Calibration, Not Channel: A Controlled Study of What Governs Foundation-Model Open-Set Recognition

Anonymous authors
Paper under double-blind review

Abstract

Open-set recognition (OSR) classifies inputs from known categories and rejects categories unseen in training. Foundation models supply two kinds of evidence for this decision. Semantic evidence comes from vision-language text prompts. Geometric evidence comes from image-only features. We organize foundation-model OSR into five evidence-combination strategies: semantic-only, geometric-only, disagreement-hybrid, naive fusion, and learned light fusion. All five share one decision interface. We evaluate them under matched, multi-seed protocols on CIFAR-10/100 and real ImageNet-200, alongside classical and recent post-hoc OOD detectors. We frame the question as two hypotheses. The Channel Hypothesis says the evidence channel, and how elaborately it is combined, drive rejection. The Calibration Hypothesis says the scoring and calibration of the evidence drive it. Our verdict is conditional. The backbone dominates absolute reliability. Once the backbone is fixed, the remaining variation follows calibration, not the raw channel or more elaborate fusion. Three results support this. First, a rejector learned directly from raw features is the worst and least-calibrated detector on every backbone, while calibrated *learned light fusion* is the best on every vision-language backbone. Second, calibrated disagreement is statistically equivalent to naive fusion (TOST, ± 1 pt, $p < 10^{-11}$); disagreement is thus a special case of fusion, not a distinct signal. Third, density scores help on DINOv2 but weaken on vision-language features. The ImageNet-200 study adds ReAct, ASH, ViM, GEN, and a matched NegLabel detector. None overturns the ordering, and learned light fusion still leads every vision-language cell. A legacy head-to-head only corroborates. The mechanism is calibration of the evidence on a fixed backbone, not learned capacity or any single channel.

Keywords: open set recognition; out-of-distribution detection; foundation models; vision-language models; CLIP; OpenOOD; calibration; reject option.

1 Introduction

A recognition system in the open world must eventually do something its training objective never asked of it. It must decide that an input belongs to *none* of the classes it was taught. A biodiversity classifier trained on a fixed catalogue of species will be shown an organism outside that catalogue. A medical triage model will meet a condition absent from its label set. A closed-set classifier cannot voice this "none of the above" verdict. It assumes every test input belongs to a training class. That assumption fails once new categories, ambiguous inputs, dataset shifts, or adversarial inputs appear. Open set recognition (OSR) removes the assumption. It asks a system to balance empirical risk on known classes against the risk of labeling unknown inputs as known (Scheirer et al., 2013; Geng et al., 2021). Deep methods such as OpenMax add explicit unknown rejection (Bendale & Boult, 2016). The neighboring out-of-distribution (OOD) literature develops post-hoc scores: maximum softmax probability, ODIN, Mahalanobis distance, energy, and outlier exposure (Hendrycks & Gimpel, 2017; Liang et al., 2018; Lee et al., 2018; Liu et al., 2020; Hendrycks et al., 2019).

Foundation models create a new opportunity and a new failure mode. Vision-language models such as CLIP, ALIGN, SigLIP, SigLIP2, and AltCLIP map images and text into aligned representation spaces (Radford et al., 2021; Jia et al., 2021; Zhai et al., 2023; Tschannen et al., 2025; Chen et al., 2023). Image-only

foundation models such as DINOv2 provide strong visual features without text supervision (Oquab et al., 2023). For open set classification, these models can be used in at least two distinct ways. A semantic branch compares images against known-class text prompts; a geometric branch estimates whether an image embedding lies near known-class feature regions. Fusion methods combine the two. The open question is not whether foundation models help, but *what* controls rejection once they are in hand.

We frame that question as two competing hypotheses. Naming the mechanism under test follows the field's tradition (Scheirer et al., 2013). The Channel Hypothesis holds that rejection is governed by which channel controls the decision (semantic vs. geometric) and by how elaborately the channels are combined. The Calibration Hypothesis holds instead that, once the backbone is fixed, rejection is governed by how the evidence is scored and calibrated. The two mechanisms it names are branch-scale normalization and density-aware scoring. It is largely indifferent to the raw channel or to the sophistication of the combiner. The two hypotheses make opposite predictions about where engineering effort should go. They can be separated by experiments that hold the backbone, features, and protocol fixed. This paper tests both. The verdict is clear. Under the dominant confound of the backbone, the strong Channel Hypothesis fails and the Calibration Hypothesis holds—hence *calibration, not channel*.

This manuscript makes two contributions. First, it organizes the design space into five approach families under a single decision interface. It positions them on a common axis, the Calibration–Capacity Spectrum (CCS), from low-capacity fixed-rule scores to a small learned combiner, with calibration of the evidence as the orthogonal lever (Figure 1). Second, it runs a controlled, multi-seed *mechanism* study rather than a leaderboard. The study adjudicates the Channel and Calibration Hypotheses by isolating what governs rejection when the channel, features, and protocol are held fixed. We stay strictly evidence-bounded. We report only experiments with recorded result files, and we flag missing method-dataset cells as coverage gaps rather than infer them.

What "channel" means, and why the thesis is not a tautology. We separate two axes the literature often conflates. The *backbone* is the pretrained model that produces features (e.g., CLIP vs. DINOv2). The *channel* is the *type* of evidence read from a backbone: semantic text-similarity versus geometric feature structure. Our controlled comparisons establish a clear ordering, *backbone* $\gg$ *calibration* $>$ *channel combination*. The backbone moves rejection AUROC the most (Table 3), so we hold it fixed and study the second and third terms. Leaderboard comparisons silently confound those terms with the first. The within-backbone claim is not the tautology "a better-ranking score ranks better." It is tied to two separable, manipulable mechanisms: branch-scale normalization (Section 6.1) and density-aware scoring. We change each while the channel, features, and backbone stay fixed, and each moves AUROC across its full range. The same disagreement evidence is below chance when summed on raw scales and competitive once standardized. The active ingredient is therefore calibration, not the evidence channel itself.

Contributions.

1. **A unifying taxonomy—the Calibration–Capacity Spectrum (CCS).** We connect classical OSR/OOD scoring to foundation-model evidence and express five strategies—semantic-only, geometric-only, disagreement-hybrid, naive fusion, learned light fusion—through a single open-set decision interface, positioning them on the CCS (Figure 1) rather than as unrelated methods.

2. **Central result: calibrated evidence—not the raw channel or learned capacity—governs rejection.** We test this on *identical* frozen features with *class-disjoint* auxiliary unknowns (released code, five backbones), through two falsifications. First, a reject head learned directly from *raw* features is the *worst*, least-calibrated detector on every backbone, whereas the calibrated learned light fusion—fusing already-calibrated post-hoc signals—is the *best* on every vision-language backbone (ECE 2.8–11%; Sections 5.4, 5.5). What helps is therefore calibrated evidence, not learned capacity. Second, the calibrated disagreement-hybrid score is algebraically nested in fusion and statistically *equivalent* to naive fusion (TOST on the paired effect, ± 1 pt, $p < 10^{-11}$). This quantified negative result shows that branch disagreement is a normalization-sensitive special case of fusion, not a distinct rejection signal.

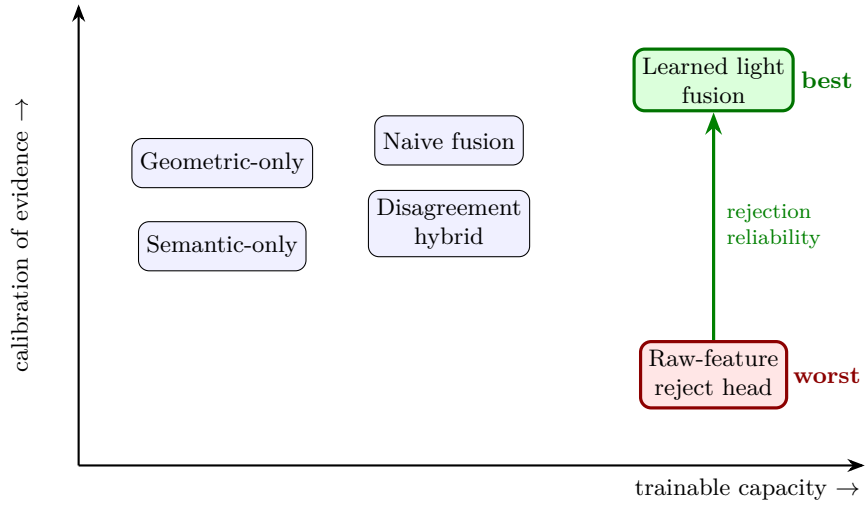

Figure 1: The Calibration–Capacity Spectrum (CCS). The five evidence-combination strategies (blue) placed by trainable capacity (horizontal) and calibration of the evidence (vertical). With the backbone held fixed, rejection reliability rises with *calibration*—the vertical lever—not with capacity: at the same high capacity, the calibrated learned light fusion is the best detector on every vision-language backbone, whereas a rejector learned directly from *raw* features is the worst. Channel choice (semantic vs. geometric) shifts a strategy little along either axis; the backbone—not shown—moves *absolute* AUROC the most.

3. **A fixed-backbone analysis that decouples channel from architecture.** The across-backbone spread (e.g., ALIGN ≈60 vs. AltCLIP ≈95 AUROC) dwarfs across-method differences (Table 3), so any cross-backbone comparison is confounded; with each backbone fixed, learned light fusion is the best of the five strategies on all ten vision-language cells and disagreement tracks naive fusion.

4. **A transparent head-to-head against published detectors, reported but *not* load-bearing.** Under our exact protocol, six classical post-hoc detectors (MSP, MaxLogit, Energy, ODIN, Mahalanobis, KNN) and MCM place a class-conditional Mahalanobis density on DINOv2 first (macro 95.46), and the score function alone spans about two AUROC points on one fixed head. Because this comparison relies on a legacy harness not part of this release and pairs DINOv2 with AltCLIP across methods, we treat it as corroboration only; the reproducible reimplementation qualifies the exact numbers (post-hoc scores closer, a simple energy score marginally best).

5. **Validation at ImageNet-200 scale.** A matched multi-seed study on real ImageNet-200 features (five backbones, three seeds, the same class-disjoint protocol) reproduces these qualitative conclusions at 200-class scale and folds in the newer ReAct, ASH, ViM, and GEN detectors plus a same-backbone NegLabel comparison on all four vision-language backbones (Section 5.7). The raw-feature head is again among the worst, simple logit scores lead the post-hoc family, learned light fusion is the best detector on every vision-language backbone—ahead of NegLabel in all four matched cells—calibrated disagreement again equals naive fusion, and density remains backbone-dependent. We release the feature-extraction and analysis code for the same-backbone and ImageNet-200 studies.

2 Background and Survey

2.1 Open Set Recognition and OOD Detection

In open set recognition the training data cover a known label set $\mathcal{Y}_K$, while the test data may include disjoint unknowns $\mathcal{Y}_U$. The predictor returns a known label or rejects via a thresholded knownness score $s(x)$ (formalized in Section 3). Open-set rejection recurs as a calibration problem across three eras. Classical

methods bounded open-space risk by reserving mass for the unknown (Scheirer et al., 2013). OpenMax recalibrated penultimate activations (Bendale & Boult, 2016) and spawned a large post-hoc OOD literature: MSP (Hendrycks & Gimpel, 2017), ODIN (Liang et al., 2018), Mahalanobis density (Lee et al., 2018), energy (Liu et al., 2020), outlier exposure (Hendrycks et al., 2019), and newer activation/feature-hybrid scores (ReAct (Sun et al., 2021), ASH (Djurisic et al., 2023), ViM (Wang et al., 2022), GEN (Liu et al., 2023)), unified by Yang et al. (2024). Foundation models change the substrate: the evidence is pre-trained and multi-channel. They do not change the question, because closed-set confidence is uncalibrated outside the training support. We re-run the classical family under one matched protocol and add ReAct/ASH/ViM/GEN in the ImageNet-200 study. Recent vision-language detectors that need their own training or prompts remain a coverage limitation (Section 7).

Because "calibration" is overloaded, we fix four senses and state which we measure: (a) ranking-level (does a score order ID above unknowns; AUROC/FPR95, what all our numbers measure); (b) branch-scale normalization (z-scoring each branch by its ID statistics so neither dominates a fusion); (c) density-aware scoring (proximity to a known-class distribution—Mahalanobis or k-NN—rather than raw confidence); and (d) probability calibration (do confidences match accuracy; ECE/reliability, measured in the reproducible same-backbone study). Our thesis concerns (a)–(c): on a fixed backbone, score design and branch-scale normalization control ranking-level separation more than the raw channel does. Tying it to the *separable* mechanisms (b) and (c)—tested by comparisons that hold the backbone and channel fixed—keeps it from collapsing into the tautology "a better-ranking score ranks better." Because OSR must preserve known-class accuracy while rejecting unknowns, we report closed-set accuracy, open-set accuracy, AUROC, and FPR95 together.

2.2 Benchmarks and Evaluation Risk

Benchmark design strongly affects apparent progress: OpenOOD standardizes splits and OOD categories (Yang et al., 2022), and NINCO showed that common ImageNet OOD sets contain in-distribution objects (Bitterwolf et al., 2023)—a contamination risk amplified by foundation models' broad pretraining, which blurs the boundary between known, semantically close unknown, and far unknown. We accordingly keep two non-interchangeable protocols separate: a compact class-split OSR protocol (known/unknown class subsets, thresholds fit on known data) and an OpenOOD v1.5 probe.

2.3 Foundation Models as Open Set Evidence

Vision-language models provide *semantic* comparators (image-text similarity to class prompts), powerful but sensitive to class names, prompts, and class order; image-only models provide *geometric* comparators (embeddings, light heads, prototypes), robust to text brittleness but less semantically specific. Prompt-learning and negative-prompt detectors—LoCoOp (Miyai et al., 2023), CLIPN (Wang et al., 2023), negative-label prompting (Jiang et al., 2024)—keep the semantic channel active alongside strong image backbones, and strong zero-shot classifiers do not automatically yield strong rejectors (Miller et al., 2024), reinforcing our separation of recognition accuracy from rejection quality. The useful lens is thus not which model is best but which evidence channel is trusted; Table 1 summarizes the five we study.

3 Five Approach Families

We present the five strategies as one escalating sequence along the horizontal (capacity) axis of the Calibration–Capacity Spectrum (Figure 1): single-channel scoring (semantic, geometric), then three ways to *connect* the channels (disagreement, naive fusion, learned light fusion). Each step adds trainable capacity while the decision interface stays fixed, so differences isolate fusion and calibration rather than model size. A recurring thread (Section 6.1) is the spectrum's vertical axis: every multi-channel score must first be placed on a common, calibrated scale. Table 1 summarizes the five.

Common decision interface. Let $\mathcal{C}_K = \{1, \ldots, K\}$ be the known classes and $\mathcal{C}_U$ the disjoint unknowns, unavailable at closed-set fitting. For each test image x a method returns a known-class prediction $\hat{y}(x) \in \mathcal{C}_K$

Table 1: Five foundation-model open set classification paradigms implemented in the repository.

Paradigm	Evidence used	Main strength	Main risk
Semantic-only	Image-text similarity to known-class prompts.	Strong zero-shot semantic grounding; minimal training.	Prompt and class-name/order sensitivity; weak if text branch is misaligned.
Geometric-only	Image embeddings, linear heads, margins, energy, or prototypes.	Works with image-only models; strong when embeddings cluster known classes.	Reject score depends on feature geometry and light-head fit.
Disagreement hybrid	Per-branch calibrated agreement/confidence between semantic and geometric branches.	Interpretable signal that is competitive once branch scores are normalized to a common scale.	Fails if branch scores are fused on incompatible scales; both branches can still agree on a wrong known label.
Naive fusion	Fixed weighted average of normalized semantic and geometric scores.	Simple, auditable, no auxiliary unknown training required.	Fixed weights cannot adapt to dataset-specific score reliability.
Learned light fusion	Lightweight reject head over branch scores, margins, entropy, JS divergence, and agreement.	Best available compact OSR performance; explicit calibration.	Requires auxiliary unknown/calibration data; can overfit the auxiliary unknown distribution.

and a scalar *knownness* score $s(x)$ (larger = more known); it decides $f(x) = \hat{y}(x)$ if $s(x) \geq \tau$, else unknown, with τ fit on held-out known-validation data rather than on test unknowns. For methods without auxiliary-unknown training, τ is a lower percentile of known-validation scores,

$$\tau = Q_p(\{s(x_i) : x_i \in \mathcal{D}_K^{val}\}), \qquad p = 5 \text{ in the compact scripts.} \tag{1}$$

This separates the two questions usually conflated: closed-set accuracy (from $\hat{y}$ on accepted knowns) and rejection quality (the ranking/thresholding of s).

Semantic-only. The sole evidence is image–text similarity in a vision-language space. Per class c, a text prototype t_c is the L2-normalized mean of prompt embeddings; with normalized image embedding $v(x)$, the semantic logit is $z_c^s(x) = \alpha\, v(x)^\top t_c$ (logit scale α), the prediction $\hat{y}_s = \arg\max_c z_c^s$, and the knownness score $s_s = \max_c p_c^s$, where p^s is the softmax (contrastive VLMs) or sigmoid (sigmoid VLMs) of z^s. Only light parameters (adapter or logit scale) are optionally trained; the encoder is frozen. This tests whether semantic alignment alone defines the known boundary; rejection is sensitive to prompt wording, class names, and semantic overlap—a visually unknown class linguistically close to a known prompt can score high.

Geometric-only. Text is dropped at rejection time: a linear head on frozen features $h(x)$ gives logits $z^g(x) = Wh(x)+b$, with class means μ_c estimated from training embeddings. We use four knownness scores—max-logit $\max_c z_c^g$, energy $\log \sum_c e^{z_c^g}$, margin $z_{(1)}^g - z_{(2)}^g$, and (negative) prototype distance $-\min_c d(h(x), \mu_c)$. This is the text-free control: it can beat semantic scoring when class names are weak or the encoder already clusters knowns, but geometric scores are not probability-calibrated and unknowns may lie near known clusters.

Disagreement hybrid. The first way to connect the channels rewards agreement and penalizes divergence. Because the semantic score is a posterior in $[0, 1]$ while the geometric max-logit is unbounded, each branch is first standardized by its in-distribution statistics $(\tilde{s}_s, \tilde{s}_g)$. The calibrated score is

$$s_d(x) = \underbrace{w_s\, \tilde{s}_s(x) + w_g\, \tilde{s}_g(x)}_{\text{calibrated weighted sum (= naive fusion)}} + w_a \mathbb{1}[\hat{y}_s(x) = \hat{y}_g(x)] - w_j\ \mathrm{JS}(p^s(x), p^g(x)), \tag{2}$$

with $(w_s, w_g, w_a, w_j) = (0.40, 0.40, 0.30, 0.60)$ and JS the Jensen–Shannon divergence. An earlier formulation summed *raw* branch quantities; the unbounded geometric magnitude (larger for confident knowns) then anti-correlated with unknownness and *inverted* the detector (Section 6.1). Standardization removes this. The falsifiable claim—that disagreement is a *distinct* signal—is rejected on compact CIFAR (Sections 5.1, 6.1): the first two terms are exactly a calibrated weighted sum, and the agreement and divergence terms add negligible ranking information.

Naive fusion. A simpler connection drops agreement modeling and averages the two standardized scores, $s_f(x) = \lambda\, \tilde{s}_s(x) + (1-\lambda)\tilde{s}_g(x)$ with $\lambda = 0.5$, predicting from the averaged posteriors $p^f = \lambda p^s + (1-\lambda)p^g$. Adding no trainable rejector, it is a strict capacity control for the learned variant: a gain over both branches indicates complementarity, and a fixed λ cannot adapt across datasets, backbones, or unknown types.

Learned light fusion. The data choose the combination. A low-dimensional fusion vector $\phi(x) = [s_s, s_g, m_s, m_g, H(p^s), H(p^g), s_{\text{proto}}, |s_s - s_g|, \text{JS}(p^s, p^g), \nVdash[\hat{y}_s = \hat{y}_g]]$ (H = posterior entropy) feeds a small reject head $q_\theta(x) = P(\text{ID} \mid x) = \sigma(r_\theta(\phi(x))/T)$, with temperature T and threshold fit on a held-out calibration split. It is the most expressive of the five at small capacity, recasting disagreement as one calibrated cue among several. Its risk is auxiliary-unknown overfitting, so we evaluate it with *class-disjoint* auxiliaries and a transfer split (Section 5.5).

4 Experimental Materials

4.1 Implementation and Reproducibility

No number in this manuscript is hand-entered or estimated; every reported value is either regenerated by released code or read from a recorded result file. We are explicit about three provenance tiers, because they differ in how completely they can be reproduced. (i) Fully reproducible from released code: the reproducible same-backbone studies (Sections 5.4, 5.5) and the matched ImageNet-200 study (Section 5.7). Their feature-extraction and analysis code is released and regenerates every cell—AUROC, FPR95, ECE, ablation, λ sweep, and figure—from the cached features. (ii) Traced to recorded result files: the compact-CIFAR five-strategy matrix (Section 5.1) and the external-baseline head-to-head (Section 5.3). The original harness emitted one record per method, dataset, seed, and backbone over $\{\text{CIFAR-10}, \text{CIFAR-100}\}$ and three seeds; a deterministic post-processing pipeline aggregates these into the mean $\pm$ std tables, the paired equivalence test, and the figures, with each cell carrying provenance to its source record. That original harness is *not* part of the present code snapshot, so these numbers are auditable against the recorded files but not re-executable end-to-end here. (iii) Not recoverable: a few original-harness hyperparameters—the Mahalanobis covariance shrinkage and the KNN neighbour count k—are fixed inside that harness and cannot be read off the snapshot (Section 4.3). The released code, configurations, and result files are provided in an anonymized repository.[1]

4.2 Datasets, Splits, and Backbones

The compact OSR study evaluates CIFAR-10 and CIFAR-100 (Krizhevsky & Hinton, 2009) under a class-disjoint known/unknown protocol: 50% of classes are held as known and the rest as unknown (5/5 on CIFAR-10, 50/50 on CIFAR-100). Each of three seeds $\{42, 1, 2\}$ resamples the known/unknown class partition, so reported mean $\pm$ std reflects variation over both optimization and the class split. Within the known classes, 10% of training data is held out as a validation/calibration split used to fit per-branch statistics and thresholds; test unknowns are never used for fitting. The evaluated foundation-model families are CLIP, ALIGN, AltCLIP, SigLIP, SigLIP2 (vision-language) and DINOv2 (image-only) (Radford et al., 2021; Jia et al., 2021; Chen et al., 2023; Zhai et al., 2023; Tschannen et al., 2025; Oquab et al., 2023). Because DINOv2 has no text encoder, it supports only the geometric and post-hoc detectors; the semantic, disagreement, naive, and learned strategies are defined only on vision-language backbones. The reported backbone per method-dataset is selected by mean known-validation accuracy, never by test performance. ImageNet-1K is evaluated only through the supplementary single-run OpenOOD v1.5 probe (Russakovsky et al., 2015; Yang

[1]Repository URL withheld for double-blind review; it will be included in the camera-ready version.

et al., 2022); ImageNet-200 is additionally evaluated under a matched multi-seed same-backbone protocol (Section 5.7), with that protocol's data construction described there.

4.3 Protocol and Hyperparameters

Light head: Geometric and fusion strategies train a linear head (optionally a residual adapter with hidden width 512) on frozen features with AdamW (learning rate 10^{-3}, weight decay 10^{-4}), batch size 64, for three epochs; the foundation encoder stays frozen. Thresholding: Methods without auxiliary-unknown training set the reject threshold at the 5th percentile of known-validation knownness scores ($p = 5$ in Eq. 1); learned light fusion instead selects its threshold on the calibration split by maximizing balanced known/unknown acceptance accuracy over a grid. Semantic branch and MCM: Class text prototypes use an ensemble of templates (e.g., "a photo of a {name}", "a close-up photo of a {name}") mean-pooled per class, with class names taken from the dataset metadata; MCM is zero-shot over the same prompts. Post-hoc baselines: The six post-hoc detectors share one linear head per backbone and differ only in the score function; Mahalanobis uses a class-conditional Gaussian with per-class means and a shared (tied) covariance estimated from known-class training features, following Lee et al. (2018), and KNN uses the negative distance to the k-th nearest known-training feature (Sun et al., 2022). The exact covariance shrinkage/regularization and the KNN neighbour count k are fixed inside the released baseline harness, which is not part of the present snapshot, so we report them as not recoverable here. Learned-light-fusion auxiliary unknowns: The reject head of learned light fusion is trained on known versus auxiliary-unknown features, with the auxiliary set split internally (20% held out for calibration). In the reproducible same-backbone CIFAR and ImageNet-200 studies, auxiliary and evaluation unknown classes are generated disjoint by construction and the emitted split artifacts are part of the release; these are the leakage-free results used for the main mechanism claim. In the original compact five-strategy matrix, the harness is auditable through recorded result files but not re-executable from this snapshot, so those legacy learned-light-fusion numbers are treated as corroborating rather than load-bearing for leakage-sensitive claims. OpenOOD coverage: The OpenOOD v1.5 probe is single-run and partial (Section 5.6).

4.4 Metrics

Closed-set accuracy measures known-class classification accuracy after accepting known samples. Open-set accuracy measures whether known inputs receive the correct known label and unknown inputs are rejected. AUROC is computed for unknown detection using the negative knownness score, and FPR95 measures the false positive rate for unknown samples when true positive rate on known samples is 95%. Lower FPR95 is better.

5 Results

5.1 Compact OSR Results

Table 2 reports the compact OSR results as mean $\pm$ std over three seeds, selecting the reported backbone per method-dataset by mean validation accuracy. Figure 2 visualizes the same AUROC summary with seed standard deviation as error bars.

The compact results support four observations. (1) Learned light fusion is strongest of the five: mean AUROC 96.35 ± 0.21 (CIFAR-10) and 91.11 ± 0.07 (CIFAR-100), macro 93.73, with the lowest FPR95 (18.73, 47.93) and smallest variance—+3.70/+5.49 AUROC over semantic-only and +12.2/+17.2 open-set accuracy—at a small closed-set cost (89.4/78.5 vs. ≈94/≈90 for single branches). (2) Geometric-only with DINOv2 is a clear second (94.15 ± 2.40 / 90.68 ± 0.40, macro 92.42), exceeding both fusions on CIFAR-100: rejection here is governed more by feature geometry than by prompts. (4) Semantic-only is the weakest single approach (AltCLIP 92.64 / 85.62, macro 89.13), with the highest FPR95.

(3) Calibrated disagreement is statistically equivalent to naive fusion. Standardized, disagreement reaches 94.53 ± 1.23 / 86.03 ± 0.86, tracking naive fusion (94.53 / 86.26). Across all 30 matched runs (5 backbones $\times$ 2 datasets $\times$ 3 seeds, spanning 58–95% AUROC), the paired difference $\Delta = \mathrm{AUROC}_{\mathrm{disag}} - \mathrm{AUROC}_{\mathrm{naive}}$

Table 2: Compact open-set recognition results, mean ± std over three seeds (42, 1, 2). For each method the backbone is chosen by mean validation accuracy (no test-set selection). All values are percentages; higher is better except FPR95 (lower is better). Best per dataset in **bold**.

Method	Backbone	Closed Acc.	Open Acc.	AUROC ↑	FPR95 ↓
CIFAR-10					
Learned light fusion	altclip	89.37	**91.64 ± 0.57**	**96.35 ± 0.21**	**18.73 ± 1.26**
Naive fusion	altclip	94.15	83.81 ± 3.79	94.53 ± 1.23	29.93 ± 6.65
Disagreement hybrid	altclip	94.14	83.82 ± 3.78	94.53 ± 1.23	29.91 ± 6.64
Geometric-only	dinov2	94.44	88.22 ± 3.55	94.15 ± 2.40	20.49 ± 6.25
Semantic-only	altclip	94.26	79.42 ± 1.57	92.64 ± 0.49	38.73 ± 2.60
CIFAR-100					
Learned light fusion	altclip	78.50	**82.35 ± 0.12**	**91.11 ± 0.07**	47.93 ± 0.61
Geometric-only	dinov2	90.22	74.28 ± 0.44	90.68 ± 0.40	**44.68 ± 1.46**
Naive fusion	altclip	89.30	67.07 ± 0.90	86.26 ± 0.72	54.19 ± 0.16
Disagreement hybrid	altclip	89.25	66.69 ± 0.99	86.03 ± 0.86	55.32 ± 0.69
Semantic-only	altclip	89.30	65.20 ± 0.91	85.62 ± 0.24	59.88 ± 1.49

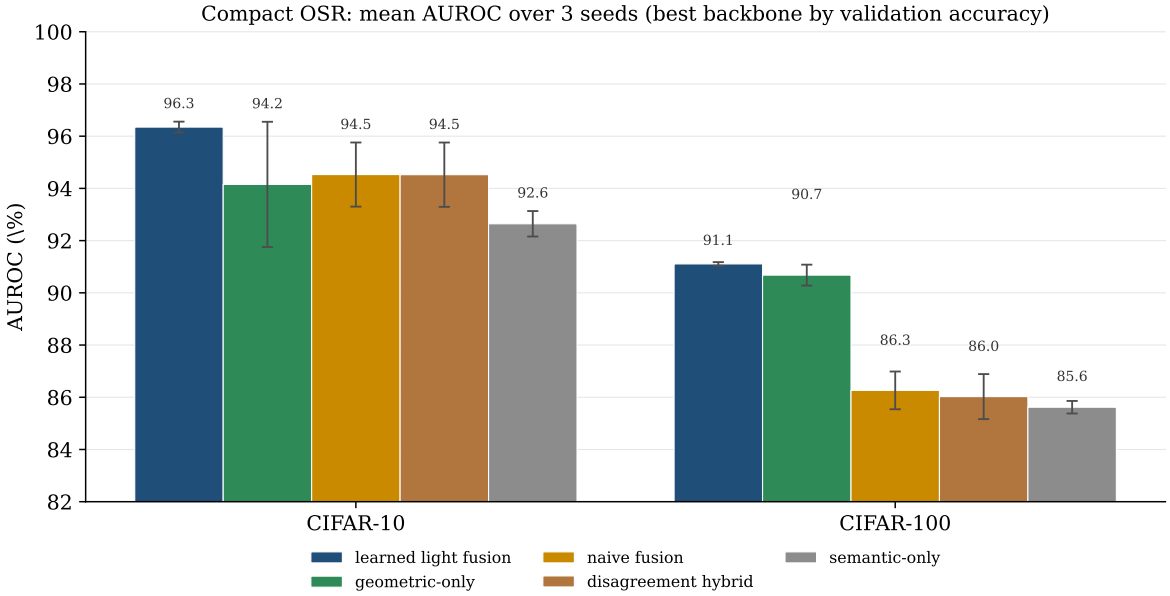

Figure 2: Compact OSR mean AUROC over three seeds by method and dataset; error bars show seed standard deviation. The reported backbone per method-dataset is selected by mean validation accuracy.

has mean -0.31 pt (95% CI $[-0.44, -0.19]$), and a TOST procedure rejects non-equivalence within a ± 1 pt margin ($p < 10^{-11}$); the small significant negative mean (paired t-test $p < 10^{-3}$) makes disagreement, if anything, marginally *worse*. This is not accidental: the first two terms of s_d (Eq. 2) *are* a calibrated weighted sum, so the agreement and Jensen–Shannon terms add no separating power—falsifying the hypothesis that disagreement is a *distinct* rejection signal on compact CIFAR.

5.2 Fixed-Backbone Analysis: Decoupling Channel from Architecture

The compact results above, like the head-to-head below, select each method's best backbone independently, so a cross-method comparison silently varies the backbone. To separate the effect of channel and combiner from the effect of backbone, Table 3 holds the backbone fixed and compares all five strategies on the same features.

Table 3: Fixed-backbone analysis: 3-seed mean AUROC of all five strategies with the *backbone held fixed*, decoupling evidence channel and combiner capacity from backbone selection. Best strategy per backbone row in **bold**. The text-free DINOv2 row admits only geometric scoring; the weak ALIGN row admits no geometric-only result. Within every vision-language backbone learned light fusion is the best of the five, and disagreement-hybrid tracks naive fusion; the across-row spread (e.g., ALIGN vs AltCLIP) far exceeds the across-method spread, so cross-backbone comparisons are confounded.

Dataset	Backbone	Semantic	Geometric	Disagree.	Naive	Learned
CIFAR-10	CLIP	83.34	85.11	86.11	86.11	**89.89**
	AltCLIP	92.64	90.94	94.53	94.53	**96.35**
	SigLIP	84.49	87.26	90.25	90.30	**94.16**
	SigLIP2	90.27	91.49	92.75	92.77	**96.17**
	ALIGN	59.11	–	58.51	58.59	**61.64**
	DINOv2	–	**94.15**	–	–	–
CIFAR-100	CLIP	77.42	78.75	76.90	77.40	**85.00**
	AltCLIP	85.62	87.85	86.03	86.26	**91.11**
	SigLIP	79.88	84.03	84.43	84.96	**87.27**
	SigLIP2	81.71	84.80	84.18	84.84	**90.31**
	ALIGN	60.17	–	59.03	60.06	**66.12**
	DINOv2	–	**90.68**	–	–	–

Three controlled patterns emerge. First, within every vision-language backbone learned light fusion is best of the five, on all ten backbone–dataset cells. Calibrated fusion does help once the backbone is fixed. The Mahalanobis density that looks superior in the head-to-head (Section 5.3) owes that to its DINOv2 backbone, not to being parameter-free. Second, calibrated disagreement tracks naive fusion in every cell, reproducing the equivalence of Section 5.1 backbone by backbone. Third, the across-backbone spread dwarfs the across-method spread. The best-to-worst method gap within a fixed backbone is typically 5–10 AUROC points. Changing the backbone moves a fixed method far more: learned light fusion spans 61.6 on ALIGN to 96.4 on AltCLIP for CIFAR-10. The consequence is decisive. Any comparison that varies the backbone is dominated by the backbone. These within-backbone rankings corroborate the thesis. Their load-bearing, re-executable form is the same-backbone studies (Sections 5.4, 5.5).

5.3 Head-to-Head Against Published Detectors

To check competitiveness with established detectors, we run six published post-hoc detectors (MSP (Hendrycks & Gimpel, 2017), MaxLogit (Vaze et al., 2022), Energy (Liu et al., 2020), ODIN (Liang et al., 2018), Mahalanobis (Lee et al., 2018), KNN (Sun et al., 2022)) and MCM (Ming et al., 2022) under the *identical* compact protocol (Table S8, Appendix). Two points matter. On a shared frozen DINOv2 head a class-conditional Mahalanobis density is strongest (macro 95.46, ahead of KNN 94.49 and Energy 93.66), and the score function alone—channel, features, and head held fixed—spans about two AUROC points on CIFAR-10 (from ≈96.4 for the max-softmax/max-logit/energy family to 98.25 for Mahalanobis), isolating calibration from representation. This comparison is *not load-bearing*: it relies on a legacy harness not in this release, and our reproducible reimplementation (Section 5.4) qualifies the exact numbers—there post-hoc scores cluster more tightly and a simple energy score, not the density, is marginally best—so we treat it as corroboration and place the load-bearing claim on the same-backbone studies below.

5.4 A Reproducible Same-Backbone Study

We ran a self-contained study to make the post-hoc comparison fully reproducible and leakage-free. It also tests whether *learning* a rejector directly from features helps with the backbone fixed. We use the *identical* frozen features of five backbones over three seeds. We compare five post-hoc scores (MSP, MaxLogit, Energy, Mahalanobis, KNN) against a *raw-feature learned reject head*—a logistic rejector trained directly on features, distinct from learned light fusion. The head is trained on known versus *class-disjoint* auxiliary unknowns, so it is measured for transfer (Table S3, Appendix; the scaled ImageNet-200 version is Table 4). Three findings

follow. First, the raw-feature head is the *worst* detector on all ten cells. It trails the best post-hoc score by up to 28 points (SigLIP/CIFAR-10 66.6 vs. 94.2; even DINOv2 93.2 vs. 99.3) and is the worst-calibrated (ECE ≈28% vs. ≈22% for max-softmax). Learning directly from raw features transfers poorly even with the backbone fixed. Second, the post-hoc scores are close, and a simple logit score always leads (Energy on every CIFAR-10 backbone), never a density. Third, density is backbone-dependent. Mahalanobis and KNN are competitive only on text-free DINOv2 (98–99 on CIFAR-10) and fall behind on the vision-language backbones (SigLIP/CIFAR-100 Mahalanobis 70.4 vs. MaxLogit 82.9). On a fixed backbone, what governs rejection is whether evidence is scored by a calibrated post-hoc rule rather than learned from raw features—not which channel or density.

5.5 Same-Backbone: Calibrated Fusion vs. Post-hoc Scoring

The study above shows a *raw-feature* reject head fails. It does not test the calibrated learned light fusion. We close that gap with a second reproducible study on the four vision-language backbones and identical frozen features. We compare the five post-hoc scores, the semantic branch, naive fusion, disagreement, and the calibrated *learned light fusion*. All use the same features, split, three seeds, and *class-disjoint* auxiliary unknowns, so learned light fusion is leakage-free here too (Table S4, Appendix). Learned light fusion is the best detector on all eight cells. It beats the best post-hoc score by 1.0–3.9 AUROC points (AltCLIP/CIFAR-10 97.17 vs. 96.18 Energy; CLIP/CIFAR-100 82.21 vs. 78.38 MSP), with the lowest FPR95 throughout. This holds on identical features with class-disjoint auxiliaries, so the advantage is neither a backbone artifact nor leakage. A rejector learned from *raw* features is worst; one from *calibrated post-hoc signals* is best. What helps is calibrated evidence, not capacity. Learned light fusion is also the best-calibrated (ECE 2.8–11% vs. 17–22% for the semantic branch and ≈28% for the raw-feature head). Densities are again weak on these vision-language features. A drop-one-feature ablation shows the prototype-distance feature matters most (−1.8 AUROC), while the agreement and Jensen–Shannon terms contribute almost nothing. This matches the disagreement≈naive equivalence and is stable across a λ sweep (Tables S6, S7, Appendix).

5.6 OpenOOD Far-OOD Grid

A third, single-run protocol—OpenOOD v1.5 *far-OOD*, where unknowns are external datasets rather than held-out ID classes (Table S9, Appendix)—adds two boundary observations. Disagreement again equals naive fusion to within ≈1 pt (CIFAR-10 97.22 vs. 97.51), a third protocol reproducing the equivalence of Section 5.1; but learned light fusion does *not* transfer its lead here (AltCLIP CIFAR-10 93.64 vs. naive 97.51), because its reject head was tuned on class-disjoint OSR unknowns—its advantage is real but *protocol-scoped*, a boundary we state rather than hide. The absolute channel ordering is also protocol-dependent (on ImageNet-1K a geometric postprocessor falls to 56.67 while semantic prompts reach 89.82). *Which* score wins shifts with protocol; the within-backbone mechanism stays stable, as the matched ImageNet-200 study confirms next.

5.7 Matched Multi-Seed ImageNet-200 Study

The supplementary probe above is single-run and partial. To test the calibration thesis beyond compact CIFAR, we ran the same reproducible same-backbone protocol on *real ImageNet* at 200-class scale. The ID training pool is real ImageNet-1K *training* images for the 200 OpenOOD ImageNet-200 classes (150 per class, streamed from the gated ImageNet-1K release). The held-out ID test set is the corresponding ImageNet-200 *validation* images, disjoint from training. Class names are recovered from the standard ImageNet-1K vocabulary, so the semantic branch is correctly grounded. We apply the identical class-disjoint OSR protocol as on CIFAR: 100 known / 100 unknown classes resampled over three seeds, with the unknown half split into class-disjoint auxiliary and test unknowns. On the *identical* frozen features of five backbones (CLIP, AltCLIP, SigLIP, SigLIP2, DINOv2) we evaluate nine post-hoc score functions, a raw-feature learned reject head, and the five evidence-combination strategies. The nine include the newer ReAct (Sun et al., 2021), ASH (Djurisic et al., 2023) (pruning variant), ViM (Wang et al., 2022), and GEN (Liu et al., 2023) detectors. On the vision-language backbones we add a same-backbone NegLabel detector (Jiang et al., 2024), matched under the identical split, seeds, and frozen features (Tables 4, 5). We use 150 training images per class, a

Table 4: Same-backbone controlled study: AUROC in percent (mean over 3 seeds; higher is better) of 9 post-hoc score functions and a raw-feature learned reject head computed on the *identical* frozen image features of each backbone. The raw-feature reject head—a logistic rejector trained directly on the features—is distinct from learned light fusion and is trained on known vs. *class-disjoint* auxiliary unknowns (no overlap with the test unknowns). Each row is one backbone×dataset; the best detector per row is in **bold**.

Dataset	Backbone	MSP	MaxLogit	Energy	GEN	ReAct	ASH-P	ViM	Mahalanobis (density)	KNN (density)	Raw-feature reject head
imagenet200	altclip	**95.65**	95.46	93.47	93.67	93.05	92.06	88.64	80.89	71.81	70.26
	clip	90.69	**90.82**	88.60	88.83	86.31	86.50	84.16	73.54	68.11	68.22
	dinov2	96.61	96.78	**96.93**	96.93	96.89	96.91	94.55	95.06	94.45	80.13
	siglip	**95.12**	95.11	94.02	94.15	92.37	93.49	90.38	76.74	73.45	72.33
	siglip2	**96.52**	96.51	95.42	95.53	94.03	95.07	92.20	80.70	78.67	71.17

Table 5: Matched same-backbone comparison (AUROC %, mean over 3 seeds; higher better) of post-hoc scores, the semantic branch, and the NegLabel vision-language detector against the calibrated *learned light fusion*, all on the *identical* frozen image features of each vision-language backbone, with the same split, seeds, and class-disjoint auxiliary unknowns. Best per row in **bold**.

Dataset	Backbone	MSP	MaxLogit	Energy	Maha.	KNN	Semantic	NegLabel	Naive	Disag.	Learned light fusion
imagenet200	clip	90.69	90.82	88.60	73.54	68.11	92.52	93.33	93.55	93.53	**94.26**
	altclip	95.65	95.46	93.47	80.89	71.81	95.33	95.04	96.43	96.52	**97.06**
	siglip	95.12	95.11	94.02	76.74	73.45	94.08	94.60	95.30	95.29	**96.08**
	siglip2	96.52	96.51	95.42	80.70	78.67	93.82	92.82	96.37	96.06	**96.94**

subset of the full ∼1300 that already saturates a linear probe on frozen features; extending to the full train split is straightforward but bandwidth-bound.

Every qualitative conclusion from compact CIFAR reappears at ImageNet-200 scale. (i) The raw-feature reject head is again among the worst detectors on every backbone (AUROC 68.2–80.1), trailing the best post-hoc score by 17–25 points (e.g. CLIP 68.2 vs. 90.8; DINOv2 80.1 vs. 96.9): learning a rejector directly from raw features still transfers poorly with the backbone fixed and class-disjoint auxiliaries. (ii) On the vision-language backbones simple logit scores lead—MSP/MaxLogit reach 90.8–96.5, ahead of the energy family (Energy, GEN, ReAct, ASH-P) and well ahead of the densities. (iii) The newer activation/energy detectors do not overturn the ordering: ReAct, ASH-P, GEN, and ViM cluster with Energy and never beat MSP/MaxLogit on a vision-language backbone, while on the text-free DINOv2 head Energy and GEN are marginally best (96.9) among a tight group (96.8–96.9) that also includes the Mahalanobis density (95.1). (iv) Density remains backbone-dependent: Mahalanobis is strong only on DINOv2 (95.1) and weak on the vision-language features (74–81), and KNN is weak throughout (68–79). The headline pattern—a rejector learned from raw features is worst, score design dominates, and no single channel or density wins everywhere—thus holds at ImageNet-200 scale on real ImageNet training data, not only on CIFAR.

The matched fusion comparison (Table 5) closes the same-backbone story on the vision-language backbones. The calibrated learned light fusion is the best detector on all four. It reaches AUROC 94.3 (CLIP), 97.1 (AltCLIP), 96.1 (SigLIP), and 96.9 (SigLIP2). It beats the best post-hoc score by 0.4–3.4 points, every density, the semantic branch, and the matched NegLabel detector. It is also the best calibrated in the emitted ECE artifacts (ECE 1.8–2.9% vs. 4.7–17.1% for the semantic branch). With real class names, the *semantic* branch is a strong baseline (92.5–95.3 AUROC). This corrects the degenerate semantic ImageNet-200 row from the supplementary probe (Section 5.6). The earlier failure was a class-name/grounding artifact, not a property of semantic evidence.

The strongest matched vision-language test is against NegLabel. Under the same split, seeds, and frozen features, NegLabel is itself a strong baseline (92.8–95.0 AUROC). It beats the class-conditional densities and rivals the simple logit scores. It nevertheless trails learned light fusion in every same-backbone cell (93.3 vs. 94.3 on CLIP, 95.0 vs. 97.1 on AltCLIP, 94.6 vs. 96.1 on SigLIP, 92.8 vs. 96.9 on SigLIP2). The central ordering therefore survives a matched comparison against a dedicated vision-language detector, not only against generic post-hoc scores. As on CIFAR, *calibrated disagreement is statistically indistinguishable from naive fusion* (e.g. AltCLIP 96.52 vs. 96.43; SigLIP2 96.06 vs. 96.37). A λ sweep is flat near its

mid-range optimum, with disagreement tracking it. A drop-one-feature ablation shows the semantic and entropy features carry the most weight, while the agreement and Jensen–Shannon terms are negligible. The ImageNet-200 study thus reproduces both pillars of the paper on real ImageNet training data, all five strategies, three seeds. Calibrated learned light fusion is the best combiner on vision-language backbones. And what governs rejection is calibration of the evidence, not channel, detector novelty, or learned capacity.

An ImageNet-200 *far-OOD* check on the same features (all 200 classes ID; unknowns are the mean over NINCO, SSB-hard, Textures, iNaturalist, and OpenImage-O; Table S10, Appendix) reaches the same qualitative conclusion: simple logit/energy scores and calibrated fusion are strongest (SigLIP2 MaxLogit 94.98 AUROC / 18.10 FPR95), naive fusion and calibrated disagreement again track each other (e.g. AltCLIP 93.89 vs. 93.89), densities transfer poorly (Mahalanobis 78.87–84.99), and a CLIP-only core NegLabel row reaches 93.16 without overturning the strongest SigLIP/SigLIP2 scores.

5.8 Positioning Against Published Results

Our load-bearing comparison is the matched same-backbone analysis (Sections 5.4–5.7); for external context only, Table S13 (Appendix) locates the study against published numbers under *non-comparable* protocols (e.g. max-logit 93.6 on CIFAR-10 (Vaze et al., 2022) vs. our learned light fusion 96.35 at a 5/5 split; our geometric OpenOOD CIFAR-10 96.58 near the strongest OpenOOD v1.5 far-OOD entry 96.6 (Zhang et al., 2024)), read as order-of-magnitude reference points. Importantly, NegLabel (Jiang et al., 2024) is not merely such a reference point: it is *matched* on all four vision-language backbones (92.8–95.0 AUROC, Table 5)—beating the densities, rivalling the simple logit scores, yet trailing learned light fusion on every cell. Only *LoCoOp* (Miyai et al., 2023) and *CLIPN* (Wang et al., 2023), which need prompt-training and released checkpoints outside our frozen-feature harness, remain non-comparable; we defer a matched comparison rather than approximate it. Our claims are scoped to frozen-feature post-hoc, prompt-ensemble, and light-fusion detectors, where the calibration mechanism is cleanly isolable; whether prompt-trained negative-label methods change the within-backbone ordering is a stated open question (Section 7).

6 Discussion

We can now return a verdict on the two hypotheses of Section 1. The strong *Channel Hypothesis* is rejected. Which channel controls the decision, and how elaborately channels are combined, explains little once the backbone is fixed. The *Calibration Hypothesis* is supported. How the evidence is scored and calibrated explains the within-backbone variation. Three comparisons converge on this: the score-function spread on a single DINOv2 head (Section 5.3), the within-backbone ranking of the five strategies (Section 5.2), and the 200-class replication (Section 5.7). The backbone remains the dominant confound. It moves AUROC far more than any score choice (Table 3). The hierarchy is therefore *backbone* $\gg$ *calibration* $>$ *channel combination*, and cross-backbone headline numbers are not clean comparisons. Our contribution isolates the second and third terms. On shared features, a raw-feature rejector is worst, calibrated learned light fusion is best on vision-language backbones, and a calibrated density is strongest on text-free DINOv2. Calibration of the evidence is the mechanism that unifies these cases.

6.1 Why Branch-Score Calibration Matters

Disagreement is attractive because it is interpretable. Its reliability, however, hinges entirely on scale. The semantic knownness score is a posterior in $[0, 1]$. The geometric maximum-logit score is unbounded and *grows* with confidence on known inputs. Summing raw score and margin differences therefore lets the geometric magnitude dominate. The aggregate "disagreement" then becomes larger for confident known samples than for unknowns, and the detector inverts—below chance on CIFAR. Standardizing each branch by its in-distribution statistics removes the artifact. It recovers a detector on par with naive fusion (CIFAR-10 AUROC 94.53, CIFAR-100 86.03). Two facts make the claim precise. **(i) Direction:** the identical evidence moves from below chance to competitive by a change of scale alone, so branch-scale normalization spans the full quality range. **(ii) Sufficiency:** once standardized, the disagreement score is statistically equivalent to a plain calibrated average (Table S2). Its leading terms *are* that average (Eq. 2), so the agreement

and divergence machinery is redundant on compact CIFAR. Disagreement is thus a normalization-sensitive special case of fusion, not a distinct signal. We have not seen this stated explicitly in the foundation-model OSR literature.

6.2 Practical Recommendations

For a deployable foundation-model open set classifier, the evidence suggests, in order: **(1)** start with a simple post-hoc score (energy, max-logit, or max-softmax) on your backbone—they led on most backbones, need no text branch or auxiliary unknowns, and are cheapest; **(2)** keep a semantic-only and a geometric-only baseline, since their gap reveals whether text prompts help, plus naive fusion as the minimal combination; **(3)** add a class-conditional density or learned light fusion only when it wins on held-out validation *and* a transfer check (density helped mainly on DINOv2; a rejector learned from raw features transferred poorly), and standardize each branch before any disagreement fusion, as uncalibrated fusion can invert the detector. These are scoped to compact CIFAR and ImageNet-200: the within-backbone ordering (raw-feature rejector worst, simple post-hoc leading, learned light fusion best on vision-language backbones) is stable across both, whereas the *absolute* geometric-versus-semantic ranking is protocol-dependent, so the first step is better read as "calibrate a simple score *and* test it against semantic scoring on the target benchmark" than as a fixed channel winner.

7 Threats to Validity, Limitations, and Future Work

Our study spans three seeds, matched five-method coverage, leakage-aware backbone selection, fixed-backbone analysis, an explicit equivalence test, and a matched ImageNet-200 replication. Five boundaries scope the conclusions and define the next steps. We frame them as threats to validity: detector and scale coverage bound *external* validity (i–ii); original-harness audits bound *internal* validity (iii); benchmark hygiene bounds *construct* validity (iv); and the text-free backbone corner bounds the *scope of generalization* (v). **(i) Detector coverage.** The compact head-to-head covers the classical post-hoc family (MSP, MaxLogit, Energy, ODIN, Mahalanobis, KNN) and MCM; the newer ReAct/ASH/ViM/GEN detectors are evaluated in the ImageNet-200 study (Section 5.7), where they cluster with the energy family and do not overturn the ordering. Coverage of recent vision-language detectors is stronger but still incomplete: NegLabel is matched across all four vision-language backbones in the same-backbone OSR comparison (Table 5) and appears as a CLIP-only core row in the far-OOD grid (Table S10); LoCoOp (Miyai et al., 2023) and CLIPN (Wang et al., 2023), which require prompt-training and author checkpoints outside our frozen-feature harness, remain outside the matched comparison, and NegLabel's grouping variant is omitted. **(ii) Scale.** The matched ImageNet-200 study uses real ImageNet-1K *training* images (150 per class, a subset of the full $\sim$1300 for tractable feature extraction) with the validation split as held-out ID test; the full train split and full ImageNet-1K (1000-class) coverage remain open. **(iii) Original-harness audits.** The compact five-strategy results come from a harness not in this snapshot, so those legacy numbers are auditable from recorded result files but not re-executable end-to-end here; the leakage-sensitive learned-light-fusion claim is instead supported by the released same-backbone studies, where auxiliary and evaluation unknown classes are disjoint by construction and ECE artifacts are emitted (ECE 2.8–11% on CIFAR and 1.8–2.9% on ImageNet-200). The legacy Mahalanobis shrinkage and KNN k also remain not recoverable from the snapshot. **(iv) Benchmark hygiene.** Semantic evaluation on large label spaces needs a class-name/order audit—the probe's degenerate ImageNet-200 semantic row was such an artifact, corrected here—and NINCO/SSB-style splits should be preferred for far-OOD evaluation. **(v) Backbone corner.** DINOv2 is text-free, so learned light fusion and the multimodal far-OOD grid cannot be compared to the strong DINOv2 density on that backbone; an image-only fusion variant would close this.

8 Conclusion

Our controlled study isolates what moves rejection reliability in foundation-model OSR. The lesson is scoped but sharp. *With the backbone held fixed, how the evidence is scored and calibrated matters more than which raw channel is used or how elaborately channels are combined.* Two reproducible same-backbone studies

make the mechanism precise on identical features with class-disjoint auxiliaries. A rejector learned from *raw* features is the worst, least-calibrated detector. Calibrated learned light fusion is best on every vision-language backbone (ECE 2.8–11% on CIFAR, 1.8–2.9% on ImageNet-200). A calibrated density is strongest on text-free DINOv2. Calibrated disagreement is statistically equivalent to naive fusion (TOST, ± 1 pt, $p < 10^{-11}$). The ImageNet-200 replication adds ReAct, ASH, ViM, GEN, and a matched NegLabel detector. None overturns the ordering, and learned light fusion still leads every vision-language cell. The scope is explicit. The backbone moves AUROC more than any score choice, so cross-backbone headline comparisons stay confounded. The *absolute* geometric-versus-semantic ranking is protocol-dependent, while the within-backbone calibration thesis is stable across scale. Remaining steps are ImageNet-1K train-scale coverage and direct comparison with prompt-trained LoCoOp and CLIPN. The practical takeaway is simple. Treat foundation-model OSR first as a calibration problem on a fixed backbone. Start with simple post-hoc scores, and add a density or learned fusion only when it clears that bar.

References

Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1563–1572, 2016. doi: 10.1109/CVPR.2016.173.

Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *International Conference on Machine Learning*, 2023.

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. Altclip: Altering the language encoder in clip for extended language capabilities. In *Findings of the Association for Computational Linguistics*, 2023.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *International Conference on Learning Representations*, 2023.

Chuanxing Geng, Sheng-Jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3614–3631, 2021. doi: 10.1109/TPAMI.2020.2981604.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning*, 2021.

Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided OOD detection with pretrained vision-language models. In *International Conference on Learning Representations*, 2024.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

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.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

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, pp. 21464–21475, 2020.

Xixi Liu, Yaroslava Lochman, and Christopher Zach. GEN: Pushing the limits of softmax-based out-of-distribution detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Dimity Miller, Niko Sünderhauf, Alex Kenna, and Keita Mason. Open-set recognition in the age of vision-language models. In *European Conference on Computer Vision (ECCV)*, 2024.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems*, 2022.

Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. LoCoOp: Few-shot out-of-distribution detection via prompt learning. In *Advances in Neural Information Processing Systems*, 2023.

Maxime Oquab, Timothee Darcet, Theo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013. doi: 10.1109/TPAMI.2012.256.

Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, 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, pp. 20827–20840, 2022.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Henaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? In *International Conference on Learning Representations*, 2022.

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

Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. CLIPN for zero-shot OOD detection: Teaching CLIP to say no. In *IEEE/CVF International Conference on Computer Vision*, 2023.

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*, volume 35, 2022.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

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 (DMLR)*, 2024.

A Supplementary Material

This appendix collects the tables referenced from the main text, grouped by the study they support. The tables are placed exactly where introduced. Every value comes from the same released code or recorded artifacts as the main-text tables.

A.1 Compact CIFAR five-strategy study

Table S1 is the per-method summary behind Section 5.1, and Table S2 reports the disagreement≈naive equivalence test (TOST).

Table S1: Macro-average compact OSR over CIFAR-10/100 (each method's per-dataset best backbone, three seeds). Best in **bold**.

Method	Datasets	Mean AUROC ↑	Mean FPR95 ↓
Learned light fusion	2	**93.73**	33.33
Geometric-only	2	92.42	**32.59**
Naive fusion	2	90.40	42.06
Disagreement hybrid	2	90.28	42.62
Semantic-only	2	89.13	49.31

Table S2: Paired equivalence of calibrated disagreement-hybrid and naive fusion. $\Delta = \mathrm{AUROC}_{\mathrm{disag}} - \mathrm{AUROC}_{\mathrm{naive}}$ (percentage points) on matched runs (same dataset, backbone, seed; mean over 3 seeds). Across backbones spanning 58–95% AUROC the aggregate paired effect is statistically equivalent within a ± 1.0 pt margin by TOST; this is a statement about the mean difference, not a pointwise guarantee—all but one seed-mean cell fall within ± 1.0 pt, the exception being the low-performing ALIGN/CIFAR-100 cell at -1.02 pt. Disagreement is, if anything, marginally worse.

Dataset	Backbone	Naive	Disag.	Δ
cifar10	clip	86.11	86.11	+0.00
cifar10	siglip	90.30	90.25	-0.05
cifar10	siglip2	92.77	92.75	-0.02
cifar10	altclip	94.53	94.53	-0.01
cifar10	align	58.59	58.51	-0.09
cifar100	clip	77.40	76.90	-0.51
cifar100	siglip	84.96	84.43	-0.53
cifar100	siglip2	84.84	84.18	-0.66
cifar100	altclip	86.26	86.03	-0.24
cifar100	align	60.06	59.03	-1.02

$n = 30$ paired runs; mean $\Delta = -0.313$ (95% CI [-0.44, -0.19]); $\max|\Delta| = 1.20$ pt.
TOST equivalence (± 1.0 pt) $p = 5.2e-12$; paired t-test $p = 0.00$; Wilcoxon $p = 0.00$.

A.2 Reproducible same-backbone study

Tables S3 and S4 are the CIFAR same-backbone post-hoc and fusion comparisons (Sections 5.4, 5.5). Table S5 reports per-detector ECE, Table S6 the learned-light-fusion feature ablation, and Table S7 the naive-fusion λ sweep.

Table S3: Same-backbone controlled study: AUROC in percent (mean over 3 seeds; higher is better) of five post-hoc score functions and a raw-feature learned reject head computed on the *identical* frozen image features of each backbone. The raw-feature reject head—a logistic rejector trained directly on the features—is distinct from learned light fusion and is trained on known vs. *class-disjoint* auxiliary unknowns (no overlap with the test unknowns). Each row is one backbone×dataset; the best detector per row is in **bold**.

Dataset	Backbone	MSP	MaxLogit	Energy	Mahalanobis (density)	KNN (density)	Raw-feature reject head
CIFAR-10	altclip	95.11	96.03	**96.18**	90.77	86.96	75.86
	clip	91.54	93.27	**93.56**	90.24	90.81	70.80
	dinov2	98.83	99.26	**99.33**	98.95	98.79	93.20
	siglip	92.09	93.88	**94.17**	90.97	89.87	66.56
	siglip2	94.02	95.78	**96.13**	93.46	93.53	70.43
CIFAR-100	altclip	**85.48**	84.85	82.93	74.32	66.47	67.45
	clip	**78.38**	77.89	75.80	71.02	68.43	64.77
	dinov2	91.07	91.70	**91.84**	88.91	88.68	74.67
	siglip	82.42	**82.94**	82.31	70.37	68.75	65.94
	siglip2	82.77	**83.44**	82.80	71.38	69.56	66.28

Table S4: Matched same-backbone comparison (AUROC %, mean over 3 seeds; higher better) of post-hoc scores and the semantic branch against the calibrated *learned light fusion*, all on the *identical* frozen image features of each vision-language backbone, with the same split, seeds, and class-disjoint auxiliary unknowns. Best per row in **bold**.

Dataset	Backbone	MSP	MaxLogit	Energy	Maha.	KNN	Semantic	Naive	Disag.	Learned light fusion
CIFAR-10	clip	91.54	93.27	93.56	90.24	90.81	90.05	93.52	93.66	**95.20**
	altclip	95.11	96.03	96.18	90.77	86.96	94.55	96.25	96.10	**97.17**
	siglip	92.09	93.88	94.17	90.97	89.87	91.91	93.74	93.59	**95.53**
	siglip2	94.02	95.78	96.13	93.46	93.53	93.00	95.19	95.26	**97.28**
CIFAR-100	clip	78.38	77.89	75.80	71.02	68.43	77.98	79.49	79.33	**82.21**
	altclip	85.48	84.85	82.93	74.32	66.47	83.99	85.83	85.80	**87.15**
	siglip	82.42	82.94	82.31	70.37	68.75	79.27	82.30	82.00	**85.03**
	siglip2	82.77	83.44	82.80	71.38	69.56	79.65	82.73	82.58	**85.31**

Table S5: Probability calibration (expected calibration error, %, mean over 3 seeds; lower is better) of the raw-feature learned reject head vs. max-softmax on the same frozen features. On CIFAR-10 and on DINOv2 the raw-feature learned reject head is markedly *worse* calibrated (ECE ≈28 vs. ≈22 on CIFAR-10); on the remaining CIFAR-100 cells the two are comparable. We report ECE only for the two detectors that emit a probability $P(\text{ID})$.

Dataset	Backbone	MSP ECE ↓	Learned reject ECE ↓
CIFAR-10	clip	22.78	27.84
	altclip	23.82	28.08
	siglip	23.42	28.45
	siglip2	22.6	28.4
	dinov2	20.11	28.2
CIFAR-100	clip	18.01	10.79
	altclip	14.01	11.23
	siglip	11.33	11.85
	siglip2	11.73	11.94
	dinov2	2.75	16.62

Table S6: Learned-light-fusion ablation: change in AUROC (mean over the four vision-language backbones and three seeds, CIFAR-100) when each fusion feature is removed. Negative means the feature helps. Confirms the rejector uses meaningful signals, not a single artifact.

Dropped feature	ΔAUROC
s sem	+0.00
s geo	-0.00
margin sem	-0.06
margin geo	-0.23
H sem	-0.15
H geo	+0.07
proto	-1.76
|s sem-s geo|	-0.00
JS	+0.04
agree	+0.01

Table S7: Naive-fusion λ sweep (AUROC %, mean±std over the four vision-language backbones and three seeds): λ=0 is geometric-only, λ=1 semantic-only. The calibrated disagreement score (last column) tracks naive fusion near its optimum, so the disagreement≈naive equivalence is not an artifact of fixing λ=0.5.

Dataset	λ=0.0	λ=0.25	λ=0.5	λ=0.75	λ=1.0	Disag.
CIFAR-10	94.74	95.16	94.68	93.92	92.38	94.65
CIFAR-100	82.28	82.88	82.59	81.64	80.22	82.43

A.3 Head-to-head against published detectors

Table S8 is the legacy compact head-to-head of Section 5.3, reported as corroboration only.

Table S8: Head-to-head comparison of our five foundation-model strategies against published post-hoc OOD detectors and the MCM vision-language detector, all run under the *same* compact OSR protocol (50/50 split, frozen foundation features, 3 epochs, mean ± std over seeds 42/1/2, backbone chosen by mean validation accuracy). AUROC and FPR95 in percent; best AUROC per column in **bold**.

Method	CIFAR-10 AUROC ↑	CIFAR-100 AUROC ↑	CIFAR-10 FPR95 ↓	CIFAR-100 FPR95 ↓	Macro AUROC
Mahalanobis (Lee et al., 2018)	**98.25 ± 0.29**	**92.67 ± 0.12**	7.69 ± 1.29	38.29 ± 0.47	**95.46**
KNN (Sun et al., 2022)	97.42 ± 0.36	91.56 ± 0.23	13.72 ± 2.21	43.49 ± 1.21	94.49
Learned light fusion (ours)	96.35 ± 0.21	91.11 ± 0.07	18.73 ± 1.26	47.93 ± 0.61	93.73
Energy (Liu et al., 2020)	96.42 ± 1.17	90.91 ± 0.80	13.41 ± 3.14	41.41 ± 0.45	93.66
ODIN (Liang et al., 2018)	96.68 ± 0.51	89.94 ± 0.76	15.72 ± 4.38	45.17 ± 1.68	93.31
MaxLogit (Vaze et al., 2022)	96.41 ± 1.10	89.96 ± 0.74	14.42 ± 3.30	45.05 ± 1.65	93.19
Geometric-only (ours)	94.15 ± 2.40	90.68 ± 0.40	20.49 ± 6.25	44.68 ± 1.46	92.42
MSP (Hendrycks & Gimpel, 2017)	96.42 ± 0.54	88.04 ± 0.59	18.28 ± 4.45	50.82 ± 0.31	92.23
Naive fusion (ours)	94.53 ± 1.23	86.26 ± 0.72	29.93 ± 6.65	54.19 ± 0.16	90.40
Disagreement hybrid (ours)	94.53 ± 1.23	86.03 ± 0.86	29.91 ± 6.64	55.32 ± 0.69	90.28
Semantic-only (ours)	92.64 ± 0.49	85.62 ± 0.24	38.73 ± 2.60	59.88 ± 1.49	89.13
MCM (Ming et al., 2022)	93.35 ± 0.00	84.03 ± 0.00	27.17 ± 0.02	69.88 ± 0.00	88.69

A.4 Far-OOD grids

Tables S9 and S10 give the CIFAR and ImageNet-200 far-OOD grids (Sections 5.6, 5.7).

Table S9: Single-run OpenOOD v1.5 *far-OOD* grid for all five strategies (AUROC / FPR95 in percent; higher AUROC, lower FPR95 better). This is a *different* protocol from the class-disjoint same-backbone study—unknowns are external far-OOD datasets, not held-out ID classes—and is single-run; methods use their selected backbone (AltCLIP for the vision-language strategies, DINOv2 for geometric-only). Best AUROC per column in **bold**.

Strategy	Backbone	CIFAR-10 AUROC / FPR95	CIFAR-100 AUROC / FPR95
Semantic-only	AltCLIP	97.12 / 16.10	–
Geometric-only	DINOv2	90.58 / 30.34	**89.43** / 29.57
Naive fusion	AltCLIP	**97.51** / 14.46	82.66 / 46.28
Disagreement hybrid	AltCLIP	97.22 / 15.40	81.59 / 48.91
Learned light fusion	AltCLIP	93.64 / 25.56	79.73 / 53.68

Table S10: ImageNet-200 OpenOOD-style far-OOD grid over five external OOD sets (NINCO, SSB-hard, Textures, iNaturalist, OpenImage-O). Cells report AUROC / FPR95 in percent; higher AUROC and lower FPR95 are better. The ID side uses real ImageNet-1K train features for the 200 ImageNet-200 classes (150 images per class, 30,000 total) and ImageNet-200 validation as ID test. Best AUROC per method column is in **bold**.

Backbone	MSP	MaxLogit	Energy	Maha.	KNN	Semantic	Naive	Disag.	NegLabel
CLIP	90.18 / 41.18	90.69 / 36.88	89.92 / 39.89	78.87 / 89.88	75.16 / 92.32	90.54 / 36.84	91.99 / 31.00	91.97 / 30.64	**93.16 / 32.07**
AltCLIP	93.21 / 24.75	93.32 / 23.97	92.59 / 27.87	84.96 / 78.87	80.20 / 87.17	**92.63 / 30.98**	93.89 / 22.84	93.89 / 22.85	–
SigLIP	94.27 / 22.57	94.73 / 20.45	94.59 / 20.31	83.75 / 79.09	80.52 / 81.61	92.00 / 35.29	**94.24 / 25.04**	**94.25 / 24.42**	–
SigLIP2	**94.66 / 19.63**	**94.98 / 18.10**	**94.82 / 19.18**	**84.99 / 75.32**	**82.29 / 78.00**	90.98 / 37.83	94.21 / 21.24	93.77 / 22.63	–

A.5 Single-run OpenOOD v1.5 probe

Tables S11 and S12 report the single-run OpenOOD probe.

Table S11: Best available OpenOOD v1.5 far-OOD summary per approach and ID dataset. Values are percentages; lower FPR95 is better.

Method	Dataset	Backbone	ID Acc.	AUROC	FPR95
disagreement_hybrid	imagenet	clip	49.24	69.57	68.88
disagreement_hybrid	imagenet200	align	77.83	75.26	58.06
geometric_only	cifar10	altclip	58.56	96.58	19.20
geometric_only	cifar100	siglip	50.24	89.50	37.75
geometric_only	imagenet	altclip	61.60	56.67	92.81
geometric_only	imagenet200	siglip	86.89	68.41	85.42
semantic_only	imagenet	altclip	72.41	89.82	39.70
semantic_only	imagenet200	clip	0.62	66.51	75.04

Table S12: Macro-average over each approach's available OpenOOD best results.

Method	Datasets	Mean AUROC	Mean FPR95
disagreement_hybrid	2	72.42	63.47
geometric_only	4	77.79	58.80
semantic_only	2	78.17	57.37

A.6 External reference points

Table S13 lists representative published numbers under non-comparable protocols, for order-of-magnitude context only.

Table S13: Published results from prior work, reported for *positioning only*. Protocols, backbones, and unknown-set definitions differ across rows and from our compact protocol (Table 2); the numbers are therefore **not directly comparable** to ours or to each other and are shown to locate foundation-model OSR/OOD detection in the literature. AUROC and FPR95 are percentages.

Method	Protocol (ID / unknown)	Backbone	AUROC ↑	FPR95 ↓
Standard class-split OSR (closest protocol to ours)				
Softmax (MSP) (Vaze et al., 2022)	CIFAR-10 OSR (6/4 split)	VGG32 (scratch)	67.7	–
MLS (Vaze et al., 2022)	CIFAR-10 OSR (6/4 split)	VGG32 (scratch)	93.6	–
OpenHybrid (Vaze et al., 2022)	CIFAR-10 OSR (6/4 split)	VGG32 (scratch)	95.0	–
MLS (Vaze et al., 2022)	TinyImageNet OSR	VGG32 (scratch)	83.0	–
Zero-shot vision-language far-OOD detection				
MCM (Ming et al., 2022)	ImageNet-1K / far-OOD	CLIP ViT-B/16	90.8	42.7
CLIPN (Wang et al., 2023)	ImageNet-1K / far-OOD	CLIP ViT-B/16	93.1	31.1
NegLabel (Jiang et al., 2024)	ImageNet-1K / far-OOD	CLIP ViT-B/16	94.2	25.4
OpenOOD v1.5 (far-OOD, averaged over MNIST/SVHN/Textures/Places365)				
MSP (Zhang et al., 2024)	CIFAR-10 / far-OOD	ResNet-18	90.7	–
KNN (Zhang et al., 2024)	CIFAR-10 / far-OOD	ResNet-18	93.0	–
RotPred (best) (Zhang et al., 2024)	CIFAR-10 / far-OOD	ResNet-18	96.6	–
MSP (Zhang et al., 2024)	CIFAR-100 / far-OOD	ResNet-18	77.8	–
KNN (Zhang et al., 2024)	CIFAR-100 / far-OOD	ResNet-18	82.4	–