Backbone-Equated Diffusion OOD via Sparse Internal Snapshots

Yadang Alexis Rouzoumka[1,2] **Jean Pinsolle**[1] **Eugénie Terreaux**[2]

Christèle Morisseau[2] **Jean-Philippe Ovarlez**[1,2] **Chengfang Ren**[1]

[1]**SONDRA, CentraleSupélec, Université Paris-Saclay, 91190 Gif-sur-Yvette, France**
[2]DEMR, ONERA, Université Paris-Saclay, 91120 Palaiseau, France
yadang-alexis.rouzoumka@centralesupelec.fr

Abstract

Fair comparison between diffusion-based OOD detectors is challenging, as conclusions can vary with backbone choice, corruption parameterization, and test-time budget. We address this issue through a *Mutualized Backbone-Equated* (MBE) protocol that aligns canonical corruption levels and logical test-time cost across diffusion backbones. Within this setting, we introduce *Canonical Feature Snapshots* (CFS), a family of detectors that probes a frozen diffusion backbone using only a tiny number of native internal states at canonical low-noise levels. On a controlled CIFAR-scale benchmark, a single late-decoder snapshot is already highly competitive, while adding a deep encoder snapshot improves the average worst-shift AUROC on both backbones at the same one-forward cost. Matched reconstruction and non-diffusion representation controls, together with controlled SCOPED comparisons, support the conclusion that the gain is tied to learned sparse diffusion representations rather than hidden test-time budget or a high-capacity downstream head. We further provide a local diagnostic theory explaining these observations through conditional encoder-decoder complementarity, diagonal-score separation, and low-noise corruption stability. The official implementation is available at https://github.com/RouzAY/cfs-diffusion-ood/.

1 Introduction

Out-of-distribution (OOD) detection asks whether a model can identify inputs that fall outside the regime where its predictions should be trusted. In diffusion models, once the backbone is frozen, the same image can be probed across a structured family of corruption levels, producing a hierarchy of hidden states before the final denoising output.

Most diffusion-based OOD methods probe this backbone in *output space*: through reconstructions, denoising residuals, trajectories, output-side geometry, or posterior-consistency scores [Mahmood et al., 2021, Graham et al., 2023, Liu et al., 2023, Gao et al., 2023, Heng et al., 2024, Rouzoumka et al., 2026, Barkley et al., 2026, Shoushtari et al., 2026]. This has produced useful detectors, but it leaves open a more basic question: Within a frozen diffusion backbone, where does the most useful information for OOD detection reside?

We argue that this question is obscured by two coupled issues. The first is *protocol confounding*: in diffusion OOD, conclusions can depend strongly on checkpoint family (e.g., DDPM vs. EDM), corruption coordinates, and test-time budget, much as protocol choices already matter in post-hoc OOD benchmarking more broadly [Yang et al., 2022, Zhang et al., 2024]. The second is representation mismatch: output-space summaries rely on compressed readouts and may miss discriminative

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

information still present in internal states, consistent with the growing view of diffusion models as representation learners [Yang and Wang, 2023, Luo et al., 2023, Yu et al., 2025].

We study this question under a controlled evaluation setting. First, we introduce a *Mutualized Backbone-Equated* (MBE) protocol that aligns checkpoint-family policy, canonical corruption levels, and logical test-time cost across diffusion backbones. Second, within this setting, we propose *Canonical Feature Snapshots* (CFS), a family of OOD detectors that probes a frozen diffusion backbone through a tiny number of aligned internal activations.

On a controlled CIFAR-scale benchmark, the late-decoder $\mathrm{CFS}_{\mathrm{dec}}(1\times1)$ variant is already highly competitive. Adding a deep encoder snapshot in $\mathrm{CFS}(1\times2)$ slightly improves average AUROC on improved-diffusion but slightly reduces it on EDM, while improving $\mathrm{AvgWorstAUROC}$ on both backbones at the same one-forward cost. These results indicate that useful relative-OOD signal is concentrated in a small number of native internal snapshots.

To explain these trends, we develop a measurable local-testing view of sparse diffusion probing. It yields three diagnostic principles: conditional encoder-decoder complementarity, diagonal-score separation, and low-noise corruption stability. The theory is local rather than universal, but it produces directly estimable quantities for selecting hooks, levels, and encoder-decoder pairings, and we show empirically that these diagnostics track downstream OOD behavior across both improved-diffusion and EDM backbones.

Contributions.

- **A controlled protocol and sparse detector.** We formulate diffusion OOD comparison under a *Mutualized Backbone-Equated* (MBE) protocol, and introduce *Canonical Feature Snapshots* (CFS), a detector family that probes frozen diffusion backbones through a tiny number of aligned internal activations.
- **Evidence that the OOD signal is highly concentrated.** Under MBE, a single late-decoder snapshot is already highly competitive, while a paired encoder-decoder readout improves $\mathrm{AvgWorstAUROC}$ on both primary backbones at the same one-forward cost.
- **A local diagnostic explanation.** We develop a measurable local-testing view that explains conditional encoder-decoder complementarity, diagonal-score separation, and low-noise stability, and show that its diagnostics track empirical behavior.

2 Related Work and Positioning

Post-hoc OOD detection and protocol sensitivity. A large literature studies post-hoc OOD scores for pretrained discriminative models, including Mahalanobis distances, energy scores, activation shaping, virtual-logit matching, and nearest-neighbor geometry [Lee et al., 2018, Liu et al., 2020, Sun et al., 2021, Wang et al., 2022, Sun et al., 2022, Lu et al., 2025]. Benchmarking efforts such as OpenOOD and OpenOOD v1.5 show how strongly protocol choices can affect conclusions [Yang et al., 2022, Zhang et al., 2024]. These concerns are even sharper in diffusion OOD, where checkpoint family, corruption coordinates, and test-time budget introduce additional confounders.

Diffusion OOD in output space. Most diffusion-based OOD methods probe denoisers through output-space quantities or output-derived summaries. MSMA aggregates multiscale score norms [Mahmood et al., 2021]; DDPM-OOD and LMD use denoising, reconstruction, or inpainting behavior [Graham et al., 2023, Liu et al., 2023]; likelihood-based diffusion OOD has also been studied [Goodier and Campbell, 2023]; DiffGuard adds conditional guidance [Gao et al., 2023]; and DIFFPATH summarizes denoising trajectories from a single unconditional diffusion backbone [Heng et al., 2024]. Other methods use output-side geometric or consistency structure: GEPC measures transformation-induced posterior-consistency violations in denoiser outputs [Rouzoumka et al., 2026], while SCOPED and EigenScore study output-side geometry or uncertainty [Barkley et al., 2026, Shoushtari et al., 2026]. Despite their differences, these methods share the same basic viewpoint: the OOD signal is extracted primarily from denoiser outputs, reconstructions, trajectories, consistency residuals, or output-side geometry.

Diffusion models as representation learners. A parallel literature views diffusion models as representation learners. Pretrained diffusion backbones provide useful internal features for downstream

Table 1: **Positioning relative to diffusion-based OOD detection.** Our distinction is not merely that internal features can help, but that under a shared-source, backbone-equated protocol, a tiny number of sparse native frozen snapshots already captures a strong relative-OOD signal.

Method	Primary probe	Native internal features	Reconstruction / guidance	Shared-source protocol
MSMA [Mahmood et al., 2021]	multiscale output descriptor	✗	✗	✗
DDPM-OOD [Graham et al., 2023]	reconstruction / manifold	✗	✓	✗
LMD [Liu et al., 2023]	inpainting / reconstruction	✗	✓	✗
DiffGuard [Gao et al., 2023]	guided output probing	✗	✓	✗
DIFFPATH [Heng et al., 2024]	denoising trajectory	✗	✗	✗
GEPC [Rouzoumka et al., 2026]	output consistency	✗	✗	✗
SCOPED [Barkley et al., 2026]	score geometry	✗	✗	✗
EIGENSCORE [Shoushtari et al., 2026]	covariance geometry	✗	✗	✗
DLSR [Yang et al., 2024]	feature reconstruction	✓	✓	✗
CFS (ours)	sparse native snapshots	✓	✗	✓

tasks [Yang and Wang, 2023], and semantically meaningful descriptors can be consolidated from multi-layer and multi-timestep states [Luo et al., 2023]. Related work also suggests encoder/decoder asymmetry: Faster Diffusion reports that encoder activations vary less across denoising steps than decoder activations [Li et al., 2024]. While not an OOD result, this supports the idea that different internal states play different functional roles. REPA further argues that strong hidden representations are important for generation quality [Yu et al., 2025]. Recent OOD work also supports representation-space modeling beyond raw-pixel likelihoods [Ding et al., 2025, Järve et al., 2025]. These works motivate our representation-first viewpoint.

Positioning. We study diffusion OOD under a *Mutualized Backbone-Equated* protocol and ask: under a shared-source, backbone-equated, budget-accounted comparison, how much relative-OOD signal is already present in a tiny number of sparse native frozen snapshots? CFS answers this question with a deliberately minimal detector family: no reconstruction module, no guidance machinery, no reverse-path recursion, and no high-capacity downstream head. Table 1 summarizes CFS positioning.

3 Mutualized Backbone-Equated Protocol (MBE)

OOD benchmarking has shown that evaluation details can dominate perceived progress [Yang et al., 2022, Zhang et al., 2024]. In diffusion OOD, this problem is amplified by additional degrees of freedom, including checkpoint family, corruption parameterization, and hidden test-time budget. We therefore evaluate all methods under a shared protocol designed to control these sources of variation.

3.1 Canonical corruption and cross-backbone alignment

A first ingredient of MBE is a shared corruption view across backbone families. Diffusion backbones evaluate corrupted versions of an input across a family of noise levels [Ho et al., 2020, Nichol and Dhariwal, 2021, Karras et al., 2022], but their native interfaces differ: some expose discrete timesteps t, while others use continuous noise scales such as σ. To compare OOD detectors across backbones, we therefore work with a common canonical corruption parameterization.

For a clean sample $\mathbf{x}_0$, we write corruption as

$$\mathbf{x}_\lambda = a(\lambda)\mathbf{x}_0 + b(\lambda)\boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, I), \tag{1}$$

and use the canonical coordinate, with a slight abuse of notation,

$$\lambda := \log \frac{a(\lambda)^2}{b(\lambda)^2}, \tag{2}$$

i.e., logSNR. Large λ corresponds to cleaner observations and small λ to noisier ones.

For improved-diffusion checkpoints, $a_t = \sqrt{\bar{\alpha}_t}$, $b_t = \sqrt{1-\bar{\alpha}_t}$, and $\lambda_t = \log \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}$, so a desired canonical level is matched to the nearest native timestep in logSNR space. EDM-style backbones instead expose a continuous noise input; our adapter maps each canonical λ to the appropriate backbone-specific model input and returns coefficients satisfying Eq. (1). Canonicalization matters because many diffusion OOD methods are intrinsically multilevel: without a shared corruption coordinate, methods may be compared at levels that are not actually matched in corruption strength.

Table 2: **Operational glossary.** Terms used throughout the method and experiments.

Term	Operational definition
Canonical level	A common noise scale shared across diffusion models.
Native level	The timestep or noise value used by a model to match a canonical level.
Hook	A fixed point in the model where an intermediate output is captured without changing the model.
Snapshot	A pooled intermediate output captured from one model location in a single forward pass.
CFS(1×1)	One noise level and one snapshot, here from the late decoder.
CFS(1×2)	One noise level and two snapshots, typically a deep encoder and a late decoder.
Slot	The model region from which a snapshot is taken, such as a late encoder or decoder block.
Head	The ID-only rule used to score pooled features, such as a diagonal score or KNN.
Logical cost	Counts of required model operations, together with measured runtime.

Prespecified operating point and ID-only selection. For one-level CFS, the canonical probing level $\lambda = 5$ was fixed *a priori*, before inspecting the reported OOD test results. It is a low-noise point that preserves most image content while still invoking a non-trivial denoising computation. Multi-level baselines are evaluated on the same shared canonical interval $[-10, 5]$, mapped into each backbone's native conditioning. This choice is separate from hook selection: the ID-only stability proxy in Section 4 ranks candidate model locations after the canonical level has been fixed; it does not select λ. Feature statistics, hook ranking, ID-bank fitting, and score calibration use ID data only. Canonical-level sweeps are reported only as sensitivity analyses.

3.2 Controlled comparison under MBE

Under MBE, a method is evaluated under the same source-family policy, preprocessing, canonical corruption levels, ID/OOD splits, and logical test-time cost as its competitors. The purpose of MBE is scientific comparison, not per-backbone peak tuning. A diffusion OOD method can otherwise appear stronger for reasons unrelated to its scoring rule: a better-matched checkpoint, a different input normalization, a different corruption coordinate, or a larger hidden test-time budget.

Concretely, MBE routes all methods through the same canonical corruption semantics and shared adapter interface, and evaluates them under the same split policy and logical budget accounting. It does not force identical native implementations. Detailed canonicalization, adapter outputs, and baseline implementation taxonomy are deferred to Appendix C and Appendix D.

For discrete backbones, mapping a continuous logSNR grid to native timesteps can produce duplicates. We therefore distinguish the candidate grid resolution K_{grid} from the number of effective canonical levels K_c actually used by a method; Appendix C.3–C.4 gives the precise construction.

We report logical test-time cost by keeping distinct primitive model operations explicit,

$$\text{Cost}_m = \#F_m\, F + \#J_m\, J,$$

where F denotes one backbone forward evaluation and J one Jacobian-vector operation. We therefore write costs such as $1F$, $10F$, or $2F + 2J$, and complement logical counts with measured runtime where relevant rather than treating different primitives as numerically equivalent.

4 Method: Canonical Feature Snapshots (CFS)

4.1 Canonical Feature Snapshots

We ask whether a tiny number of native frozen internal activations already captures useful OOD signal once protocol confounders are removed. Let $P_\star$ denote the source distribution used to train a frozen diffusion checkpoint, and let P and Q denote the evaluation ID and OOD datasets. We do not treat the checkpoint as an OOD oracle for $P_\star$; instead, we use it as a frozen representation map and define OOD relative to the evaluation reference bank P.

A CFS instance is specified by a small set of canonical levels Λ, a small set of native internal hooks $\mathcal{H}$, and an ID-only scoring head on the resulting pooled slot descriptors. For each $\lambda \in \Lambda$, we form

$$\mathbf{x}_\lambda = a(\lambda)\mathbf{x}_0 + b(\lambda)\boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

run one forward pass through the frozen diffusion backbone, and pool each selected activation $\mathbf{u}_{\lambda,h}(\mathbf{x}_\lambda)$ into a descriptor

$$\mathbf{z}_{\lambda,h}(\mathbf{x}_0) := \text{Pool}(\mathbf{u}_{\lambda,h}(\mathbf{x}_\lambda))\,.$$

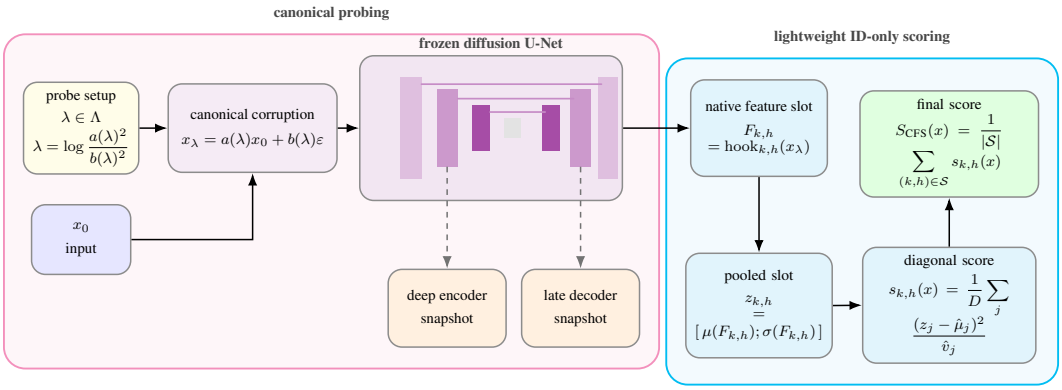

Figure 1: **Overview of CFS.** The input is corrupted at a canonical level, processed by a frozen diffusion U-Net, probed through a small number of native internal snapshots, and scored with a lightweight ID-only head.

The sparse representation is

$$\mathbf{\Phi}(\mathbf{x}_0) = \left[\mathbf{z}_{\lambda,h}(\mathbf{x}_0)\right]_{(\lambda,h)\in\mathcal{S}}, \qquad \mathcal{S} \subseteq \Lambda \times \mathcal{H}.$$

OOD detection then reduces to fitting ID-only slot statistics on $\mathbf{\Phi}(P)$ and scoring deviation from that reference geometry.

The paper focuses on two operating points: $\text{CFS}(1\times2)$, which uses one deep encoder and one late decoder hook at a single low-noise level, and $\text{CFS}_{\text{dec}}(1\times1)$, a decoder-only variant. Because all retained hooks at a fixed level are captured within the same backbone forward pass, logical cost depends only on the number of selected levels:

$$\#F_{\text{CFS}} = |\Lambda|, \qquad \#J_{\text{CFS}} = 0.$$

Figure 1 summarizes this pipeline.

4.2 A local diagnostic view

Our theory is local and diagnostic rather than universal: it asks which frozen internal statistics should be most useful for testing membership in the evaluation ID reference distribution, and focuses on quantities that can be measured directly on held-out data.

Fix a canonical level λ, and let $\mathbf{z}_{d,\lambda}$ and $\mathbf{z}_{e,\lambda}$ denote selected late-decoder and deep-encoder snapshots. We use the local testing approximation

$$\mathbf{z}_\lambda \mid H_0 \sim \mathcal{N}(\boldsymbol{\mu}_\lambda, \mathbf{\Sigma}_\lambda), \qquad \mathbf{z}_\lambda \mid H_1 \sim \mathcal{N}(\boldsymbol{\mu}_\lambda + \mathbf{\Delta}_\lambda, \mathbf{\Sigma}_\lambda), \tag{3}$$

where $H_0 : \mathbf{x}_0 \sim P$, $H_1 : \mathbf{x}_0 \sim Q$, and $\mathbf{z}_\lambda = [\mathbf{z}_{d,\lambda}^\top, \mathbf{z}_{e,\lambda}^\top]^\top$. This is a local model on pooled representations, not a global image model.

Theorem 1 (Conditional encoder-decoder complementarity)**.** *Under Eq. (3), assume $\mathbf{\Sigma}_\lambda \succ 0$. Paired separation equals decoder separation plus a nonnegative conditional encoder residual. Hence $\text{Sep}_{\text{pair}}(\lambda) \geq \text{Sep}_{\text{dec}}(\lambda)$, with equality when the encoder shift is fully explained by the decoder shift. The full expression is given in Appendix B.3.*

This is an oracle separation result under the joint covariance model. The practical CFS score uses estimated diagonal statistics and uniform slot averaging. The benefit of adding an encoder is therefore evaluated empirically.

Proposition 1 (Low-noise corruption stability)**.** *Under a first-order mean-square expansion with bounded local feature sensitivity, corruption-induced within-image feature variance vanishes as the noise amplitude tends to zero. Its leading-order expression depends on both the injected noise variance and the local feature sensitivity. The assumptions and proof are given in Appendix B.4.*

This supports low-noise probing as a means of controlling stochastic feature instability under the stated assumptions.

Finally, the local Gaussian model provides a separation interpretation for diagonal scoring through the noncentrality diagnostic κ_λ/d_λ. We empirically examine this diagnostic and the content-to-instability ratio $R_h(\lambda)$ in relation to sparse-probe quality. The theoretical analysis and empirical diagnostics are presented in Appendices B.2 and B.7.

4.3 Hook selection and lightweight scoring

We hook *block outputs* rather than arbitrary leaf submodules. This improves portability across architectures and avoids unstable low-level activations. The local theory suggests two priorities: late decoder states are the primary candidates, and a deep encoder hook, when used, should be interpreted as complementary residual information.

Within each structural region, we rank a small shortlist of candidate hooks using an ID-only feature-variation proxy. The implementation supports normalization by within-image corruption variance when multiple corruption repeats are used. In the reported configuration, hook selection uses one corruption repeat and therefore ranks candidates by between-image feature variance alone. The exact estimator and configuration are given in Appendix J.1. Hook selection is performed after the canonical probing level has been fixed and does not select λ.

For a selected slot (k, ℓ), let

$$\mathbf{F}_{k,\ell}(\mathbf{x}_0) \in \mathbb{R}^{C_{k,\ell} \times H_{k,\ell} \times W_{k,\ell}}$$

denote the hooked feature map. We use the pooled descriptor

$$\mathbf{z}_{k,\ell}(\mathbf{x}_0) = [\mathrm{Mean}_{\mathrm{sp}}(\mathbf{F}_{k,\ell}(\mathbf{x}_0))^\top, \mathrm{Std}_{\mathrm{sp}}(\mathbf{F}_{k,\ell}(\mathbf{x}_0))^\top]^\top \in \mathbb{R}^{D_{k,\ell}}, \tag{4}$$

where $\mathrm{Mean}_{\mathrm{sp}}$ and $\mathrm{Std}_{\mathrm{sp}}$ are channel-wise spatial mean and standard deviation. Using only ID-train data, we fit diagonal statistics $(\hat{\boldsymbol{\mu}}_{k,\ell}, \hat{\mathbf{v}}_{k,\ell})$. The slot score is

$$s_{k,\ell}(\mathbf{x}_0) = \frac{1}{D_{k,\ell}} \sum_{j=1}^{D_{k,\ell}} \frac{\left(\mathbf{z}_{k,\ell}^{(j)}(\mathbf{x}_0) - \hat{\boldsymbol{\mu}}_{k,\ell}^{(j)}\right)^2}{\hat{\mathbf{v}}_{k,\ell}^{(j)}}, \tag{5}$$

and the final score is

$$S_{\mathrm{CFS}}(\mathbf{x}_0) = \frac{1}{|\mathcal{S}|} \sum_{(k,\ell)\in\mathcal{S}} s_{k,\ell}(\mathbf{x}_0). \tag{6}$$

We intentionally use a lightweight diagonal score: CFS is meant to test the quality of sparse frozen representations, not to rely on a high-capacity downstream classifier. Alternative ID-only heads are reported in Appendix E.9. We examine $\hat{\kappa}_\lambda$ and $\hat{R}_h(\lambda)$ as empirical diagnostics of sparse-probe quality.

5 Experiments

5.1 Setup

Our main benchmark is CIFAR-scale, where protocol control is the cleanest. We use

$$\mathcal{I}_{\mathrm{small}} = \{\mathrm{CIFAR\text{-}10}, \mathrm{SVHN}, \mathrm{CelebA32}\},$$
$$\mathcal{O}_{\mathrm{small}} = \{\mathrm{CIFAR\text{-}10}, \mathrm{SVHN}, \mathrm{CelebA32}, \mathrm{CIFAR\text{-}100}, \mathrm{DTD}\}.$$

For each $i \in \mathcal{I}_{\mathrm{small}}$, all datasets in $\mathcal{O}_{\mathrm{small}} \setminus \{i\}$ are treated as OOD, yielding 12 ID→OOD pairs per backbone.

We evaluate two diffusion backbone families through the same adapter and canonical corruption interface: improved-diffusion backbones with discrete timesteps and native ε-prediction, and EDM-family backbones with continuous noise conditioning and one-shot $\hat{x}_0$ estimation. In the main benchmark, all methods use a shared CIFAR-10 checkpoint family; we then repeat the same benchmark with a shared CelebA32 checkpoint family for source-family robustness. A checkpoint-controlled ImageNet-scale stress test, including the full per-shift breakdown, is reported in Appendix I.

5.2 Baselines and metrics

All baselines are evaluated under the shared MBE pipeline. Our main comparators are MSMA [Mahmood et al., 2021], DIFFPATH [Heng et al., 2024], DDPM-OOD [Graham et al., 2023], GEPC [Rouzoumka et al., 2026], and SCOPED [Barkley et al., 2026]. We also include a matched one-forward reconstruction control that reads only the final denoised output from the same forward computation used by CFS. EIGENSCORE is retained in external positioning rather than the low-budget controlled table because its reported evaluation requires roughly $300F$. Detailed baseline taxonomy, native-rerun checks, and implementation choices are deferred to Appendix C and Appendix D.

We report both one-forward sparse operating points throughout: $\mathrm{CFS}_{\mathrm{dec}}(1\times1)$, the minimal late-decoder detector, and $\mathrm{CFS}(1\times2)$, the paired encoder-decoder variant. In the primary benchmark, their average-AUROC ordering depends on the backbone, while the paired variant has higher $\mathrm{AvgWorstAUROC}$ on both backbones. Richer variants such as $\mathrm{CFS}(2\times2)$ and $\mathrm{CFS}(2\times4)$ are studied in Appendix E.1.

All methods output OOD-high scores. Our primary metric is AUROC; full FPR95 breakdowns are reported in the appendix. Across pairs, we summarize performance by

$$\mathrm{AvgAUROC}_{m,b} = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} \mathrm{AUROC}_{m,b}(p), \tag{7}$$

$$\mathrm{AvgWorstAUROC}_{m,b} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \min_{o\in\mathcal{O}(i)} \mathrm{AUROC}_{m,b}(i\to o). \tag{8}$$

We report logical cost using explicit primitive counts such as $1F$, $10F$, and $2F+2J$.

Appendix E.12 provides an architecture-transfer sanity check on U-ViT, where CFS is applied to early/middle/late transformer block snapshots rather than U-Net encoder/decoder maps.

6 Results

Claim scope. Our claim is strictly about controlled comparison under MBE: when source-family policy, canonical corruption semantics, and logical test-time cost are matched, sparse native snapshot probing is stronger than the output-space alternatives we evaluate at comparable or lower logical cost.

6.1 Main CIFAR-scale comparison under MBE

Table 3 reports the main controlled CIFAR-scale comparison under MBE.

The two one-forward CFS variants exhibit a backbone-dependent average-performance trade-off. The paired readout has slightly higher $\mathrm{AvgAUROC}$ on improved-diffusion (0.887 versus 0.886), whereas decoder-only has higher $\mathrm{AvgAUROC}$ on EDM (0.919 versus 0.917). The paired readout has higher $\mathrm{AvgWorstAUROC}$ on both backbones: 0.799 versus 0.793 on improved-diffusion, and 0.814 versus 0.809 on EDM. At the same $1F$ cost, both sparse readouts are substantially stronger than final reconstruction MSE, showing that the advantage comes from where the same denoising computation is read rather than from additional test-time budget. They also remain stronger than the controlled SCOPED baseline despite its $2F+2J$ cost. Budget-matched ablations are reported in Appendix E.1.

DLSR is the closest prior in spirit, but it lies outside our MBE protocol because it introduces an additional learned feature-reconstruction module. Since DLSR is only available on its native published evaluation pairs, we report a comparison on that DLSR-native subset in Table 26 of Appendix H.1.

6.2 Theory diagnostics: measurable quantities predict sparse-probe quality

We next test whether the local-testing theory yields measurable quantities that predict sparse-probe quality. It does (see Table 4).

First, the diagonal noncentrality diagnostic $\hat{\kappa}_\lambda(\mathcal{S})/d$ is strongly aligned with downstream OOD performance across both backbone families: candidate sparse probes with larger estimated diagonal separation consistently yield larger AUROC. This supports the interpretation of the diagonal score as

Table 3: **Main comparison under MBE on the CIFAR-scale benchmark.** IDs: CIFAR-10, SVHN, CelebA32; each is evaluated against the other source datasets plus CIFAR-100 and DTD, yielding 12 pairs per backbone. All rows use the same source-family policy, preprocessing, canonical corruption semantics, and split policy. The reconstruction control uses the same one-forward denoising computation as CFS but reads only the final output.

Method	Backbone	AvgAUROC ↑	AvgWorstAUROC ↑	Cost ↓	Notes
Matched recon. MSE	improved	0.720	0.618	**1F**	final denoised output only
Matched recon. MSE	EDM	0.777	0.635	**1F**	final denoised output only
MSMA	improved	0.792	0.688	10F	multiscale output descriptor
MSMA	EDM	0.796	0.689	10F	multiscale output descriptor
DIFFPATH	improved	0.778	0.642	10F	recursive path statistic
DIFFPATH	EDM	0.792	0.635	10F	recursive path statistic
DDPM-OOD	improved	0.550	0.316	364F	reconstruction / manifold
DDPM-OOD	EDM	0.559	0.320	364F	reconstruction / manifold
GEPC	improved	0.616	0.546	8F	output-space consistency
GEPC	EDM	0.774	0.600	8F	output-space consistency
SCOPED	improved	0.773	0.647	2F+2J	score-curvature geometry
SCOPED	EDM	0.589	0.559	2F+2J	score-curvature geometry
CFS$_{\text{dec}}(1\times1)$	improved	<u>0.886</u>	<u>0.793</u>	**1F**	one late-decoder snapshot
CFS$_{\text{dec}}(1\times1)$	EDM	**0.919**	<u>0.809</u>	**1F**	one late-decoder snapshot
CFS(1×2)	improved	**0.887**	**0.799**	**1F**	encoder + decoder snapshots
CFS(1×2)	EDM	<u>0.917</u>	**0.814**	**1F**	encoder + decoder snapshots

Table 4: **Empirical validation of the measurable theory diagnostics.** Spearman correlations are computed across candidate sparse probes and ID→OOD pairs. The diagnostics are not used to fit the detector; they only test whether the measured behavior of sparse hooks and canonical levels agrees with the local theory.

Diagnostic	Backbone	Spearman ρ
$\hat{\kappa}_\lambda(\mathcal{S})/d$ vs AUROC	improved	0.923
$\hat{\kappa}_\lambda(\mathcal{S})/d$ vs AUROC	EDM	0.938
$\hat{R}_h(\lambda)$ vs Avg AUROC	improved	0.862
$\hat{R}_h(\lambda)$ vs Avg AUROC	EDM	0.760

a local detector. Representative diagnostic scatter plots for $\hat{\kappa}_\lambda(\mathcal{S})/d$ versus AUROC are deferred to Appendix B.7.

Second, the content-to-instability ratio $\hat{R}_h(\lambda)$ is also strongly predictive of hook quality. Moreover, the within-image corruption variance decreases sharply as $b(\lambda)^2$ decreases, in agreement with Proposition 1, supporting the low-noise advantage. Appendix B.7 reports the complementary $\hat{R}_h(\lambda)$ and low-noise stability plots.

Transfer to a diffusion transformer. The same picture is visible on U-ViT. At a matched one-forward budget, final reconstruction reaches 0.778 AvgAUROC, a late transformer snapshot reaches 0.891, and an early+late pair reaches 0.896. Across U-ViT blocks and canonical levels, the diagonal-separation diagnostic correlates strongly with pairwise AUROC (Spearman $\rho = 0.879$, $p = 1.40 \times 10^{-109}$, $n = 336$), while the content-to-instability ratio correlates with average AUROC ($\rho = 0.731$, $p = 2.75 \times 10^{-15}$, $n = 84$). Injected diffusion variance is also positively, though more moderately, associated with within-image feature variance ($\rho = 0.400$, $p = 1.63 \times 10^{-4}$, $n = 84$). This is architectural-transfer evidence, not a full transformer MBE benchmark; the complete breakdown is in Appendix E.12.

6.3 Focused ablations

The appendix provides matched controls for hidden test-time budget, clean versus low-noise probing, non-diffusion and random-weight representations, hard shifts and image complexity, head choice, and

Table 5: **External positioning outside MBE (single-checkpoint setting only).** Prior-method results are taken from the respective papers or officially reported artifacts. Full results are in Appendix H.

Method	Avg. AUROC ↑	Cost ↓
SCOPED-CelebA	0.893	2F+2J
GEPC-CelebA	0.910	8F
DiffPath-6D-CelebA	0.931	10F
DiffPath-6D-ImageNet	0.850	10F
CFS(1×2)**-CelebA (ours)**	<u>0.935</u>	**1F**
CFS(1×2)**-ImageNet (ours)**	**0.962**	**1F**

Table 6: **Source-family robustness on the CIFAR-scale benchmark.** Same protocol as Table 3, but with a shared CelebA32 source family.

	Improved		EDM		Cost
Method	AvgAUROC ↑	AvgWorstAUROC ↑	AvgAUROC ↑	AvgWorstAUROC ↑	
MSMA	0.881	0.809	0.790	0.673	10F
DiffPath	0.829	0.743	0.777	0.651	10F
DDPM-OOD	0.581	0.362	0.575	0.325	364F
GEPC	0.749	0.650	0.745	0.648	8F
SCOPED	0.806	0.639	0.758	0.623	2F+2J
CFS$_{\text{dec}}(1\times1)$	<u>0.907</u>	<u>0.823</u>	<u>0.914</u>	<u>0.827</u>	**1F**
CFS(1×2)	**0.908**	**0.826**	**0.929**	**0.845**	**1F**

stochastic stability. Across these controls, the central conclusion is narrower but unchanged: sparse learned diffusion representations expose strong OOD information at very low cost, while difficult near-OOD shifts remain genuine failure cases.

Hook-selection robustness. The ID-only hook-selection proxy does not need to recover the exact oracle pair to be effective. On improved-diffusion, the selected pair reaches 0.886 AvgAUROC versus 0.895 for the best admissible pair, while on EDM, the pairwise landscape exhibits a broad plateau of strong pairs. This indicates that CFS(1×2) is not driven by brittle hook cherry-picking; see Appendix E.2.

Clean-input and representation controls. A matched clean-endpoint control shows that sparse diffusion features are already strong without injected corruption; the low-noise probe improves AvgAUROC by 0.002–0.026, while AvgWorst is backbone-dependent. Matched random, supervised, and SimCLR representations using the same pooling and ID-only head are weaker than learned diffusion features; the strongest SimCLR two-readout control reaches $0.852/0.761$ Avg/AvgWorst versus $0.887/0.799$ and $0.917/0.814$ for paired CFS on improved-diffusion and EDM. Full controls are reported in Appendices E.5 and E.6.

Difficult shifts and image complexity. For CIFAR-10 as ID, the hardest shifts remain CelebA32 and CIFAR-100. A seven-statistic complexity-only predictor reaches only 0.430 and 0.541 AUROC on these pairs, while CFS decoder reaches $0.534/0.688$ on improved-diffusion and $0.444/0.700$ on EDM. Residualizing the CFS score against the ID-only complexity predictor lowers aggregate performance and does not repair either difficult pair. Simple image complexity is therefore neither sufficient to explain CFS nor a confound whose removal resolves its failures; full results are in Appendix E.7.

External positioning outside MBE. Table 5 provides a non-protocol-matched positioning against previously reported diffusion-OOD results. These comparisons are included for external positioning.

6.4 Source-family robustness

In Table 6, we repeat the same CIFAR-scale benchmark with a shared CelebA32 source family to test whether the ranking persists after changing the frozen source representation.

Changing the shared source family does not remove the sparse-feature advantage. In this alternative setting the paired readout is stronger than decoder-only on both average and worst-case metrics, reinforcing the narrower interpretation that encoder complementarity is source- and regime-dependent

rather than uniformly beneficial. Across both source families and both backbones, controlled SCOPED remains behind the best CFS variant at higher logical cost.

7 Discussion and Limitations

The controlled results support an accuracy-cost claim rather than universal robustness. A single late-decoder snapshot already provides strong relative-OOD discrimination. Adding a deep encoder snapshot slightly improves average AUROC on improved-diffusion but slightly reduces it on EDM, while improving $\mathrm{AvgWorstAUROC}$ on both primary backbones. Matched one-forward reconstruction and frozen-representation controls further indicate that the gain is not explained by extra denoising budget, arbitrary architecture, or the lightweight scoring head alone.

The CIFAR-scale benchmark is intentional: most diffusion-OOD comparators studied here were developed and evaluated in this regime, where source policy, corruption coordinates, and compute can be equated cleanly. The checkpoint-controlled ImageNet-scale experiment shows that the sparse-feature observation persists beyond 32×32 inputs, but near-OOD semantic shifts remain difficult. In particular, CFS is not reliable on every near or semantically overlapping shift, and the CIFAR-10 to CelebA32/CIFAR-100 cases remain genuine failure modes rather than being hidden by aggregate metrics.

Limitations. CFS requires internal access to the diffusion backbone and is therefore less black-box than output-space scores. Its portability is structural rather than representational: canonical noise levels can be aligned across backbones, but the most informative states remain architecture- and source-dependent. Our current local theory and controlled evaluation concern pixel-space diffusion. Although CFS can be applied operationally to frozen latent-diffusion states, its interpretation does not transfer automatically: the score may reflect both the autoencoder representation and the diffusion backbone, and informative locations may change under latent compression, patching, or multiscale attention. The ImageNet-scale experiment is therefore not a complete validation of high-resolution latent diffusion. Extending the analysis to latent-space diffusion is an important direction for future work. Given its $1F$ cost, combining CFS with a complementary specialist for subtle near-OOD shifts, invoked only for uncertain cases, is another promising direction that we do not evaluate here.

8 Conclusion

We introduced MBE, a protocol for controlled cross-backbone diffusion OOD evaluation, and CFS, a family of minimal representation-space detectors based on sparse native snapshots. Across the primary controlled benchmark, a single late-decoder snapshot is already highly competitive, while an encoder-decoder pair improves $\mathrm{AvgWorstAUROC}$ on both backbones at the same one-forward cost. The average-AUROC ordering of these variants depends on the backbone. Matched reconstruction, non-diffusion, and random-weight controls, controlled SCOPED comparisons, and transformer diagnostics support the conclusion that learned sparse diffusion representations expose useful relative-OOD structure far upstream of elaborate output-space probes. The remaining difficult near-OOD and latent-diffusion regimes define the main scope of the current result.

Acknowledgments and Disclosure of Funding

Part of this work was supported by ANR-ASTRID NEPTUNE 3 (ANR-23-ASM2-0009). The authors declare no competing interests.

References

Brett Barkley, Preston Culbertson, and David Fridovich-Keil. Scoped: Score-curvature out-of-distribution proximity evaluator for diffusion. In *International Conference on Learning Representations*, 2026.

Yifan Ding, Arturas Aleksandrauskas, Amirhossein Ahmadian, Jonas Unger, Fredrik Lindsten, and Gabriel Eilertsen. Revisiting likelihood-based out-of-distribution detection by modeling representations. In *Image Analysis*, Lecture Notes in Computer Science, pages 166–179, 2025.

Ruiyuan Gao, Chenchen Zhao, Lanqing Hong, and Qiang Xu. DIFFGUARD: Semantic mismatch-guided out-of-distribution detection using pre-trained diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1579–1589, 2023.

Joseph Goodier and Neill D. F. Campbell. Likelihood-based out-of-distribution detection with denoising diffusion probabilistic models. *arXiv preprint arXiv:2310.17432*, 2023.

Mark S. Graham, Walter H. L. Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Denoising diffusion models for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2948–2957, 2023.

Alvin Heng, Alexandre H. Thiery, and Harold Soh. Out-of-distribution detection with a single unconditional diffusion model. In *Advances in Neural Information Processing Systems*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Joonas Järve, Karl Kaspar Haavel, and Meelis Kull. Probability density from latent diffusion models for out-of-distribution detection, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Senmao Li, Taihang Hu, Joost van de Weijer, Fahad Shahbaz Khan, Tao Liu, Linxuan Li, Shiqi Yang, Yaxing Wang, Ming-Ming Cheng, and Jian Yang. Faster diffusion: Rethinking the role of the encoder for diffusion model inference. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, pages 21464–21475, 2020.

Zhenzhen Liu, Jin Peng Zhou, Yufan Wang, and Kilian Q. Weinberger. Unsupervised out-of-distribution detection with diffusion inpainting. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22528–22538, 2023.

Shuo Lu, Yingsheng Wang, Lijun Sheng, Lingxiao He, Aihua Zheng, and Jian Liang. Out-of-distribution detection: A task-oriented survey of recent advances. *ACM Computing Surveys*, 2025.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ahsan Mahmood, Junier Oliva, and Martin Andreas Styner. Multiscale score matching for out-of-distribution detection. In *International Conference on Learning Representations*, 2021.

Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171, 2021.

Yadang Alexis Rouzoumka, Jean Pinsolle, Eugénie Terreaux, Christèle Morisseau, Jean-Philippe Ovarlez, and Chengfang Ren. GEPC: Group-equivariant posterior consistency for out-of-distribution detection in diffusion models. In *Forty-third International Conference on Machine Learning*, 2026.

Shirin Shoushtari, Yi Wang, Xiao Shi, Salman Asif, and Ulugbek S. Kamilov. Eigenscore: Ood detection using posterior covariance in diffusion models. In *International Conference on Learning Representations*, 2026.

Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20827–20840, 2022.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4921–4930, 2022.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2022.

Xingyi Yang and Xinchao Wang. Diffusion model as representation learner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18938–18949, 2023.

Ying Yang, De Cheng, Chaowei Fang, Yubiao Wang, Changzhe Jiao, Lechao Cheng, Nannan Wang, and Xinbo Gao. Diffusion-based layer-wise semantic reconstruction for unsupervised out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. *Journal of Data-centric Machine Learning Research*, 2024. Dataset Certification.

A Appendix Roadmap

This appendix supports the paper's central claim: under a backbone-equated protocol, a frozen diffusion checkpoint can be used as a relative-OOD representation map, and a tiny number of sparse low-noise internal snapshots already captures strong useful signal.

It is organized as follows:

- Section B: proofs and supplementary validations for the main-paper theory, including conditional encoder-decoder complementarity, low-noise corruption stability, diagonal-score separation, canonical-level matching, and the interpretation of the hook-selection proxy.
- Sections C–D: shared protocol, canonicalization, logical cost accounting, and harmonized baseline implementations.
- Section E: focused ablations testing budget, hook robustness, canonical-level robustness, pooling, head choice, bank size, and seed stability.
- Sections F and G: full CIFAR-scale pairwise results under the primary and alternative source-family policies.
- Section H: external positioning against prior reported diffusion-based CIFAR-scale results outside the controlled MBE protocol.
- Section I: checkpoint-controlled large-scale results on ImageNet200 and ImageNet1K using a single official ImageNet-64 improved-diffusion backbone.
- Section J: implementation and configuration details, including the main CFS hyperparameters, ID-only head configurations, determinism settings, profiling protocol, and representative compute-resource reporting.
- Section K: broader impact and responsible-use discussion.

Unless noted otherwise, appendix ablations are centered on $\mathrm{CFS}(1 \times 2)$, the primary main-paper operating point. We retain $\mathrm{CFS}_{\mathrm{dec}}(1\times1)$ only in targeted controls where the compression result is directly relevant.

B Theory Appendix

This appendix supplies the formal statements, proofs, and supplementary validations supporting the main paper theory. The main text relies directly on two load-bearing claims: conditional encoder-decoder complementarity and low-noise corruption stability. We additionally formalize here the diagonal-score separation result deferred from the main text, a canonical-level stability bound under discretization, and a formal interpretation of the hook-selection proxy.

B.1 Notation

Fix a frozen diffusion backbone and evaluation domains P and Q, corresponding to the relative ID and OOD distributions. For a canonical corruption level λ,

$$\mathbf{x}_\lambda = a(\lambda)\,\mathbf{x}_0 + b(\lambda)\,\boldsymbol{\varepsilon}\,, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})\,.$$

For a selected late decoder hook and a selected deep encoder hook,

$$\mathbf{z}_\lambda(\mathbf{x}_0) = \begin{bmatrix} \mathbf{z}_{d,\lambda}(\mathbf{x}_0) \\ \mathbf{z}_{e,\lambda}(\mathbf{x}_0) \end{bmatrix}\,,$$

and the local model is

$$\mathbf{z}_\lambda(\mathbf{x}_0) \mid H_0 \sim \mathcal{N}(\boldsymbol{\mu}_\lambda, \boldsymbol{\Sigma}_\lambda)\,, \qquad \mathbf{z}_\lambda(\mathbf{x}_0) \mid H_1 \sim \mathcal{N}(\boldsymbol{\mu}_\lambda + \boldsymbol{\Delta}_\lambda, \boldsymbol{\Sigma}_\lambda)\,,$$

with $H_0 : \mathbf{x}_0 \sim P$ and $H_1 : \mathbf{x}_0 \sim Q$.

B.2 Deferred statement from the main text

The following result is used in the interpretation of the diagonal CFS score, but is deferred here to keep the main text focused.

Let

$$\mathbf{D}_\lambda = \mathrm{diag}(\mathbf{\Sigma}_\lambda), \qquad d_\lambda = \dim(\mathbf{z}_\lambda),$$

and consider the oracle diagonal score

$$s^\circ_\lambda(\mathbf{z}) := \frac{1}{d_\lambda} (\mathbf{z} - \boldsymbol{\mu}_\lambda)^\top \mathbf{D}_\lambda^{-1} (\mathbf{z} - \boldsymbol{\mu}_\lambda), \qquad \kappa_\lambda := \mathbf{\Delta}_\lambda^\top \mathbf{D}_\lambda^{-1} \mathbf{\Delta}_\lambda .$$

Theorem 2 (Diagonal-score separation under the local model)**.** *Under Eq. (3),*

$$\mathbb{E}_{H_0}[s^\circ_\lambda] = 1, \qquad \mathbb{E}_{H_1}[s^\circ_\lambda] = 1 + \frac{\kappa_\lambda}{d_\lambda} .$$

Moreover, the variance of s°_λ is controlled by the correlation structure of $\mathbf{\Sigma}_\lambda$, yielding the detection-power bound proved below.

B.3 Proof of Theorem 1

Proof. Write

$$\mathbf{\Delta}_\lambda = \begin{bmatrix} \mathbf{\Delta}_{d,\lambda} \\ \mathbf{\Delta}_{e,\lambda} \end{bmatrix}, \qquad \mathbf{\Sigma}_\lambda = \begin{bmatrix} \mathbf{\Sigma}_{dd,\lambda} & \mathbf{\Sigma}_{de,\lambda} \\ \mathbf{\Sigma}_{ed,\lambda} & \mathbf{\Sigma}_{ee,\lambda} \end{bmatrix} .$$

The paired Mahalanobis separation is

$$\mathrm{Sep}_{\mathrm{pair}}(\lambda) = \mathbf{\Delta}_\lambda^\top \mathbf{\Sigma}_\lambda^{-1} \mathbf{\Delta}_\lambda ,$$

while the decoder-only separation is

$$\mathrm{Sep}_{\mathrm{dec}}(\lambda) = \mathbf{\Delta}_{d,\lambda}^\top \mathbf{\Sigma}_{dd,\lambda}^{-1} \mathbf{\Delta}_{d,\lambda} .$$

Let

$$\mathbf{\Sigma}_{e|d,\lambda} = \mathbf{\Sigma}_{ee,\lambda} - \mathbf{\Sigma}_{ed,\lambda} \mathbf{\Sigma}_{dd,\lambda}^{-1} \mathbf{\Sigma}_{de,\lambda} ,$$

be the Schur complement of the decoder block. By the block inverse formula,

$$\mathbf{\Sigma}_\lambda^{-1} = \begin{bmatrix} \mathbf{\Sigma}_{dd,\lambda}^{-1} + \mathbf{\Sigma}_{dd,\lambda}^{-1} \mathbf{\Sigma}_{de,\lambda} \mathbf{\Sigma}_{e|d,\lambda}^{-1} \mathbf{\Sigma}_{ed,\lambda} \mathbf{\Sigma}_{dd,\lambda}^{-1} & -\mathbf{\Sigma}_{dd,\lambda}^{-1} \mathbf{\Sigma}_{de,\lambda} \mathbf{\Sigma}_{e|d,\lambda}^{-1} \\ -\mathbf{\Sigma}_{e|d,\lambda}^{-1} \mathbf{\Sigma}_{ed,\lambda} \mathbf{\Sigma}_{dd,\lambda}^{-1} & \mathbf{\Sigma}_{e|d,\lambda}^{-1} \end{bmatrix} .$$

Substituting this expression into $\mathbf{\Delta}_\lambda^\top \mathbf{\Sigma}_\lambda^{-1} \mathbf{\Delta}_\lambda$ and collecting terms gives

$$\mathrm{Sep}_{\mathrm{pair}}(\lambda) = \mathrm{Sep}_{\mathrm{dec}}(\lambda) + \mathrm{Res}_{e|d}(\lambda) ,$$

where

$$\mathrm{Res}_{e|d}(\lambda) = \left(\mathbf{\Delta}_{e,\lambda} - \mathbf{\Sigma}_{ed,\lambda} \mathbf{\Sigma}_{dd,\lambda}^{-1} \mathbf{\Delta}_{d,\lambda} \right)^\top \mathbf{\Sigma}_{e|d,\lambda}^{-1} \left(\mathbf{\Delta}_{e,\lambda} - \mathbf{\Sigma}_{ed,\lambda} \mathbf{\Sigma}_{dd,\lambda}^{-1} \mathbf{\Delta}_{d,\lambda} \right) .$$

Since $\mathbf{\Sigma}_{e|d,\lambda} \succ 0$, the residual term is nonnegative. Therefore

$$\mathrm{Sep}_{\mathrm{pair}}(\lambda) \geq \mathrm{Sep}_{\mathrm{dec}}(\lambda) .$$

Equality holds if and only if the conditional residual vanishes, namely

$$\mathbf{\Delta}_{e,\lambda} = \mathbf{\Sigma}_{ed,\lambda} \mathbf{\Sigma}_{dd,\lambda}^{-1} \mathbf{\Delta}_{d,\lambda} .$$

This proves the decomposition and the claimed nonnegativity. □

B.4 Proof of Proposition 1

Proof. Fix $\mathbf{x}_0$ and write

$$\mathbf{x}_\lambda = a(\lambda)\,\mathbf{x}_0 + b(\lambda)\,\boldsymbol{\varepsilon}\,.$$

By the assumed first-order mean-square expansion of $\boldsymbol{\phi}_{\lambda,h}$ around $a(\lambda)\mathbf{x}_0$, we have

$$\boldsymbol{\phi}_{\lambda,h}(\mathbf{x}_\lambda) = \boldsymbol{\phi}_{\lambda,h}(a(\lambda)\mathbf{x}_0) + b(\lambda)\,\mathbf{J}_{\lambda,h}(\mathbf{x}_0)\,\boldsymbol{\varepsilon} + \mathbf{r}_{\lambda,h}(\mathbf{x}_0,\boldsymbol{\varepsilon})\,,$$

with

$$\mathbb{E}\Big[\|\mathbf{r}_{\lambda,h}(\mathbf{x}_0,\boldsymbol{\varepsilon})\|_2^2 \,\Big|\, \mathbf{x}_0\Big] = o\left(b(\lambda)^2\right)\,.$$

Therefore,

$$\mathbf{z}_{\lambda,h}(\mathbf{x}_0) - \boldsymbol{\phi}_{\lambda,h}(a(\lambda)\,\mathbf{x}_0) = b(\lambda)\,\mathbf{J}_{\lambda,h}(\mathbf{x}_0)\,\boldsymbol{\varepsilon} + \mathbf{r}_{\lambda,h}(\mathbf{x}_0,\boldsymbol{\varepsilon})\,.$$

Taking squared norms and conditional expectations gives

$$\mathbb{E}\Big[\big\|\mathbf{z}_{\lambda,h}(\mathbf{x}_0) - \boldsymbol{\phi}_{\lambda,h}(a(\lambda)\,\mathbf{x}_0)\big\|_2^2 \,\Big|\, \mathbf{x}_0\Big] = b(\lambda)^2\mathbb{E}\Big[\|\mathbf{J}_{\lambda,h}(\mathbf{x}_0)\boldsymbol{\varepsilon}\|_2^2\Big] + o\left(b(\lambda)^2\right)\,.$$

Since $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$,

$$\mathbb{E}\Big[\|\mathbf{J}_{\lambda,h}(\mathbf{x}_0)\,\boldsymbol{\varepsilon}\|_2^2\Big] = \operatorname{tr}\big(\mathbf{J}_{\lambda,h}(\mathbf{x}_0)\,\mathbf{J}_{\lambda,h}(\mathbf{x}_0)^\top\big)\,.$$

Hence

$$\mathbb{E}\Big[\big\|\mathbf{z}_{\lambda,h}(\mathbf{x}_0) - \boldsymbol{\phi}_{\lambda,h}(a(\lambda)\,\mathbf{x}_0)\big\|_2^2 \,\Big|\, \mathbf{x}_0\Big] = b(\lambda)^2\operatorname{tr}\big(\mathbf{J}_{\lambda,h}(\mathbf{x}_0)\,\mathbf{J}_{\lambda,h}(\mathbf{x}_0)^\top\big) + o\big(b(\lambda)^2\big)\,.$$

Assume also that $\|\mathbf{J}_{\lambda,h}(\mathbf{x}_0)\|_F$ remains bounded as $b(\lambda) \to 0$. Then Cauchy–Schwarz makes the cross term $o(b(\lambda)^2)$. Since the linear noise term has zero mean, Jensen's inequality and the remainder assumption give

$$\big\|\mathbb{E}[\mathbf{z}_{\lambda,h} \mid \mathbf{x}_0] - \boldsymbol{\phi}_{\lambda,h}(a(\lambda)\mathbf{x}_0)\big\|_2^2 \le \mathbb{E}[\|\mathbf{r}_{\lambda,h}\|_2^2 \mid \mathbf{x}_0] = o(b(\lambda)^2).$$

The conditional variance therefore has the same leading-order expression as the second moment around the clean reference. This proves the result. □

B.5 Proof of Theorem 2

Proof. Let

$$\mathbf{D}_\lambda = \operatorname{diag}(\boldsymbol{\Sigma}_\lambda), \qquad \mathbf{R}_\lambda = \mathbf{D}_\lambda^{-1/2}\boldsymbol{\Sigma}_\lambda\mathbf{D}_\lambda^{-1/2}, \qquad \boldsymbol{\delta}_\lambda = \mathbf{D}_\lambda^{-1/2}\boldsymbol{\Delta}_\lambda\,.$$

Define the whitened-but-correlated variable

$$\mathbf{y} = \mathbf{D}_\lambda^{-1/2}(\mathbf{z} - \boldsymbol{\mu}_\lambda)\,.$$

Under H_0, we have

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0},\mathbf{R}_\lambda)\,,$$

and under H_1,

$$\mathbf{y} \sim \mathcal{N}(\boldsymbol{\delta}_\lambda,\mathbf{R}_\lambda)\,.$$

The oracle diagonal score can be written as

$$s_\lambda^\circ(\mathbf{z}) = \frac{1}{d_\lambda}\,\mathbf{y}^\top\,\mathbf{y}\,.$$

Because $\mathbf{R}_\lambda$ is a correlation matrix, $\operatorname{tr}(\mathbf{R}_\lambda) = d_\lambda$. Hence

$$\mathbb{E}_{H_0}[\mathbf{y}^\top\mathbf{y}] = \operatorname{tr}(\mathbf{R}_\lambda) = d_\lambda\,,$$

so

$$\mathbb{E}_{H_0}[s_\lambda^\circ] = 1\,.$$

Under H_1,

$$\mathbb{E}_{H_1}\left[\mathbf{y}^\top\mathbf{y}\right] = \operatorname{tr}(\mathbf{R}_\lambda) + \boldsymbol{\delta}_\lambda^\top\,\boldsymbol{\delta}_\lambda = d_\lambda + \kappa_\lambda\,,$$

where

$$\kappa_\lambda = \boldsymbol{\Delta}_\lambda^\top \, \mathbf{D}_\lambda^{-1} \, \boldsymbol{\Delta}_\lambda = \boldsymbol{\delta}_\lambda^\top \, \boldsymbol{\delta}_\lambda \, .$$

Therefore

$$\mathbb{E}_{H_1}[s_\lambda^\circ] = 1 + \frac{\kappa_\lambda}{d_\lambda} \, .$$

For a Gaussian quadratic form $\mathbf{y}^\top \, \mathbf{A} \, \mathbf{y}$ with $\mathbf{A} = \mathbf{I}$, the variance satisfies

$$\mathrm{Var}\left(\mathbf{y}^\top \, \mathbf{y}\right) = 2 \, \mathrm{tr}\left(\mathbf{R}_\lambda^2\right) + 4 \, \boldsymbol{\delta}_\lambda^\top \, \mathbf{R}_\lambda \, \boldsymbol{\delta}_\lambda \, ,$$

with the second term absent under H_0. Dividing by d_λ^2 gives

$$\mathrm{Var}_{H_0}(s_\lambda^\circ) = \frac{2}{d_\lambda^2} \, \mathrm{tr}\left(\mathbf{R}_\lambda^2\right) \, ,$$

and

$$\mathrm{Var}_{H_1}(s_\lambda^\circ) = \frac{2}{d_\lambda^2} \, \mathrm{tr}\left(\mathbf{R}_\lambda^2\right) + \frac{4}{d_\lambda^2} \, \boldsymbol{\delta}_\lambda^\top \, \mathbf{R}_\lambda \, \boldsymbol{\delta}_\lambda \, .$$

Finally, apply Cantelli's inequality under H_1 to the random variable s_λ°. For any $\tau < \mathbb{E}_{H_1}[s_\lambda^\circ]$,

$$\Pr_{H_1}[s_\lambda^\circ \le \tau] \le \frac{\mathrm{Var}_{H_1}(s_\lambda^\circ)}{\mathrm{Var}_{H_1}(s_\lambda^\circ) + \left(\mathbb{E}_{H_1}[s_\lambda^\circ] - \tau\right)^2} \, .$$

Taking complements and substituting the mean gives

$$\Pr_{H_1}[s_\lambda^\circ > \tau] \ge 1 - \frac{\mathrm{Var}_{H_1}(s_\lambda^\circ)}{\mathrm{Var}_{H_1}(s_\lambda^\circ) + \left(1 + \kappa_\lambda/d_\lambda - \tau\right)^2} \, .$$

This proves the theorem. □

B.6 Canonical-level matching sanity check

This diagnostic tests the implementation role of canonical logSNR matching. For a discrete backbone, a requested canonical level λ_{ref} must be mapped to a native timestep t, whose effective logSNR can be different from λ_{ref} and is denoted by λ_t.

The following simple bound explains why such mismatches can matter: if the internal descriptor and the resulting slot score vary smoothly with the canonical level, then the logSNR mismatch induces controlled score drift.

Fix a selected hook h. Assume that, on the evaluation domain $\mathcal{X}$, the pooled descriptor $\mathbf{z}_{\cdot,h}(\mathbf{x})$ is L_h-Lipschitz in the canonical level:

$$\|\mathbf{z}_{\lambda,h}(\mathbf{x}) - \mathbf{z}_{\lambda',h}(\mathbf{x})\|_2 \le L_h|\lambda - \lambda'| \, , \qquad \forall \mathbf{x} \in \mathcal{X}, \ \lambda \text{ and } \lambda' \text{ two corruption levels.}$$

Assume also that the corresponding oracle slot score $s_{\cdot,h}^\circ$ is M_h-Lipschitz in its feature argument over the attained range. Then

$$|s_{\lambda,h}^\circ(\mathbf{x}) - s_{\lambda',h}^\circ(\mathbf{x})| \le M_h L_h |\lambda - \lambda'|.$$

Consequently, for a discrete backbone using a native matched level λ_t and a requested canonical level λ_{ref}, by averaging these bounds over the selected hooks in the oracle CFS score :

$$|S_{\mathrm{CFS},\lambda_{\mathrm{ref}}}^\circ(\mathbf{x}) - S_{\mathrm{CFS},\lambda_t}^\circ(\mathbf{x})| \le \frac{1}{|\mathcal{S}|} \sum_{h \in \mathcal{S}} M_h L_h |\lambda_{\mathrm{ref}} - \lambda_t| \, .$$

The same reasoning can be applied to AUROC score difference.

This diagnostic is only meaningful for discrete backbones (it is detailed in Appendix C.4), so we report it on improved-diffusion. Figure 2 shows two complementary views. First, larger effective canonical mismatches tend to induce larger OOD score drift. Second, sufficiently large mismatches can also produce measurable AUROC degradation relative to the matched logSNR-uniform reference policy. The relation is not perfectly linear, which is expected under discrete timestep collisions and pair-dependent difficulty, but the global trend supports the role of canonical logSNR matching as a genuine methodological requirement rather than a cosmetic alignment choice.

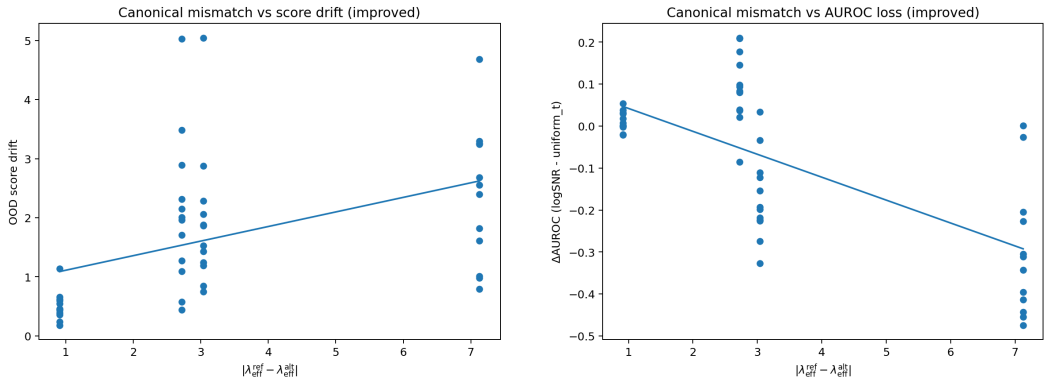

Figure 2: **Canonical-level matching sanity check on improved-diffusion. Left:** standardized score drift versus effective logSNR mismatch $|\lambda_{\text{ref}} - \lambda_t|$. **Right:** AUROC degradation relative to the matched logSNR policy. Large mismatches can induce score drift and degrade AUROC, supporting logSNR matching as an implementation requirement rather than a cosmetic alignment choice.

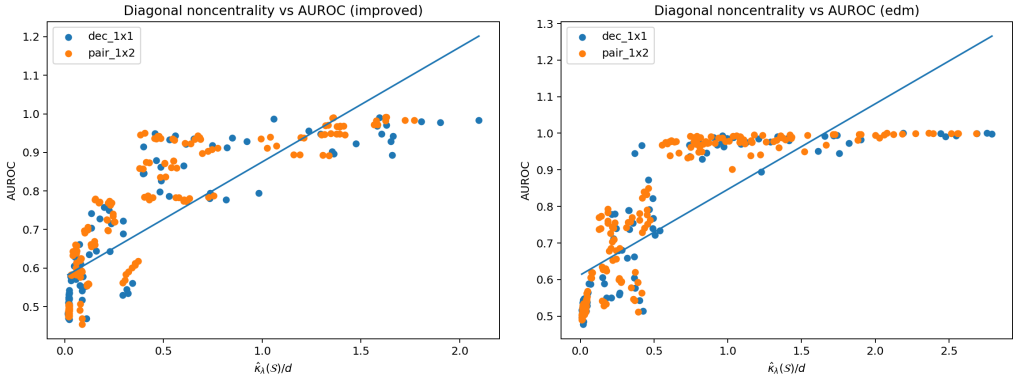

Figure 3: **Empirical validation of diagonal separation.** The estimated diagonal noncentrality $\hat{\kappa}_\lambda(\mathcal{S})/d$ is strongly aligned with downstream AUROC on both improved-diffusion and EDM backbones.

B.7 Empirical diagnostic protocol

The theory of Section 4 yields two measurable quantities used in our empirical diagnostics: the diagonal noncentrality $\hat{\kappa}_\lambda(\mathcal{S})$ and the content-to-instability ratio $\hat{R}_h(\lambda)$. These quantities are estimated on held-out data and compared directly against downstream OOD performance.

For $\hat{\kappa}_\lambda(\mathcal{S})$, we study rank correlation with pairwise AUROC across candidate sparse probes and ID→OOD pairs. This diagnostic is summarized in the main paper through Table 4, with representative scatter plots deferred to Figure 3. It confirms this prediction on both improved-diffusion and EDM backbones. Across candidate sparse probes and ID→OOD pairs, larger values of $\hat{\kappa}_\lambda(\mathcal{S})/d$ are strongly aligned with larger AUROC, supporting the interpretation of the diagonal score as a structured local detector rather than as an arbitrary lightweight classifier.

For $\hat{R}_h(\lambda)$, we study rank correlation with average AUROC across candidate hooks, and we additionally measure within-image corruption variance as a function of $b(\lambda)^2$. These validations are reported in Figure 4.

B.8 Interpretation

Taken together, the appendix theory supports three concrete conclusions about sparse diffusion probing.

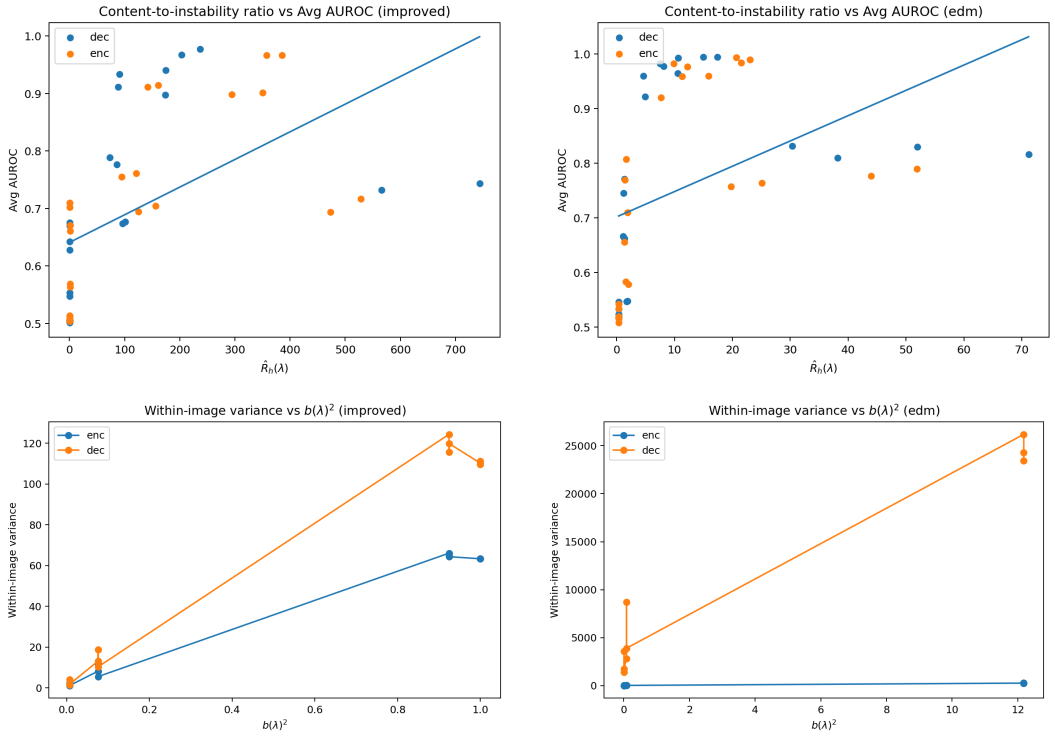

Figure 4: **Empirical validation of low-noise stability and the content-to-instability diagnostic. Top row:** $\hat{R}_h(\lambda)$ versus downstream Avg AUROC across candidate hooks for improved-diffusion (left) and EDM (right). **Bottom row:** within-image corruption variance versus $b(\lambda)^2$ for the selected encoder and decoder hooks on improved-diffusion (left) and EDM (right). Across both backbone families, larger content-to-instability ratios align with stronger downstream performance, and lower-noise canonical probing reduces corruption-induced variability.

First, the paired encoder-decoder separation decomposes into a decoder term plus a nonnegative conditional encoder residual. This supports using a late decoder snapshot as the primary sparse probe, while interpreting encoder snapshots as complementary rather than primary. Second, the diagonal CFS score is governed by a measurable diagonal noncentrality parameter κ_λ, so the score can be analyzed as a structured local detector rather than as an arbitrary lightweight head. Third, low-noise canonical probing reduces corruption-induced instability: under a local smoothness approximation, within-image feature variance scales as $b(\lambda)^2$, and the empirical content-to-instability diagnostic captures this effect.

When included, the canonical-matching diagnostic further supports that logSNR alignment is not merely a bookkeeping device: large discrete-level mismatches can induce score drift and degrade AUROC.

The theory remains local rather than universal. It does not claim that sparse internal probing dominates every conceivable output-space detector. Its purpose is narrower: to explain and predict which sparse internal probes should work best under a backbone-equated protocol.

C Experimental Protocols, Splits, and Canonicalization

This section groups the protocol elements shared across methods: dataset splits, canonical level construction, backbone-specific corruption mappings, adapter outputs, and logical cost accounting.

C.1 Small-scale benchmark protocol

The small-scale benchmark uses the three ID datasets

$$\mathcal{I}_{\text{small}} = \{\text{CIFAR-10}, \text{SVHN}, \text{CelebA32}\},$$

and the OOD pool

$$\mathcal{O}_{\text{small}} = \{\text{CIFAR-10}, \text{SVHN}, \text{CelebA32}, \text{CIFAR-100}, \text{DTD}\}.$$

For each ID dataset, all remaining datasets in the pool are treated as OOD, producing 12 ID→OOD pairs per backbone.

For each ID dataset, we define:

- an ID-fit split used to fit ID-only statistics and density heads;
- an ID-test split used for in-distribution evaluation;
- external OOD-test splits for all OOD datasets.

These splits are fixed and shared across methods.

Relative-OOD interpretation. Let $P_\star$ denote the training distribution of the frozen checkpoint, and let P denote the chosen evaluation ID dataset. Our detector should be interpreted as performing OOD detection *relative to the evaluation reference bank* P, not as an absolute test of membership in $P_\star$. This distinction is especially important in cross-source or transfer settings.

C.2 ImageNet200 / ImageNet1K protocol

ImageNet200 and ImageNet1K are used as checkpoint-controlled large-scale benchmarks under a single official ImageNet-64 improved-diffusion backbone. Since the official training split is not used in our setup, we adopt deterministic disjoint val-only protocols. For ImageNet200, we use a class-stratified split of the validation set into:

- **ID-fit / bank split:** used to fit ID statistics;
- **ID-test split:** the complementary subset used for in-distribution evaluation.

For ImageNet1K, we use the same val-only deterministic split policy to define ID-fit and ID-test subsets.

In both benchmarks, the OOD suite contains two near-OOD sets, **NINCO** and **SSB-hard**, together with one far-OOD texture-shift set, **Textures**.

Checkpoint-controlled protocol. All large-scale results in this appendix use the same official ImageNet-64 improved-diffusion checkpoint (`imagenet64_uncond_100M_1500K.pt`). This removes cross-family variability and isolates method differences under a single shared-source backbone.

C.3 Canonical level construction: K_{grid} vs. K_c

Purpose. Discrete backbones require mapping continuous canonical levels to discrete timesteps, which may produce duplicates. We separate:

- K_{grid}: candidate grid resolution used to build a dense set of possible canonical levels;
- K_c: number of selected levels actually used by the method.

Construction. We construct a candidate grid

$$\{\lambda_i^{\text{grid}}\}_{i=1}^{K_{\text{grid}}} \subset [\lambda_{\min}, \lambda_{\max}]\,.$$

For discrete backbones, each candidate level is mapped to the nearest available timestep in logSNR space. When `unique=true`, repeated timestep assignments are removed, and the final selected set contains at most K_c distinct levels. For continuous backbones, the selected levels are evenly spaced directly in canonical space.

Interpretation. Increasing K_{grid} improves mapping fidelity on discrete backbones without necessarily increasing test-time cost, because only the final selected K_c levels are used downstream.

Remark 1. *The separation between K_{grid} and K_c is important for fair backbone matching: selection fidelity and evaluation cost should not be conflated.*

C.4 Backbone-specific realization of canonical corruption

C.4.1 Improved-diffusion (VP-DDPM)

For improved-diffusion checkpoints, the native forward corruption process is

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

where $\bar{\alpha}_t$ is the cumulative noise-schedule coefficient of the checkpoint.

Hence, the corresponding corruption coefficients in the shared canonical form

$$\mathbf{x}_\lambda = a(\lambda)\,\mathbf{x}_0 + b(\lambda)\,\boldsymbol{\varepsilon},$$

are

$$a_t = \sqrt{\bar{\alpha}_t}, \qquad b_t = \sqrt{1-\bar{\alpha}_t},$$

and the induced discrete canonical logSNR is

$$\lambda_t = \log \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}.$$

A desired canonical level λ is therefore mapped to the nearest native timestep in logSNR space:

$$t(\lambda) = \arg\min_t |\lambda_t - \lambda|.$$

When several candidate canonical levels map to the same native timestep, duplicates are removed by the unique-level construction described in Appendix C.3.

C.4.2 EDM family

EDM-style models expose a continuous noise variable, but the exact native parameterization depends on the checkpoint preconditioning family. Rather than matching checkpoints through their native interface directly, we canonicalize them through the effective corruption ratio

$$\tilde{\sigma}(\lambda) := \frac{b(\lambda)}{a(\lambda)} = \exp(-\lambda/2),$$

which follows from

$$\lambda = \log \frac{a(\lambda)^2}{b(\lambda)^2}.$$

The adapter, therefore, maps each canonical level λ to the checkpoint-specific model input corresponding to the same effective ratio $\tilde{\sigma}(\lambda)$, and returns coefficients (a, b) such that

$$\mathbf{x}_\lambda = a\,\mathbf{x}_0 + b\,\boldsymbol{\varepsilon}.$$

Different EDM-family preconditionings may realize the same $\tilde{\sigma}$ with different coefficient pairs (a, b). For example, a VE/EDM-style realization uses

$$a = 1, \qquad b = \tilde{\sigma},$$

whereas a VP-style realization can be written as

$$a = (1+\tilde{\sigma}^2)^{-1/2}, \qquad b = \tilde{\sigma}(1+\tilde{\sigma}^2)^{-1/2}.$$

In both cases,

$$\frac{b}{a} = \tilde{\sigma},$$

so the same canonical λ corresponds to the same effective corruption strength even though the native checkpoint interface differs.

C.4.3 Shared canonical semantics

Although improved-diffusion and EDM differ in interface and preconditioning, all methods operate through the same canonical corruption abstraction. This allows meaningful cross-backbone comparisons at matched canonical levels.

C.5 Shared adapter outputs and reconstruction rules

All harmonized methods are routed through a shared adapter layer. At each canonical level, the adapter provides:

- a denoised estimate $\hat{\mathbf{x}}_0$,
- optional access to intermediate activations,
- and, when needed, an $\hat{\varepsilon}$ estimate recovered through

$$\hat{\varepsilon} = \frac{\mathbf{x} - a\,\hat{\mathbf{x}}_0}{b}\,. \tag{9}$$

This ensures that all output-space baselines and internal-feature methods use a compatible corruption semantics, regardless of the native output conventions of the original implementation.

C.6 Implementation details specific to CFS

The conceptual design of CFS is described in Section 4. Here we report only implementation-level details needed for exact reproduction.

C.6.1 Admissible hook search and shortlist size

A candidate hook is admissible only if a dry forward pass returns a 4D tensor of shape $B \times C \times H \times W$. In practice, we restrict the search to stage-level encoder and decoder blocks and keep a small shortlist within each structural region before applying the ID-only proxy. For every experiment, we report:

- the number of admissible encoder candidates,
- the number of admissible decoder candidates,
- the shortlist size retained in each region,
- the final selected encoder and decoder module names.

C.6.2 ID-only probe configuration

The proxy uses a small ID-only probe set. For exact reproducibility, we report:

- the number of ID probe images,
- the number of corruption repeats per image,
- whether the same noise draw is reused across canonical levels,

C.6.3 Exact pooled-slot construction

For each retained slot, we apply channel-wise spatial mean and standard deviation pooling, yielding a descriptor of dimension $2C_{k,\ell}$. We report, for each backbone family and benchmark:

- the selected canonical levels,
- the selected encoder and decoder block names,
- the resulting feature-map shapes,
- the pooled descriptor dimensions.

C.6.4 Diagonal score details

Each pooled slot is modeled independently with ID-only diagonal statistics. For exact reproduction, we report:

- whether the slot features are standardized before fitting,
- whether slot scores are averaged uniformly or reweighted.

D Implementation Taxonomy and Baseline Specifications

D.1 Implementation taxonomy

We implement these methods to preserve the core probe and score logic of the original method while routing it through our shared adapter and canonicalization pipeline.

Methods in this paper.

- MSMA: harmonized faithful port;
- DIFFPATH: harmonized faithful port;
- DDPM-OOD: harmonized faithful port;
- GEPC: harmonized faithful port;
- SCOPED: independent paper-faithful implementation under native and MBEsettings;
- CFS: proposed internal representation-space method.

High-cost comparator retained for external positioning. EIGENSCOREis not included in the low-budget controlled table because its published evaluation requires roughly $300F$ per image. We therefore report it in external positioning, while SCOPEDis now included directly under MBE.

Official repositories used as starting points. When available, our harmonized implementations were initialized from the official public repositories of the corresponding methods: MSMA[1], DiffPath[2], DDPM-OOD[3], and GEPC[4]. These repositories were then adapted to the shared adapter, canonicalization, and logical-budget interface of MBE. No public vision implementation of SCOPEDwas available, so we implemented its paper-described vision detector independently and validate it both in its native reference setting and under MBE.

D.2 Common harmonized interface

All baselines are evaluated through the same adapter interface, the same input normalization to $[-1, 1]$, and the same canonical corruption semantics. For a selected canonical level λ_k with coefficients (a_k, b_k), we explicitly corrupt the clean image as

$$\mathbf{x}_k = a_k\,\mathbf{x}_0 + b_k\,\boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})\,.$$

From the frozen backbone, we then recover native denoising quantities $\hat{\mathbf{x}}_{0,k}$ and $\hat{\boldsymbol{\varepsilon}}_k$. For improved-diffusion backbones, $\hat{\varepsilon}_k$ is extracted natively from the model output using the correct timestep scaling and mean-parameterization conventions, and $\hat{\mathbf{x}}_{0,k}$ is recovered from the same forward pass. For EDM-style backbones, the adapter returns $\hat{\mathbf{x}}_{0,k}$ through one denoising call and reconstructs

$$\hat{\boldsymbol{\varepsilon}}_k = \frac{\mathbf{x}_k - a_k\,\hat{\mathbf{x}}_{0,k}}{b_k}\,.$$

Thus, one backbone evaluation at one level counts as one logical forward pass, even if both $\hat{\mathbf{x}}_0$ and $\hat{\boldsymbol{\varepsilon}}$ are recovered from that same call.

D.3 MSMA (harmonized faithful port)

MSMA is the closest baseline in our suite to a harmonized faithful port. Its core logic is preserved: build a multiscale descriptor from denoiser-derived quantities across K_c levels, then fit an ID-only density head on the resulting feature vectors.

Feature construction. For each canonical level λ_k, we compute

$$f_k(\mathbf{x}_0) = \|\hat{\boldsymbol{\varepsilon}}_k\|_2\,,$$

where $\hat{\boldsymbol{\varepsilon}}_k$ is obtained through the shared native adapter interface. The final multiscale descriptor is

$$F(\mathbf{x}_0) = \big[f_1(\mathbf{x}_0), \ldots, f_{K_c}(\mathbf{x}_0)\big] \in \mathbb{R}^{K_c}\,.$$

[1] https://github.com/ahsanMah/msma
[2] https://github.com/clear-nus/diffpath
[3] https://github.com/marksgraham/ddpm-ood
[4] https://github.com/RouzAY/gepc-diffusion

Head and scoring. We fit an ID-only density model on $F(x)$, after optional feature standardization. Our implementation supports:

- a diagonal Gaussian head,
- a Gaussian mixture model (GMM),
- a k-nearest-neighbor distance head.

The final score is always OOD-high: negative log-likelihood for Gaussian/GMM heads, or distance for the KNN head.

What is preserved and what is adapted. The preserved part is the multiscale descriptor and the density score. The adapted part is the use of a common adapter interface, canonical logSNR levels, and a shared improved/EDM implementation.

D.4 DiffPath (harmonized faithful port)

DiffPath is implemented as a faithful port for path-based diffusion OOD detection. Rather than reproducing a native implementation verbatim, we preserve the key idea: summarize a *multilevel denoising path* and fit an ID-only density model on the resulting path statistics.

Recursive path construction. Levels are ordered from clean to noisy. We initialize

$$\mathbf{x}_1 = a_1\,\mathbf{x}_0 + b_1\,\boldsymbol{\varepsilon}\,,$$

and for each level k compute $(\hat{\mathbf{x}}_{0,k}, \hat{\boldsymbol{\varepsilon}}_k)$ from the current state $\mathbf{x}_k$. A scalar path statistic q_k is then extracted from $\hat{\boldsymbol{\varepsilon}}_k$ using one of the following reductions:

$$q_k \in \Big\{\mathrm{mean}(\hat{\boldsymbol{\varepsilon}}_k),\ \mathrm{mean}(|\hat{\boldsymbol{\varepsilon}}_k|),\ \mathrm{mean}(\hat{\boldsymbol{\varepsilon}}_k^2)\Big\}\,.$$

We then propagate the path recursively:

$$\mathbf{x}_{k+1} = a_{k+1}\,\hat{\mathbf{x}}_{0,k} + b_{k+1}\,\hat{\boldsymbol{\varepsilon}}_k\,.$$

Feature variants. We use two harmonized feature families. The 1D variant computes

$$\phi_{\mathrm{1d}}(x) = \sqrt{\frac{1}{K_c-1}\sum_{k=1}^{K_c-1}\left(\frac{q_{k+1}-q_k}{\Delta\lambda_k}\right)^2}\,,$$

The 6D variant computes low-order moments of both the path values $Q = (q_1, \ldots, q_{K_c})$ and their level-wise differences $\Delta Q = (q_2 - q_1, \ldots, q_{K_c} - q_{K_c-1})$:

$$\phi_{\mathrm{6d}}(x) = \big[\,\mathrm{mean}(Q),\ \mathrm{mean}(Q^2),\ \|Q\|_3,\ \mathrm{mean}(|\Delta Q|),\ \mathrm{mean}((\Delta Q)^2),\ \|\Delta Q\|_3\big].$$

Head and scoring. We fit either a 1D KDE for the 1D feature or a diagonal Gaussian model for the 6D feature. Scores are OOD-high via negative log-density.

D.5 DDPM-OOD (harmonized faithful port)

Our DDPM-OOD baseline is implemented as a harmonized *multi-start reconstruction*. It preserves the main scoring logic of DDPM-OOD: reconstruction error should be evaluated from several noisy starting points, normalized using ID statistics *per start*, and then aggregated into a single OOD score.

Canonical reverse schedule and starting points. We first build a canonical level set

$$\lambda_1 > \lambda_2 > \cdots > \lambda_{K_c}\,,$$

ordered from clean to noisy, and reverse it into a noisy-to-clean reconstruction schedule. A subset of starting points is then selected by subsampling this reverse schedule.

Multi-start reconstruction. For a selected starting level $\lambda^{(s)}$, we generate

$$\mathbf{x}_s = a_s\,\mathbf{x}_0 + b_s\,\boldsymbol{\varepsilon}\,, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})\,,$$

and reconstruct deterministically along the corresponding reverse suffix. If the suffix levels are denoted

$$\lambda^{(s)} = \lambda_{j_1}, \lambda_{j_2}, \ldots, \lambda_{j_m}\,,$$

then at each step, we compute $(\hat{x}_{0,j_r}, \hat{\varepsilon}_{j_r})$ from the current state and propagate toward the next cleaner canonical level:

$$\mathbf{x}_{j_{r+1}} = a_{j_{r+1}}\,\hat{\mathbf{x}}_{0,j_r} + b_{j_{r+1}}\,\hat{\boldsymbol{\varepsilon}}_{j_r}\,.$$

The final clean reconstruction is the terminal $\hat{\mathbf{x}}_0$ at the end of the suffix.

Per-start reconstruction errors and ID normalization. For each start s, we compute the reconstruction error

$$m_s(\mathbf{x}_0) = \frac{1}{d}\,\|\hat{\mathbf{x}}_{0,s} - \mathbf{x}_0\|_2^2\,.$$

This yields a vector of per-start errors

$$M(\mathbf{x}_0) = \big[m_1(\mathbf{x}_0), \ldots, m_S(\mathbf{x}_0)\big]\,.$$

For each start s, we fit ID-only normalization statistics and form

$$z_s(\mathbf{x}) = \frac{m_s(\mathbf{x}) - c_s}{\sigma_s}\,.$$

Aggregation. The final scalar score is obtained by aggregating the per-start normalized deviations:

$$S(\mathbf{x}) = \mathrm{Agg}\left(z_1(\mathbf{x}), \ldots, z_S(\mathbf{x})\right),$$

where Agg denotes mean, median, or sum.

D.6 GEPC (harmonized faithful port)

Our GEPC baseline is implemented as a MBE-adapted output-consistency faithful port. The central idea is preserved: if the denoiser behaves approximately equivariantly under a small discrete transformation group, then in-distribution inputs should exhibit stronger posterior consistency than OOD inputs.

Canonical corruption and transformed outputs. For each selected canonical level λ_k, we first form

$$\mathbf{x}_k = a_k\,\mathbf{x}_0 + b_k\,\boldsymbol{\varepsilon}\,, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})\,,$$

then evaluate the denoiser on both $\mathbf{x}_k$ and transformed copies $g \cdot \mathbf{x}_k$, where g belongs to a small discrete group such as flips or 180° rotation. From each output we recover $\hat{\mathbf{x}}_{0,k}$ and $\hat{\boldsymbol{\varepsilon}}_k$, and build the score proxy

$$\hat{\mathbf{s}}_k = -\frac{\hat{\boldsymbol{\varepsilon}}_k}{b_k}\,.$$

Consistency features. We compare the reference output to the transformed-and-brought-back outputs $g^{-1} \cdot \hat{\mathbf{s}}_k(g \cdot \mathbf{x}_k)$ and $g^{-1} \cdot \hat{\mathbf{x}}_{0,k}(g \cdot \mathbf{x}_k)$. This yields several OOD-high consistency features, including a normalized score discrepancy, a cosine consistency score, and a normalized $\hat{\mathbf{x}}_0$-consistency score.

Level-wise calibration and aggregation. At each level, these raw consistency features are calibrated with an ID-only head, typically KDE or z-score normalization. Feature scores are then aggregated within each level, and finally across canonical levels using mean, weighted mean, or trimmed mean aggregation.

Table 7: **MBE, native-rerun, and reported aggregates on nine common pairs.** The table is a sanity check, not a like-for-like replacement of the controlled 12-pair benchmark. Reported rows may use different checkpoints, training policies, preprocessing, calibration, or test samples.

Method	MBE Avg.	MBE AvgWorst	Native Avg.	Native AvgWorst	Reported Avg.	Reported AvgWorst	Cost
MSMA	0.876	0.810	0.861	0.789	0.928	0.834	10F
DIFFPATH	0.834	0.754	0.918	0.842	0.918	0.842	10F
DDPM-OOD	0.588	0.369	0.586	0.361	0.742	0.657	364F
SCOPED	0.795	0.639	0.791	0.687	0.893	0.787	2F+2J
GEPC	0.757	0.650	0.910	0.818	0.910	0.818	8F

D.7 SCOPED: native and MBE implementations

Our independent SCOPED implementation follows the paper-described two-level vision detector: two score-curvature evaluations, Hutchinson estimation of the required Jacobian quantity, per-level ID-only KDE calibration, the prescribed bounded signed statistic, and two-level aggregation. Its logical cost is $2F + 2J$. For the native sanity check, we restore the reference DDPM levels $t = 1$ and $t = 300$. Under MBE, those reference levels are converted to canonical logSNR once and then mapped into the native conditioning of improved-diffusion or EDM; raw timestep indices are never compared across unrelated schedules.

D.8 Native-rerun sanity checks for harmonized baselines

To separate implementation fidelity from the effect of backbone-equated constraints, Table 7 distinguishes three regimes on the nine ID–OOD pairs common to the prior benchmarks: **MBE**, using a shared checkpoint and canonical cross-backbone levels; **Native rerun**, using the same shared checkpoint and samples but restoring each method's native timesteps, score, and aggregation; and **Reported**, taken from the method-specific setting in prior work.

For DIFFPATH and GEPC, native reruns recover the reported aggregates, so their lower MBE values arise from the shared canonical-level constraints rather than failed reproduction. For MSMA, the main gap to the reported result is consistent with the reported ID-specific checkpoint policy versus our shared-checkpoint evaluation. DDPM-OOD changes negligibly between native and MBE, so its remaining gap to the reported value is not caused by canonicalization. For SCOPED, native and MBE are also close; the remaining gap to the published 0.893 aggregate is disclosed rather than attributed to MBE, since no official vision implementation is available for exact verification.

E Focused Ablations

We restrict appendix ablations to controls that directly test the paper's central claim: the gain of CFS comes from *where* the frozen diffusion backbone is probed, rather than from hidden budget, brittle hook choices, or downstream head complexity.

E.1 Pareto budget analysis and budget-matched comparisons

A standard concern in diffusion OOD comparisons is hidden compute. We therefore report a budget-aware comparison in which logical test-time cost is made explicit. The goal is simple: if output-space baselines are given a matched or larger budget, does the representation-space advantage persist?

More precisely, we vary the number of selected canonical levels while keeping Monte Carlo test $\mathrm{MC}_{\mathrm{test}} = 1$, so that the dominant logical cost scales with the number of backbone evaluations. For CFS, $K_c \times K_s$ denotes the number of selected canonical levels and retained stage slots. Since all retained hooks at a given level are extracted in the same forward pass, logical cost depends on K_c, not on K_s.

Table 8 shows that the best one-forward operating points are already sparse CFS variants. In particular, $\mathrm{CFS}_{\mathrm{dec}}(1 \times 1_{\mathrm{low}})$ and $\mathrm{CFS}(1 \times 2_{\mathrm{low}})$ dominate budget-matched output-space baselines while using only one backbone evaluation per image. Richer CFS variants improve representation size without changing the backbone cost at fixed K_c, but the gains beyond the strongest one-forward operating

Table 8: **Budget-aware comparison under MBE.** We vary the logical test-time budget while keeping $\mathrm{MC}_{\mathrm{test}} = 1$. This disentangles representation quality from hidden multilevel accumulation. For CFS, $K_c \times K_s$ denotes the number of selected canonical levels and retained stage slots, respectively. Because all retained hooks at a given canonical level are extracted within the same backbone forward pass, the logical cost depends on K_c only, not on K_s.

Budget (#F/img)	Method	Improved-diffusion AvgAUROC ↑	Improved-diffusion AvgWorstAUROC ↑	EDM AvgAUROC ↑	EDM AvgWorstAUROC ↑
1	$\mathrm{CFS}_{\mathrm{enc}}\ (1 \times 1_{\mathrm{low}})$	0.849	0.773	0.900	0.805
1	$\mathrm{CFS}_{\mathrm{dec}}\ (1 \times 1_{\mathrm{low}})$	<u>0.886</u>	<u>0.793</u>	**0.919**	<u>0.809</u>
1	CFS $(1 \times 2_{\mathrm{low}})$	**0.887**	**0.799**	<u>0.917</u>	**0.814**
1	CFS $(1 \times 2_{\mathrm{high}})$	0.497	0.492	0.498	0.495
1	$\mathrm{CFS}_{\mathrm{enc}}\ (1 \times 1_{\mathrm{high}})$	0.500	0.496	0.498	0.494
1	$\mathrm{CFS}_{\mathrm{dec}}\ (1 \times 1_{\mathrm{high}})$	0.495	0.486	0.498	0.495
2	CFS $(2 \times 1_{\mathrm{enc}})$	0.839	0.761	0.880	0.781
2	CFS $(2 \times 1_{\mathrm{dec}})$	0.853	0.762	0.906	0.800
2	CFS (2×2)	0.868	0.779	0.902	0.804
2	CFS (2×4)	0.873	0.783	0.900	0.794
2	MSMA	0.781	0.652	0.770	0.645
2	DIFFPATH	0.802	0.633	0.778	0.630
4	CFS 4×2	0.843	0.763	0.873	0.791
4	MSMA	0.785	0.669	0.790	0.673
4	DIFFPATH	0.766	0.642	0.787	0.631
8	MSMA	0.779	0.675	0.798	0.682
8	DIFFPATH	0.779	0.640	0.792	0.634

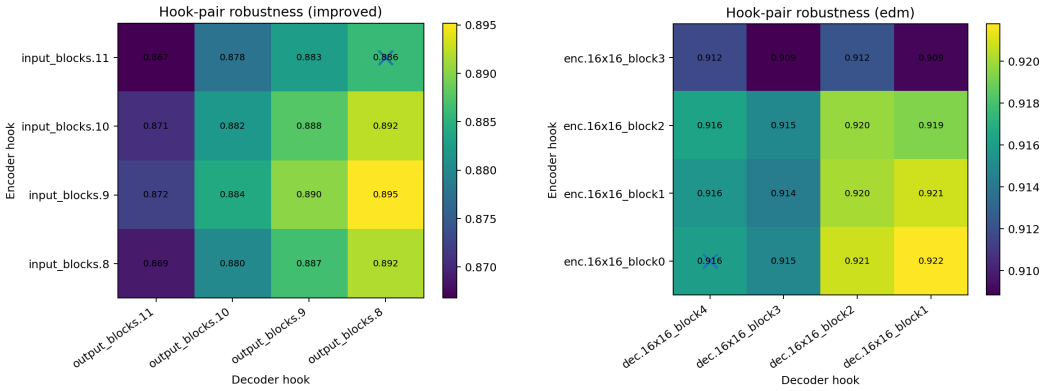

Figure 5: **Corresponding hook-pair heatmaps for CFS(1×2).** Each cell reports the Avg AUROC of one admissible pair at the low-noise canonical level used in the main paper. The marker denotes the final pair obtained from the ID-side proxy shortlist. For the improved backbone, the selected pair falls in the same high-performing basin as the oracle pair. For EDM, the dominant pattern is a broad plateau of strong pairs, indicating low sensitivity to the exact hook choice even when the proxy is less predictive.

points are modest. The key conclusion is therefore not merely that CFS is accurate, but that its gain does not come from hidden multilevel accumulation.

E.2 Hook-pair robustness and proxy validation

A natural concern is that the encoder-decoder pair selected by CFS(1×2) could be fragile or cherry-picked. To test this, we evaluate multiple admissible encoder and decoder candidates within the same structural regions and report (i) the pairwise performance landscape, (ii) the relation between the ID-side pair proxy and final pair performance, and (iii) a complementary region-wise proxy validation. All results are computed at the fixed low-noise canonical level used in the main paper.

Figure 5 supports robustness more strongly than exact pair recovery. For the improved backbone, the landscape is structured rather than spiky: the proxy-selected pair reaches 0.886 Avg AUROC, versus

Table 9: **Region-wise proxy-selected modules versus empirical oracle modules for CFS(1×2).** For each backbone and region, we compare the downstream $\mathrm{AvgAUROC}$ obtained with the proxy-selected module to that obtained with the best admissible module in hindsight, while keeping the opposite region fixed to its proxy-selected choice. Small gaps indicate that the proxy-selected module lies close to the empirical oracle without using any OOD labels.

Backbone	Region	Proxy-selected $\mathrm{AvgAUROC}\uparrow$	Oracle $\mathrm{AvgAUROC}\uparrow$	Gap $\downarrow$
Improved-diffusion	Encoder	0.8861	0.8952	0.0091
Improved-diffusion	Decoder	0.8861	0.8861	0.0000
EDM	Encoder	0.9161	0.9163	0.0002
EDM	Decoder	0.9161	0.9218	0.0057

0.895 for the oracle admissible pair, so the proxy lands in the correct high-performing basin even without identifying the exact optimum. For EDM, the heatmap is flatter, and the main conclusion is different: several nearby pairs perform similarly well, so the method is not driven by a single brittle choice.

Table 9 shows that, on improved-diffusion, the proxy is more informative on the decoder side than on the encoder side, matching the stronger decoder-column structure in the heatmap. On EDM, the proxy is weaker as a regional ranker, but the oracle gaps remain small. Overall, the selection rule does not need to find the exact best pair; it only needs to place CFS(1×2) in a stable, high-performing part of the network without exhaustive search.

E.3 Canonical-level robustness

The main paper identifies low-noise canonical probing as the strongest operating regime. We now test whether this effect reflects a stable region or a narrowly tuned choice of canonical level.

Since the final method uses single-level variants, we sweep the canonical level λ and evaluate two representative detectors: $\mathrm{CFS}_{\mathrm{dec}}(1\times1)$ and CFS(1×2). The key question is whether performance remains strong over a neighborhood of low-noise levels, rather than peaking at a single hand-picked value.

Figure 6 shows the same qualitative pattern on both backbones. For strongly negative λ, both detectors remain close to chance, indicating that highly corrupted canonical levels are not useful. Performance then rises rapidly as λ moves toward lower-noise regimes, before flattening into a broad plateau for positive λ. Thus, the benefit of low-noise probing is not confined to one operating point.

There are, however, mild backbone-specific differences. On improved-diffusion, CFS(1×2) is most helpful in the transition regime, whereas on the final high-λ plateau the gap to decoder-only probing becomes small. EDM shows an even flatter plateau once λ becomes positive. Overall, the sweeps suggest that the canonical level is the primary driver, whereas richer representation composition mainly improves the intermediate regime and worst-case robustness.

E.4 CFS family analysis

The role of representation composition is already visible in the budget-matched rows of Table 8, so we do not duplicate those results. At fixed logical cost, the comparison between CFS$(2 \times 1_{\mathrm{enc}})$, CFS$(2 \times 1_{\mathrm{dec}})$, and CFS$(2 \times 2)$ shows how performance changes when using encoder-only, decoder-only, or combined encoder-decoder snapshots across two canonical levels.

Decoder-only probing already captures most of the gain once the canonical level reaches the strong low-noise regime. Encoder-decoder composition is most useful before the final plateau is reached and for worst-case robustness. The main claim is therefore not that composition is always necessary, but that it provides a meaningful robustness margin beyond an already strong decoder-only baseline.

E.5 Matched clean-input control

We test whether CFS merely behaves as a clean frozen-feature detector by keeping the checkpoint, selected hooks, pooling, ID-only head, ID bank, test samples, and $1F$ cost fixed while replacing the

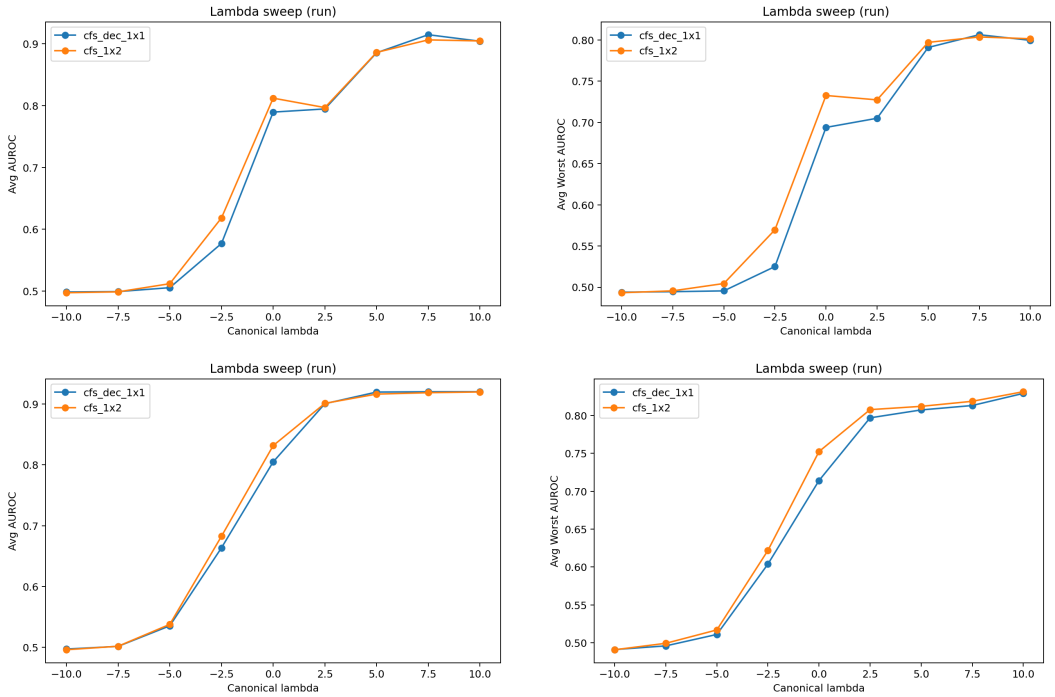

Figure 6: **Canonical-level robustness for single-level CFS variants. Top:** improved-diffusion backbone. **Bottom:** EDM backbone. **Left:** $\mathrm{AvgAUROC}$. **Right:** $\mathrm{AvgWorstAUROC}$. Across both backbones, performance is near chance for strongly negative λ, rises sharply through the intermediate regime, and then enters a broad high-performing plateau at positive λ. This supports the claim that the low-noise advantage is robust rather than tied to a single finely tuned canonical level.

Table 10: **Matched clean-endpoint control.** Same frozen representation and ID-only score; only the input/conditioning endpoint changes.

Variant	Backbone	Low-noise Avg./AvgWorst	Clean Avg./AvgWorst
$\mathrm{CFS}_{\mathrm{dec}}(1\times1)$	Improved	0.938 / 0.857	0.912 / 0.828
$\mathrm{CFS}(1\times2)$	Improved	0.935 / 0.855	0.910 / 0.817
$\mathrm{CFS}_{\mathrm{dec}}(1\times1)$	EDM	0.920 / 0.808	0.914 / 0.816
$\mathrm{CFS}(1\times2)$	EDM	0.917 / 0.813	0.915 / 0.820

low-noise corrupted input by the clean endpoint. The low-noise arm below is a dedicated high-logSNR diagnostic at $\lambda = 7$; it is not used to replace the prespecified $\lambda = 5$ main MBEresults.

Sparse diffusion features are already strong at the clean endpoint. Low-noise probing improves AvgAUROC in all four comparisons (by 0.002–0.026), while AvgWorst is backbone-dependent: the low-noise intervention improves both Improved readouts but the clean endpoint remains slightly stronger on EDM. The corruption therefore does not create the OOD signal, although it can improve average separation.

E.6 Matched non-diffusion and random-weight controls

We compare CFS with frozen representations using the same ID samples, channel-wise mean/std pooling, matching late/two readouts, and the same diagonal ID-only score. This controls the pooling and downstream scoring rule, allowing comparison of the resulting frozen representations.

Randomizing the diffusion U-Net weights while preserving its architecture and hooks also causes a large drop: decoder/pair AvgAUROC falls to $0.690/0.713$ on Improved and $0.784/0.787$ on EDM (AvgWorst $0.638/0.662$ and $0.675/0.680$, respectively). Thus arbitrary pooled activations or architecture alone do not explain CFS; the learned diffusion representation contributes materially.

Table 11: **Matched frozen-representation controls.** The strongest generic alternative is SimCLR, but it remains below learned diffusion representations under the same low-capacity ID-only score.

Representation / training	Slot	Avg AUROC	AvgWorst
ResNet-18 random	late	0.647	0.578
ResNet-18 random	two	0.668	0.590
ResNet-18 CIFAR-10 supervised	late	0.710	0.660
ResNet-18 CIFAR-10 supervised	two	0.808	0.730
ResNet-50 CIFAR-10 SimCLR	late	0.745	0.697
ResNet-50 CIFAR-10 SimCLR	two	0.852	0.761
CFS Improved	late	0.886	0.793
CFS Improved	two	0.887	0.799
CFS EDM	late	0.919	0.809
CFS EDM	two	0.917	0.814

Table 12: **Image-complexity analysis on the difficult CIFAR-10 shifts.** Complexity alone is weak, while removing its predictable component from CFS decreases aggregate performance without repairing the hard pairs. These results come from a separate control run and may differ slightly from the primary benchmark results.

Backbone	Detector	C10→CelebA	C10→C100	Avg.	AvgWorst
–	Complexity only	0.430	0.541	0.759	0.620
Improved	CFS decoder	0.534	0.688	0.886	0.791
Improved	Decoder, complexity-regressed	0.380	0.521	0.813	0.685
Improved	CFS paired	0.519	0.635	0.886	0.797
Improved	Paired, complexity-regressed	0.381	0.525	0.821	0.707
EDM	CFS decoder	0.444	0.700	0.919	0.807
EDM	Decoder, complexity-regressed	0.338	0.559	0.887	0.756
EDM	CFS paired	0.462	0.667	0.916	0.812
EDM	Paired, complexity-regressed	0.358	0.558	0.886	0.764

E.7 Difficult shifts and image-complexity residualization

To test whether CFS primarily proxies simple visual complexity, we compute seven low-level statistics: total variation, gradient energy, Laplacian energy, grayscale entropy, pixel standard deviation, high-frequency Fourier energy, and lossless compression size. Using ID data only, a Ridge model predicts the CFS score from these statistics. Let $\mathbf{c}(x)$ be the complexity vector and $s_{\mathrm{CFS}}(x)$ the original score. We fit

$$\widehat{s}_{\mathrm{comp}}(x) = f_{\mathrm{Ridge}}(\mathbf{c}(x)), \qquad r(x) = s_{\mathrm{CFS}}(x) - \widehat{s}_{\mathrm{comp}}(x),$$

and evaluate both complexity-only detection and an ID-calibrated two-sided score derived from $r(x)$. Here "complexity-regressed" refers only to this residualization; it is unrelated to reconstruction residuals or residual network blocks.

Complexity alone is substantially weaker globally and near chance on both difficult pairs. Residualization lowers AvgAUROC by 0.030–0.073 and AvgWorst by 0.048–0.106 without resolving either shift. We therefore treat CIFAR-10 to CelebA/CIFAR-100 as genuine limitations rather than as artifacts of a simple complexity confound.

E.8 Pooling rule

We compare mean-only, std-only, and mean+std pooling for CFS(1×2). This tests whether first-order channel averages suffice or whether within-channel dispersion carries additional useful signal once both encoder and decoder snapshots are retained. Results are presented in Table 13.

Combining mean and standard-deviation pooling is clearly beneficial on improved-diffusion, where it gives the best average and worst-case performance. On EDM, mean+std gives the best average AUROC, while mean-only is marginally better in AvgWorstAUROC. Overall, dispersion features help the main operating point, but their gain is not perfectly uniform across criteria.

Table 13: **Pooling rule ablation for CFS(1×2).** We report both backbone families separately to test whether the gain of the primary detector relies only on first-order channel averages or also on within-channel dispersion captured by standard deviation pooling.

Pooling	Improved-diffusion		EDM	
	AvgAUROC ↑	AvgWorstAUROC ↑	AvgAUROC ↑	AvgWorstAUROC ↑
Mean only	0.865	0.779	0.908	**0.815**
Std only	0.868	0.738	0.905	0.769
Mean + Std	**0.887**	**0.799**	**0.916**	0.814

Table 14: **Head sensitivity under the same sparse representation for CFS(1×2).** We report both backbone families separately. The goal is to test whether the usefulness of the primary detector comes mainly from the selected representation rather than from a particular downstream head.

Head	Improved-diffusion		EDM	
	AvgAUROC ↑	AvgWorstAUROC ↑	AvgAUROC ↑	AvgWorstAUROC ↑
Diagonal (main paper)	<u>0.887</u>	0.799	0.917	0.814
Shrinkage covariance	0.876	<u>0.823</u>	<u>0.921</u>	<u>0.862</u>
KNN	**0.910**	**0.862**	**0.950**	**0.904**
GMM-light	0.860	0.784	0.912	0.784

E.9 Diagonal score versus alternative ID-only scores

The main paper uses a lightweight diagonal score so that the comparison stays focused on representation quality rather than downstream classifier capacity. We now test whether the strong performance of CFS(1×2) persists under stronger but still ID-only scores.

Table 14 shows that the sparse CFS(1×2) representation remains strong across several ID-only heads, which supports the view that the main signal comes from the selected features rather than from a specialized classifier. At the same time, the exact ranking is not head-invariant: KNN is strongest on both backbones, and shrinkage covariance also improves worst-case robustness over the diagonal head.

We therefore interpret the diagonal score as a conservative evaluation choice rather than as the empirically strongest one. Its role in the main paper is to keep the detector lightweight and to avoid conflating feature quality with additional head capacity.

E.10 ID-fit bank size sensitivity

Because CFS(1×2) is calibrated from ID-only reference statistics, practical performance may depend on the size of the ID-fit bank. We therefore vary the number of ID-fit samples used to estimate the slot statistics and report the resulting performance.

Table 15 shows that CFS(1×2) is only weakly sensitive to the ID-fit bank size on both backbones. Even a bank of 100 ID samples already performs very close to the larger-bank regime. There is therefore no evidence that the main operating point depends on an unusually large calibration set.

E.11 Statistical stability across random seeds

Finally, we report repeated runs that vary the stochastic seed in the evaluation pipeline. The goal is not to estimate every source of variance exhaustively, but to check whether the main conclusions remain stable under the stochastic components that most directly affect the reported scores.

Tables 16 and 17 report seed stability for both CFS(1×2) and CFS$_{\mathrm{dec}}(1\times1)$.

The small standard deviations indicate stability with respect to the stochastic components varied in these repeated runs. The qualitative ordering agrees with the primary benchmark: on improved-diffusion, the paired variant has slightly higher average AUROC and higher AvgWorstAUROC; on EDM, decoder-only has higher average AUROC, while the paired variant has higher AvgWorstAUROC.

Table 15: **Sensitivity to the ID-fit bank size for CFS(1×2).** Each entry reports the mean over three random seeds. The goal is to assess whether the sparse two-slot representation reaches stable performance with moderate ID-only calibration data.

ID-fit bank size	Improved-diffusion		EDM	
	AvgAUROC ↑	AvgWorstAUROC ↑	AvgAUROC ↑	AvgWorstAUROC ↑
100	0.8873	0.8031	0.9177	0.8207
500	0.8871	0.8006	0.9164	0.8150
1000	0.8863	0.7982	0.9161	0.8117
Full	0.8866	0.7979	0.9164	0.8127

Table 16: **Statistical stability across random seeds on improved-diffusion.** We report mean $\pm$ standard deviation over repeated runs. The very small deviations indicate that the comparison between the two sparse CFS variants is not driven by stochastic fluctuations.

Method	AvgAUROC ↑	AvgWorstAUROC ↑	FPR95 ↓
$\text{CFS}_{\text{dec}}(1\times1)$	0.88566 ± 0.00008	0.79103 ± 0.00015	0.35678 ± 0.00016
$\text{CFS}(1\times2)$	0.88609 ± 0.00002	0.79721 ± 0.00003	0.33546 ± 0.00024

E.12 Architecture-transfer sanity check

Although our main experiments use U-Net-style diffusion backbones, CFS only requires access to block-level internal activations at canonical corruption levels. As a sanity check, we apply the same sparse snapshot construction to a transformer-based diffusion backbone, U-ViT. Unlike convolutional U-Nets, U-ViT keeps a token sequence of nearly constant size across its transformer blocks; its U-shaped structure comes from input blocks, a middle block, output blocks, and long skip connections rather than from explicit spatial downsampling and upsampling. We therefore replace the U-Net encoder/decoder hook taxonomy by early, middle, and late transformer-block snapshots, and pool token features using channel-wise mean and standard deviation over image tokens.

The purpose is to test whether the sparse-snapshot principle transfers beyond convolutional U-Nets. All runs use the same low-noise canonical level, with target $\lambda = 5.0$, effective $\lambda = 5.0329$, native U-ViT timestep $t = 21$, and $b^2 = 6.48 \times 10^{-3}$. The hooked U-ViT activations have shape $B \times 257 \times 512$, corresponding to 256 image tokens plus one time token; CFS pools the image tokens only. We also include an output-space baseline using the same U-ViT checkpoint and the same canonical level: it scores images by the final reconstruction MSE $\|\hat{\mathbf{x}}_0 - \mathbf{x}_0\|^2$ with an ID-only one-dimensional density fit.

Tables 18 and 20 support the architecture-transfer interpretation: sparse internal snapshots remain informative even when spatial U-Net feature maps are replaced by token-level transformer states. Importantly, the gain does not come merely from using a low-noise U-ViT forward: the output reconstruction baseline reaches only 0.778 AvgAUROC and 0.653 AvgFPR95, whereas CFS-late improves to 0.891 AvgAUROC and 0.213 AvgFPR95 at the same one-forward cost. The paired early+late variant slightly improves AvgAUROC, while the late snapshot gives the best AvgFPR95, suggesting that late transformer blocks already concentrate most of the useful signal in this setting. We emphasize that this result should not be read as a new U-ViT OOD benchmark: performance is not uniform across all source datasets, and the CIFAR-10-as-ID case remains the main failure mode.

We additionally evaluate whether the diagnostics used for U-Net sparse probes transfer across transformer blocks. We use six U-ViT blocks plus the first+last pair over canonical levels $\{-7.5, -2.5, 2.5, 5\}$, with the same token mean/std pooling and ID-only diagonal score.

The first two relationships transfer strongly to the transformer setting. The third is positive and significant but more moderate, indicating that scalar injected noise does not by itself determine transformer feature variability; block depth, attention, residual pathways, and token mixing also contribute.

Table 17: **Statistical stability across random seeds on EDM.** We report mean ± standard deviation over repeated runs. Again, the deviations are extremely small, showing that the reported trade-off between the two sparse CFS variants is stable across seeds.

Method	AvgAUROC ↑	AvgWorstAUROC ↑	FPR95 ↓
$\text{CFS}_{\text{dec}}(1\times1)$	0.91950 ± 0.00009	0.80730 ± 0.00027	0.18149 ± 0.00021
$\text{CFS}(1\times2)$	0.91618 ± 0.00009	0.81223 ± 0.00005	0.19132 ± 0.00012

Table 18: **Architecture-transfer sanity check on U-ViT.** CFS probes early, middle, and late U-ViT block outputs instead of U-Net encoder/decoder hooks. All variants use one canonical low-noise level, one backbone forward, token mean/std pooling, and an ID-only diagonal score. The output baseline uses the same checkpoint, level, and one-forward budget, but scores only the final reconstruction error. Averages are computed over 12 ordered pairs: for each ID dataset in {CIFAR-10, SVHN, CelebA32}, we evaluate against the other two source datasets plus CIFAR-100 and DTD as OOD.

Backbone	Method	Region / signal	AvgAUROC ↑	AvgFPR95 ↓	Cost ↓
U-ViT-CIFAR10	Output recon. MSE	output	0.778	0.653	1F
U-ViT-CIFAR10	CFS-early	`in_blocks.2`	0.872	0.356	1F
U-ViT-CIFAR10	CFS-mid	`mid_block`	0.874	0.292	1F
U-ViT-CIFAR10	CFS-late	`out_blocks.2`	<u>0.891</u>	**0.213**	1F
U-ViT-CIFAR10	CFS-early+late	`in_blocks.2+out_blocks.2`	**0.896**	<u>0.238</u>	1F

F Extended CIFAR-Scale Results

This section provides the full pairwise breakdown behind the compact summaries shown in the main paper. Its role is descriptive rather than argumentative: it exposes where the main representation-space advantage comes from, pair by pair, under the same controlled protocol.

We begin with the full pairwise breakdown for the primary shared-source benchmark, i.e. the CIFAR10-source checkpoint family used in the main paper. Tables 21 and 22 report AUROC and FPR95 for all ID→OOD pairs. To keep the appendix readable, we expose pairwise operating-point behavior mainly for the two main sparse variants, $\text{CFS}_{\text{dec}}(1\times1)$ and $\text{CFS}(1\times2)$.

The full tables make two patterns clear. First, the main-paper summaries are not driven by one isolated pair: the sparse CFS variants are strong across a broad set of ID→OOD configurations, especially on compact IDs and texture-driven shifts. Second, the decoder-only and paired encoder-decoder variants remain very close on many pairs, which reinforces the compression result already visible in the main tables.

G Full Source-Family Robustness Results

This section tests whether the main representation-space pattern is tied to one checkpoint family or survives a controlled change in the frozen source representation.

We report the full pairwise breakdown for the alternative CelebA32-source checkpoint family. Tables 23 and 24 provide the detailed AUROC and FPR95 results. The goal is to test whether the source-family change modifies only absolute performance or also the qualitative ranking between representation-space and output-space methods.

A source-family change modifies the representation bias of the frozen checkpoint. The relevant question is therefore not whether absolute numbers remain identical, but whether the ranking and the representation-versus-output-space contrast remain qualitatively stable.

The full tables support the same broad conclusion as the compact main-paper summary: changing the source family shifts absolute numbers but does not erase the main pattern. Sparse representation-space probing remains highly competitive, and the contrast between CFS and broader output-space baselines survives the change in frozen source representation.

Table 19: **Diagnostic transfer on U-ViT.** Spearman correlations use the same CIFAR-10 U-ViT checkpoint and ID banks as the architecture-transfer experiment.

Diagnostic relation	ρ	n
$\hat{\kappa}/d$ vs pairwise AUROC	0.879	336
$\hat{R}$ vs average AUROC	0.731	84
b^2 vs within-image feature variance	0.400	84

Table 20: **Per-source breakdown of the U-ViT architecture-transfer sanity check.** Each cell reports the average over the four OOD datasets associated with the corresponding ID source. The output baseline uses final reconstruction MSE, while CFS variants use internal token snapshots. The main failure mode is CIFAR-10 as ID, especially against CelebA32 and CIFAR-100.

Method	Metric	ID=SVHN	ID=CIFAR-10	ID=CelebA32
Output recon. MSE	AvgAUROC ↑	0.927	0.594	0.813
Output recon. MSE	AvgFPR95 ↓	0.467	0.803	0.687
CFS-early	AvgAUROC ↑	0.945	**0.722**	0.949
CFS-early	AvgFPR95 ↓	0.238	0.638	0.191
CFS-mid	AvgAUROC ↑	0.979	0.652	<u>0.991</u>
CFS-mid	AvgFPR95 ↓	0.086	0.747	0.044
CFS-late	AvgAUROC ↑	**0.996**	0.686	**0.992**
CFS-late	AvgFPR95 ↓	**0.016**	<u>0.587</u>	**0.036**
CFS-early+late	AvgAUROC ↑	<u>0.985</u>	<u>0.719</u>	0.984
CFS-early+late	AvgFPR95 ↓	<u>0.072</u>	**0.575**	<u>0.066</u>

H External Positioning Against Prior Reported Diffusion Results

This section provides external context by comparing CFS to previously reported diffusion-based CIFAR-scale results.

Table 25 shows that the sparse representation-space advantage of CFS is not confined to the controlled MBE setting. Even outside backbone-equated evaluation, CFS remains competitive with or stronger than previously reported single-checkpoint diffusion results, while using only one backbone evaluation per image.

H.1 Positioning against DLSR on its native published settings

DLSR is the closest prior in spirit, but it studies a different object: learned feature reconstruction rather than sparse probing of native frozen states. It also falls outside our MBE protocol. Since DLSR is only available on the ID/OOD pairs supported by its published setup, we report in Table 26 a positioning comparison on the DLSR-native overlapping subset, without treating this table as part of the main MBE claim.

DLSR is therefore evaluated only on pairs supported by the original protocol/repo. The full cross-ID benchmark would require extending the original DLSR training and evaluation pipeline beyond the published setup.

I Checkpoint-Controlled Large-Scale Results

This appendix reports a checkpoint-controlled large-scale comparison on ImageNet200 and ImageNet1K using a single official ImageNet-64 improved-diffusion checkpoint. These experiments are not meant to replace the controlled CIFAR-scale evidence in the main paper. Their role is narrower: to test whether the same representation-first signal remains informative in a substantially harder large-scale regime under a fixed backbone.

In all experiments below, all methods use the same official ImageNet-64 improved-diffusion checkpoint (`imagenet64_uncond_100M_1500K.pt`). We evaluate on two near-OOD regimes, **NINCO** and **SSB-hard**, and one far-OOD regime, **Textures**. Unless stated otherwise, both CFS variants use

Table 21: **Full per-pair AUROC on the CIFAR-scale benchmark under the primary CIFAR10-source policy.** Each entry corresponds to one ID→OOD pair. The same fixed shared-source protocol is used for all methods. All metrics are reported with three decimal places.

		CIFAR10 as ID				SVHN as ID				CelebA32 as ID			
Method	Backbone	SVHN	CelebA32	CIFAR100	DTD	CIFAR10	CelebA32	CIFAR100	DTD	CIFAR10	SVHN	CIFAR100	DTD
MSMA	improved	0.646	0.438	0.579	0.769	0.916	0.925	0.935	0.936	0.710	0.950	0.762	0.936
MSMA	EDM	0.663	0.490	0.606	0.796	0.937	0.962	0.936	0.941	0.640	0.933	0.738	0.908
DIFFPATH	improved	0.886	0.371	0.550	0.683	0.940	0.965	0.946	0.914	0.640	0.955	0.688	0.797
DIFFPATH	EDM	0.832	0.316	0.561	0.780	0.903	0.960	0.891	0.924	0.697	0.970	0.763	0.911
DDPM-OOD	improved	0.132	0.527	0.538	0.433	0.894	0.947	0.904	0.756	0.476	0.059	0.522	0.415
DDPM-OOD	EDM	0.124	**0.586**	0.568	0.431	0.907	0.965	0.926	0.802	0.441	0.035	0.516	0.409
GEPC	improved	0.584	0.487	0.494	0.600	0.754	0.580	0.595	0.584	0.736	0.659	0.571	0.747
GEPC	EDM	0.917	0.317	0.506	0.680	0.963	0.954	0.855	0.798	0.799	0.962	0.686	0.848
CFS$_{\text{dec}}(1\times1)$	improved	0.925	0.540	0.687	0.941	0.964	0.897	0.873	0.893	0.983	0.973	0.966	0.990
CFS$_{\text{dec}}(1\times1)$	EDM	**0.969**	0.451	**0.697**	**0.968**	**0.997**	**0.993**	0.986	**0.991**	**0.993**	**0.999**	**0.990**	**0.999**
CFS(1×2)	improved	0.888	0.524	0.635	0.932	0.966	0.947	0.907	0.917	0.981	0.984	0.966	0.993
CFS(1×2)	EDM	**0.969**	0.468	0.665	0.959	0.996	0.990	**0.989**	**0.991**	0.988	**0.999**	0.986	0.998

Table 22: **Full per-pair FPR95 on the CIFAR-scale benchmark under the primary CIFAR10-source policy.** Each entry corresponds to one ID→OOD pair. Lower is better. All metrics are reported with three decimal places.

		CIFAR10 as ID				SVHN as ID				CelebA32 as ID			
Method	Backbone	SVHN	CelebA32	CIFAR100	DTD	CIFAR10	CelebA32	CIFAR100	DTD	CIFAR10	SVHN	CIFAR100	DTD
MSMA	improved	0.828	0.966	0.925	0.780	0.416	0.310	0.367	0.387	0.872	0.273	0.825	0.412
MSMA	EDM	0.839	0.945	0.908	0.782	0.339	0.169	0.356	0.380	0.904	0.266	0.841	0.574
DIFFPATH	improved	0.335	0.969	0.931	0.882	0.381	0.185	0.370	0.590	0.897	0.188	0.868	0.745
DIFFPATH	EDM	0.435	0.979	0.927	0.785	0.641	0.206	0.704	0.539	0.891	0.139	0.840	0.553
DDPM-OOD	improved	0.996	0.853	0.948	0.996	0.548	0.229	0.523	0.969	0.993	1.000	0.993	1.000
DDPM-OOD	EDM	0.996	**0.809**	0.932	0.993	0.468	0.134	0.387	0.900	0.998	1.000	0.996	1.000
GEPC	improved	0.826	0.948	0.949	0.847	0.714	0.863	0.874	0.892	0.793	0.832	0.903	0.712
GEPC	EDM	0.330	0.977	0.947	0.902	0.241	0.261	0.672	0.851	0.801	0.188	0.856	0.683
CFS$_{\text{dec}}(1\times1)$	improved	0.269	0.835	**0.803**	0.271	0.230	0.431	0.523	0.487	0.095	0.126	0.149	0.048
CFS$_{\text{dec}}(1\times1)$	EDM	**0.093**	0.868	0.849	**0.133**	**0.011**	**0.033**	0.064	**0.043**	**0.036**	**0.003**	**0.046**	**0.003**
CFS(1×2)	improved	0.371	0.842	0.855	0.329	0.205	0.216	0.409	0.404	0.109	0.080	0.157	0.037
CFS(1×2)	EDM	0.098	0.862	0.861	0.177	0.021	0.042	**0.052**	0.048	0.057	0.004	0.065	0.007

a single low-noise canonical level and therefore retain a logical test-time cost of $1F + 0J$, whereas MSMA and DIFFPATH use $10F + 0J$.

Table 27 reports the resulting large-scale comparison. In addition to the primary main-paper operating point CFS(1×2), we also include the ultra-sparse decoder-only companion CFS$_{\text{dec}}(1\times1)$.

These numbers constitute a checkpoint-controlled stress test of whether the representation-first signal remains informative outside the controlled CIFAR-scale regime

This checkpoint-controlled large-scale view yields a clear pattern. First, near-OOD ImageNet-scale detection remains difficult for all methods, especially on SSB-hard. Second, Textures is substantially more discriminative and reveals a clear advantage for sparse CFS probes over MSMA and DIFFPATH. Third, under this official improved-backbone setting, the sparse decoder-only variant CFS$_{\text{dec}}(1\times1)$ is consistently slightly stronger than CFS(1×2) on both ImageNet200 and ImageNet1K, while preserving the same $1F$ logical cost.

Interestingly, in the ImageNet-scale checkpoint-controlled setting, CFS$_{\text{dec}}(1\times1)$ is slightly stronger than CFS(1×2). The result suggests that, in this harder large-scale regime, the conditional encoder residual of Theorem 1, is either small or not reliably exploitable by the lightweight diagonal score, so the late decoder snapshot remains the most robust sparse probe.

The NINCO and SSB-hard results indicate that large-scale semantic OOD remains difficult for all diffusion-based methods under this fixed single-backbone evaluation setting. These experiments use a pixel-space ImageNet-64 diffusion backbone; they should not be interpreted as validation of latent-space diffusion, where the autoencoder and latent geometry introduce additional factors.

J Implementation and Configuration Details

This section reports the implementation details needed to reproduce the main CFS runs and the corresponding ablations. All methods are evaluated through the shared MBE adapter described in

Table 23: **Full per-pair AUROC on the CIFAR-scale benchmark under the alternative CelebA32-source policy.** This is the detailed source-family robustness breakdown complementary to the main-paper summary. All metrics are reported with three decimal places.

		CIFAR10 as ID				SVHN as ID				CelebA32 as ID			
Method	Backbone	SVHN	CelebA32	CIFAR100	DTD	CIFAR10	CelebA32	CIFAR100	DTD	CIFAR10	SVHN	CIFAR100	DTD
MSMA	improved	0.766	0.859	0.562	0.781	0.937	0.976	0.944	0.934	0.956	0.930	0.953	0.973
MSMA	EDM	0.659	0.278	0.536	0.705	0.928	0.937	0.931	0.891	0.851	0.984	0.860	0.916
DiffPath	improved	**0.952**	0.641	0.538	0.700	**0.979**	0.946	**0.970**	0.913	0.779	0.915	0.787	0.827
DiffPath	EDM	0.817	0.350	0.524	0.726	0.936	0.957	0.933	0.922	0.680	0.900	0.709	0.864
DDPM-OOD	improved	0.094	0.201	0.482	0.349	0.928	0.761	0.914	0.741	0.841	0.251	0.816	0.595
DDPM-OOD	EDM	0.069	0.242	0.503	0.348	0.955	0.914	0.950	0.809	0.750	0.097	0.741	0.520
GEPC	improved	0.533	**0.975**	0.508	0.643	0.516	0.963	0.515	0.577	0.929	0.944	0.928	0.957
GEPC	EDM	0.728	0.545	0.533	0.654	0.910	0.966	0.913	0.833	0.577	0.925	0.590	0.765
CFS$_{\text{dec}}(1\times1)$	improved	0.827	0.943	0.604	0.865	0.884	**0.997**	0.905	0.867	**0.999**	**1.000**	**0.998**	**1.000**
CFS$_{\text{dec}}(1\times1)$	EDM	0.858	0.953	0.605	0.863	0.935	0.991	0.950	0.962	0.941	0.976	0.947	0.985
CFS(1×2)	improved	0.828	0.907	0.602	0.878	0.903	0.993	0.916	0.881	0.997	**1.000**	0.996	0.999
CFS(1×2)	EDM	0.879	0.957	**0.618**	**0.887**	0.955	0.993	0.966	**0.974**	0.963	0.989	0.968	0.993

Table 24: **Full per-pair FPR95 on the CIFAR-scale benchmark under the alternative CelebA32-source policy.** This is the detailed source-family robustness breakdown complementary to the main-paper summary. Lower is better. All metrics are reported with three decimal places.

		CIFAR10 as ID				SVHN as ID				CelebA32 as ID			
Method	Backbone	SVHN	CelebA32	CIFAR100	DTD	CIFAR10	CelebA32	CIFAR100	DTD	CIFAR10	SVHN	CIFAR100	DTD
MSMA	improved	0.627	0.520	0.941	0.788	0.380	0.121	0.335	0.423	0.233	0.298	0.251	0.135
MSMA	EDM	0.768	0.990	0.941	0.872	0.362	0.261	0.355	0.593	0.618	0.069	0.600	0.460
DiffPath	improved	**0.118**	0.848	0.948	0.875	**0.057**	0.312	0.141	0.674	0.849	0.293	0.845	0.820
DiffPath	EDM	0.351	0.975	0.940	0.813	0.399	0.181	0.434	0.470	0.886	0.235	0.877	0.668
DDPM-OOD	improved	0.999	0.986	0.962	0.999	0.370	0.681	0.456	0.960	0.768	0.998	0.837	0.999
DDPM-OOD	EDM	1.000	0.977	0.953	0.999	0.220	0.280	0.243	0.897	0.901	1.000	0.925	1.000
GEPC	improved	0.922	0.193	0.948	0.864	0.937	0.202	0.943	0.914	0.175	0.137	0.188	0.135
GEPC	EDM	0.858	0.918	0.937	0.906	0.425	0.151	0.433	0.757	0.929	0.481	0.910	0.781
CFS$_{\text{dec}}(1\times1)$	improved	0.532	0.154	0.918	0.599	0.484	**0.014**	0.438	0.580	**0.004**	**0.001**	**0.007**	**0.001**
CFS$_{\text{dec}}(1\times1)$	EDM	0.378	0.143	**0.903**	0.545	0.294	0.023	0.231	0.158	0.146	0.065	0.140	0.053
CFS(1×2)	improved	0.538	0.218	0.917	0.557	0.385	0.028	0.371	0.487	0.013	0.002	0.017	0.002
CFS(1×2)	EDM	0.354	**0.133**	0.909	**0.501**	0.178	0.017	**0.140**	**0.112**	0.101	0.037	0.098	0.033

Appendix C. Unless stated otherwise, images are normalized to $[-1, 1]$, the test-time Monte Carlo count is one, and all reported main-paper methods use OOD-high scores.

J.1 CFS probing configuration

Table 28 reports the sparse probing configuration used for the primary one-level CFS runs. The same canonical probing setup is used for CFS(1×2), while decoder-only variants restrict the retained region to the decoder side. The value $\lambda = 5$ was fixed before inspecting OOD test performance. The ID-only hook proxy only chooses the model location within the admissible structural regions; it does not tune this canonical level.

ID-only hook-selection estimator. For each candidate hook h, let $\mathbf{z}_{r,i,h} \in \mathbb{R}^{D_h}$ denote the pooled descriptor of ID probe image i, using corruption repeat r. When several canonical levels are selected, descriptors are averaged across those levels before computing the proxy; the same noise draw is reused across levels within each repeat. For the main one-level configuration, no level averaging is needed.

Let N be the number of probe images and R the number of corruption repeats. Define

$$\bar{\mathbf{z}}_{i,h} = \frac{1}{R}\sum_{r=1}^{R} \mathbf{z}_{r,i,h}, \qquad \bar{\mathbf{z}}_{h} = \frac{1}{N}\sum_{i=1}^{N} \bar{\mathbf{z}}_{i,h}.$$

The between-image and within-image statistics are

$$\widehat{B}_h = \frac{1}{ND_h}\sum_{i=1}^{N} \|\bar{\mathbf{z}}_{i,h} - \bar{\mathbf{z}}_{h}\|_2^2,$$

$$\widehat{W}_h = \frac{1}{RND_h}\sum_{r=1}^{R}\sum_{i=1}^{N} \|\mathbf{z}_{r,i,h} - \bar{\mathbf{z}}_{i,h}\|_2^2.$$

Table 25: **External positioning against prior reported diffusion-based results on the CIFAR-scale benchmark.** Rows above the final block are taken from prior work. We separate: (i) prior diffusion-based methods reported in ID-specific settings, and (ii) single-checkpoint diffusion methods using one frozen checkpoint across multiple ID/OOD pairs.

Method	CIFAR-10 as ID				SVHN as ID				CelebA32 as ID				Avg.	Cost
	SVHN	CelebA	C100	DTD	C10	CelebA	C100	DTD	C10	SVHN	C100	DTD		
Diffusion-based, ID-specific / reported from prior work														
NLL	0.091	0.574	0.521	0.609	0.990	0.999	**0.992**	0.983	0.814	0.105	0.786	0.809	0.689	1000F
IC	0.921	0.516	0.519	0.553	0.080	0.028	0.100	0.174	0.485	0.972	0.510	0.559	0.451	1000F
MSMA	0.957	**1.000**	0.615	0.986	0.976	0.995	0.980	0.996	0.910	0.996	0.927	0.999	0.945	10F
DDPM-OOD	0.390	0.659	0.536	0.598	0.951	0.986	0.945	0.910	0.795	0.636	0.778	0.773	0.746	350F
LMD	**0.992**	0.557	0.604	0.667	0.919	0.890	0.881	0.914	0.989	**1.000**	0.979	0.972	0.865	10^4F
EigenScore	0.810	0.873	**0.880**	–	**0.992**	0.994	0.982	–	0.965	0.888	0.944	–	0.925	300F
Diffusion-based, single-checkpoint / outside MBE *(mostly reported from prior work)*														
SCOPED-CelebA	0.814	0.940	0.477	–	0.971	0.996	0.959	–	0.925	0.994	0.962	–	0.892	2F+2J
GEPC-CelebA	0.842	0.999	0.554	–	0.880	**1.000**	0.897	–	**1.000**	**1.000**	**1.000**	–	0.908	8F
DiffPath-6D-CelebA	0.910	0.897	0.590	0.923	0.939	0.979	0.953	0.981	0.998	**1.000**	0.998	0.999	0.931	10F
DiffPath-6D-ImageNet	0.856	0.502	0.580	0.841	0.943	0.964	0.954	0.969	0.807	0.981	0.843	0.964	0.850	10F
CFS(1×2)**-CelebA (ours)**	0.901	0.939	0.614	0.900	0.952	0.998	0.960	0.957	0.999	**1.000**	0.999	**1.000**	0.935	**1F**
CFS(1×2)**-ImageNet (ours)**	0.962	0.994	0.629	**0.996**	0.981	**1.000**	0.982	**1.000**	0.999	**1.000**	0.999	0.998	**0.962**	**1F**

Method	CIFAR-10 as ID				SVHN as ID				CelebA32 as ID				Avg. overlap	Cost
	SVHN	CelebA	C100	DTD	C10	CelebA	C100	DTD	C10	SVHN	C100	DTD		
DLSR (+MSE)	0.973	–	**0.875**	**1.000**	–	–	–	–	–	–	–	0.999	0.962	–
DLSR (+LR)	0.982	–	0.872	**1.000**	–	–	–	–	–	–	–	0.985	0.960	–
DLSR (+MFsim)	**0.989**	–	0.856	**1.000**	–	–	–	–	–	–	–	**1.000**	0.961	–
CFS(1×2)**-ImageNet (diag)**	0.962	0.994	0.629	0.996	0.981	**1.000**	0.982	**1.000**	**0.999**	**1.000**	**0.999**	0.998	0.962	**1F**
CFS(1×2)**-ImageNet (kNN)**	0.977	**1.000**	0.717	**1.000**	**0.998**	**1.000**	**0.998**	**1.000**	**0.999**	**1.000**	**0.999**	0.999	**0.974**	**1F**

Table 26: **Positioning against DLSR on its native published evaluation pairs.** DLSR lies outside the MBE protocol and uses an additional learned feature-reconstruction module, so we do not report a backbone-forward-equivalent logical cost for it. For our methods, Cost denotes the backbone-forward logical cost. Averages are computed over the non-missing overlapping entries of each method.

Candidates are ranked in descending order of

$$q_h = \frac{\widehat{B}_h}{\widehat{W}_h + \varepsilon}.$$

These estimators use population-normalized empirical variances, matching `unbiased=False` in the implementation.

The reported configuration uses $N = 32$, $R = 1$, and $\varepsilon = 10^{-6}$. Consequently, $\widehat{W}_h = 0$, and selection reduces to ranking between-image feature variance. This single-repeat configuration does not estimate corruption instability for hook selection. Encoder and decoder candidates are ranked separately within structural shortlists of at most four candidates per region. The paired variant retains one candidate from each region; decoder-only retains one decoder candidate.

For the main results, each selected slot is scored with the lightweight diagonal statistic in Eq. (5). Table 29 reports the corresponding ID-only head configuration. For the head-sensitivity ablation in Appendix E.9, we additionally evaluate stronger ID-only heads on the same sparse representation.

Pooling. The main-paper pooling rule is the one defined in Eq. (4): channel-wise spatial mean concatenated with channel-wise spatial standard deviation. Mean-only and std-only variants are treated as pooling ablations in Appendix E.8, not as the default CFS configuration.

J.2 Compute profiling protocol

We profile wall-clock inference cost using the same benchmark runner as the reported experiments in Table 30. For each method, profiling is performed after ID-only fitting and measures the full scoring path: canonical corruption, backbone evaluations, hook or output extraction, and score computation. We use 5 warm-up batches and 50 measured batches, with CUDA synchronization before and after

Table 27: **Checkpoint-controlled large-scale comparison on the official ImageNet-64 improved-diffusion backbone.** All methods use the same official improved-diffusion checkpoint (`imagenet64_uncond_100M_1500K.pt`). We report AUROC on NINCO, SSB-hard, and Textures, together with averaged AUROC, AUPR, and FPR95 over these three OOD sets. The logical test-time cost of the displayed CFS variants is $1F$; for MSMA and DIFFPATH it is $10F$.

ID	Method	NINCO ↑	SSB-hard ↑	Textures ↑	Avg. AUROC ↑	Avg. AUPR ↑	Avg. FPR95 ↓	Cost
ImageNet200	MSMA	0.6121	0.5244	0.7804	0.6390	0.9651	0.8917	$\underline{10F}$
ImageNet200	DIFFPATH	0.5874	0.5052	0.6325	0.5750	0.9525	0.9333	$\underline{10F}$
ImageNet200	**CFS**$_{\mathrm{dec}}(1\times1)$	**0.6549**	<u>0.5357</u>	**0.9093**	**0.7000**	**0.9739**	**0.7444**	**1F**
ImageNet200	**CFS**(1×2)	<u>0.6492</u>	**0.5361**	<u>0.8969</u>	<u>0.6941</u>	<u>0.9732</u>	<u>0.7578</u>	**1F**
ImageNet1K	MSMA	0.5972	0.5100	0.7790	0.6287	0.8513	0.8871	$\underline{10F}$
ImageNet1K	DIFFPATH	0.5913	0.5117	0.6359	0.5796	0.8130	0.9187	$\underline{10F}$
ImageNet1K	**CFS**$_{\mathrm{dec}}(1\times1)$	**0.6613**	**0.5337**	**0.9081**	**0.7010**	**0.8902**	**0.7401**	**1F**
ImageNet1K	**CFS**(1×2)	<u>0.6490</u>	<u>0.5247</u>	<u>0.8948</u>	<u>0.6895</u>	<u>0.8855</u>	<u>0.7591</u>	**1F**

Table 28: **Sparse CFS probing configuration.** This configuration specifies the canonical level, hook policy, pooling rule, Monte Carlo settings, and ID-only hook-selection proxy used for the main sparse probing runs.

Parameter	Value
Number of selected canonical levels K_c	1
Candidate canonical grid K_{grid}	1
Explicit canonical levels	$[5.0]$
Pooling rule	channel-wise mean + standard deviation
Dynamics features	disabled
Hook policy	`sparse_ed_id`
Region mode for CFS(1×2)	encoder + decoder
Retained encoder hooks	1
Retained decoder hooks	1
Fit-time Monte Carlo count $\mathrm{MC}_{\mathrm{fit}}$	1
Test-time Monte Carlo count $\mathrm{MC}_{\mathrm{test}}$	1
Maximum fit batches	64
Internal batch size	64
ID probe images for hook selection	32
Corruption repeats for hook selection	1
Maximum shortlisted candidates per region	4
Clamping	disabled
Numerical floor ε	10^{-6}

the measured region. Peak memory is measured with `torch.cuda.max_memory_allocated`. All timings are reported at the same batch size used in the benchmark.

We distinguish logical cost from measured runtime. Logical cost counts backbone evaluations per image, $\#F$; measured runtime additionally includes feature pooling, density-head evaluation, data movement, and method-specific aggregation overhead.

All profiles in Table 30 were obtained on a single NVIDIA GeForce RTX 4060 Laptop GPU with batch size 128, and PyTorch 2.10.0. For DDPM-OOD with EDM, we report the logical cost but omit wall-clock profiling because the $364F$ scoring path is prohibitively expensive; the improved-diffusion profile already illustrates this order-of-magnitude cost.

K Broader Impact and Responsible Use

This work studies post-hoc OOD detection for frozen diffusion backbones. A potential positive impact is improved reliability and auditability of vision systems: sparse internal probes may help identify inputs that fall outside an evaluation reference distribution, while keeping test-time cost low and making protocol confounders explicit.

The main negative risk is over-reliance. An OOD score is not a certificate of safety, correctness, or fairness, and false negatives may create unwarranted confidence in high-stakes settings. This

Table 29: **Default ID-only head configuration for the main CFS results.** The main paper uses the diagonal score to keep the detector lightweight and to avoid conflating representation quality with downstream head capacity.

Parameter	Value
Head type	diagonal Gaussian / diagonal Mahalanobis score
Slot statistics	ID-only mean and diagonal variance
Variance floor	10^{-6}
Feature pooling	channel-wise mean + standard deviation
Slot aggregation	uniform average over selected slots
KNN head	not used in main results
GMM head	not used in main results
Shrinkage covariance head	not used in main results

Table 30: **Representative compute profile on the CIFAR-scale benchmark.** Timings are measured after ID-only fitting using the full scoring path with 5 warm-up batches and 50 measured batches. GPU-hours estimate the fit plus evaluation cost for one ID-vs-OOD benchmark run under the reported split sizes.

Method	Backbone	Logical cost	Fit time (s)	ms/img	Peak mem. (GB)	GPU-hours
MSMA	improved	10F	160.8	19.52	0.66	0.41
DIFFPATH	improved	10F	163.1	19.70	0.66	0.42
DDPM-OOD	improved	364F	5851.2	724.83	0.66	15.29
GEPC	improved	8F	129.0	16.26	0.67	0.34
$\mathrm{CFS}_{\mathrm{dec}}(1\times1)$	improved	1F	26.5	3.23	0.78	0.07
$\mathrm{CFS}(1\times2)$	improved	1F	26.9	3.20	0.79	0.07
MSMA	EDM	10F	506.3	64.12	0.84	1.35
DIFFPATH	EDM	10F	473.9	57.89	0.84	1.22
DDPM-OOD	EDM	364F	–	–	–	–
GEPC	EDM	8F	381.7	48.23	0.85	1.02
$\mathrm{CFS}_{\mathrm{dec}}(1\times1)$	EDM	1F	53.3	6.42	0.85	0.14
$\mathrm{CFS}(1\times2)$	EDM	1F	49.5	5.88	0.87	0.12

is especially important for sensitive visual domains such as medical imaging, biometric analysis, surveillance, or safety-critical autonomy, where dataset shift, demographic imbalance, or acquisition bias may produce failures not captured by the evaluation protocol.

Our method is intended as a diagnostic and benchmarking tool rather than a standalone deployment safeguard. In practical use, it should be combined with domain-specific validation, calibrated thresholds, uncertainty auditing, and human oversight where decisions may affect people.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the scope of the work: a backbone-equated protocol for diffusion OOD, the CFS sparse internal-probing family, controlled CIFAR-scale results, and a local diagnostic theory. The Results section explicitly states the claim scope under MBE, and the limitations are discussed in Section 7.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 7 discusses limitations, including the need for internal backbone access, architecture/source dependence of internal features, difficult multimodal ID regimes, difficult ImageNet-scale near-OOD regimes, and the local rather than universal nature of the theory.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: The main theoretical statements are stated in Section 4 with their assumptions, and full proofs are provided in Appendix B, including conditional encoder-decoder complementarity, low-noise corruption stability, diagonal-score separation, canonical-level stability, and the hook-proxy interpretation.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The paper specifies the controlled MBE protocol, datasets, ID/OOD pair construction, backbone families, canonical corruption semantics, logical cost accounting, baseline taxonomy, and implementation details in Sections 3–5 and Appendices C–D. Full pairwise results and ablations are reported in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: We provide an anonymized supplementary code package containing the implementation of MBE, CFS, the harmonized baselines, configuration files, and scripts needed to reproduce the main tables and ablations. The experiments use public datasets and public diffusion checkpoints; the supplementary material specifies the required data preparation, checkpoint access, and run commands.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The experimental setting is described in Sections 3 and 5, including datasets, ID/OOD splits, backbones, corruption alignment, baselines, metrics, and logical cost. Additional details on canonicalization, adapters, baseline implementations, hook selection, and large-scale protocols are provided in Appendices C–D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: Appendix E.11 reports repeated-seed stability with mean and standard deviation for the primary sparse CFS variants, and Table 4 reports Spearman correlations for the diagnostic theory. The main tables are deterministic protocol summaries, while seed sensitivity is provided for the primary claims.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Table 30 reports representative compute profiles for the methods listed, including logical cost, inference time, peak GPU memory, and estimated GPU-hours. Wall-clock profiling is not reported for SCOPED or EDM DDPM-OOD. All experiments were conducted on a single NVIDIA GeForce RTX 4060 Laptop GPU with batch size 128 using PyTorch 2.10.0.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work uses public datasets and pretrained checkpoints for post-hoc OOD detection research and does not involve human-subject experiments, private data collection, or deployment decisions. The submission preserves the distinction between diagnostic evaluation and deployment-ready safety guarantees.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Appendix K discusses positive impacts, such as improved reliability and auditability of vision systems, and negative risks, such as false confidence in high-stakes settings or use in sensitive visual domains.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not introduce or release a new high-risk pretrained generative model, scraped dataset, or deployment system. It studies post-hoc OOD scoring on existing public checkpoints and datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The paper and supplementary code package explicitly credit the original repositories, datasets, pretrained checkpoints, and baseline implementations used in the experiments, including OpenAI Improved Diffusion, NVIDIA EDM, U-ViT, OpenOOD, DiffPath, and EigenScore. The supplementary material and repository documentation specify the original sources and access instructions for all third-party assets. The released package does not redistribute external checkpoints or datasets and is intended to comply with the corresponding licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes].

    Justification: The paper introduces a new code package for CFS and MBE-based diffusion OOD evaluation. The package documentation specifies the repository structure, supported methods, configuration files, dataset preparation, evaluation commands, metrics, and compute-cost accounting. It does not introduce new datasets or pretrained checkpoints. All external checkpoints used in the experiments are documented separately and are retrieved from their original sources or prior public repositories; the released package does not redistribute third-party checkpoints.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, user studies, annotation labor, or experiments with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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

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

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or human-subject research, so IRB or equivalent approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: The core methodology, experiments, and theoretical results do not involve LLMs as an important, original, or non-standard component. Any use of LLMs for writing, editing, or formatting does not affect the scientific method or results and therefore does not require declaration under the checklist wording.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.