DIVERGENCE-AWARE FLOW LIKELIHOODS FOR INDUSTRIAL ANOMALY LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work on flow matching for visual anomaly detection has adopted Gaussian endpoint scores and density proxies that avoid evaluating the full flow likelihood. These choices bypass the costly divergence integral, raising the question of when this term can be omitted and what information it contributes to detection and localization. We introduce FlowAD, which learns a flow-matching density over frozen foundation-model patch features. Its likelihood score separates into Gaussian endpoint energy and integrated divergence, the volume-change correction. We estimate divergence at inference using forward-only finite-difference Hutchinson probes. Controlled comparisons of endpoint-only, correction-only and full scores show that the correction improves localization even when image-level detection declines: on VisA, adding it to single-layer C-RADIO endpoint scores raises P-AUPRO by 4.92 percentage points while image AUROC falls by 0.66 points. Localization gains also appear on MVTec AD and with a second backbone, but the correction's additional benefit depends on how layers are combined. In a separate fixed-model-bank control on VisA, multi-layer fusion reduces its mean P-AUPRO gain from 4.87 to 1.15 percentage points and introduces category-level losses. We further prove that, for Gaussian data with positive-definite covariance and an independent standard Gaussian base under linear interpolation, the population-optimal flow-matching field is affine and its divergence is input-independent, so endpoint scoring preserves the full likelihood ranking. These findings establish a sufficient condition for omitting divergence integration and show that the correction's empirical value depends on the task and layer configuration.

1 INTRODUCTION

Unsupervised industrial anomaly detection learns normal appearance from anomaly-free training images and identifies deviations at test time (Bergmann et al., 2019; Zou et al., 2022; Wang et al., 2024). This one-class setting requires both image-level detection and pixel-level localization: an inspection system must determine whether a product is defective and delineate the affected regions. Both tasks depend on an anomaly score that separates defects from normal variation; localization requires this distinction within an image.

Pretrained visual representations provide a feature space in which to model normality. Patch-based methods characterize normal feature embeddings through nearest-neighbor retrieval, as in PatchCore (Roth et al., 2022), or distribution modeling, as in PaDiM (Defard et al., 2021). Normalizing-flow methods such as FastFlow (Yu et al., 2021) learn feature densities through invertible transformations and use negative log-likelihood for anomaly scoring. This formulation connects industrial anomaly detection to generative modeling, with the learned feature density defining the scoring rule.

Flow Matching (Lipman et al., 2023) offers a simulation-free way to train continuous normalizing flows. Evaluating their likelihood, however, still requires integrating both the velocity field and its divergence, the trace of the field's Jacobian (Chen et al., 2018; Grathwohl et al., 2019). The resulting negative log-likelihood combines a base-density term, which scores the transformed feature under a Gaussian prior, and a volume-change correction obtained by integrating the divergence, which accounts for local expansion and contraction along the flow. Recent anomaly detectors avoid this costly likelihood computation: MATCH omits the divergence term (Kruse et al., 2026), while WT-Flow uses a density proxy (Li et al., 2026). These approaches raise a question about the scoring rule

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

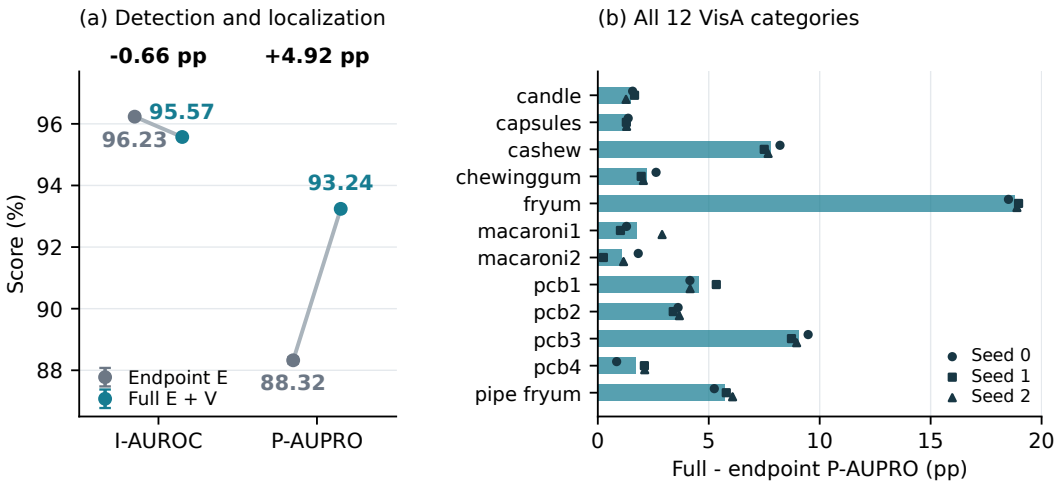

Figure 1: **Integrated divergence adds localization information on VisA.** Fixed C-RADIOv4-H single-layer features and trained fields, all 12 categories and three model seeds. **(a)** Gaussian endpoint E versus full score $E + V$: mean $\pm$ seed SD, with differences in percentage points. **(b)** Every category's P-AUPRO gain; bars average seeds and markers show individual seeds. P-AUPRO integrates to FPR 0.3 and is checked against the official implementation.

itself: *when does the volume-change correction carry anomaly information beyond the Gaussian endpoint, and what is lost by omitting it?*

We study this question through FlowAD, which learns a normal-data density over frozen foundation-model patch features. Its score separates into Gaussian endpoint energy E and an integrated divergence correction V. We estimate divergence at scoring time with forward-only finite-difference Hutchinson probes, enabling comparisons of endpoint-only, correction-only and full scores on the same learned flow. This decomposition isolates the contribution of volume change to anomaly ranking, with the feature representation and trained field held fixed.

In learned industrial feature densities, the correction contributes to localization even when image-level detection does not improve. On VisA, adding it to single-layer C-RADIO endpoint scores raises P-AUPRO from 88.32 to 93.24 and pixel AP from 42.53 to 46.43, while I-AUROC decreases from 96.23 to 95.57 (Figure 1). The P-AUPRO gain is positive in all 12 categories for each of three model seeds. Localization also improves on MVTec AD and in a DINOv3 control on VisA with one trained field per category. Five-layer fusion reduces the mean correction gain and introduces category-level losses within one fixed VisA model bank.

Contributions.

- We show that the divergence integral bypassed by recent flow-based anomaly detectors carries localization information beyond Gaussian endpoint scores. With single-layer C-RADIO features and learned fields fixed, retaining this correction raises P-AUPRO by 4.92 percentage points on VisA and 1.08 on MVTec AD, averaged over three model seeds.
- We prove that endpoint scoring preserves full likelihood rankings in a Gaussian regime. The divergence correction is constant across inputs, so integrating it adds no discriminative information. This holds for the population-optimal field under linear interpolation between an independent standard Gaussian base and Gaussian data with positive-definite covariance.
- We introduce FlowAD, a flow-matching method that detects and localizes anomalies by combining divergence-aware likelihoods of frozen foundation-model patch features across layers. Its five-layer score exceeds PatchCore by 2.46 P-AUPRO points on VisA under common evaluation. In a three-category MVTec AD audit on an A800, forward-only finite differences use about 41% of JVP native-patch scoring time and 22% less peak allocated memory, with full-score P-AUPRO differences below 0.009 points under shared probes.

2 Related Work

Unsupervised industrial anomaly detection. Industrial anomaly detectors differ in how they turn normal training data into a test-time score. UniAD and Dinomaly use feature reconstruction (You et al., 2022; Guo et al., 2025), while DRAEM learns a reconstruction embedding and discriminative boundary from simulated defects (Zavrtanik et al., 2021). Feature-distribution methods include PaDiM's spatially indexed Gaussian models (Defard et al., 2021) and PatchCore's nearest-neighbor distances to a subsampled memory bank (Roth et al., 2022). FlowAD also models normal features, but fits a continuous density to pooled patch tokens. This formulation allows us to examine which terms of a learned density contribute to localization, with the representation and fitted model held fixed.

Flow-based density estimation. Normalizing flows make feature density tractable through invertible transformations. FastFlow uses two-dimensional flows on pretrained features (Yu et al., 2021); CS-Flow and MSFlow incorporate cross-scale interactions (Rudolph et al., 2022; Zhou et al., 2025). Continuous normalizing flows instead evolve features through an ODE, with log-density changes determined by the velocity field's divergence (Chen et al., 2018). FFJORD makes this computation scalable with stochastic trace estimation (Grathwohl et al., 2019). Flow Matching trains the velocity field through simulation-free regression (Lipman et al., 2023); evaluating its likelihood still requires integration of the field and its divergence. FlowAD combines this training objective with finite-difference Hutchinson evaluation to study the divergence term in frozen patch-feature scoring.

Flow- and diffusion-based anomaly scoring. Recent methods use several alternatives to evaluating a full flow likelihood. WT-Flow learns time-reversed dynamics with a local worst-transport construction and uses a Gaussian density proxy, enabling one-step localization (Li et al., 2026). MATCH learns a multi-view Flow-Matching model and omits the divergence for efficiency; its Real-IAD ablation reports negligible gains from including this term (Kruse et al., 2026). InvAD uses DDIM inversion and scores the resulting latent variables against the known prior (Sakai et al., 2026). Flow Mismatching instead aggregates discrepancies between predicted velocities and the velocities of affine paths toward a test image (Chen et al., 2026). These choices motivate our paired comparison of endpoint energy E, volume correction V, and their sum on the same learned trajectories. We examine their effects separately on detection and localization using C-RADIO and DINOv3 features. In the multi-view setting, Multi-Flow (Kruse & Rosenhahn, 2025) and MATCH explicitly model view structure; our Real-IAD experiments evaluate an unconditional per-view patch density.

Likelihood-based out-of-distribution detection. The usefulness of likelihood for anomaly detection depends on the representation and the scoring rule. Generative models can assign high likelihood to out-of-distribution inputs (Nalisnick et al., 2019), and flow architectures can favor shared local image statistics over semantic differences (Kirichenko et al., 2020). Density rankings also depend on the chosen parameterization (Le Lan & Dinh, 2021). Our study fixes the feature representation and asks when the change-of-variables correction changes anomaly rankings. The Gaussian proposition gives sufficient conditions for identical rankings (Sec. 4); paired experiments measure the correction's contribution in trained fields on industrial features.

3 Method

Overview. FlowAD learns a transport from $\mathcal{N}(0, I)$ to normal frozen patch features. Reverse integration gives endpoint energy E and divergence correction V; we compare E, V and $E + V + c$ on the same numerical path. Each score yields a localization map whose maximum is the image score. Independently trained layers can be fused.

3.1 Flow-Matching Likelihoods on Patch Tokens

Patch tokens from frozen foundation models. We model the distribution of frozen features. A frozen ViT backbone maps an input image $x \in \mathbb{R}^{3\times H\times W}$ to a grid of patch tokens $F \in \mathbb{R}^{P\times D}$ ($P = HW/16^2$ for patch size 16); a single token $f \in \mathbb{R}^D$ is the unit FlowAD scores, and the normal tokens pooled over the training set constitute the empirical normal-data distribution p_{data}. We use

C-RADIOv4-H (NVIDIA Research, 2026) for the primary experiments and DINOv3-L/16 (Siméoni et al., 2025) for the representation control.

Flow-Matching likelihoods. We train a velocity field $v_\theta(z,t) : \mathbb{R}^D \times [0,1] \to \mathbb{R}^D$ with the simulation-free Flow-Matching objective (Lipman et al., 2023)

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0,x_1,t}\left[\frac{1}{D}\left\|v_\theta((1-t)x_0 + tx_1, t) - (x_1 - x_0)\right\|_2^2\right], \tag{1}$$

where $x_1 \sim p_{\text{data}}$ and $x_0 \sim \mathcal{N}(0, I)$ are independent. Training samples t independently and uniformly from $[10^{-4}, 1-10^{-4}]$; the population proposition considers the full linear path on $[0,1]$. At inference, the probability-flow ODE $\dot{z}_t = v_\theta(z_t, t)$ defines a continuous-time normalizing flow Φ with learned density p_θ, approximating p_{data}. Integrating it from t=1 down to t=0 with initial condition $z_1 = f$ recovers the base point $z_0 = \Phi^{-1}(f)$, and the change of variables gives the exact log-likelihood (Chen et al., 2018; Grathwohl et al., 2019).

FlowAD. FlowAD scores a test patch by the negative log-likelihood the trained density assigns it. The change-of-variables identity decomposes this score into a Gaussian-base endpoint term and an integral of the field's divergence, the trace of its Jacobian:

$$-\log p_\theta(f) \;=\; \underbrace{\tfrac{1}{2}\|z_0\|^2 + c}_{\text{endpoint term}} \;-\; \underbrace{\int_1^0 \operatorname{tr}\!\left(\tfrac{\partial v_\theta}{\partial z}\right) dt}_{\text{trace integral}}, \tag{2}$$

computed by integrating the reverse ODE toward the base while accumulating the trace integral. Here, $c = \frac{D}{2}\log 2\pi$ is the normalization constant of the base distribution. We denote the endpoint energy by $E = \frac{1}{2}\|z_0\|^2$ and the signed correction by $V = -\int_1^0 \operatorname{tr}(\partial v_\theta/\partial z)\, dt$; thus the full score is $E + V + c$, with c irrelevant to within-model ranking. We evaluate this identity numerically using ODE discretization, Hutchinson probing and finite differences.

Network architecture. The velocity field v_θ is a residual MLP of 4 blocks with a hidden multiplier of 2.0 (29M parameters at D=1280), with a sinusoidal time embedding modulating every block through per-channel FiLM scales and shifts (Perez et al., 2018). The MLP operates on individual frozen patch tokens. We train one velocity field per category and selected layer. Training uses normal patches only, with AdamW for 30 epochs and batches of 4096 tokens (Appendix A).

3.2 Hutchinson Likelihood with a Finite-Difference Trace

Eq. 2 requires the divergence, or Jacobian trace $\operatorname{tr}\big(\partial v_\theta/\partial z\big)$ at each of the K_{ODE} reverse Euler steps. Exact evaluation costs D backward passes per patch, intractable at D=1280 over millions of test patches; the standard remedy is the Hutchinson trace estimator with S=5 Gaussian probes ε_s (Grathwohl et al., 2019).

Forward-only trace evaluation. We implement Hutchinson probing with ordinary network evaluations, using a *forward-only finite-difference* estimate of the Jacobian-vector product:

$$J(z,t)\,\varepsilon \;\approx\; \frac{v_\theta(z + h\varepsilon,\, t) \;-\; v_\theta(z,\, t)}{h}. \tag{3}$$

Each probe requires one extra forward pass beyond the velocity evaluation, with no automatic differentiation. The resulting per-step trace estimate is

$$\widehat{\operatorname{tr}}(J) \;=\; \frac{1}{S}\sum_{s=1}^{S} \varepsilon_s^\top \left(\frac{v_\theta(z + h\varepsilon_s, t) - v_\theta(z,t)}{h}\right), \tag{4}$$

costing $K_{\text{ODE}}(1+S)$ forward passes per patch, with peak activation memory of the order of one chunked forward pass, plus states and accumulators. We use $K_{\text{ODE}} = 20$, draw fresh probes at every step, and accumulate $\widehat{V} = -\sum_k \Delta t\, \widehat{\operatorname{tr}}(J_k)$ with $\Delta t = -1/K_{\text{ODE}}$. The probes leave the Euler path unchanged.

Step size h. The implementation sets $h_i = \alpha \max(1, \|f_i\|_2)$ from the initial test token, with $\alpha = 10^{-3}$, and keeps it fixed along that token's trajectory. This scales the perturbation with the input norm without per-backbone retuning. Approximation accuracy depends on the field's local variation over this perturbation scale.

Bias and memory. The one-sided difference's expected bias is $\mathcal{O}(h^2)$, the first-order term cancelling under the Gaussian probes' odd moments (Appendix A). Scoring in chunks of B patches (Algorithm 1), peak working memory is $\mathcal{O}(B\,D + |v_\theta|)$ for a fixed-depth field of $|v_\theta|$ parameters and hidden width proportional to D. This bound concerns chunk-local integration; cached inputs and retained output maps can still grow with dataset size. Fixed-model sensitivity and automatic-differentiation JVP controls assess numerical accuracy alongside scoring costs in Sec. 5.4.

3.3 Multi-Layer Score Fusion

We extract tokens from a fixed set of deep blocks $\ell \in \mathcal{L}$ (configuration in Appendix B), rescale each layer by one train-split scalar $\sqrt{D}/\mathbb{E}\|x\|_2$ applied to train and test alike before fitting, so that no depth dominates by feature magnitude, and train an independent velocity field per layer, yielding score maps $s_\ell \in \mathbb{R}^{H\times W}$. This scalar preserves relative token radii within each layer and is used for the deep-5 model bank, including its L31-only ablation.

At inference we fuse these maps by standardized summation, $S_{\text{final}} = |\mathcal{L}|^{-1} \sum_{\ell\in\mathcal{L}} \text{zscore}(s_\ell)$, using one mean and sample standard deviation per layer and category, computed over all pixels of all upsampled, smoothed test maps. The standard deviation is floored at 10^{-8}. This fusion uses unlabeled test-set statistics and yields a standardized detector score. Single-layer LL scores estimate negative log-densities in nats and bypass this standardization. We report both PRO and pixel AP because fusion can improve region coverage while reducing pixel precision.

Alternative scoring method. For comparison, TrajBif scores the negative endpoint spread of eight reverse trajectories with independent noise at each step. Its sign was chosen in exploratory evaluations with anomaly labels and is fixed across reported configurations. We compare it with likelihood using the same trained fields; Appendix F gives the construction, parameters and diagnostic, and Figure A2 shows both scoring methods.

4 When Can Divergence Integration Be Omitted?

Endpoint scoring preserves the full likelihood ranking whenever the integrated divergence correction V is input-independent. Under the following Gaussian conditions, the population-optimal Flow-Matching field has exactly this property.

Proposition 1 (Gaussian zero-information)**.** *For independent $x_0 \sim \mathcal{N}(0, I)$ and $x_1 \sim \mathcal{N}(\mu, \Sigma)$ with $\Sigma \succ 0$, the population-optimal Flow-Matching field under $x_t = (1-t)x_0 + tx_1$ is affine. Its divergence is therefore spatially uniform, so the trace integral is identical for every input. The trace carries zero discriminative information and the endpoint, equivalent to half the squared Mahalanobis distance, preserves the full likelihood ranking. (Proof in Appendix C.)*

Proof sketch. Gaussian conditioning gives $v^*(z,t) = \mu + B_t(z - t\mu)$, where

$$C_t = (1-t)^2 I + t^2\Sigma, \qquad B_t = [t\Sigma - (1-t)I]C_t^{-1}.$$

Since $\partial_z v^* = B_t$ and $\text{tr}(B_t) = \frac{1}{2}\partial_t \log\det C_t$, integration yields $V = \frac{1}{2}\log\det\Sigma$ and $E = \frac{1}{2}(f-\mu)^\top\Sigma^{-1}(f-\mu)$. Thus V adds the same constant to every endpoint score, leaving rankings unchanged. Identically distributed perturbations also yield input-independent endpoint dispersion (Appendix C).

Implications for ranking. For two patches with $E(f) > E(g)$, adding the correction reverses their order if and only if $V(g) - V(f) > E(f) - E(g)$. Even small corrections can therefore reorder nearly tied patches. If $|V(f) - b| \le \delta$ uniformly over the scored patches for a constant b, correction differences are at most 2δ, so every endpoint gap greater than 2δ preserves its ordering. The 0.57% magnitude share reported by Kruse et al. (2026) measures a different property. A paired

Table 1: **Fixed-model trace decomposition.** All categories; three-seed mean ± SD of dataset means (percent). E: Gaussian endpoint; V: trace correction. P-AUPRO integrates to FPR 0.3, independently checked against the official implementation. Bold: best mean per dataset.

Dataset	Score	I-AUROC	P-AUROC	P-AP	P-AUPRO
VisA	E	$\mathbf{96.23} \pm 0.07$	98.73 ± 0.02	42.53 ± 0.13	88.32 ± 0.10
VisA	V	94.67 ± 0.09	99.20 ± 0.00	$\mathbf{46.57} \pm 0.06$	$\mathbf{93.31} \pm 0.03$
VisA	$E + V$	95.57 ± 0.03	$\mathbf{99.25} \pm 0.00$	46.43 ± 0.04	93.24 ± 0.03
MVTec AD	E	$\mathbf{98.64} \pm 0.04$	98.13 ± 0.01	64.76 ± 0.12	94.55 ± 0.05
MVTec AD	V	98.12 ± 0.04	98.37 ± 0.01	65.86 ± 0.20	95.52 ± 0.04
MVTec AD	$E + V$	98.48 ± 0.03	$\mathbf{98.44} \pm 0.01$	$\mathbf{66.12} \pm 0.09$	$\mathbf{95.63} \pm 0.02$

comparison of E, V and $E + V$ measures whether retaining the correction improves detection or localization rankings.

Interpreting learned fields. In learned fields, a trace gain can reflect feature geometry, departure from the population optimum, or numerical approximation. Separating these sources requires geometry interventions and approximation controls.

5 Experiments

Evaluation setup. We evaluate on MVTec AD (15 categories) (Bergmann et al., 2019), VisA (12) (Zou et al., 2022), and multi-view Real-IAD (30 categories, five views, scored per image) (Wang et al., 2024). We report image AUROC (I-AUROC), pixel AUROC, pixel AP, and region P-AUPRO capped at 30% FPR. Our frozen C-RADIOv4-H backbone at 448^2 yields a 28×28 grid of $D{=}1280$ tokens; DINOv3-L/16 (Siméoni et al., 2025) is used for the second-representation decomposition. FlowAD maps are upsampled and Gaussian-blurred, with the maximum pixel score used for detection. MVTec AD and VisA main-table FlowAD results are three-model-seed means; Real-IAD uses one checkpoint per category and configuration. The DINOv3 decomposition uses one newly trained model per VisA category. Per-configuration seed spreads are in Table A12.

5.1 Paired Component Comparisons

Integrated divergence adds localization information. With a trained single-layer field held fixed, we compare the Gaussian endpoint $E = \frac{1}{2}\|z_0\|^2$, the integrated divergence correction V, and their sum. Table 1 extends the VisA motivation to MVTec AD and includes the trace-only diagnostic. Full minus endpoint improves P-AUPRO by 4.92 points on VisA and 1.08 on MVTec AD, and pixel AP by 3.90 and 1.36 points, respectively. Across seed-averaged category means, P-AUPRO improves in 12/12 VisA and 14/15 MVTec AD categories (the exception is `transistor`). I-AUROC decreases by 0.66 and 0.16 points. The correction therefore contributes localization information beyond the endpoint, with a different effect on image detection. Appendix B details the paired protocol.

Full likelihood and correction-only scoring. On VisA, V slightly exceeds $E + V$ in P-AUPRO (93.31 versus 93.24) and pixel AP (46.57 versus 46.43), while $E + V$ has higher image AUROC (95.57 versus 94.67). On MVTec AD, the full score exceeds V in all four metrics. These results support retaining the correction for localization; the endpoint's additional value depends on the task, dataset, and layer configuration. We use $E + V$ as the primary score because, together with the known base constant, it estimates the learned negative log-density. The correction alone is an informative alternative localization score. The following controls examine whether these findings persist across representations and layer configurations.

Replication on DINOv3. The localization gain persists on DINOv3-L/16 features (Table 2). We freshly extract the normalized final-block tokens at 448^2 and train one field per category on the official VisA normal training split, using seed 0 and the final checkpoint after 30 epochs. All 12 categories are included, with no layer search or test-based checkpoint selection. Within each model,

Table 2: **DINOv3 trace decomposition on VisA.** All 12 categories, one newly trained model per category (seed 0); category-macro means in percent. The normalized final-block features, fitted field, trajectory and postprocessing are fixed within each category. Bold: best per metric.

Score	I-AUROC	P-AUROC	P-AP	P-AUPRO
E	91.43	98.01	43.48	85.73
V	92.55	99.23	**51.21**	**93.46**
$E + V$	**92.67**	**99.24**	50.92	93.45
Full minus E (pp)	$+1.24$	$+1.23$	$+7.44$	$+7.72$

E, V, and $E + V$ share the same reverse trajectory, probes and postprocessing; the estimator uses the same $K = 20$, $S = 5$, and relative difference scale 10^{-3}. Appendix B.1 gives the extraction, training and per-category results.

Adding V improves mean P-AUPRO by 7.72 points, with gains in all 12 categories. Pixel AP increases by 7.44 points and image AUROC by 1.24 points; pixel AP still falls in one category and image AUROC in four. Correction-only scoring reaches 93.46 P-AUPRO and 51.21 pixel AP, versus 93.45 and 50.92 for the full score. As in the single-layer C-RADIO control, the endpoint adds little to these localization metrics beyond V alone.

Region coverage at matched false-positive rates. At a fixed 1% pixel false-positive rate, adding V increases mean defect-region coverage from 47.54% to 62.81%; at 5%, coverage rises from 76.51% to 89.06%. All 12 categories improve at both operating points. Correction-only coverage is 63.11% and 89.02%, respectively. These post-hoc diagnostics reuse the same single-seed scores and match false-positive rates using test masks; Appendix B.2 gives the thresholds and complete results.

5.2 Layer Fusion and the Marginal Correction Gain

Table 3: **Component ablation within the five-layer model bank.** VisA, all 12 categories and 2162 test images; category-macro means in percent. One existing model bank (60 checkpoints); 1L is its L31 model. All three five-layer heads share fusion scales computed from the full score. Full denotes the endpoint plus correction and base constant before centering. Bold: best within each layer setting.

Layers	Score	I-AUROC	P-AUROC	P-AP	P-AUPRO
1L	E	**96.27**	98.76	42.66	88.37
1L	V	94.49	99.20	46.31	**93.36**
1L	Full	95.61	**99.25**	**46.33**	93.24
1L	Full$-E$ (pp)	-0.66	+0.49	+3.67	+4.87
5L	E	**97.41**	**99.24**	43.48	94.04
5L	V	94.86	98.98	41.28	94.74
5L	Full	96.72	99.21	**44.48**	**95.19**
5L	Full$-E$ (pp)	-0.69	-0.03	+1.00	+1.15

We test how fusion changes the correction's incremental value using one C-RADIO five-layer model bank; its L31 model retains train-radius normalization and differs from the independently trained single-layer models in Sec. 5.1. The L31 scores are unstandardized. The five-layer E, V and full scores share the full score's fusion scales, preserving the component comparison after fusion: each component is centered by its own mean and divided by the full map's standard deviation. Thus the fused endpoint and correction sum to the fused full score up to floating-point error; separate component scales would change their relative layer weights.

Across all 12 VisA categories, fusion raises mean endpoint P-AUPRO by 5.67 points and full-score P-AUPRO by 1.95 points relative to its L31 model. The full-minus-endpoint gain therefore falls from 4.87 to 1.15 points. At five layers, adding V also increases pixel AP by 1.00 points but decreases image AUROC by 0.69 points; P-AUPRO improves in 7/12 categories. Appendix B.3 gives the normalization, fixed-weight comparison, and complete category results for these 60 checkpoints. The endpoint also becomes more useful alongside the correction: full minus correction-only P-

Method	MVTec AD				VisA				Real-IAD			
	I-AUC	P-AUC	AP	PRO	I-AUC	P-AUC	AP	PRO	I-AUC	P-AUC	AP	PRO
(a) Recent methods: original reports, native protocols												
MSAttnFlow (2025)	99.7	98.9	—	—	98.1	99.1	—	—	—	—	—	—
Multi-Flow (2025)[MV]	—	—	—	—	—	—	—	—	90.27	96.47	—	87.91
WT-Flow (2026)	99.05	97.95	—	—	96.52	98.90	—	—	—	—	—	—
InvAD (2026)[MC]	99.0	97.5	46.5	92.7	96.9	99.1	39.2	92.9	—	—	—	—
MATCH (2026)[VC]	97.46	97.21	—	93.88	—	—	—	—	91.17	99.24	32.68	94.76
(b) Independent reproduction: native backbone and evaluation												
U-Flow (2024)	97.48	97.96	53.37	84.56	95.17	98.96	40.50	83.46	—	—	—	—
(c) Our C-RADIO evaluation: common backbone, input size and metrics												
PatchCore (multi)	**99.20**	98.29	65.02	**96.48**	**97.01**	98.84	42.87	92.73	**92.02**	**99.27**	**41.67**	96.16
FlowAD-LL (1L)	98.48	98.44	**66.12**	95.63	95.57	**99.25**	**46.43**	93.24	88.91	98.96	39.33	94.73
FlowAD-LL (5L)	98.30	98.46	63.26	96.21	96.72	99.21	44.43	95.19	89.62	99.14	40.29	**96.18**
FlowAD-TB (5L)	98.09	**98.49**	64.00	96.16	95.75	99.22	43.98	**95.23**	88.26	99.11	40.80	96.02

Table 4: **Recent density/flow methods and our C-RADIO evaluation.** Category means (%); higher is better. I-AUC/P-AUC: image/pixel AUROC; AP: pixel AP; PRO: pixel AUPRO at 30% FPR. (a) Original reports, native protocols; (b) U-Flow reproduction, original protocol; (c) frozen C-RADIOv4-H, 448^2 inputs, common metric code, and PRO verified against the official implementation on all three datasets. **Bold**: best within (c). Layer selection, fusion and image aggregation vary across rows. PatchCore: our full-bank cosine 1-NN adaptation (Appendix D.2). LL: likelihood; TB: TRAJBIF; 1L: single layer; 5L: fixed deep-5, test-set score standardization. MC: multi-class; MV: cross-view input; VC: Real-IAD view conditioning. Dash: unverified or unavailable. Protocols and sources: Appendix D.1.

AUPRO changes from -0.12 points at L31 to $+0.45$ after fusion, with pixel-AP differences of $+0.02$ and $+3.20$ points.

5.3 BENCHMARK COMPARISONS

Comparisons on frozen features. Block (c) of Table 4 uses a common frozen C-RADIO backbone, input size, and metric implementation; layer selection, fusion and image aggregation vary across configurations. PatchCore denotes our cosine 1-NN adaptation with all training patches and three-layer averaging (Appendix D.2). Appendix D documents baseline configurations, provenance and discrepancies from self-reported numbers.

Recent-method context. Block (a) of Table 4 reports recent methods under their native protocols, including multi-class training and view-aware configurations; block (b) retains an independently evaluated U-Flow reproduction with its native backbone and original evaluation protocol. Appendix D.1 identifies the source tables and settings. Our separate DINOv3-L reproduction of Dinomaly (Guo et al., 2025) appears in the extended appendix comparisons (Table A9).

Detection and localization. FlowAD's clearest localization gain over PatchCore is on VisA (Table 4(c)). Deep-5 likelihood reaches 95.19 P-AUPRO, exceeding PatchCore's 92.73 by 2.46 points; deep-5 TRAJBIF reaches 95.23, a 2.50-point gain. Single-layer likelihood has the highest pixel AP in this block, 46.43 versus 42.87 for PatchCore. On MVTec AD, PatchCore leads P-AUPRO at 96.48, versus 96.21 for deep-5 likelihood and 96.16 for deep-5 TRAJBIF. Single-layer likelihood instead improves pixel AP from 65.02 to 66.12. PatchCore retains the highest image AUROC on both datasets.

On all 30 Real-IAD categories, deep-5 likelihood and PatchCore are nearly tied in P-AUPRO (96.18 versus 96.16), while deep-5 TRAJBIF reaches 96.02. PatchCore leads image AUROC (92.02 versus

89.62 for deep-5 likelihood) and pixel AP (41.67 versus 40.29). Evaluation preserves original image labels and pixel masks, including positive images with empty masks.

Layer fusion trades pixel precision for region coverage. The component ablation in Sec. 5.2 isolates the correction's marginal gain. Table A13 instead compares LL and TRAJBIF at L31 and five layers within each fixed bank across three model seeds. It retains the bank normalization and unstandardized L31 scores described there; each method uses its own test-score scales for five-layer fusion (Sec. 3.3). For LL, fusion increases P-AUPRO by 0.65 points on MVTec AD and 1.95 on VisA, while decreasing pixel AP by 2.54 and 1.86 points. Deep-5 LL and TRAJBIF differ by less than 0.06 P-AUPRO points on both datasets.

Auxiliary density comparison. On identical single-layer scoring features, the fitted FlowAD models exceed Gaussian P-AUPRO by 1.01 points on MVTec AD and 1.07 on VisA. Table A11 and Appendix D.3 report this separate single-checkpoint comparison. Statistical estimators are refitted; the reused neural checkpoints have incompletely established training inputs and seeds, and fitting and selection budgets differ.

5.4 NUMERICAL ACCURACY AND SCORING COST

Numerical accuracy. A fixed-model audit of three MVTec AD categories varies the step count, probe count and difference scale: the largest P-AUPRO change is 0.052 points under the official-reference-checked metric (Appendix A.1). A paired automatic-differentiation JVP check on these three categories preserves the correction's image-score rankings ($\rho \geq 0.99998$); the largest full-score P-AUPRO difference from finite differences is 0.0085 points (Appendix A.2). Both branches retain Euler discretization and Hutchinson probe error.

Cost of retaining the correction. An independent three-category comparison on a shared A800 measures endpoint-only scoring at 57.3–57.4 ms per image, FD full scoring at 345.9–348.2 ms and forward-JVP full scoring at 840.8–844.4 ms, using 32 fixed images per category. FD takes about 41% of the JVP time in this native-patch scope, with peak allocated memory of 487.9 versus 627.9 MiB. Appendix A.3 reports the matched numerical settings, three cyclic repetitions and the separate transfer-and-postprocessing scope.

Model state and fused scoring. FlowAD's retained model state is fixed by the architecture and layer count: approximately 120 MB for one layer and 0.6 GB for deep-5. Training the single-layer field takes a median 44 s per MVTec AD category on one L20, excluding feature extraction. In the paired eight-image cached-feature timing probe, LL requires 279 ms per image at one layer and 1417 ms with five layers; TRAJBIF requires 392 and 1969 ms, respectively. These shared-L20 measurements include scoring, transfers, smoothing and fusion, but exclude feature extraction, feature normalization and model loading. Appendix E gives the timing protocol.

6 CONCLUSION

The integrated divergence contributes to anomaly localization beyond what Gaussian endpoint scoring captures. In paired single-layer C-RADIO evaluations, retaining it improves P-AUPRO by 4.92 points on VisA and 1.08 on MVTec AD across three model seeds, with small decreases in image AUROC. Strong correction-only localization on VisA and the smaller gains under five-layer fusion, including category-level losses in the fixed-bank control, show that the two likelihood components contribute differently across scoring tasks and layer configurations.

Our theory identifies a case in which divergence integration can be omitted: for Gaussian data with positive-definite covariance and an independent standard Gaussian base under linear interpolation, the population-optimal field has input-independent divergence, so endpoint and full-likelihood rankings coincide. FlowAD estimates divergence in learned fields with forward-only finite-difference Hutchinson probes, enabling the paired comparisons above. The observed localization gains and their reduction under fusion motivate faster correction estimation and fusion rules that preserve its contribution to localization.

AI USE STATEMENT

Generative AI assisted literature review, experiment planning, coding and code review, result aggregation, visualization, and writing. Authors verified cited sources, mathematics, executed code, measurements, and AI-assisted text and artifacts, and take responsibility for this work.

REPRODUCIBILITY STATEMENT

Appendix A and Sec. 5 specify the model, estimator, training, preprocessing, and evaluation protocols. The appendix includes proofs, pseudocode, per-head and per-configuration results, and baseline provenance. Anonymized code and configurations will accompany submission.

REFERENCES

Samet Akcay, Dick Ameln, Ashwin Vaidya, Barath Lakshmanan, Nilesh Ahuja, and Utku Genc. Anomalib: A deep learning library for anomaly detection. In *IEEE International Conference on Image Processing (ICIP)*, 2022.

Kilian Batzner, Lars Heckler, and Rebecca König. EfficientAD: Accurate visual anomaly detection at millisecond-level latencies. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD – a comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, 2019.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *NeurIPS*, 2018.

Shengzhe Chen, Mehrdad Moradi, Kamran Paynabar, and Hao Yan. Flow mismatching: Unsupervised anomaly detection via velocity discrepancies in flow matching models, 2026. arXiv:2605.23070.

Thomas Defard, Aleksandr Setkov, Angélique Loesch, and Romaric Audigier. PaDiM: A patch distribution modeling framework for anomaly detection and localization. In *ICPR International Workshops and Challenges*, 2021.

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations (ICLR)*, 2017.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *ICLR*, 2019.

Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. CFLOW-AD: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022.

Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *CVPR*, 2025.

Zhengnan Hu, Xiangrui Zeng, Yiqun Li, Zhouping Yin, Erli Meng, Ziyu Wei, Leyan Zhu, and Zitian Wang. MSAttnFlow: Normalizing flow for unsupervised anomaly detection with multi-scale attention. *Pattern Recognition*, 161:111220, 2025. doi: 10.1016/j.patcog.2024.111220.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Why normalizing flows fail to detect out-of-distribution data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Mathis Kruse and Bodo Rosenhahn. Multi-Flow: Multi-view-enriched normalizing flows for industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3933–3944, June 2025. doi: 10.1109/CVPRW67362.2025.00378.

Mathis Kruse, Melissa Schween, and Bodo Rosenhahn. MATCH: Flow matching for multi-view anomaly detection. In *European Conference on Computer Vision (ECCV)*, 2026. arXiv:2606.24375.

Charline Le Lan and Laurent Dinh. Perfect density models cannot guarantee anomaly detection. *Entropy*, 23(12):1690, 2021. doi: 10.3390/e23121690.

Liangwei Li, Lin Liu, Hanzhe Liang, Juanxiu Liu, Jing Zhang, Ruqian Hao, Xiaohui Du, Yong Liu, and Pan Li. Time-reversed flow matching with worst transport in high-dimensional latent space for image anomaly detection, 2026. URL https://arxiv.org/abs/2508.05461v3. arXiv:2508.05461v3.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *International Conference on Learning Representations (ICLR)*, 2019.

NVIDIA Research. C-RADIOv4: Distilled vision foundation model. https://huggingface.co/nvidia/C-RADIOv4-H, 2026.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, 2022.

Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-based defect detection. In *WACV*, 2022.

Shunsuke Sakai, Xiangteng He, Chunzhi Gu, Leonid Sigal, and Tatsuhito Hasegawa. InvAD: Inversion-based reconstruction-free anomaly detection with diffusion models. In *CVPR*, 2026. arXiv:2504.05662.

Robin Tibor Schirrmeister, Yuxuan Zhou, Tonio Ball, and Dan Zhang. Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Matías Tailanian, Álvaro Pardo, and Pablo Musé. U-Flow: A U-shaped normalizing flow for anomaly detection with unsupervised threshold. *Journal of Mathematical Imaging and Vision*, 66(4):678–696, 2024. doi: 10.1007/s10851-024-01193-y.

Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, et al. Real-IAD: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *CVPR*, 2024.

Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. FastFlow: Unsupervised anomaly detection and localization via 2D normalizing flows, 2021. arXiv:2111.07677.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRAEM: A discriminatively trained reconstruction embedding for surface anomaly detection. In *ICCV*, 2021.

Jiangning Zhang, Haoyang He, Zhenye Gan, Qingdong He, Yuxuan Cai, Zhucun Xue, Yabiao Wang, Chengjie Wang, Lei Xie, and Yong Liu. ADer: A comprehensive benchmark for multi-class visual anomaly detection. *arXiv preprint arXiv:2406.03262*, 2024.

Yixuan Zhou, Xing Xu, Jingkuan Song, Fumin Shen, and Heng Tao Shen. MSFlow: Multiscale flow-based framework for unsupervised anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):2437–2450, 2025. doi: 10.1109/TNNLS.2023.3344118.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. SPot-the-Difference self-supervised pre-training for anomaly detection and segmentation. In *ECCV*, 2022.

Algorithm 1 HUTCHINSON scoring (per category).

Require: trained v_θ; test patch tokens $\{f_i\}_{i=1}^N$; reverse Euler steps K_{ODE}, probes S, chunk size B, scale $\alpha{=}10^{-3}$.
Ensure: log-likelihood estimate $\widehat{\ell}_i \approx \log p_\theta(f_i)$ for each patch.

for each chunk $\mathcal{C}$ of B patches (all lines below act row-wise) **do**
  $z \leftarrow [f_i]_{i\in\mathcal{C}}$, $\delta \log p \leftarrow 0$
  $h_i \leftarrow \alpha \max(1, \|f_i\|_2)$ for $i \in \mathcal{C}$ {fixed along each path}
  for $k = 0, \ldots, K_{\mathrm{ODE}}{-}1$ **do**
    $t \leftarrow 1 - k/K_{\mathrm{ODE}}$, $dt \leftarrow -1/K_{\mathrm{ODE}}$
    $v \leftarrow v_\theta(z, t)$
    for $s = 1, \ldots, S$ **do**
      $\varepsilon_s \sim \mathcal{N}(0, I)$ {fresh for each step and probe}
      $\delta \log p \mathrel{+}= dt \cdot \frac{1}{S}\, \varepsilon_s^\top\big((v_\theta(z + h\varepsilon_s, t) - v)/h\big)$
    end for
    $z \leftarrow z + dt \cdot v$ {Euler step toward $t{=}0$}
  end for
  $\widehat{\ell}_i \leftarrow -\frac{1}{2}\|z_i\|^2 - \frac{D}{2}\log 2\pi + \delta \log p_i$, $i \in \mathcal{C}$
  free z, v on GPU
end for

A IMPLEMENTATION AND ESTIMATOR DETAILS

Training configuration. Velocity fields are trained for 30 epochs with AdamW (learning rate 10^{-3} cosine-decayed to 10^{-5}, weight decay 10^{-4}, gradient clipping 1.0). An epoch is one shuffled pass over *all* cached patches of the category in batches of 4096: at 448^2 every image contributes 784 tokens, so a category holds 0.05–0.31M patches on MVTec AD, 0.35–0.71M on VisA and 0.65–0.99M on Real-IAD, giving 0.4–2.3k, 2.6–5.2k and 4.8–7.3k optimizer steps. Features are extracted once and cached in `fp16`; training upcasts to `fp32` and uses neither mixed precision nor TF32. The quoted wall-clock time covers training alone: the median on one L20 is 44 s across the 15 MVTec AD categories and 165 s on VisA, using C-RADIOv4-H features. Scoring integrates the reverse ODE with $K_{\mathrm{ODE}}{=}20$ Euler steps and $S{=}5$ finite-difference Hutchinson probes. In the common-feature comparison, Gaussian and GMM use covariance regularization $0.01\,\mathrm{tr}(\Sigma)/D$ times the identity; the GMM uses $K = 5$ full-covariance components and 50 EM steps; the coupling flow has eight coupling blocks with hidden width 1280; the Parzen/KDE baseline retains at most 100,000 training patches and sets its bandwidth from median nearest-neighbor distances in a training subsample of at most 20,000 patches. Statistical refits use seed 0; neural checkpoints are reused without retraining. Training and model-selection budgets differ across estimators. The reported scorer uses sequential forward-only probes, avoiding automatic differentiation during scoring.

Scoring cost. The shared-L20 head/layer timings in Appendix E include feature transfers, scoring, smoothing and fusion. The separate A800 endpoint/FD/JVP comparison in Appendix A.3 holds the models, selected inputs, chunk size and numerical settings fixed across implementations. Retained model state is fixed by architecture and layer count; measured peak allocation also includes the inputs and working tensors specified for each timing scope.

Finite-difference bias. Hutchinson is unbiased for exact Jacobian–vector products. For a fixed Gaussian probe, the one-sided difference (3) has $\mathcal{O}(h)$ error; in expectation its first-order term contracts three probe components and vanishes by central symmetry. The expected trace estimate has the expansion

$$\mathbb{E}\big[\widehat{\mathrm{tr}}(J)\big] \;=\; \mathrm{tr}(J) + \frac{h^2}{2}\,\Delta\big(\nabla{\cdot}v_\theta\big) + \mathcal{O}(h^4), \tag{5}$$

where Δ acts on the scalar divergence field. Thus one-sided probing is second-order accurate in expectation. The expansion assumes sufficient smoothness and integrability for the Gaussian expectation; it does not include floating-point or ODE error. Because $h_i = \alpha \max(1, \|f_i\|_2)$ is fixed from each initial token, the leading bias along its path is $\frac{\alpha^2 \max(1,\|f_i\|_2)^2}{2}\Delta(\nabla{\cdot}v_\theta)(z_t, t)$. It varies

Change	Min. image ρ	$\lvert\Delta\rvert$ I-AUC	$\lvert\Delta\rvert$ P-AUC	$\lvert\Delta\rvert$ AP	$\lvert\Delta\rvert$ PRO
$K = 10$	0.99925	0.287	0.009	0.047	0.046
$K = 40$	0.99965	0.061	0.004	0.123	0.025
$S = 1$	0.99893	0.205	0.004	0.074	0.015
$S = 10$	0.99968	0.000	0.001	0.079	0.012
$\alpha = 3 \times 10^{-4}$	0.99998	0.020	0.002	0.004	0.007
$\alpha = 3 \times 10^{-3}$	0.99975	0.160	0.014	0.027	0.052

Table A1: **Fixed-model numerical sensitivity.** Each row changes one setting from $K = 20$, $S = 5$, $\alpha = 10^{-3}$. Values are the minimum image-score Spearman correlation and maximum absolute metric changes (percentage points) across `screw`, `capsule` and `metal_nut`; different columns may attain their maxima in different categories. All six variants are shown. PRO is integrated to FPR 0.3 and checked against the official implementation.

across inputs and integration times, so ranking stability is an empirical question, not a consequence of unbiased exact-JVP probing.

A.1 Fixed-Model Numerical Sensitivity

The estimated full-score rankings are stable over the tested numerical grid (Table A1). We use all 407 test images from `screw`, `capsule` and `metal_nut`, with one frozen checkpoint and feature cache per category. Starting from $K = 20$, $S = 5$, $\alpha = 10^{-3}$, we separately vary $K \in \{10, 40\}$, $S \in \{1, 10\}$ and $\alpha \in \{3 \times 10^{-4}, 3 \times 10^{-3}\}$, for 21 category/configuration runs including the baselines. Scoring uses float32, chunks of 4096 patches, and seed 0 reset for each configuration. Models, cached tokens, sample order and postprocessing are fixed, with no retraining or parameter selection.

The minimum image-score Spearman correlation is 0.99893. The largest absolute changes are 0.287 points in image AUROC, 0.014 in pixel AUROC, 0.123 in pixel AP and 0.052 in P-AUPRO. The numerical grid was fixed before scoring. Reported PRO values are a post-hoc re-evaluation of all saved maps using integration through FPR 0.3, replacing a threshold-grid calculation that omitted the final interval. The experiment measures full-score sensitivity on these three fixed models. The separate correction is examined directly in the shared-probe comparison below.

A.2 Shared-Probe Finite-Difference versus JVP Control

Replacing finite differences with automatic-differentiation Jacobian–vector products preserves the localization gain on the three tested models (Table A2). We use all 407 test images from `screw`, `capsule`, and `metal_nut`, with one existing single-layer C-RADIO checkpoint and its frozen feature cache per category. The branches share every reverse-Euler state and Gaussian probe ($K = 20$, $S = 5$, scoring seed 0, float32, TF32 disabled, chunk size 4096). The FD branch uses $h_i = 10^{-3} \max(1, \|f_i\|_2)$; the JVP branch evaluates the same directional derivative using forward-mode automatic differentiation. We save E, V_{FD}, V_{JVP}, and both full scores, with identical upsampling and smoothing.

Correction-only image-score Spearman correlations exceed 0.99998; native-patch correlations exceed 0.99994. The full-score P-AUPRO differences are below 0.0085 points, and the JVP full score retains positive gains over E in each category. Native correction values have relative L_2 differences of 0.728–1.020%, so the agreement is stronger for rankings than for absolute correction values. Relative L_2 uses the JVP values as the reference. Both branches retain Euler discretization, stochastic trace estimation and floating-point error; the control isolates the additional effect of finite differences.

Analytic-field checks verify the sign and finite-difference bias, and the FD branch reproduces the production scorer bitwise on the first two test images of each category. All 15 category/head PRO values agree with the official implementation within 4×10^{-11} on the $[0, 1]$ scale. The experiment fixes the provided checkpoints and caches; historical backbone weights and physical layer identities are not fully authenticated.

Table A2: **Shared-probe FD/JVP comparison.** One fixed model per category. ρ_V correlates correction-only image scores; relative L_2 compares native patch corrections against JVP (percent). ΔPRO is FD minus JVP in percentage points; the final column is full JVP minus E.

Category	N	ρ_V	Rel. L_2	ΔPRO V	ΔPRO full	JVP gain
screw	160	0.999982	0.728	$+0.0073$	$+0.0084$	$+6.340$
capsule	132	0.999990	1.020	$+0.0031$	$+0.0009$	$+1.013$
metal_nut	115	0.999984	0.775	$+0.0011$	-0.0011	$+1.128$

A.3 Independent Endpoint, FD and Forward-JVP Costs

Table A3 measures the three implementations independently on the same frozen single-layer fields used in Appendix A.2. For each of `screw`, `capsule` and `metal_nut`, we select the first 16 normal and first 16 anomalous test images in canonical path order, giving 96 images. Cache and checkpoint hashes match the accepted paired experiment. This fixed subset is used only for resource measurement; Table A2 uses all 407 images for the accuracy comparison.

Table A3: Independent scoring costs on a shared A800 80 GB. Each category uses 32 fixed images. Time is the median of three cyclically ordered repetitions; memory is the maximum peak PyTorch allocated memory, including model state and resident inputs. Pipeline includes transfers and pixel postprocessing.

		Time (ms/image)		Peak allocated (MiB)	
Category	Scorer	Native	Pipeline	Native	Pipeline
screw	E	57.3	58.3	427.9	305.4
	FD	345.9	348.6	487.9	365.4
	JVP	840.8	846.3	627.9	505.4
capsule	E	57.4	58.1	427.9	305.4
	FD	347.7	347.9	487.9	365.4
	JVP	843.8	846.1	627.9	505.4
metal_nut	E	57.4	58.2	427.9	305.4
	FD	348.2	348.8	487.9	365.4
	JVP	844.4	847.9	627.9	505.4

All scorers use float32, $K = 20$, chunks of 4096 patches and two CPU threads, with TF32 and autocast disabled. FD and JVP use $S = 5$ Gaussian probes with seed 0 reset before each invocation; endpoint scoring uses no probes. FD uses $h_i = 10^{-3} \max(1, \|f_i\|_2)$. Forward-mode JVP reuses the first probe's primal velocity for the Euler step. Model parameters require no gradients, and all branches run under `no_grad`, retaining no reverse-mode graph. The implementation uses PyTorch 2.6.0 with CUDA 11.8 on a shared NVIDIA A800 80 GB. GPU process snapshots are retained; the measurements do not assume exclusive device access.

The *native* scope starts with all 32 images' features resident on GPU and returns native GPU maps. The *pipeline* scope streams CPU features in chunks, then upsamples to 448^2, applies Gaussian smoothing ($\sigma = 4$), takes the image maximum and copies image and pixel scores to CPU. Both exclude feature extraction and normalization, cache/model loading, training, metric calculation and file I/O. Native input residency contributes to its allocated-memory baseline. Reported peaks include model state, resident inputs and working tensors; the raw record also separates baseline and incremental allocation, and records allocator-reserved memory.

Each category and scope receives one warmup invocation per method, followed by three repetitions in cyclic orders: E/FD/JVP, FD/JVP/E, and JVP/E/FD. CUDA is synchronized around each wall-clock timer. After warmup cleanup, the allocator cache is emptied and peak statistics reset before each timed invocation. We report median time and maximum peak allocated memory. Repeated outputs are bitwise identical. Analytic and multi-chunk checks pass; on the first two selected images of every category, each independent branch reproduces the accepted joint scorer bitwise, and FD also reproduces the production implementation.

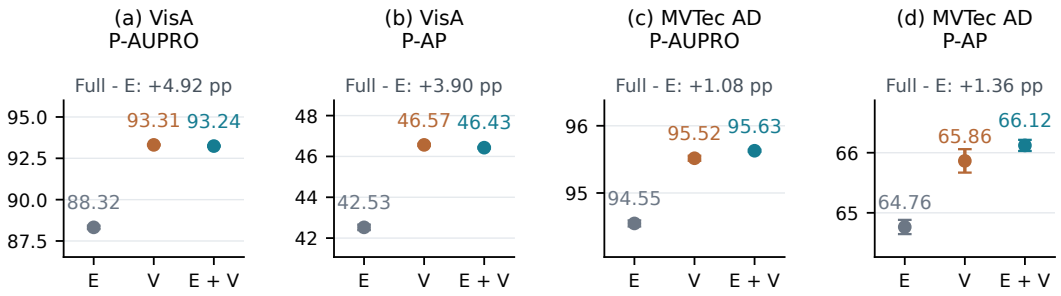

Figure A1: **Endpoint, trace and full-score localization.** Gaussian endpoint E, trace correction V, and full score $E + V$ on **(a,b)** VisA and **(c,d)** MVTec AD. Points and error bars show category-macro means and SD across three model seeds; axes are in percent. Fixed single-layer features, trained fields and postprocessing isolate the scoring terms. Both datasets use P-AUPRO at FPR 0.3; the full four-metric results are in Table 1.

B Paired Trace Decomposition

Paired evaluation. Figures 1 and A1 use all 12 VisA and all 15 MVTec AD categories, each with three existing model seeds. We hold fixed the C-RADIOv4-H final-layer unnormalized features, trained field, test order and 20-step reverse Euler path. The full score uses five Hutchinson probes. The endpoint is replayed without retraining; a synthetic check and a full-category replay check verify bitwise agreement with the joint decomposition. The trace-only diagnostic is the native float32 residual $V = (E + V) - E$. Each head uses identical bilinear resizing from 28^2 to 448^2, Gaussian smoothing ($\sigma = 4$, kernel 17, reflect padding), and the maximum of its own pixel map for image scoring. Every category/head is checked against the official PRO implementation at an absolute tolerance of 10^{-7}, with integration to FPR 0.3.

Full-minus-endpoint P-AUPRO gains for the three model seeds are $4.90/4.83/5.02$ points on VisA and $1.05/1.07/1.13$ on MVTec AD. Descriptive paired category-bootstrap 95% intervals are $[2.58, 8.09]$ and $[0.41, 2.04]$ points, respectively, from 20,000 resamples after averaging seeds within each category. These intervals quantify across-category variation with the three model seeds held fixed. All 36 VisA category/seed pairs improve P-AUPRO; MVTec AD has 14/15 positive category means, with `transistor` negative. Pixel AP improves in 10/12 VisA category means. Trace-only slightly exceeds full on VisA P-AUPRO and AP, while its image AUROC is lower.

Layer configuration. The fixed deep-5 configuration uses blocks $\{15, 19, 23, 27, 31\}$. Table A13 compares these five layers with L31 from that same model bank. The component decomposition above uses a separately trained final-layer model with full-path integration.

B.1 DINOv3 Representation Control

Data and feature extraction. We use the official VisA one-class split, including all 12 categories, 8,659 normal training images and 2,162 test images. File identities and labels are checked against the official split CSV; image bytes and resized binary masks are checked against a manifest generated from the original VisA files. Features are freshly extracted with the frozen DINOv3 ViT-L/16 backbone (Siméoni et al., 2025), using physical block 23 (zero-based), final LayerNorm, and no train-radius rescaling. Images are bicubically resized to 448×448 and ImageNet-normalized, yielding 28×28 tokens of dimension 1024. Extraction uses float32 with autocast and TF32 disabled; tokens are stored in float16 and upcast to float32 for training and scoring. Backbone, code, split, and cache hashes are recorded for each category.

Training and paired scoring. Each category receives a new four-residual-block velocity MLP with hidden width 2048 and time-embedding dimension 256. Seed 0 is set before model initialization and again at training entry. We use the normal training patches only, the 30-epoch optimization schedule in Appendix A, and the last checkpoint, without test-based selection. The three heads share a 20-step reverse Euler trajectory and five Gaussian probes per step, with scoring seed 0, float32

Table A4: **DINOv3 per-category results.** P-AUPRO values in percent; the four rightmost columns report full minus endpoint in percentage points. All heads use the same seed-0 field per category.

Category	PRO E	PRO V	PRO full	ΔI-AUC	ΔP-AUC	ΔAP	ΔPRO
candle	96.76	97.37	97.40	−1.27	+0.53	+4.91	+0.64
capsules	92.49	96.27	96.35	+1.90	+0.79	+0.01	+3.86
cashew	86.11	96.27	96.07	+5.18	+0.27	+10.23	+9.96
chewinggum	88.16	91.25	91.13	+0.34	+0.15	+1.15	+2.97
fryum	75.93	88.14	87.96	+3.76	+1.80	+16.22	+12.02
macaroni1	89.50	95.84	95.97	+1.53	+1.19	−3.37	+6.47
macaroni2	82.24	92.24	92.55	+2.89	+4.89	+4.00	+10.32
pcb1	89.87	95.66	95.70	−1.01	+0.14	+2.46	+5.83
pcb2	82.12	91.20	91.23	−0.03	+0.74	+6.02	+9.11
pcb3	73.25	87.61	87.57	−0.88	+1.44	+28.08	+14.31
pcb4	79.49	93.60	93.59	+1.83	+2.56	+8.61	+14.09
pipe_fryum	92.83	96.00	95.91	+0.66	+0.28	+11.01	+3.08

computation, and chunks of 4096 patches. The finite-difference scale is $h_i = 10^{-3}\max(1, \|f_i\|_2)$, fixed along each trajectory. We accumulate the signed V directly and compute $E = \|z_0\|^2/2$ and $E + 512\log(2\pi) + V$. All heads use the same bilinear upsampling, $\sigma = 4$ Gaussian smoothing, and image maximum. There is no layer fusion, test-score standardization, or sign selection.

Training and scoring are distributed across two A800 workers (eight categories) and one L20 worker (four categories), with identical backbone weights and software versions (PyTorch 2.6.0, torchvision 0.21.0, timm 1.0.26). Each category's heads are paired on the same device and checkpoint. These runs test the decomposition within DINOv3; absolute differences from the historical C-RADIO results also reflect different fitted models and training provenance. Cross-seed variability is not estimated.

Results and numerical checks. Table A4 reports every category, including negative changes. P-AUPRO and pixel AUROC improve in 12/12 categories, pixel AP in 11/12, and image AUROC in 8/12. The correction-only and full P-AUPRO means differ by less than 0.005 points. For each category, the first two test images reproduce the production full-score implementation bitwise. All 36 category/head PRO results are checked against the official implementation; the maximum absolute discrepancy is 2.23×10^{-10} on the $[0, 1]$ scale.

B.2 Region Coverage at Matched False-Positive Rates

The correction improves defect-region coverage at the tested false-positive rates on all 12 DINOv3–VisA categories (Table A5). This post-hoc diagnostic reuses all 2162 test images and the seed-0 E, V and full scores from Appendix B.1, with unchanged upsampling and smoothing. We fix the operating-point grid at 1%, 5%, 10% and 30% before computing this diagnostic. No retraining or checkpoint selection is performed.

Table A5: Region coverage at matched pixel FPR on DINOv3–VisA. All 12 categories, 2162 test images, one model seed. Coverage is in percent, averaged over regions within each category, then over categories. Full denotes $E + V$ (plus the base constant). Thresholds are evaluated using test masks.

Pixel FPR	E coverage	V coverage	Full coverage	Full−E (pp)
1%	47.54	63.11	62.81	+15.27
5%	76.51	89.02	89.06	+12.55
10%	86.67	95.48	95.56	+8.88
30%	96.60	99.61	99.63	+3.03

Thresholds and coverage. For each category and head, the negative reference contains every GT-negative pixel in all test images, including backgrounds of anomalous images, as in P-AUPRO. We

predict anomaly when the score strictly exceeds an empirical order-statistic threshold chosen so that FPR does not exceed the target. Ties are kept together; across all reported combinations, the largest target–achieved FPR gap is 8.40×10^{-7} on the $[0, 1]$ scale. These thresholds use test ground truth to define evaluation operating points; deployment calibration would require a separate reference set. Coverage is the detected fraction of each 8-connected defect region, averaged equally over regions within a category and then equally over the 12 VisA categories. Scores retain their original scale.

Per-category behavior and supplementary cases. Table A6 reports every VisA category. The full score improves coverage over E at every listed operating point, while V alone remains close to full scoring in the macro-average. We additionally diagnose the existing seed-0 C-RADIO `screw` and `transistor` scores on all 160 and 100 test images, respectively, using the same protocol. These two categories are reported separately: `transistor` loses 4.08 coverage points at 5% FPR and 2.44 at 10%, despite improving at 1%. The correction's value can therefore depend on the operating point as well as the category.

Table A6: Per-category full-minus-endpoint region coverage (percentage points) at four matched pixel FPR targets. The first block covers all VisA categories on DINOv3; the second contains two supplementary C-RADIO MVTec AD cases, excluded from the VisA macro-average. All rows use seed 0.

Category	FPR 1%	FPR 5%	FPR 10%	FPR 30%
DINOv3–VisA (all 12 categories)				
candle	+2.54	+1.09	+0.90	+0.03
capsules	+12.27	+7.54	+4.34	+0.95
cashew	+23.16	+18.88	+12.20	+1.75
chewinggum	+12.72	+7.20	+0.78	+0.62
fryum	+13.28	+13.79	+16.15	+6.75
macaroni1	+4.86	+11.87	+8.88	+2.76
macaroni2	+28.75	+16.58	+10.33	+3.90
pcb1	+18.44	+11.12	+6.96	+1.05
pcb2	+14.52	+10.07	+10.33	+6.74
pcb3	+23.33	+19.91	+13.41	+7.46
pcb4	+13.29	+26.77	+19.42	+4.33
pipe_fryum	+16.08	+5.79	+2.92	+0.00
C-RADIO–MVTec AD (two supplementary cases)				
screw	+6.93	+7.95	+7.33	+3.84
transistor	+8.34	-4.08	-2.44	+0.28

Validation. Image order and labels are checked against the saved scoring records; VisA source masks and resized binary masks match the official manifest hashes. Independent summation of per-image TP/FP counts and per-region coverage reproduces all 168 category/head/operating-point results. The evidence file also retains anomalous-pixel recall, precision, and FPR separately for normal images and anomalous-image backgrounds. The diagnostic uses one fitted model per category; it adds operating-point resolution to the paired evaluation rather than another independent model replication.

B.3 Component Ablation within the Five-Layer Model Bank

Fixed models and inputs. We evaluate all 12 VisA categories and 2162 test images using one existing five-layer model bank, comprising 60 checkpoints at blocks $\{15, 19, 23, 27, 31\}$. Each layer uses its normal-training-cache radius normalization. The 1L condition is the L31 model from this same bank; all checkpoints are held fixed, with no retraining or selection. Current checkpoint and cache hashes are recorded. Complete historical initialization and feature-extraction receipts are unavailable; current hashes identify the fixed assets used for this comparison.

Paired scoring and shared fusion weights. We recompute E, V, and the full score on a shared reverse trajectory and Gaussian probes, with $K = 20$, $S = 5$, relative difference scale 10^{-3}, float32 arithmetic, and chunks of 4096 patches. Scoring seeds are $20260924 + \ell$ for layer indices $\ell =$

$0, \ldots, 4$. All heads use bilinear resizing to 448^2 and Gaussian smoothing ($\sigma = 4$). Let $M_h^{(\ell)}$ be the resulting map for head h, with mean $\mu_h^{(\ell)}$ over all test pixels in a category. The five-layer score is

$$S_h^{(5)} = \frac{1}{5} \sum_{\ell=0}^{4} \frac{M_h^{(\ell)} - \mu_h^{(\ell)}}{\sigma_F^{(\ell)}}, \qquad h \in \{E, V, F\}, \tag{6}$$

where $\sigma_F^{(\ell)}$ is the full map's sample standard deviation, clamped below at 10^{-8}. Thus all three heads retain the same positive fusion weights from the full-score protocol; E and V receive no separately fitted scales. Centering removes category-wise constants, giving $S_E^{(5)} + S_V^{(5)} \simeq S_F^{(5)}$ up to floating-point error. The 1L heads use their unstandardized L31 maps. Image scores are the maximum of the final pixel map. The five-layer statistics use unlabeled test scores, and the fused outputs are detector scores.

The full score passes bitwise production-scoring checks on two images per checkpoint, and full fusion reproduces the production aggregation bitwise. All 72 category/condition P-AUPRO values pass the independent official implementation at tolerance 10^{-7}. Historical scoring RNG states are unavailable; this fresh paired scoring pass is reported separately from the three-bank benchmark means.

Table A7: Per-category P-AUPRO (percent) for the same-bank component ablation. The six columns use the same conditions as Table 3.

Category	1L E	1L V	1L Full	5L E	5L V	5L Full
candle	96.05	97.95	97.91	97.59	97.89	97.96
capsules	95.40	96.61	96.58	97.94	97.73	97.97
cashew	86.41	95.34	94.57	95.54	98.04	97.73
chewinggum	86.35	89.47	88.92	92.02	92.59	92.54
fryum	66.93	87.55	85.74	83.32	91.67	90.99
macaroni1	97.14	96.90	97.32	98.51	97.12	97.81
macaroni2	92.27	93.20	93.95	97.27	95.87	96.84
pcb1	93.04	96.51	96.84	97.49	96.57	97.34
pcb2	87.62	91.11	91.51	92.80	90.21	92.10
pcb3	82.47	91.74	91.01	91.27	93.71	94.35
pcb4	86.54	87.14	88.66	89.73	87.52	89.05
pipe_fryum	90.25	96.76	95.83	94.96	97.93	97.58

Table A8: Per-category P-AP (percent) for the same-bank component ablation. The six columns use the same conditions as Table 3.

Category	1L E	1L V	1L Full	5L E	5L V	5L Full
candle	30.11	37.61	36.02	27.54	33.66	32.04
capsules	53.54	48.10	51.32	61.27	48.08	56.56
cashew	77.49	80.02	80.90	67.98	61.60	66.83
chewinggum	80.95	81.04	81.32	71.62	75.02	73.43
fryum	38.90	54.46	50.13	45.31	51.58	50.52
macaroni1	17.62	18.47	19.31	21.58	20.26	23.27
macaroni2	11.24	13.42	13.49	15.86	18.03	19.04
pcb1	86.50	88.28	90.29	81.91	62.89	79.95
pcb2	11.18	21.99	17.53	14.29	23.46	18.98
pcb3	8.08	25.69	18.96	19.25	34.49	29.92
pcb4	38.99	27.79	36.40	42.34	19.94	31.43
pipe_fryum	57.34	58.86	60.32	52.81	46.37	51.82

Component and fusion effects. Fusion increases mean P-AUPRO by 5.67, 1.38, and 1.95 points for E, V, and full, respectively. The full-minus-E gain decreases in all 12 categories, from 4.87 to 1.15 points on average; the difference between these marginal gains is -3.72 points. At five layers the correction improves P-AUPRO in 7/12 categories and pixel AP in 7/12. P-AUPRO falls in `macaroni1`, `macaroni2`, `pcb1`, `pcb2`, and `pcb4`; the largest pixel-AP loss is 10.91 points

on pcb4. Fusion itself improves full-score P-AUPRO in all 12 categories but reduces its mean pixel AP by 1.85 points.

The endpoint becomes more useful relative to correction-only scoring after fusion. At L31, full minus V is -0.12 P-AUPRO and $+0.02$ pixel-AP points; at five layers it is $+0.45$ and $+3.20$ points. The fused full score also exceeds fused V in mean image and pixel AUROC. These comparisons use the fixed full-score fusion scales; complete four-metric category results accompany the source supplement.

C Proof of the Gaussian Zero-Information Proposition

Setting. Normal features $x_1 \sim \mathcal{N}(\mu, \Sigma)$ with $\Sigma \succ 0$; base $x_0 \sim \mathcal{N}(0, I_D)$ drawn independently; linear interpolant $x_t = (1-t)\,x_0 + t\,x_1$. The exact minimizer of the FM objective over measurable fields is the conditional-expectation field (Lipman et al., 2023)

$$v^*(z,t) \;=\; \mathbb{E}\big[x_1 - x_0 \,\big|\, x_t = z\big]. \tag{7}$$

Step 1: the optimal field is affine. (x_0, x_1) is jointly Gaussian, so $(x_t,\, x_1 - x_0)$ is jointly Gaussian, and the conditional expectation of one jointly Gaussian vector given another is affine in the conditioning variable:

$$v^*(z,t) = \mu + B_t\,(z - t\mu), \qquad B_t = \mathrm{Cov}(x_1 - x_0,\, x_t)\,C_t^{-1}, \tag{8}$$

with $C_t := \mathrm{Cov}(x_t) = (1-t)^2 I + t^2\Sigma \succ 0$ and $\mathrm{Cov}(x_1 - x_0,\, x_t) = t\Sigma - (1-t)I$.

Step 2: spatially uniform divergence. The Jacobian $\partial v^*/\partial z = B_t =: A_t$ does not depend on z, so $\mathrm{tr}\, A_t$ is the same at every spatial position.

Step 3: the trace integral is input-independent. Along any reverse trajectory the integrand $\mathrm{tr}\big(\partial v^*/\partial z\big)(z_t, t) \;=\; \mathrm{tr}\, A_t$ depends on t only, so the trace integral is one constant for every input. Its value follows per eigenvalue λ of Σ: with $a_t = \big(t\lambda - (1-t)\big)/\big((1-t)^2 + t^2\lambda\big)$,

$$\frac{d}{dt}\log\big((1-t)^2 + t^2\lambda\big) = 2a_t \;\Longrightarrow\; \int_0^1 a_t\,dt = \tfrac{1}{2}\log\lambda, \tag{9}$$

so $\int_0^1 \mathrm{tr}\, A_t\,dt = \frac{1}{2}\log\det\Sigma$.

Step 4: the endpoint is the Mahalanobis distance. The marginal field transports $\mathcal{N}(0, I)$ to $\mathcal{N}(\mu, \Sigma)$ exactly, so the change of variables (Eq. 2) with the constant trace integral of Step 3 gives $-\log p(x) = \frac{1}{2}\|z_0(x)\|^2 + \frac{D}{2}\log 2\pi + \frac{1}{2}\log\det\Sigma$. Matching the Gaussian negative log-likelihood forces $\|z_0(x)\|^2 = (x-\mu)^\top\Sigma^{-1}(x-\mu)$: the endpoint is the Mahalanobis distance and, the trace being constant, a sufficient statistic for likelihood scoring. □

Corollary (TrajBif). For the affine field, the difference δ_t between two reverse trajectories obeys $\dot\delta = A_t\delta$, and a Langevin perturbation obeys $d\delta = A_t\delta\,dt + \sigma\,dW$; in both cases the propagator depends on t only. The endpoint-spread distribution is therefore identical for every input, and the TrajBif statistic of Appendix F carries zero discriminative information in the Gaussian regime, exactly as the trace does.

D Standardized Evaluation Protocol and Baseline Comparisons

Table A10 archives the numbers the original publications report for themselves, verified against each paper's own tables. Table 4 separates recent original reports in (a), our native-backbone reproduction in (b), and our C-RADIO evaluation in (c). Table A9 provides additional independent reproductions of classic methods. Protocol differences can produce substantial gaps between original reports and reproductions, through documented mechanisms: (1) under the Anomalib library's default re-runs (Akcay et al., 2022), FastFlow falls from its reported 99.4 to 96.3 (best backbone; 94.4 with the paper's CaiT), CFLOW-AD from 98.3 to 96.2, PatchCore from 99.1 to 98.0, and

Method	Source	I-AUC	P-AUC	PRO
MVTec AD, per-class, independent re-runs				
PaDiM	Anomalib (ours)	89.6	96.5	—
CFLOW-AD	Anomalib (ours)	72.8[a]	96.3	—
U-Flow	Anomalib (ours)	97.5	98.0	84.6
FastFlow	Anomalib (ours)	90.0[b]	95.2	—
CS-Flow	Anomalib (ours)	97.4	84.5	—
EfficientAD-M	Anomalib	97.5	—	—
PatchCore	Anomalib (ours)	98.2	98.0	—
SimpleNet	EffAD re-eval.	98.2	—	—
RD4AD	Anomalib (ours)	98.4	97.4	—
Dinomaly	ours (DINOv3-L)	99.6	98.8	—
VisA, per-class, independent re-runs				
PaDiM	Anomalib (ours)	81.0	96.0	—
CFLOW-AD	Anomalib (ours)	81.5[a]	97.8	—
U-Flow	Anomalib (ours)	95.2	99.0	83.5
FastFlow	Anomalib (ours)	82.4	93.8	—
CS-Flow	Anomalib (ours)	90.6	87.9	—
PatchCore	Anomalib (ours)	90.1	98.0	—
Real-IAD, multi-view split, benchmark-sourced				
PatchCore	Real-IAD benchmark	89.4	—	92.5
SimpleNet	Real-IAD benchmark	88.5	—	—
Dinomaly	Dinomaly paper	92.0	—	95.1

Table A9: **Independently reproduced numbers for the classic methods** (Anomalib (Akcay et al., 2022) defaults, WideResNet-50 rows; SimpleNet from the peer-reviewed re-evaluation that disables its test-set model selection (Batzner et al., 2024)). Except for the separately evaluated U-Flow reproduction, historical PRO values from local reproductions are omitted until re-evaluated; their image and pixel AUROC values are retained. U-Flow retains its original evaluation protocol (Appendix D.1). The remaining *ours* entries use the local evaluation pipeline. We report independent reproductions because self-reported metrics from original publications are typically 1–5 points higher (documented gaps in Appendix D). Dashes indicate unavailable or withdrawn PRO values. [a]: CFLOW-AD's image-level score significantly degrades when using default library settings (with several categories performing near or below chance level) while its pixel metrics remain sound; the original selects the input resolution per class, the sharpest reproduction gap we measure. [b]: FastFlow exhibits substantial performance drops on two categories under default configurations (`hazelnut` 59.5, `screw` 66.4 I-AUROC), consistent with its documented configuration sensitivity. The Real-IAD block is *benchmark-sourced*, not re-run by us: the classical methods are not published on this benchmark individually, so we quote the dataset paper's own single-class evaluation and the Dinomaly paper's single-class table, both on the multi-view split we use. Our common-backbone Real-IAD rows are in Table A15.

EfficientAD-M from 99.1 to 97.5, while RD4AD's 98.5 reproduces exactly; (2) SimpleNet's official training loop evaluates on the test set after every epoch and reports the maximum (model selection on the test set); disabling it yields 98.2 versus the reported 99.6 (Batzner et al., 2024); (3) CFLOW-AD selects its input resolution per class (Gudovskiy et al., 2022), and PatchCore's often-quoted 99.6 is a three-backbone ensemble (the single model reports 99.1) (Roth et al., 2022); (4) benchmark authors note that the absence of standardized setups "can lead to potential biases . . . resulting in erroneous conclusions" (Zhang et al., 2024). These gaps motivate separating native-protocol context from the controlled comparisons in Table 4(c) and Table A11. The former shares a backbone, input size and metric code; the latter fixes the scoring features and postprocessing while reusing neural checkpoints. Our DINOv3-L adaptation of Dinomaly reaches 99.60 image AUROC on MVTec AD; its backbone differs from the original paper's DINOv2 setting. Its Real-IAD evaluation required a numerical fix to the decoder normalization. The evaluation codebase and per-method preprocessing, training budgets, and metric provenance will be released upon acceptance; native-protocol reproductions are documented separately from the common-metric evaluations.

Method	Backbone / protocol	I-AUC	P-AUC	PRO
MVTec AD				
UniAD	EffNet-b4, 224^2, MC	96.5	96.8	—
RD4AD	WRN-50, 256^2	98.5	97.8	93.9
CFLOW-AD	WRN-50, res. per class	98.3	98.6	94.6
FastFlow	CaiT-M48, 448^2	99.4	98.5	—
SimpleNet	WRN-50, 224^2	99.6	98.1	—
EfficientAD-M	PDN distilled	99.1	—	93.5
PatchCore	WRN-50, 224^2	99.1	98.1	93.4
Dinomaly	DINOv2-B, MC	99.6	98.4	94.8
WT-Flow	WRN-50, 512^2	99.1	98.0	—
InvAD	EffNet-B4, 256^2, MC	99.0	97.5	92.7
VisA				
EfficientAD-M	PDN distilled	98.1	—	94.0
Dinomaly	DINOv2-B, MC	98.7	98.7	94.5
InvAD	EffNet-B4, 256^2, MC	96.9	99.1	92.9
Real-IAD				
Dinomaly	DINOv2-B, MC	89.3	98.8	93.9
MATCH	WRN-50, 256^2	91.2	99.2	94.8

Table A10: **Self-reported numbers from the original publications**, archived for reference (MC = one multi-class model). RD4AD and SimpleNet results are from Deng & Li (2022) and Liu et al. (2023), respectively. Backbones, resolutions, class protocols, seeds, and PRO implementations all differ, and several of these numbers exceed their independent reproductions by 1–5 points (Table A9); see the discussion in this appendix. —: not reported.

D.1 Recent-Method Protocols and Sources

MSAttnFlow (Hu et al., 2025) contributes publisher-reported image and pixel AUROC on MVTec AD and VisA; its AP and PRO have not been verified. Multi-Flow (Kruse & Rosenhahn, 2025) uses EfficientNet-B5 at 768^2, cross-view message passing and foreground masking. Its Real-IAD values come from Table 1 and Sec. 4.4 of the original report (five runs); we do not mix in AP from another paper's re-evaluation.

WT-Flow (Li et al., 2026) uses WRN50 with single-class, single-scale training; Tables I/II of the April 2026 v3 report use 512^2 inputs except `transistor` at 256^2, with 20 detection steps and one localization step. InvAD (Sakai et al., 2026) contributes Table 2 throughout: multi-class training, EfficientNet-B4, 256^2 inputs, and three inversion steps with prior and deviation scores. MATCH (Kruse et al., 2026) uses WRN50 at 256^2 with five Euler steps. Its Real-IAD values come from Table 1 with view embeddings, and its MVTec AD values from Appendix C, Table 7; the separate multi-class Real-IAD setting is excluded. WT-Flow and MATCH use endpoint-based scores without the divergence term in their reported default configurations.

U-Flow (Tailanian et al., 2024) is our complete 15/12-category MVTec AD/VisA reproduction with Anomalib 2.3.0 defaults, evaluated separately from the C-RADIO reruns. Its native backbone and evaluation protocol are retained; its PRO values have not undergone the official-reference check used for block (c). Original-paper metrics in Table 4(a) retain their reported precision and metric implementations and are not used for a common-protocol ranking.

D.2 PatchCore Adaptation on C-RADIO Features

Our PatchCore adaptation uses three cached blocks, configured as zero-based indices $\{16, 24, 31\}$, and retains every normal training patch without coreset subsampling. Each layer scores a test patch by cosine distance to its nearest training token after per-token L2 normalization ($k = 1$). Each layer's grid is bilinearly resized to 448^2 and Gaussian-smoothed with $\sigma = 4$, kernel size 17 and reflect padding. Pixel maps are averaged across layers. Image scores are computed as the mean of the three per-layer map maxima; FlowAD instead takes the maximum after fusion. Thus the

Method	MVTec AD			VisA		
	I-AUC	AP	PRO	I-AUC	AP	PRO
Gaussian	97.62	62.72	94.61	92.61	43.27	92.24
GMM	96.99	60.52	94.63	91.27	40.56	91.51
KDE	88.83	59.92	91.55	83.88	36.76	78.63
Coupling NF	**98.90**	44.22	79.13	**95.76**	40.23	90.41
FlowAD-LL (1L)	98.56	**66.07**	**95.62**	95.57	**46.51**	**93.31**

Table A11: **Fitted density estimators on identical scoring features.** Category means (%); bold marks the best value per column. All five methods use the same final-layer C-RADIO cache, sample order, labels, masks, 448^2 output and smoothing. PRO is checked against the official implementation at 30% FPR. Gaussian/GMM/KDE are refitted with seed 0; FlowAD and coupling NF use one existing checkpoint per category, with no neural retraining. This single-checkpoint comparison fixes scoring inputs and postprocessing, separate from the three-seed FlowAD results in Table 4. Historical neural training inputs and seeds are not established by the common rescoring cache; training and model-selection budgets differ.

common-backbone comparison also includes differences in feature normalization, selected layers and aggregation. The memory and latency comparisons refer to this full-bank implementation.

D.3 Fitted Density Estimators on Common Scoring Features

We compare a global Gaussian, GMM, Parzen/KDE, a coupling flow (Dinh et al., 2017), and FlowAD. Statistical estimators are refitted; neural estimators reuse checkpoints whose historical training inputs and seeds are not fully established. The common cache fixes scoring inputs and postprocessing.

FlowAD improves localization over Gaussian scoring. Table A11 uses identical cached final-layer features and postprocessing for all five estimators. FlowAD exceeds Gaussian P-AUPRO by 1.01 points on MVTec AD (95.62 versus 94.61) and 1.07 on VisA (93.31 versus 92.24); pixel AP improves by 3.34 and 3.24 points. Gaussian and GMM remain strong localization baselines. The tested coupling flow has the highest image AUROC in this comparison, but lower localization metrics. This comparison evaluates fitted estimators: Gaussian/GMM/KDE are refitted, while neural methods use existing checkpoints. Training and model-selection budgets differ across estimators. Estimator settings are in Appendix A; independently reproduced classical flow baselines are in Appendix D.

E FlowAD Per-Head and Per-Configuration Results

Table A12 archives each scoring head at one layer and with deep-5 score fusion for all three benchmarks, with the seed spread and additional configurations omitted from Table 4. Table A13 separately isolates head and fusion effects in the same deep-5 bank; its L31 model is not the independent single-layer model. Fusion improves PRO but lowers pixel AP for both heads on both datasets. In a small shared-L20 timing probe on eight cached-feature images, median LL/TB scoring times are $279/392$ ms per image at one layer and $1417/1969$ ms at five layers. This includes transfers, scoring, smoothing and fusion, but excludes feature extraction, normalization, model loading and training. Tables A14 and A15 retain the extended baseline comparisons, complementing the single-layer estimator comparison in Table A11.

Dataset	Configuration	I-AUC	P-AUC	P-AP	P-AUPRO
MVTec AD	Likelihood (1 layer)	98.48 ±0.032	98.44	66.12	95.63 ±0.018
	TRAJBIF (1 layer)	98.55	98.40	65.99	—
	HUTCHINSON (deep-5)	98.30 ±0.009	98.46	63.26	96.21 ±0.008
	TRAJBIF (deep-5)	98.09 ±0.027	98.49	64.00	96.16 ±0.020
VisA	HUTCHINSON (L31)	95.57 ±0.035	99.25	46.43	93.24 ±0.030
	TRAJBIF (L31)	95.26	99.27	47.18	—
	HUTCHINSON (deep-5)	96.72 ±0.010	99.21	44.43	95.19 ±0.004
	TRAJBIF (deep-5)	95.75 ±0.055	99.22	43.98	95.23 ±0.008
Real-IAD	HUTCHINSON (single)	88.91	98.96	39.33	94.73
	TRAJBIF (single)	88.40	98.95	39.86	—
	+ view conditioning	89.06	98.97	39.45	—
	HUTCHINSON (deep-5)	89.62	99.14	40.29	96.18
	TRAJBIF (deep-5)	88.26	99.11	40.80	96.02

Table A12: **FlowAD per-head and per-configuration breakdown**, with dashes omitting historical PRO values without re-evaluation, complementing Table 4: each scoring head alone, at a single layer (L31 for MVTec AD/VisA) and with deep-5 score fusion. Reported PRO uses the official-reference-checked evaluation. For MVTec AD/VisA likelihood and deep-5 rows, I-AUROC and PRO include population SD across three model-seed dataset means. Single-layer likelihoods achieve higher pixel-level precision (P-AP) on MVTec AD and VisA, whereas their deep-5 likelihood fusion yields higher region-level coverage (P-AUPRO), highlighting the trade-off discussed in Sec. 5. The single-checkpoint density comparison in Table A11 is a separate result group. The view-conditioning row is exploratory and is not used to claim a localization benefit.

Dataset	Head / layers	I-AUROC	P-AUROC	P-AP	P-AUPRO
MVTec AD	LL / 1L	98.46 ±0.073	98.43 ±0.004	65.80 ±0.070	95.56 ±0.025
	TB / 1L	97.52 ±1.227	98.35 ±0.029	65.33 ±0.854	95.09 ±0.195
	LL / 5L	98.30 ±0.009	98.46 ±0.002	63.26 ±0.033	96.21 ±0.008
	TB / 5L	98.09 ±0.027	98.49 ±0.002	64.00 ±0.116	96.16 ±0.020
VisA	LL / 1L	95.56 ±0.045	99.25 ±0.002	46.29 ±0.031	93.24 ±0.028
	TB / 1L	95.16 ±0.102	99.28 ±0.004	47.26 ±0.039	93.17 ±0.022
	LL / 5L	96.72 ±0.010	99.21 ±0.000	44.43 ±0.054	95.19 ±0.004
	TB / 5L	95.75 ±0.055	99.22 ±0.001	43.98 ±0.101	95.23 ±0.008

Table A13: **Paired head × layer-fusion ablation.** Means and population SD across three model seeds; each seed covers all 15 MVTec AD or 12 VisA categories. The 1L condition is L31 from the same deep-5 model bank, with its train-radius normalization, rather than the independently trained single-layer model in Table 4. Both heads use 20 steps; LL uses five probes and TB eight trajectories with noise scale 0.1. The 5L conditions retain test-score z-standardization. Reconstructed deep-5 maps exactly match the original arrays, and new 1L PRO values pass the official reference. Historical scoring RNG states are unavailable; this compares saved score realizations across model seeds.

Method	MVTec AD				VisA			
	I-AUC	P-AUC	P-AP	P-AUPRO	I-AUC	P-AUC	P-AP	P-AUPRO
(a) Our reproductions: native backbones; evaluation provenance below								
PaDiM	89.6	96.5	40.8	—	81.0	96.0	21.5	—
CFLOW-AD	72.8	96.3	45.0	—	81.5	97.8	27.7	—
FastFlow	90.0	95.2	40.3	—	82.4	93.8	18.2	—
CS-Flow	97.4	84.5	17.7	—	90.6	87.9	9.2	—
U-Flow	97.5	98.0	53.4	84.6	95.2	99.0	40.5	83.5
(b) Our C-RADIO evaluation: common backbone, input size and metrics								
Gaussian	97.62	98.12	62.72	94.61	92.61	99.01	43.27	92.24
PaDiM-style	91.32	97.56	51.98	—	81.48	97.76	28.36	—
GMM	96.99	98.05	60.52	94.63	91.27	98.96	40.56	91.51
KDE	88.83	95.93	59.92	91.55	83.88	97.22	36.76	78.63
Coupling NF	98.90	94.27	44.22	79.13	95.76	99.00	40.23	90.41
PatchCore	**99.20**	98.29	65.02	**96.48**	**97.01**	98.84	42.87	92.73
FlowAD-Likelihood (1 layer)	98.48	98.44	**66.12**	95.63	95.57	99.25	46.43	93.24
FlowAD-TRAJBIF (1 layer)	98.55	98.40	65.99	—	95.26	**99.27**	**47.18**	—
FlowAD-Likelihood (deep-5)	98.30	98.46	63.26	96.21	96.72	99.21	44.43	95.19
FlowAD-TRAJBIF (deep-5)	98.09	**98.49**	64.00	96.16	95.75	99.22	43.98	**95.23**

Table A14: **Extended baseline and C-RADIO comparisons**, complementing Tables 4 and A11 (%; category means). Dashes denote unavailable PRO values under the reported evaluation. Native-backbone results retain their original reproduction protocols (Appendix D.1). Block (a) contains our native-backbone reproductions; (b) shares the frozen C-RADIO backbone, 448^2 inputs, and metric code, with layer selection and image aggregation and fusion varying across configurations (Appendix D.2). Table A11 compares fitted estimators on identical single-layer scoring features, using its own single-checkpoint FlowAD row rather than the three-seed production row here. **Bold** marks the best non-PRO value within (b), including configurations omitted from the main table; PRO bolding uses only the available re-evaluated entries. CS-Flow averages all 15 MVTec AD categories. Recent original-paper reports are in Table 4(a).

Method	I-AUC	P-AUC	P-AP	P-AUPRO
CFLOW-AD†	77.0	94.8	17.6	80.4
Gaussian	—	—	—	—
PaDiM-style	—	—	—	—
GMM	—	—	—	—
KDE	—	—	—	—
Coupling NF	—	—	—	—
PatchCore	**92.02**	**99.27**	**41.67**	96.16
FlowAD-Likelihood	89.62	99.14	40.29	**96.18**
FlowAD-TRAJBIF	88.26	99.11	40.80	96.02

Table A15: **Real-IAD under common-backbone evaluation** (%; mean over the 30 categories of the multi-view split). PatchCore and both deep-5 FlowAD rows use complete four-metric re-evaluations with official-reference-checked PRO. They share the backbone, input size, labels, masks and sample order; layer selection, fusion and image aggregation differ (Appendix D.2). Source-annotated positive images with empty original masks retain their original image labels and pixel masks. Bold marks the best re-evaluated value per column. PatchCore leads image AUROC and pixel AP; deep-5 likelihood is nearly tied with it in PRO (96.18 versus 96.16). Dashes denote unavailable full-category results under this evaluation. † quoted from the source, so the protocol differs.

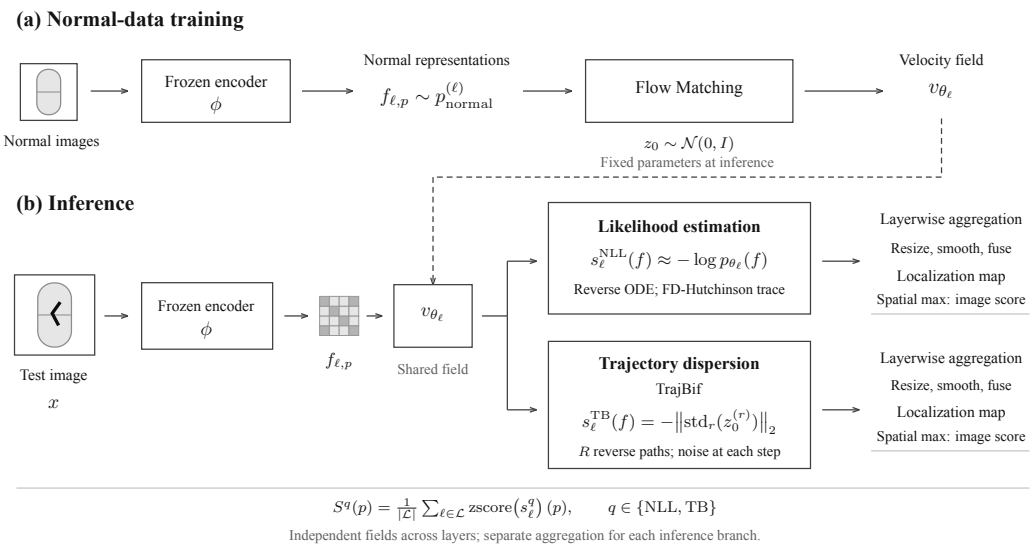

Figure A2: **FlowAD overview.** (a) Flow Matching learns one velocity field per layer and category from frozen representations of normal patches. (b) At inference, two functionals share each fitted field: likelihood estimation uses a reverse ODE and finite-difference Hutchinson probes of the velocity-field divergence; TRAJBIF measures endpoint dispersion across reverse trajectories with independent noise at each step. Each branch aggregates its own maps across layers. For either branch, interpolation and smoothing yield a localization map, whose spatial maximum is the image score. Layer fusion standardizes and averages score maps; single-layer density evaluation bypasses fusion. Image symbols are schematic.

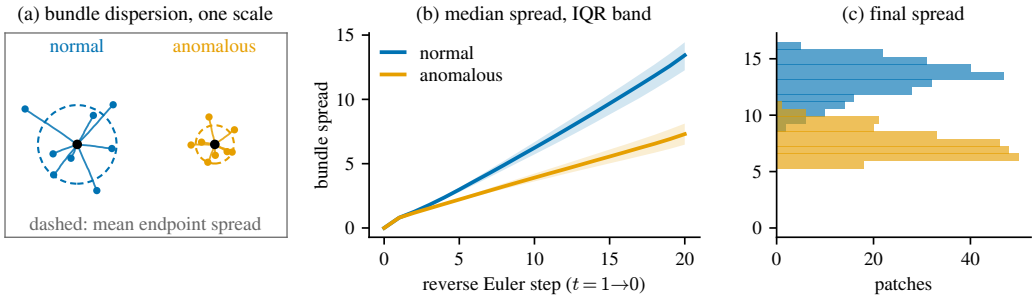

Figure A3: **TRAJBIF dispersion diagnostic** (`screw`, L31). **(a)** Each of the R=8 replicas as its deviation from the bundle centre, both bundles at one scale: the normal bundle disperses, the anomalous one stays tighter; the dashed circles are the endpoint spread the score reads. **(b)** That gap opens steadily along reverse time (median, IQR band). **(c)** The final-spread histograms of ∼250 normal and ∼250 anomalous patches barely overlap. This diagnostic covers the `screw` category at L31.

F TRAJBIF CONSTRUCTION AND DIAGNOSTIC

Figure A3 illustrates the sign convention used by the negative-spread score. Normal patches in this diagnostic have larger endpoint dispersion than anomalous patches. Benchmark comparisons of this score are in Tables 4 and A13.

Trajectory construction. Concretely, all R replicas start at the *same* unperturbed patch $z_1 = f$ and are integrated to t=0 by Euler–Maruyama with $\Delta t = -1/K_{\text{ODE}}$,

$$z^{(r)}_{t+\Delta t} = z^{(r)}_t + \Delta t\, v_\theta\big(z^{(r)}_t, t\big) + \sigma\sqrt{|\Delta t|}\,\xi^{(r)}_t,$$

with $\xi_t^{(r)} \sim \mathcal{N}(0, I)$ drawn fresh at every step and for every replica; the replicas differ only through ξ. With σ=0.1 and the same K_{ODE}=20 steps as the likelihood head, this costs $R\,K_{\mathrm{ODE}} = 160$ forward passes per patch against $K_{\mathrm{ODE}}(1{+}S) = 120$ for the trace. The score is the negated spread of the endpoint bundle, taken per coordinate across replicas and then reduced by Euclidean norm, so that larger means more anomalous:

$$s_{\textsc{TrajBif}}(f) \;=\; -\big\|\,\mathrm{std}_r\big(z_0^{(r)}\big)\,\big\|_2 \;=\; -\sqrt{\mathrm{tr}\,\mathrm{Cov}_r\big(z_0^{(r)}\big)}.$$

Linearized dispersion. Both heads read the same transport geometry, but they integrate different objects along the path. Linearizing the update rule above, the endpoint deviation accumulates every step's noise, $\delta z_0 \approx \sigma \sum_t \sqrt{|\Delta t|}\, J_{t\to 0}\, \xi_t$, where $J_{t\to 0}$ is the Jacobian of the *remaining* reverse flow, so

$$s^2_{\textsc{TrajBif}} \;\approx\; \sigma^2 \sum_t |\Delta t|\, \big\|J_{t\to 0}\big\|_F^2,$$

a sum of squared Frobenius norms of the remaining-flow Jacobians. This expression describes the linearized population covariance; finite replica counts add sampling variation. In contrast, the continuous-time divergence integral gives $\log|\det J_{\Phi^{-1}}|$, the log-volume change of the complete reverse flow. Thus TrajBif measures accumulated noise amplification, while the trace measures net log-volume change. In an affine field both have input-independent population statistics (Sec. 4).

G Additional Qualitative Results

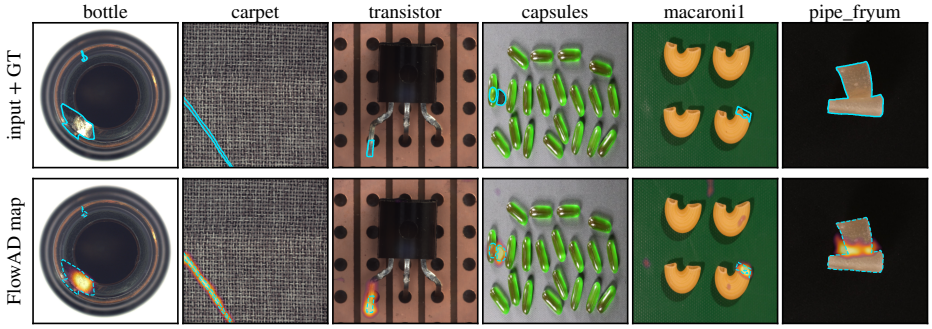

Figure A4: **Full qualitative gallery** across six object and texture categories of MVTec AD and VisA: input with ground truth outlined in cyan (top) and the FlowAD anomaly map (likelihood head, single layer; bottom) with ground truth dashed.

H SCOPE AND LIMITATIONS

Density-based anomaly scoring. FlowAD learns the density of normal patch features. Its usefulness for anomaly scoring depends on how anomalous features rank under that learned density. High-likelihood out-of-distribution inputs are a known limitation of density models (Nalisnick et al., 2019; Kirichenko et al., 2020; Schirrmeister et al., 2020); feature-space modeling and density normalization alone provide no guarantee of anomaly separation or shared thresholds across categories.

Spatial and view dependence. FlowAD models pooled patch marginals. An anomaly that preserves the marginal feature distribution while changing spatial arrangement or cross-view relations can be missed. The evaluated defect datasets leave logical anomalies and shifts in view composition as open settings for joint spatial and view-dependent models.

Fusion under deployment shifts. Multi-layer fusion uses unlabeled test-score statistics to standardize correlated layer outputs. The fused score is a detector output rather than a normalized joint density. Its robustness to changes in the test distribution remains untested; calibration on normal training data is an alternative for future evaluation.

Scoring cost and approximation. Reverse integration requires repeated network evaluations: the reported single-layer likelihood uses 120 forward passes per patch. ODE discretization, finite-difference bias and probe variance affect the numerical score. Appendix A specifies the estimator and the assumptions behind its bias expansion. The shared-probe JVP control in Appendix A.2 isolates the finite-difference contribution on three fixed models; Euler and Hutchinson errors remain in both branches. The DINOv3 and five-layer component controls each cover all VisA categories with one fitted model bank; variability across training runs remains unmeasured for these controls.