**
  Set $\mathbb{1}_{\text{eng}} \leftarrow (e \geq e_{\text{eng}})$, $\mathbb{1}_{\text{ood}} \leftarrow (e \geq e_{\text{ood}})$, and $\mathbb{1}_{\text{adv}} \leftarrow (e \geq e_{\text{adv}})$.
  if $e \geq e_{\text{eng}}$ **and** $(e - e_{\text{eng}}) \bmod 10 = 0$ **then**
    Refresh scalar τ_{id} from correctly classified LibriSeVoc validation predictions without gradient updates.
  end if
  for mini-batch $\{(x_i, y_i)\}_{i=1}^{B}$ **do**
    Compute $\mathbf{z}_s$, $\mathbf{z}_p$, $\mathbf{z}^*$, $p_{\text{fake}}(x_i)$, and $E(x_i)$.
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{base}} + \lambda_{\text{cmra}}\mathcal{L}_{\text{cmra}}$.
    if $\mathbb{1}_{\text{eng}}$ **then**
      $\mathcal{L} \leftarrow \mathcal{L} + \lambda_{\text{eng}}\mathcal{L}_{\text{energy}}$.
    end if
    if $\mathbb{1}_{\text{ood}}$ **then**
      Sample $x_i' \sim \text{Aug}(x_i)$ and compute $\mathcal{L}_{\text{ood}}$.
      $\mathcal{L} \leftarrow \mathcal{L} + \lambda_{\text{ood}}\mathcal{L}_{\text{ood}}$.
    end if
    if $\mathbb{1}_{\text{adv}}$ **then**
      Generate x_i^{adv} by PGD and compute $\mathcal{L}_{\text{adv}}(x_i)$.
      Compute $\mathcal{L}_{\text{ood}}^{\text{gate}}(x_i) = (E(\text{Aug}(x_i)) - E(x_i))^2$.
      Compute raw gate scores $q_i = \sigma(w_g\, \text{sg}[E(x_i)] + b_g)$.
      Normalize the gate to unit mini-batch mean: $g_i = q_i / \left(\frac{1}{B}\sum_{j=1}^{B} q_j\right)$.
      Compute $\mathcal{L}_{\text{rob}}$ using Eq. 9.
      $\mathcal{L} \leftarrow \mathcal{L} + \mathcal{L}_{\text{rob}}$.
    end if
    Update θ, w_g, b_g with AdamW.
  end for
end for
return trained detector parameters θ.

All phases use AdamW with cosine learning-rate decay. Optimizer settings and loss coefficients are specified in App. C.4.

B.8 Stage Diagnostics

Tab. 6 reports stage-local diagnostics computed on a fixed set of LibriSeVoc evaluation utterances immediately before and after each security-objective transition, so changes are not confounded by a different evaluation sample. Energy anchoring does not directly optimize paired clean–transform drift; the first-row change is therefore interpreted as an associated stage-local diagnostic rather than a direct effect estimate.

For the paired OOD diagnostics, x_c denotes a clean utterance and $x_a \sim \mathrm{Aug}(x_c)$ denotes its paired codec/channel-transformed version. For the codec-distribution diagnostic, $p_{\mathrm{clean}}(E)$ and $p_{\mathrm{codec}}(E)$ denote empirical distributions of the energy score for paired clean and codec-transformed utterances, respectively. The same fixed density-estimation procedure is used before and after the corresponding stage transition.

These stage-local measurements are mechanism checks rather than final-model performance metrics. In particular, the $0.31 \rightarrow 0.09$ KL reduction is measured immediately across the $\mathcal{L}_{\mathrm{ood}}$ transition and indicates reduced clean–codec distributional drift. It is distinct from final-model OOD-AUROC, which reports directional rank separation between clean and compressed scores. The complete cumulative-schedule comparison is reported in App. E.3.

Table 6: **Stage-local diagnostics on LibriSeVoc.** Each diagnostic is evaluated on the same fixed samples immediately before and after the corresponding stage transition. Lower values indicate greater stability for all three diagnostics.

Transition	**Observed property**	**Diagnostic**	**Before → after**
Base $\rightarrow +\mathcal{L}_{\mathrm{energy}}$	Paired energy-drift variation	$\mathrm{Var}[E(x_c) - E(x_a)]$	$0.089 \rightarrow 0.021$
$+\mathcal{L}_{\mathrm{energy}} \rightarrow +\mathcal{L}_{\mathrm{ood}}$	Clean–codec distributional drift	$D_{\mathrm{KL}}(p_{\mathrm{clean}}(E) \,\|\, p_{\mathrm{codec}}(E))$	$0.31 \rightarrow 0.09$
$+\mathcal{L}_{\mathrm{ood}} \rightarrow +\mathcal{L}_{\mathrm{rob}}$	Adversarial logit stability	$\|f_\theta(\mathbf{z}^*(x)) - f_\theta(\mathbf{z}^*(x_{\mathrm{adv}}))\|_2$	$0.44 \rightarrow 0.17$

B.9 Schedule-Timing Sensitivity

After fixing the default schedule before test evaluation, Tab. 7 reports a post-hoc sensitivity check on held-out LibriSeVoc: all three activation boundaries are shifted jointly by ± 10 percentage points. These runs are not used to select or retune the default schedule.

Table 7: **Post-hoc schedule-timing sensitivity on the LibriSeVoc test split.** The reported default is fixed beforehand; no row is used for tuning or model selection.

Schedule (% of epochs)	Robust AUC ↑	OOD-AUROC	ECE ↓
Early (10/30/50)	0.85	0.68	0.15
Default (20/40/60)	**0.90**	0.74	**0.10**
Late (30/50/70)	0.87	0.70	0.13

B.10 Hyperparameter Audit and Coefficient Sensitivity

Tab. 8 distinguishes learned, validation-derived, scheduled, shared, and freely specified quantities. The four ASNet-specific loss coefficients use a common default value of 1.0 and are not retuned for individual corpora, codecs, attacks, or test conditions.

Table 8: **Hyperparameter audit.** Method-related quantities are grouped by how they are determined. The four freely specified ASNet-specific loss coefficients all default to 1.0.

Category	**Parameters**	**Count**	**How set**
Learned	ECRM gate w_g, b_g	2	Joint AdamW updates through normalized $\mathcal{L}_{\mathrm{rob}}$
Validation-derived	Energy anchor τ_{id}	1	Median energy of correctly classified validation samples
Validation-derived	CMRA margin $m = 0.8$	1	Upper-tail cross-view similarity on clean validation
Shared where applicable	Learning rate, weight decay, batch size, PGD budget/step	5	Common protocol settings
Schedule constants	Stage boundaries $20/40/60\%$	3	Fixed before test evaluation; post-hoc sensitivity in Tab. 7
Base-loss coefficients	$\lambda_{\mathrm{dom}}, \lambda_{\mathrm{con}}$	2	Fixed for all runs; App. C.4
Freely specified ASNet-specific coefficients	$\lambda_{\mathrm{cmra}}, \lambda_{\mathrm{eng}}, \lambda_{\mathrm{adv}}, \lambda_{\mathrm{ood}}$	4	Unit defaults fixed before test evaluation; post-hoc sensitivity in Tab. 9

Coefficient sensitivity. After fixing the unit coefficients before test evaluation, Tab. 9 reports a post-hoc $\pm 50\%$ sensitivity check. Across these trials, Robust AUC, OOD-AUROC, and ECE vary by at most 0.04, 0.08, and 0.05, respectively; no trial is used for selection.

Table 9: **Post-hoc coefficient sensitivity on the LibriSeVoc test split.** The reported defaults are fixed beforehand; no row is used for tuning or model selection. "Range" is the maximum variation across these checks.

Trial	λ_{cmra}	λ_{eng}	λ_{adv}	λ_{ood}	Robust AUC ↑	OOD-AUROC	ECE ↓
Default	1.0	1.0	1.0	1.0	0.90	0.74	0.10
P1	0.5	1.0	1.5	0.5	0.89	0.67	0.13
P2	1.5	0.5	1.0	1.5	0.88	0.70	0.15
P3	0.5	1.5	0.5	1.0	0.86	0.68	0.13
P4	1.0	0.5	1.5	0.5	0.89	0.66	0.14
Range	—	—	—	—	**0.04**	**0.08**	**0.05**

C Extended Experimental Setup

This appendix provides the dataset composition, preprocessing, evaluation protocol, training configuration, baseline implementations, reimplementation validation, and metric definitions supporting Sec. 4. The main experiments use the controlled single-source LibriSeVoc protocol; multi-domain results are reported separately in App. D.5.

C.1 Datasets

LibriSeVoc is the sole training dataset in the main experiments. ASVspoof21, WaveFake, FakeAVCeleb, and CodecFake are used as external evaluation datasets without adaptation or corpus-specific retuning. Tab. 10 summarizes the evaluated splits and dataset statistics.

Table 10: **Dataset statistics and evaluated splits.** LibriSeVoc is the sole training dataset; ASVspoof21, WaveFake, FakeAVCeleb, and CodecFake are used for external evaluation. Dataset-specific split details are provided below.

Corpus	#Utterances	Hours	#Speakers	Languages
LibriSeVoc	92,407	243 h	2,456	EN
ASVspoof21 LA (eval)	121,552	∼80 h	121	EN
WaveFake	117,985 (fake)	196 h	2	EN, JA
FakeAVCeleb (audio)	20,000	36 h	59	EN
CodecFake (test: C1–C7, A1–A3)	421,157	–	328[†]	EN, ZH

[†]The reported 328 source speakers comprise 110 VCTK speakers and 218 AISHELL3 speakers in the official CodecFake test release.

CodecFake. CodecFake (Xie et al., 2024) is an external deepfake-audio dataset covering neural-codec and audio-language-model generation methods. We evaluate the complete official test release as provided, comprising C1–C6 neural-codec conditions, C7 with an unseen neural codec, and A1–A3 audio-language-model conditions. The release contains 421,157 utterances from 328 VCTK/AISHELL3 source speakers across English and Mandarin, disjoint from the LibriSeVoc training speakers. The official CodecFake training and development splits are not used for training, validation, checkpoint selection, threshold selection, or adaptation.

C.2 Preprocessing

All audio is converted to mono, resampled to 16 kHz, and peak-normalized. Training uses random 6 s crops. At evaluation, 6 s windows with 50% overlap are processed independently and their logits are averaged to obtain an utterance-level prediction. Spectral and prosodic feature extraction follows App. B.2. Baselines use the same segmentation and score-aggregation protocol unless their native architecture requires an equivalent fixed-length waveform input.

C.3 Evaluation Protocol

Evaluation follows the single-source protocol in Sec. 4 and the digital threat model in Sec. 2.1. All external datasets are evaluated only after model and abstention-threshold selection are complete, with no evaluation-domain adaptation or retuning. Waveform attacks operate before feature extraction; default PGD generation is specified in App. B.6, and stronger and abstention-aware attacks are specified in App. D.2.

C.4 Training Details

Default-protocol runs use AdamW with learning rate 2×10^{-4}, weight decay 10^{-4}, batch size 128, cosine decay, and an 80-epoch training budget over seeds $\{0, 1, 2\}$. The base objective is

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{det}} + \lambda_{\text{dom}}\mathcal{L}_{\text{dom}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}},$$

with $\lambda_{\text{dom}} = \lambda_{\text{con}} = 1.0$. The four ASNet-specific coefficients λ_{cmra}, λ_{eng}, λ_{adv}, and λ_{ood} also default to 1.0 and are not retuned by corpus, codec, or attack. The staged training procedure is given in App. B.7. In every ECRM-enabled run, the raw sigmoid gate scores are normalized to unit mini-batch mean exactly as in Eq. 8 before computing $\mathcal{L}_{\text{rob}}$; w_g and b_g are then updated jointly with the detector by AdamW. Coefficient and schedule sensitivity are reported in Apps. B.10 and B.9.

For the compute-matched control, Ren et al.$^{\texttt{SecAug+Adv}}$ receives an extended training horizon and corresponding cosine-decay schedule until its cumulative A100 GPU-hour budget matches Full ASNet. Its architecture, objective, data, augmentation, optimizer settings, and validation-only checkpoint-selection procedure remain unchanged.

C.5 Baseline Configurations

LCNN (Lavrentyeva et al., 2019) uses its max-feature-map convolutional architecture on log-mel inputs. RawNet2 (Tak et al., 2021) uses its waveform CNN–GRU architecture with fixed-length waveform crops. AASIST (weon Jung et al., 2021) uses its spectro-temporal graph-attention architecture. RawNetLite (Pierno et al., 2025) uses the corresponding lightweight RawNet-style detector; its multi-domain training is disabled in the single-source comparison. Ren et al. (Ren et al., 2024) is retrained on LibriSeVoc using its architecture and objective formulation under the common preprocessing and evaluation protocol.

The `SecAug` variants receive the same codec and channel operators used by ASNet's OOD-consistency stage while leaving the baseline architecture, objective, and inference procedure unchanged. Ren et al.$^{\texttt{SecAug+Adv}}$ additionally receives adversarial training. These variants distinguish additional perturbation exposure from the energy-stabilization and energy-conditioned allocation mechanisms proposed in ASNet.

Self-supervised encoders such as WavLM and XLS-R rely on external pretraining and therefore fall outside the controlled LibriSeVoc-only protocol. We exclude them from the protocol-matched ranking rather than compare heterogeneous published results obtained under different evaluation settings.

C.6 Baseline Reimplementation Validation

To check for implementation regressions, we additionally reproduce each baseline under its corresponding native published ASVspoof LA protocol and compare the resulting EER with the value reported for that same method. These are within-method reproduction checks rather than a cross-model benchmark: rows may correspond to different native evaluation configurations. This validation is separate from the controlled LibriSeVoc-only comparisons.

C.7 Metric Definitions

AUC and EER are computed at the utterance level after window-level logits are aggregated. ECE uses 15-bin equal-mass estimation on outputs before post-hoc recalibration; no temperature scaling is applied to any model.

Table 11: **Native-protocol baseline reproduction checks.** Each retrained value is compared only with the published value for the same method under its cited native ASVspoof LA protocol. Rows are not intended as a common cross-model ranking.

Model	Published EER (%)	Retrained EER (%)	Δ
RawNet2 (Tak et al., 2021)	5.13	5.21	+0.08
AASIST (weon Jung et al., 2021)	0.83	0.91	+0.08
RawNetLite (Pierno et al., 2025)	2.10	2.18	+0.08
Ren et al. (Ren et al., 2024)	1.71	1.73	+0.02

For a corpus $\mathcal{D}$, Robust AUC is the detection AUC obtained after waveform-level PGD-10 perturbation with $\epsilon = 10^{-3}$. External Robust AUC is the unweighted arithmetic mean of the corresponding values on ASVspoof21, WaveFake, FakeAVCeleb, and CodecFake. When an ablation reports LibriSeVoc Robust AUC, it refers instead to the perturbed held-out LibriSeVoc test set; the two quantities are not compared as if they were the same evaluation regime.

OOD-AUROC. For an evaluation corpus $\mathcal{D}$ and scalar uncertainty score $s(x)$, we compute OOD-AUROC by distinguishing clean utterances from fixed mp3-64 kbps counterparts:

$$\text{OOD-AUROC}(\mathcal{D}; s) = \text{AUROC}(\{(s(x), 0) : x \in \mathcal{D}\} \cup \{(s(c_{\text{mp3}}(x)), 1) : x \in \mathcal{D}\}) . \tag{17}$$

The shifted condition is positive; 0.5 denotes no directional rank separation, while values above or below 0.5 indicate opposite directions. Neither direction nor a larger value is interpreted as better. OOD-AUROC is reported only as a fixed rank-separation diagnostic, not as evidence of stabilization or unseen-OOD detection.

OOD-AUROC is rank-based, whereas $\mathcal{L}_{\text{ood}}$ penalizes paired score-magnitude drift. Small but consistently signed score changes can therefore yield substantial rank separation while maintaining small paired drift. Hence OOD-AUROC need not approach 0.5 as $\mathcal{L}_{\text{ood}}$ decreases; direct stabilization evidence is provided by App. B.8.

For model-to-model selective comparisons, each model selects its threshold using only its own clean LibriSeVoc validation scores to target 20% abstention; the threshold is then frozen before test evaluation. For ASNet, this gives $\tau = 0.22$.

Selective evaluation uses exactly the same utterance-level detector outputs as the corresponding full-set evaluation. Window-level logits are first aggregated according to App. C.2; the resulting utterance-level logit determines both $p_{\text{fake}}(x_i)$ and the class-symmetric energy $E(x_i)$. The frozen threshold then defines

$$\mathcal{K}_\tau = \{i : E(x_i) < \tau\}. \tag{18}$$

A single mask $\mathcal{K}_\tau$ is applied identically to the labels and detection scores, after which

$$AUC_{\text{kept}} = AUC(\{(y_i, p_{\text{fake}}(x_i)) : i \in \mathcal{K}_\tau\}) . \tag{19}$$

The reported clean-test AUC_{kept} values are recomputed from the same utterance-level predictions used for the full-set metrics; no alternate test condition, window-level filtering, test-set threshold tuning, or score reaggregation is used. AUC_{kept} measures discrimination within the accepted subset and is used as a supporting diagnostic rather than as a selective-risk measure. Error catch is the fraction of misclassified samples rejected by the frozen rule and is used here as a within-ASNet diagnostic. Under abstention-aware attack, UASR is the fraction of initially correct samples made both misclassified and accepted.

All main comparisons use three seeds. Reported differences are interpreted descriptively; no formal statistical-significance test is performed. Differences below 0.01 are treated as reporting-level ties.

C.8 Attack Details

Training-time adversarial examples use untargeted waveform-level PGD-2 with $\epsilon = 2 \times 10^{-3}$. The default test-time robustness evaluation uses PGD-10 at $\epsilon = 10^{-3}$ with step size 2.5×10^{-4}.

For classifier-to-classifier comparison, APGD-CE (Croce & Hein, 2020) uses 10 steps and three restarts and CW (Carlini & Wagner, 2017) uses 50 iterations; each compared model receives the

same attack configuration. We additionally evaluate BPDA-APGD (Athalye et al., 2018) because the prosodic extractor contains non-differentiable RAPT-based f_0, binary voicing, and ZCR operations. BPDA retains the deployed extractor in the forward pass and uses differentiable surrogates for these operators in the backward pass. The same BPDA configuration at $\epsilon = 10^{-3}$ is applied to ASNet and the compute-matched Ren et al.$^{\texttt{SecAug+Adv}}$ control. These are finite-budget comparative stress tests, not convergence tests. Black-box transfer PGD from a separately trained RawNet2 is reported as a mismatched-attack diagnostic and is not used to validate white-box convergence.

The separate abstention-aware evaluation instead optimizes

$$\max_{\|x'-x\|_\infty \le 2\times 10^{-3}} \mathcal{L}_{\mathrm{CE}}(f(x'), y) - \lambda E(x')$$

for $\lambda \in \{0.25, 0.5, 1, 2, 4\}$. For each λ, APGD uses 100 iterations and five restarts, yielding 25 trajectories per sample. Because this attack additionally targets the deployed rejection signal, it is used to stress-test the deployed rejection rule rather than to establish a comparative classifier ranking. The validation-selected threshold $\tau = 0.22$ is never changed. For selective metrics, an accepted misclassification is selected whenever one exists; Robust AUC is computed separately using the candidate with maximal classification loss. Results and interpretation are reported in App. D.2.

D Extended Results

This section provides supporting evaluations for the main results, including selective prediction, stronger and adaptive attacks, cross-vocoder generalization, multi-domain training, and runtime measurements.

D.1 Selective Classification and Energy Diagnostics

We evaluate two complementary aspects of selective behavior. Tab. 12 reports clean-LibriSeVoc retained-set discrimination under model-specific validation-selected thresholds, while Tab. 13 evaluates how ASNet's single validation-fixed threshold transfers across clean, external, and adversarial conditions without test-time retuning.

Table 12: **Retained-set discrimination on clean LibriSeVoc.** Each model selects its threshold only on clean LibriSeVoc validation to target 20% abstention and freezes it before test evaluation. AUC_{kept} is recomputed from exactly the retained utterance-level labels and detection scores.

Model	$\mathbf{AUC}_{\mathrm{kept}} \uparrow$
Ren et al.$^{\texttt{SecAug+Adv}}$	0.994
Ren et al.$^{\texttt{SecAug+Adv}}$ (compute-matched)	0.995
ASNet (ours)	0.998

All three systems preserve near-ceiling discrimination after validation-fixed abstention. The small difference between ASNet and the compute-matched Ren control is treated as a reporting-level tie rather than as evidence of selective superiority; the selective analysis instead focuses on error rejection and robustness of the fixed abstention rule.

Table 13: **Per-condition ASNet abstention efficacy.** The validation-fixed $\tau = 0.22$ is applied without retuning. This is a within-ASNet diagnostic, not a cross-model comparison.

Evaluation setting	**% Caught**$\uparrow$	**% Residual**$\downarrow$
Clean LibriSeVoc test (AUC $= 0.999$)	81	19
CodecFake test (AUC $= 0.739$)	66	34
Adversarial PGD-10 (Robust AUC $= 0.90$)	85	15

The larger residual on CodecFake is consistent with the greater overlap between correct and incorrect energy distributions in Fig. 3, indicating useful but incomplete transfer of the rejection rule under external shift.

D.2 Adaptive and Black-Box Attack Evaluation

We extend the main PGD-10 evaluation with stronger and mismatched inference-time attacks on LibriSeVoc: APGD-CE (Croce & Hein, 2020), CW (Carlini & Wagner, 2017), black-box transfer PGD from RawNet2, reverberation plus noise, and attacks that explicitly target the energy-based rejection rule.

Joint adaptive attack. To test whether an adversary can jointly induce classification error and avoid abstention, we optimize

$$\max_{\|x'-x\|_\infty \le 2\times 10^{-3}} \mathcal{L}_{\mathrm{CE}}(f(x'), y) - \lambda E(x').$$

For each $\lambda \in \{0.25, 0.5, 1, 2, 4\}$, APGD uses 100 iterations and five random restarts, yielding 25 trajectories per sample. The validation-selected threshold $\tau = 0.22$ remains fixed. For selective metrics, an accepted misclassification with $E(x') < \tau$ is selected whenever one exists; Robust AUC is computed separately from the candidate with maximal classification loss.

Table 14: **Abstention-aware attacks on LibriSeVoc.** Values are three-seed means. Catch is the fraction of induced errors rejected by the fixed threshold; UASR is the fraction of initially correct samples made both misclassified and accepted.

Attack	ϵ	Steps/traj.	Robust AUC↑	ASR↓	Catch↑	UASR↓
Energy-aware PGD ($\alpha_E = 0.5$)	10^{-3}	10/1	0.87	34.0%	48.0%	17.7%
Joint APGD-CE + energy suppression	2×10^{-3}	100/25	0.76	48.0%	35.0%	31.2%

The stronger joint evaluation reduces Robust AUC from 0.87 to 0.76, reduces error catch from 48.0% to 35.0%, and increases UASR from 17.7% to 31.2%. Because it also uses a larger perturbation budget, more iterations, and more search trajectories, this degradation reflects the combined stronger adaptive evaluation rather than energy suppression alone.

Table 15: **Matched finite-budget classification attacks on LibriSeVoc.** Entries report Robust AUC/Robust EER without abstention. The Ren column uses the compute-matched Ren et al.$^{\texttt{SecAug+Adv}}$ control from Tab. 1; both models use identical attack configurations. Results are finite-budget comparisons, not convergence claims. †APGD-CE uses 10 steps and three restarts. N/A: the energy-aware attack targets ASNet's deployed energy-based rejection rule, which the Ren control does not deploy.

Attack	Budget	Ren$^{\texttt{SecAug}+\mathrm{Adv}}$	ASNet (ours)
PGD white-box	$\epsilon = 10^{-3}$	0.88 / 0.19	**0.90 / 0.16**
APGD-CE†	$\epsilon = 2\times10^{-3}$	0.86 / 0.23	**0.89 / 0.20**
CW	$\epsilon = 2\times10^{-3}$	0.85 / 0.24	**0.88 / 0.21**
BPDA-APGD	$\epsilon = 10^{-3}$	0.85 / 0.25	**0.87 / 0.23**
Transfer PGD (RawNet2 → ASNet)	—	0.88 / 0.21	**0.90 / 0.18**
Reverb + noise (10 dB SNR)	—	0.92 / 0.17	**0.94 / 0.15**
Energy-aware PGD ($\alpha_E = 0.5$)	$\epsilon = 10^{-3}$	N/A	**0.87 / 0.22**

Across the reported matched finite-budget classification attacks with a corresponding Ren et al. result, ASNet has higher Robust AUC and lower Robust EER. We do not extrapolate this ordering to unconstrained or converged attack search. The stronger abstention-aware evaluation serves a different purpose: it shows that classifier robustness does not imply robustness of the energy-based rejection rule when that signal is explicitly targeted.

D.3 Cross-Vocoder Generalization

Fig. 4 stratifies EER by vocoder family under LibriSeVoc-only training. ASNet obtains average EER 20.3%, compared with 27.9% for Ren et al. and and 33.3% for RawNetLite. The largest observed difference, approximately 10 EER points, occurs for diffusion vocoders. This pattern is consistent with the proposed complementarity of spectral and prosodic evidence, but does not by itself establish that mechanism.

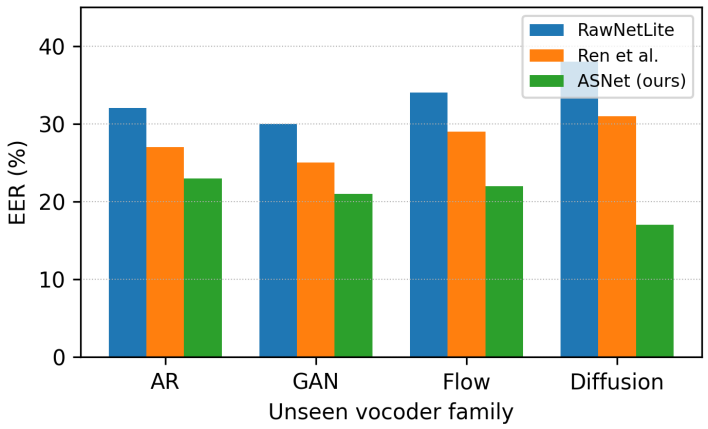


Figure 4: **Cross-vocoder EER under LibriSeVoc-only training.** ASNet obtains the lowest average EER among the compared methods, with the largest observed difference on diffusion-based vocoders.

D.4 Externally Pretrained Encoders

Large self-supervised encoders such as WavLM and XLS-R rely on external pretraining and therefore do not satisfy the controlled LibriSeVoc-only protocol used for the main comparisons. We consequently exclude them from the protocol-matched ranking rather than compare heterogeneous published results across different ASVspoof evaluation settings.

D.5 Cross-Domain and Multi-Domain Generalization

We additionally evaluate two broader training protocols following the RawNetLite compilation setting (Pierno et al., 2025). Hyperparameters remain fixed from the single-source experiments, with no evaluation-domain retuning.

Double-domain training (LibriSeVoc + ASVspoof21). Tab. 16 evaluates FakeOrReal, CodecFake, and WaveFake. ASNet obtains average AUC 0.865 and average EER 0.154, compared with $0.837/0.177$ for Ren et al. and $0.765/0.220$ for RawNetLite.

Table 16: **Double-domain training (LibriSeVoc + ASVspoof21).** All methods use fixed hyperparameters without evaluation-domain retuning. *OOD-AUROC for RawNetLite was not computed in this configuration because of evaluation-coverage limitations.

Model	FakeOrReal	CodecFake	WaveFake	Avg. AUC↑	Avg. EER↓	OOD-AUROC
RawNetLite * (Pierno et al., 2025)	**0.952 / 0.080**	0.514 / 0.357	0.828 / 0.222	0.765	0.220	—
Ren et al. (Ren et al., 2024)	0.938 / 0.109	0.621 / 0.319	0.951 / 0.103	0.837	0.177	0.58
ASNet (ours)	0.941 / 0.103	**0.673 / 0.298**	**0.982 / 0.061**	**0.865**	**0.154**	0.61

Triple-domain training (LibriSeVoc + ASVspoof21 + WaveFake). Under the broader training mixture, ASNet obtains average AUC 0.887 and average EER 0.147, compared with $0.865/0.161$ for Ren et al. and $0.744/0.284$ for RawNetLite (Tab. 17).

Table 17: **Triple-domain training (LibriSeVoc + ASVspoof21 + WaveFake).** All methods use fixed hyperparameters without evaluation-domain retuning. *OOD-AUROC for RawNetLite was not computed in this configuration because of evaluation-coverage limitations.

Model	FakeOrReal	CodecFake	ASVspoof21 test	Avg. AUC↑	Avg. EER↓	OOD-AUROC
RawNetLite * (Pierno et al., 2025)	**0.969 / 0.055**	0.643 / 0.357	0.620 / 0.440	0.744	0.284	—
Ren et al. (Ren et al., 2024)	0.957 / 0.078	0.701 / 0.285	0.938 / 0.119	0.865	0.161	0.51
ASNet (ours)	0.961 / 0.072	**0.740 / 0.262**	**0.961 / 0.106**	**0.887**	**0.147**	0.54

FakeOrReal native protocol. For completeness, Tab. 18 reports FakeOrReal performance under the RawNetLite compilation protocol. RawNetLite obtains the lowest EER (0.29%), followed by ASNet (0.32%) and Ren et al. (0.37%). This contextual experiment is not used to support the main LibriSeVoc-only claim.

Table 18: **EER on FakeOrReal under the RawNetLite compilation protocol.** This contextual evaluation is separate from the main LibriSeVoc-only comparison.

Method	**EER (%)**
RawNet2 (Tak et al., 2021)	25.1
MoE LCNN Fusion (Lavrentyeva et al., 2019)	2.74
RawNetLite (Pierno et al., 2025)	**0.29**
Ren et al. (Ren et al., 2024)	0.37
ASNet (ours)	0.32

These multi-domain experiments provide a complementary check that the observed behavior is not confined to single-source training, while remaining separate from the controlled protocol used for the paper's primary claims.

D.6 Runtime and Efficiency

Training ASNet on LibriSeVoc with batch size 128 and 6 s clips takes 10–12 hours on a single A100-80G, approximately $1.4\times$ the wall-clock cost of Ren et al. in the same hardware regime. Inference runs at $0.37\times$ real time on a single CPU core and faster than real time on a commodity GPU. ASNet uses approximately $1.15\times$ the parameters of RawNet2 and $0.9\times$ those of Ren et al. Pruning 25% of channels reduces parameter count by approximately 18% at an approximately 0.5-point AUC cost.

E Extended Ablations

This section provides supporting mechanism studies for Sec. 5.3: hold-one-out tests of the energy-stabilization objectives, the stabilization–allocation 2×2, complete cumulative-schedule controls, per-corpus robustness transfer, cross-domain mechanism ablations, cost–reliability analysis, and an independent conditioning-signal comparison.

E.1 Energy-Stabilization Hold-One-Out Ablation

Tab. 19 isolates energy anchoring and OOD consistency while holding the spectral–prosodic architecture, optimizer, training budget, robustness mechanism, and single-source protocol fixed. Removing either objective lowers Robust AUC and increases ECE relative to Full ASNet, while removing both produces the largest Robust AUC reduction. In particular, removing $\mathcal{L}_{\text{energy}}$ changes ECE from 0.10 to 0.23, consistent with direct confidence-magnitude regularization by the anchoring objective. However, the Uniform control retains stabilization and reaches ECE 0.19, showing that anchoring does not reproduce the calibration of Full ASNet by itself. We therefore attribute the final calibration profile to the complete trained system rather than to a single objective. OOD-AUROC is reported only as a shift-separation diagnostic, not as a monotonic measure of stabilization.

Table 19: **Energy-stabilization hold-one-out ablation on LibriSeVoc.** All variants use the same architecture, optimizer, training budget, and single-source protocol as Full ASNet. ✓ = active; ✗ = removed.

Variant	**Energy-Stabilization Objectives**		**AUC↑**	**Robust AUC↑**	**OOD-AUROC**	**ECE↓**
	$\mathcal{L}_{\text{energy}}$	$\mathcal{L}_{\text{ood}}$				
Full ASNet	✓	✓	**0.999**	**0.90**	0.74	**0.10**
w/o $\mathcal{L}_{\text{energy}}$	✗	✓	0.990	0.86	0.63	0.23
w/o $\mathcal{L}_{\text{ood}}$	✓	✗	0.989	0.87	0.60	0.21
w/o $\mathcal{L}_{\text{energy}}$ and $\mathcal{L}_{\text{ood}}$	✗	✗	0.986	0.82	0.64	0.20

E.2 Stabilization–Allocation Interaction

To separate stabilization from conditioned allocation, we complete the matched 2×2 crossing energy stabilization ($\mathcal{L}_{\text{energy}} + \mathcal{L}_{\text{ood}}$) with energy-conditioned versus uniform ($g_i \equiv 1$) allocation. All four cells use the same architecture, optimizer, robustness objective, activation epoch, and training budget; the Uniform variants retain $\mathcal{L}_{\text{adv}}$, $\mathcal{L}_{\text{ood}}^{\text{gate}}$, and the corresponding stage-level objectives.

Table 20: **Stabilization–allocation control on LibriSeVoc.** Entries are three-seed mean Robust AUC /ECE. The ECRM activation epoch (60%) is identical in all cells; without stabilization, no earlier security stage is active.

	Conditioned	**Uniform** ($g_i \equiv 1$)
With stabilization	**0.90 / 0.10**	0.83 / 0.19
Without stabilization	0.82 / 0.20	0.78 / 0.25

Conditioning improves Robust AUC by 0.07 with stabilization and by 0.04 without it; stabilization improves Robust AUC by 0.08 under conditioned allocation and by 0.05 under uniform allocation. The resulting descriptive difference-in-differences is $+0.03$, and ECE follows the same ordering: conditioning lowers ECE from 0.19 to 0.10 with stabilization and from 0.25 to 0.20 without it. Thus, both factors contribute, with a larger observed conditioning benefit after stabilization. Because only three seeds are available, we treat this interaction descriptively rather than as a formal significance claim.

The $1.4\times$ gate span in App. B.6 is instantaneous rather than cumulative. During the final 40% of training, the normalized weights are applied at every optimization step to both $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{ood}}^{\text{gate}}$. Stop-gradient blocks gradients from the gate weights through E, but $\mathcal{L}_{\text{ood}}^{\text{gate}}$ itself remains differentiable through the detector-derived score. The factorial control therefore establishes the end-to-end stabilization–allocation pattern; it does not attribute its magnitude to a single gradient path.

E.3 Cumulative Schedule Comparison

Tab. 21 compares all six cumulative activation schedules for energy anchoring, OOD consistency, and adversarial robustness, together with simultaneous optimization. In every sequential configuration, objectives remain active after introduction at the fixed $20/40/60\%$ boundaries. Consequently, permuting the sequence changes both activation order and cumulative exposure, so this experiment should be interpreted as a schedule-level comparison rather than a pure order intervention.

Among the tested schedules, $\mathcal{L}_{\text{energy}} \to \mathcal{L}_{\text{ood}} \to \mathcal{L}_{\text{adv}}{+}ECRM$ obtains the highest Robust AUC. OOD-AUROC is reported descriptively and is not used to rank schedules or infer $\mathcal{L}_{\text{ood}}$-induced invariance. The results are interpreted at the complete-system level rather than as an order-only intervention.

All schedule-comparison results are three-seed means and are interpreted descriptively rather than as formal significance tests.

E.4 Per-Condition Robustness and Ablation Transfer

Tab. 22 decomposes PGD-10 robustness by corpus and tests whether the principal robustness controls retain their relative effect beyond the source-domain LibriSeVoc evaluation. All attacks use $\epsilon = 10^{-3}$ with no corpus-specific adaptation or retuning.

Full ASNet exceeds the No-$\mathcal{L}_{\text{rob}}$ control by 0.05–0.09 and the Uniform control by 0.03–0.05 on every external corpus. The corresponding external-mean gains are $+0.068$ and $+0.040$, respectively, with the Full-ASNet mean of 0.650 matching Tab. 2. Thus, the robustness block retains a measurable effect under external shift, and the consistent Full-over-Uniform gap shows that the complete energy-conditioned ECRM pipeline outperforms matched uniform allocation across all evaluated external corpora. Absolute robustness nevertheless remains below the LibriSeVoc value of 0.90, exposing the remaining cross-corpus robustness gap. These three-seed differences are interpreted descriptively.

Table 21: **Cumulative schedule comparison on LibriSeVoc.** All six permutations use the same architecture, 80-epoch budget, and $20/40/60\%$ activation boundaries, with objectives remaining active after introduction; an all-simultaneous control is also included. Thus, permutations change both activation order and cumulative exposure.

Cumulative activation schedule	AUC↑	Robust AUC↑	OOD-AUROC	Diagnostic / interpretation
Eng → OOD → Adv (ours)	**0.999**	**0.90**	0.74	—
OOD → Eng → Adv	0.991	0.84	0.63	$\mathcal{L}_{\text{ood}}$ precedes energy anchoring; KL remains 0.28
Adv → Eng → OOD	0.992	0.83	0.64	ECRM is activated before energy anchoring and OOD consistency
Eng → Adv → OOD	0.990	0.85	0.61	ECRM is activated before OOD consistency; OOD-AUROC is 0.61
OOD → Adv → Eng	0.987	0.81	0.59	Adversarial robustness precedes energy anchoring
Adv → OOD → Eng	0.985	0.78	0.58	Both robustness and OOD consistency precede energy anchoring
All simultaneous	0.994	0.85	0.69	No staged activation schedule

Table 22: **Per-condition PGD-10 robustness and ablation transfer.** No-$\mathcal{L}_{\text{rob}}$ retains the full dual-view architecture, learned fusion/CMRA, and energy stabilization but removes the adversarial/ECRM robustness block. Uniform retains the same architecture and stabilization but replaces energy-conditioned weights with $g_i \equiv 1$. These rows are matched counterfactuals to Full ASNet and are distinct from the cost-oriented Lite/Mid variants in App. E.7. Entries are three-seed means; the external mean is unweighted over the four external corpora.

Dataset	Full ASNet	No $\mathcal{L}_{\text{rob}}$	Uniform	Full − No $\mathcal{L}_{\text{rob}}$	Full − Uniform
LibriSeVoc	**0.90**	0.81	0.83	**+0.09**	**+0.07**
ASVspoof21	**0.69**	0.64	0.66	**+0.05**	**+0.03**
WaveFake	**0.71**	0.65	0.67	**+0.06**	**+0.04**
FakeAVCeleb	**0.64**	0.55	0.59	**+0.09**	**+0.05**
CodecFake	**0.56**	0.49	0.52	**+0.07**	**+0.04**
External-corpus mean	**0.650**	**0.583**	**0.610**	**+0.068**	**+0.040**

E.5 CMRA Selection and Cross-Domain Mechanism Controls

CMRA margin selection. We use $m = 0.8$ in Eq. 3. The margin is selected once from the upper tail of cross-view cosine similarity on correctly classified clean LibriSeVoc validation samples and is then fixed across datasets and test conditions. Thus, CMRA introduces no corpus-specific or test-time margin retuning.

Cross-domain mechanism controls. We next test whether the reliability gains persist beyond the source-domain LibriSeVoc evaluation. Tab. 23 removes broader mechanism groups while all models remain trained only on LibriSeVoc and are evaluated without adaptation on ASVspoof21 and CodecFake. These controls are grouped counterfactuals rather than one-factor CMRA ablations: they test structured cross-view interaction, energy stabilization, and adversarial/ECRM training under the same external protocol.

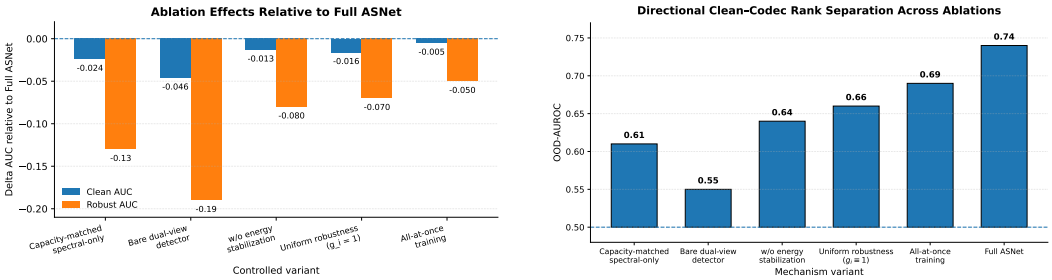
Full ASNet performs best on both external corpora, while every grouped removal lowers AUC and increases EER. Removing adversarial/ECRM training reduces AUC from 0.892 to 0.828 on ASVspoof21 and from 0.739 to 0.642 on CodecFake; removing energy stabilization yields 0.817 and 0.611, respectively. Replacing structured cross-view interaction with concatenation gives $0.801/0.603$, while the dual-view architecture alone reaches only $0.751/0.553$.

These results provide a cross-domain complement to Tab. 3: neither the additional evidence stream nor any single mechanism group accounts for the full external performance. Rather, the strongest observed result is obtained when structured cross-view interaction, energy stabilization, and adversarial/ECRM training are combined.

Table 23: **Cross-domain mechanism ablation.** AUC/EER on ASVspoof21 and CodecFake under LibriSeVoc-only training and adaptation-free evaluation. All variants retain the dual spectral–prosodic encoders; rows remove broader mechanism groups rather than isolating individual components. ✓ = active; ✗ = disabled.

Counterfactual	Active mechanism groups			ASVspoof21		CodecFake	
	Fusion/CMRA	Energy stab.	Adv./ECRM	AUC ↑	EER ↓	AUC ↑	EER ↓
Full ASNet	✓	✓	✓	**0.892**	**0.158**	**0.739**	**0.292**
No adversarial/ECRM robustness	✓	✓	✗	0.828	0.205	0.642	0.341
No learned fusion; concat only	✗	✓	✓	0.801	0.231	0.603	0.373
No energy stabilization	✓	✗	✓	0.817	0.214	0.611	0.361
Dual-view architecture only	✗	✗	✗	0.751	0.296	0.553	0.412

E.6 Ablation Visualizations



(a) ΔAUC and ΔRobust AUC vs. Full ASNet (b) Raw OOD-AUROC under clean–codec shift

Figure 5: **Mechanism-level ablation diagnostics.** Panel (a) reports changes in clean AUC and LibriSeVoc Robust AUC relative to Full ASNet. Panel (b) reports raw OOD-AUROC values for the same variants; the dashed 0.5 line denotes no directional rank separation. The bars are referenced to this neutral level only for visualization. OOD-AUROC is treated descriptively and is not used to rank variants or assess stabilization. All values correspond to Tab. 3.

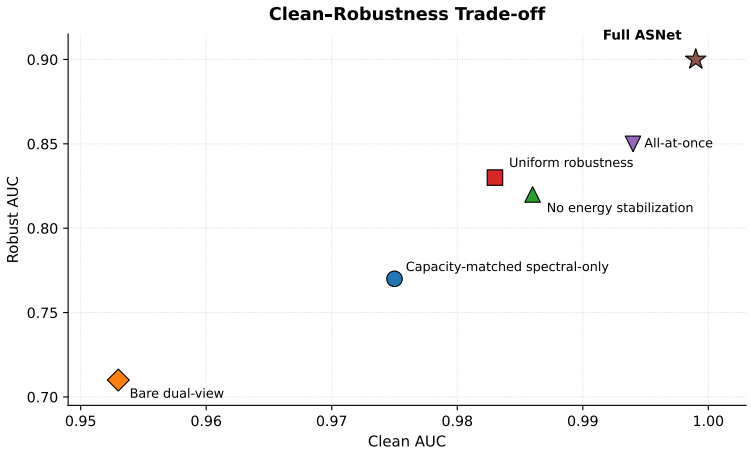


Figure 6: **Clean detection and adversarial robustness across controlled ablations.** Each point is a variant from Tab. 3: the horizontal axis reports clean AUC and the vertical axis reports LibriSeVoc Robust AUC under PGD-10 at $\epsilon = 10^{-3}$. Using the reported three-seed means, Full ASNet lies above and to the right of every tested control. This comparison is restricted to the evaluated variants.

Figs. 5 and 6 summarize the controlled ablations. Only AUC and Robust AUC are interpreted as performance metrics with a preferred direction; OOD-AUROC remains a directional diagnostic and is not used for model ranking or stabilization claims.

E.7 Complexity–Performance Trade-off

Tab. 24 reports cost-oriented simplified variants rather than matched mechanism ablations. ASNet-Mid retains prosody and energy stabilization but omits learned fusion, CMRA, and the adversarial/ECRM block, yielding lower model and training cost. By contrast, the No-$\mathcal{L}_{\mathrm{rob}}$ control in Tab. 22 retains the full dual-view architecture, learned fusion/CMRA, and energy stabilization while removing only the adversarial/ECRM robustness block. Thus, the 0.84 versus 0.81 ordering is not used to attribute a causal effect to fusion or CMRA. It instead shows that their contribution is configuration dependent rather than monotonically beneficial in isolation. Mechanism claims are therefore based on the matched controls in Tabs. 3 and 22, while Lite/Mid characterize cost–reliability operating points.

Table 24: **Simplified ASNet variants.** Cost $(P/T/I)$ denotes parameter count, training-time cost, and inference-time cost normalized by Full ASNet. In this table, "ECRM" denotes the complete adversarial/ECRM robustness block; ✗ therefore means that neither energy-conditioned nor uniform adversarial robustness training is applied. ✓ = active; ✗ = inactive.

Variant	Active Mechanism Groups					Robust AUC↑	OOD-AUROC	ECE↓	Cost (P/T/I)
	Prosody	Energy Stab.	Fusion	CMRA	ECRM				
`ASNet-Lite`	✗	✗	✗	✗	✗	0.76	0.58	0.23	$0.62/0.58/0.73$
`ASNet-Lite+P`	✓	✗	✗	✗	✗	0.81	0.60	0.21	$0.79/0.70/0.86$
`ASNet-Mid`	✓	✓	✗	✗	✗	0.84	0.64	0.17	$0.87/0.81/0.92$
Full ASNet	✓	✓	✓	✓	✓	**0.90**	0.74	**0.10**	$1.00/1.00/1.00$

E.8 Conditioning-Signal Comparison

The original validation sweep informed the choice of conditioning signal and is therefore reported only as design-time evidence. To evaluate that choice independently, we subsequently fix the architecture, objectives, optimizer, schedule, training budget, checkpoint rule, hyperparameters, and validation-derived abstention thresholds using only LibriSeVoc train/validation data, and rerun five conditioning signals with identical paired seeds. Mahalanobis statistics are also estimated using train/validation features only; neither held-out corpus is used for signal selection, threshold tuning, checkpoint selection, or adaptation.

We compare energy with logit margin, predictive entropy, Mahalanobis distance on $\mathbf{z}^*$, and a learned auxiliary confidence head. For a fixed binary detector, energy, logit margin, and predictive entropy are monotone-equivalent uncertainty rankings and therefore yield the same retained set at matched coverage, up to ties. ECRM, however, acts on score magnitude rather than rank alone, so monotone-equivalent scores need not produce the same normalized training weights. The variants are separately retrained with each signal; their differences therefore reflect training-time conditioning rather than post-hoc ranking. Tab. 25 reports the design-time validation sweep, whereas Tab. 26 reports the independent held-out comparison.

Table 25: **Design-time conditioning-signal comparison on LibriSeVoc validation.** This sweep informed model development and is reported only as design-time evidence. All variants otherwise share the same architecture, training schedule, ECRM infrastructure, and abstention rule. "Catch" is the fraction of misclassified samples rejected at 20% abstention.

Signal	Type	Robust AUC ↑	OOD-AUROC	ECE ↓	Catch↑	Residual↓
Logit margin	logit-derived	0.821	0.683	0.061	0.70	0.30
Predictive entropy	logit-derived	0.828	0.691	0.058	0.73	0.27
Mahalanobis on $\mathbf{z}^*$	non-logit	0.832	0.700	0.056	0.74	0.26
Auxiliary confidence head	non-logit (learned)	0.838	0.706	0.053	0.76	0.24
Energy	class-symmetric	**0.857**	0.734	**0.041**	**0.81**	**0.19**

Table 26: **Independent held-out conditioning-signal comparison.** All model and training choices are fixed using only LibriSeVoc train/validation data before test evaluation. Each signal defines a separately trained variant with identical paired seeds; variants differ only in the scalar conditioning/abstention signal. Thresholds target 20% abstention on LibriSeVoc validation and are transferred unchanged. "Cov." denotes realized test coverage and "Catch" the fraction of PGD-10 errors rejected. Entries are three-seed means.

Signal	Robust AUC↑	OOD-AUROC	ECE↓	Cov. (%)	Catch (%)↑
LibriSeVoc test					
Logit margin	0.84	0.68	0.17	80.4	65.1
Predictive entropy	0.85	0.70	0.16	79.8	67.6
Mahalanobis on $\mathbf{z}^*$	0.86	0.71	0.15	78.9	70.8
Auxiliary confidence head	0.88	0.73	0.14	80.2	73.9
Energy	**0.90**	0.72	**0.11**	80.5	**80.1**
ASVspoof21 (thresholds unchanged)					
Logit margin	0.61	0.61	0.24	72.4	44.1
Predictive entropy	0.63	0.64	0.22	74.0	47.8
Mahalanobis on $\mathbf{z}^*$	0.64	0.66	0.21	70.8	50.2
Auxiliary confidence head	0.65	0.69	0.18	75.1	54.6
Energy	**0.67**	0.68	**0.16**	75.6	**60.3**

On the held-out comparison, energy improves Robust AUC by 0.02 over the auxiliary confidence head on both LibriSeVoc and ASVspoof21, reduces ECE by 0.03 and 0.02, and increases PGD-error catch by 6.2 and 5.7 percentage points, respectively. OOD-AUROC differs by 0.01 between energy and the auxiliary head on each corpus and is not used to rank conditioning signals. Across the three paired seeds, all six energy-minus-auxiliary Robust AUC differences are positive (0.017–0.022); with three seeds, this is interpreted as descriptive sign consistency rather than formal statistical significance.

The dedicated energy rerun reproduces the main Robust AUC value of 0.90. Its OOD-AUROC/ECE/PGD-error-catch values are $0.72/0.11/80.1\%$, compared with $0.74/0.10/85\%$ in the final Full-ASNet evaluation. These are separately trained evaluations under the same metric definition rather than duplicate measurements from the same trained models. The held-out study is therefore used for within-study conditioning-signal comparison, not to replace the final Full-ASNet diagnostics. Thus, it supports energy as the strongest observed training-time conditioning choice among the tested variants without implying superior post-hoc uncertainty ranking.