Back to Memory: Local Normal Support and Reference Routing for 3D Anomaly Detection

Anonymous authors
Paper under double-blind review

Abstract

Memory-bank methods for 3D anomaly detection must select appropriate normal references and distinguish defects from allowable local variation. Feature-only retrieval can match defects to similar geometry in incompatible regions, while nearest-neighbor distances do not explicitly describe the variation permitted by each normal pattern. We propose BTM, a training-free framework that separates normal memory construction, reference routing, and anomaly scoring. Manifold-Aware Memory builds local normal patterns from graph-geodesic neighborhoods of fixed FPFH features, storing local subspaces, finite tangential supports, and orthogonal thicknesses. Spatial-Context Gated Routing restricts candidates by registered position and selectively reranks low-similarity queries using multiscale context. Corrected Tangent-Normal Scoring measures only deviations beyond the selected patterns' support bounds. BTM requires only normal training point clouds, without learned encoders or synthetic anomalies. It achieves P-AUROC of $95.19 \pm 0.08\%$ on Real3D-AD and $94.03 \pm 0.06\%$ on the original 40-category Anomaly-ShapeNet benchmark, leading the reported point-level comparisons on both datasets. Object-level results rank first and second, respectively, among the compared methods. Module comparisons and mechanism analyses examine the contributions and limitations of reference routing and local-support correction.

1 Introduction

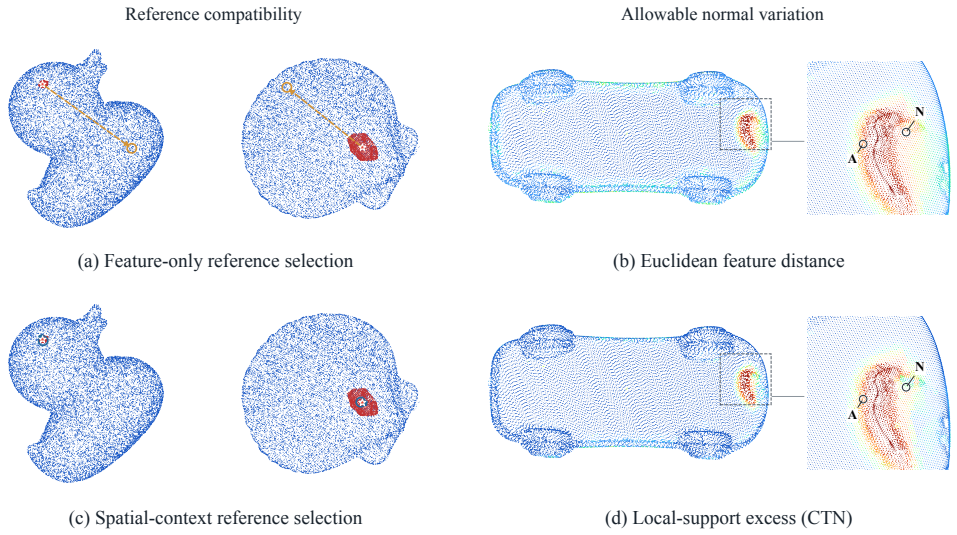
Industrial 3D anomaly detection identifies defects such as missing regions, indentations, and local deformations in point clouds and localizes the affected areas. Because defects are diverse and difficult to collect, we consider the setting where only a few normal scans are available for training (Liu et al., 2023a). The challenge is to establish reliable normal references and determine the variation they permit.

Memory-bank methods retrieve normal features to assess test points. BTF and Simple3D improve geometric representations (Horwitz & Hoshen, 2023; Cheng et al., 2026), Reg3D-AD and PointCore incorporate registration or spatial constraints (Liu et al., 2023a; Zhao et al., 2026), and the multimodal method G^2SF learns direction-adaptive local distances (Tao et al., 2025). Two issues motivate further separating reference selection from anomaly scoring.

First, feature similarity alone can select references from incompatible regions. Similar geometric structures, anomalies, and residual registration errors may produce a match that reduces the residual and masks a defect. Figure 1(a,c) shows selected Real3D-AD cases where global retrieval matches a duck's head defect to its body and a shell's defect to a distant region. Spatial-context routing instead selects nearby references. These cases motivate combining feature similarity with spatial and structural compatibility, without assuming that proximity guarantees a correct match.

Second, even a compatible reference permits direction-dependent normal variation. A nearest-reference distance alone does not express how much variation is allowed along or away from the local normal pattern. The same deviation magnitude can therefore reflect either normal variability or a defect. In Figure 1(b,d), normal-background point N and anomalous point A have nearly equal weighted feature distances, yet local-support correction reduces the response at N more strongly. References and weights are fixed across scoring rules, isolating this effect from reference selection. Together, these observations motivate going back to the memory itself: how it represents allowable normal variation and provides compatible references.

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.



(a) Feature-only reference selection (b) Euclidean feature distance

(c) Spatial-context reference selection (d) Local-support excess (CTN)

Figure 1: Two failure modes in selected Real3D-AD cases (seed 42). (a,c) Global feature retrieval selects distant references; spatial-context routing selects nearby ones. Blue: normal reference; red: registered query defects; stars/circles: query/top-1 reference. (b,d) CTN reduces normal-background responses relative to Euclidean scoring with top-3 references, weights, and color mapping fixed. N/A mark normal/anomalous points; warmer colors indicate higher scores.

We propose BTM, a training-free framework with three connected modules. Manifold-Aware Memory uses graph-geodesic neighborhoods of normal features to construct local subspaces with finite tangential support and orthogonal thickness. Spatial-Context Gated Routing selects spatial candidates after registration, with multiscale context as a selective refinement for low-similarity queries. The gate preserves high-similarity matches and focuses reranking on queries whose local feature match is less reliable. Corrected Tangent-Normal Scoring (CTN) then measures deviations beyond the selected patterns' support. The resulting memory represents local normal variation and supplies references for both point-level localization and object-level detection.

Our contributions are:

- We examine reference mismatch and unmodeled local normal variation as two failure modes of memory-based 3D anomaly detection, motivating compatible reference selection and local-support assessment.
- We introduce a unified training-free framework combining graph-geodesic normal patterns, spatial-context routing, and tangent–normal support correction.
- We report leading point-level results among the compared methods on Real3D-AD and Anomaly-ShapeNet, and examine module contributions through system comparisons, held-out representation analysis, and scoring diagnostics.

2 Related Work

2.1 3D Anomaly Detection

As in 2D anomaly detection, 3D methods broadly follow reconstruction and feature embedding approaches; extended background is provided in Appendix A.1. Reconstruction methods include IMRNet for masked geometry recovery, R3D-AD for diffusion-based correction, and PO3AD for point-wise offset prediction (Li et al., 2024; Zhou et al., 2024b; Ye et al., 2025). PASDF uses continuous SDF templates, while AARD explores geometry-aligned anomaly-aware reconstruction (Zheng et al., 2025; Wu et al., 2026).

Feature methods organize normal references for retrieval. BTF combines FPFH with nearest-neighbor memory (Horwitz & Hoshen, 2023; Rusu et al., 2009); ISMP models internal spatial structure, and Simple3D aggregates multiscale neighborhood features (Liang et al., 2025; Cheng et al., 2026). Reg3D-AD uses registration, and PointCore constrains matching by spatial proximity (Liu et al., 2023a; Zhao et al., 2026). Thus, spatially constrained retrieval is an established ingredient. In multimodal detection, G^{2}SF learns direction-adaptive local metrics for score fusion (Tao et al., 2025). BTM combines spatial-context reference selection with explicit finite-support models of normal feature variation; its score measures excess beyond the selected patterns' tangential extent and orthogonal thickness.

2.2 Local Geometry and Normal Support

Isomap approximates intrinsic proximity through shortest paths on a neighborhood graph (Tenenbaum et al., 2000), and LTSA estimates local tangent spaces to represent nonlinear manifolds (Zhang & Zha, 2004). Mixture PPCA uses multiple local linear models (Tipping & Bishop, 1999); COP measures outlierness in arbitrarily oriented local subspaces (Kriegel et al., 2012); and one-class methods estimate the support of a normal distribution (Schölkopf et al., 2001).

BTM builds on these neighborhood, subspace, and support-estimation ideas. Its mutual-neighbor graph organizes offline training neighborhoods, without embedding the full graph or using graph distance as the test anomaly score. Local PCA supplies directions, while anchor-centered statistics define finite support along and orthogonal to those directions. At inference, spatial-context routing selects references before CTN evaluates their support excess. This separates the geometry used to construct normal patterns from the criterion used to retrieve them and from the final anomaly score. Our evaluation therefore examines representation compactness together with held-out fidelity, as well as routing and scoring effects.

3 Method

3.1 Problem and Overview

For category c, we construct a scoring function from normal point clouds $D_c^{\text{tr}} = \{P_m\}_{m=1}^{M}$, with $P_m \subset \mathbb{R}^3$. An unlabeled test cloud P from the same category receives point scores $s(x)$ and an object score $S(P)$; higher values indicate stronger anomaly evidence. Figure 2 summarizes the pipeline: Manifold-Aware Memory stores local normal subspaces and supports, Spatial-Context Gated Routing selects references and weights, and CTN scores deviations beyond their supports. All three modules share a frozen normal memory and require no anomalous training samples, pseudo-anomalies, or learned encoders.

3.2 Manifold-Aware Memory

Normal Features and Anchors. Normal clouds are registered to a category reference, augmented with visible-surface views, and voxel-downsampled. Their fixed FPFH descriptors $\{\phi_j\}_{j=1}^{N} \subset \mathbb{R}^{33}$ are standardized using normal-training statistics and then ℓ_2-normalized. Test features $\phi(x)$ use the same frozen transformation. Greedy k-center sampling selects initial anchors $A = \{a_i\}$. Registration, augmentation, and sampling settings appear in Appendix A.2.

Graph-Geodesic Neighborhoods. Let $\mathcal{N}_{k_g}(j)$ contain the k_g nearest feature neighbors of ϕ_j. We build a mutual-kNN graph $G = (V, E)$ with Euclidean edge weights:

$$(i,j) \in E \iff i \in \mathcal{N}_{k_g}(j) \ \wedge \ j \in \mathcal{N}_{k_g}(i), \qquad w_{ij} = \|\phi_i - \phi_j\|_2. \tag{1}$$

For anchor a_i with training index ι_i in connected component V_i, its neighborhood contains the $q_i = \min(k_b, |V_i|)$ closest vertices by graph distance:

$$d_G(u,v) = \min_{\pi: u \rightsquigarrow v} \sum_{(p,q)\in\pi} w_{pq}, \qquad \mathcal{G}_i = \text{Top}_{q_i}^{\uparrow}\big(V_i; d_G(\iota_i, \cdot)\big). \tag{2}$$

Here, $\text{Top}_m^{\uparrow}$ selects the m smallest distances, and $d_G = +\infty$ between disconnected components. Truncated Dijkstra search includes the anchor and never crosses components. All features outside

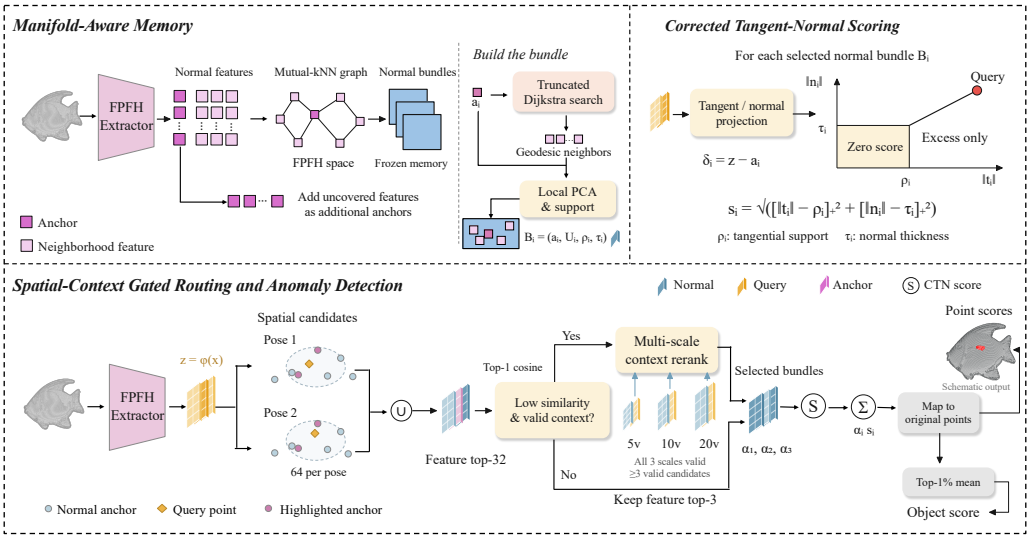


Figure 2: Overview of BTM. During memory construction, local normal patterns are built from normal point clouds. During inference, references are selected for test points, and residuals beyond their local support regions are used for point-level anomaly localization and object-level anomaly detection.

the union of the initial neighborhoods are appended as anchors in training-index order, with neighborhoods computed on the same graph. This covers the observed training features and can increase memory beyond the initial coreset size. We use $k_g = 16$ and $k_b = 64$.

Local Subspaces and Finite Support. PCA of each mean-centered neighborhood yields an orthonormal basis $U_i \in \mathbb{R}^{33\times r_i}$. The smallest rank explaining 95% of variance is constrained by numerical rank and bounds 2–16; covariance and validity details are in Appendix A.2. Directions are estimated about the neighborhood mean, but support is measured relative to the anchor. For any feature z, define

$$t_i(z) = U_i U_i^\top (z - a_i), \qquad n_i(z) = (I - U_i U_i^\top)(z - a_i). \tag{3}$$

The normal pattern $B_i = (a_i, U_i, \rho_i, \tau_i)$ stores finite tangential extent and orthogonal thickness:

$$\rho_i = Q^{\mathrm{NR}}_{0.95}(\{\|t_i(\phi_j)\|_2 : j \in \mathcal{G}_i\}), \qquad \tau_i = \frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i} \|n_i(\phi_j)\|_2, \tag{4}$$

where Q^{NR} is the nearest-rank empirical quantile. Each pattern also stores its anchor coordinates and multiscale context for routing.

3.3 Spatial-Context Gated Routing

Spatial Candidates and Similarity Gating. Test clouds are aligned to the same category reference. Let $x^{(h)}$ be a query's coordinates under retained pose h, and p_i the coordinates of anchor a_i. We merge spatial neighbors over up to $H = 2$ poses, then retain $L = 32$ feature candidates:

$$\begin{aligned} C_{\mathrm{pos}}(x) &= \bigcup_{h=1}^{H} \mathrm{Top}^{\uparrow}_{K_{\mathrm{pos}}}\big(\{1,\dots,|A|\}; \|x^{(h)} - p_i\|_2\big), \\ C(x) &= \mathrm{Top}^{\uparrow}_{L}\big(C_{\mathrm{pos}}(x); d_F(x,i)\big), \qquad d_F(x,i) = \|\phi(x) - a_i\|_2, \end{aligned} \tag{5}$$

with $K_{\mathrm{pos}} = 64$ per pose. The base route is feature top-3, $R_0(x) = \mathrm{Top}^{\uparrow}_3(C(x); d_F(x,\cdot))$. For unit-normalized features, the closest candidate's distance d_1 gives cosine similarity $\sigma(x) = \mathrm{clip}(1 - \frac{1}{2}d_1^2, -1, 1)$. A frozen category threshold θ_c gates context reranking. Low similarity flags

a potentially unreliable local match, which may arise from a defect or normal variation; it does not identify anomalous points. Context therefore serves as a selective refinement within the spatial candidate pool. Spatial filtering restricts locations but does not guarantee correct reference identity.

Context Reranking and Weights. Mean pooling and unit normalization of neighboring FPFH descriptors at radii $\{5v, 10v, 20v\}$, where v is voxel size, yield contexts $c_\ell(x)$ and $c_\ell(i)$. The fused distance is

$$d_{\text{route}}(x,i) = \tfrac{1}{2} d_F(x,i) + \tfrac{1}{6} \sum_{\ell=1}^{3} \|c_\ell(x) - c_\ell(i)\|_2. \tag{6}$$

Let $C_{\text{valid}}(x)$ contain candidates with valid context at all scales, and $v(x)$ indicate query context validity. We select

$$R(x) = \begin{cases} \text{Top}_3^{\uparrow}\big(C_{\text{valid}}(x); d_{\text{route}}(x,\cdot)\big), & \sigma(x) < \theta_c,\ v(x) = 1,\ |C_{\text{valid}}(x)| \geq 3, \\ R_0(x), & \text{otherwise.} \end{cases} \tag{7}$$

High-similarity queries and those with insufficient context retain the base route. Writing $d_i = d_F(x,i)$ for the base branch and $d_i = d_{\text{route}}(x,i)$ after reranking, normalized inverse-distance weights are

$$\alpha_i(x) = \frac{\max(d_i, \varepsilon)^{-1}}{\sum_{j \in R(x)} \max(d_j, \varepsilon)^{-1}}, \qquad \varepsilon = 10^{-6}. \tag{8}$$

Context validity and neighbor limits are specified in Appendix A.2.

3.4 Corrected Tangent-Normal Scoring

The selected pattern permits variation within its tangential support and orthogonal thickness, defining the continuous region

$$\Omega_i = \{z \in \mathbb{R}^{33} : \|t_i(z)\|_2 \leq \rho_i,\ \|n_i(z)\|_2 \leq \tau_i\}. \tag{9}$$

Using the decomposition in Eq. 3, CTN scores a query $z = \phi(x)$ only through excess beyond these bounds:

$$s_i^{\text{CTN}}(x) = \sqrt{\max\{\|t_i(z)\|_2 - \rho_i, 0\}^2 + \max\{\|n_i(z)\|_2 - \tau_i, 0\}^2}. \tag{10}$$

For an orthonormal U_i, this equals the Euclidean distance from z to Ω_i; features within both bounds score zero. Invalid patterns use the Euclidean fallback $s_i^{\text{CTN}}(x) = \|z - a_i\|_2$. The downsampled point score is

$$s(x) = \sum_{i \in R(x)} \alpha_i(x) s_i^{\text{CTN}}(x). \tag{11}$$

Each original point receives its nearest downsampled point's score. On both datasets, $S(P)$ is the mean of the $\max(1, \lceil 0.01|P| \rceil)$ highest original point scores (top-1% mean). CTN corrects local support and thickness; it is not probabilistic calibration, and training-derived bounds do not guarantee out-of-sample coverage.

4 Experiments

4.1 Experimental Setup

Data and Metrics. We evaluate all 12 categories of Real3D-AD (Liu et al., 2023a) and the original 40 categories of Anomaly-ShapeNet (Li et al., 2024), excluding its 10-category extension. Both benchmarks use only normal training point clouds. P-AUROC is primary; O-AUROC uses top-1% mean point-score aggregation. Each run first averages AUROC equally over categories, followed by the mean $\pm$ sample standard deviation across runs. Real3D-AD P-AUROC uses seeds $\{0, 1, 2, 3, 42\}$; its O-AUROC is computed from saved point scores for seeds $\{0, 1, 42\}$. Anomaly-ShapeNet uses $\{0, 1, 42\}$ for both metrics. Data alignment and seed coverage appear in Appendices A.2 and A.8.

Baselines. We compare BTM with feature-memory, reconstruction, and self-supervised methods, including the geometric-descriptor baseline Simple3D (Cheng et al., 2026). Results follow published benchmark-specific configurations; feature backbones and training procedures need not match. Appendix A.1 describes each method, and Appendix A.3 documents result sources.

4.2 Main Results

Table 1 compares BTM with published results from papers and official supplements, including benchmark-author evaluations. These are literature reports, not reruns under one common protocol; source details are in Appendix A.3.

Table 1: Main results on all 12 Real3D-AD categories and the original 40 Anomaly-ShapeNet categories (AUROC, %). Baselines are literature reports; BTM reports mean ± sample SD. Bold marks column bests; – denotes an unavailable comparable result.

Method	Real3D-AD		Anomaly-ShapeNet	
	O-AUROC	P-AUROC	O-AUROC	P-AUROC
BTF (Raw)	63.5	57.1	49.3	55.0
BTF (FPFH)	60.3	73.3	52.8	62.8
PatchCore (FPFH)	59.3	68.2	56.8	58.0
PatchCore (PointMAE)	59.4	62.0	56.2	57.7
M3DM	55.2	63.1	55.2	61.6
CPMF	–	–	55.9	57.3
Reg3D-AD	70.4	70.5	57.2	66.8
IMRNet	72.5	–	66.1	65.0
ISMP	76.7	83.6	75.7	69.1
R3D-AD	73.4	59.2	74.9	–
PO3AD	76.5	–	83.9	89.8
PASDF	80.2	74.5	–	–
Simple3D	80.4	92.3	86.0	92.9
CASL	–	–	**88.7**	89.9
AARD	82.0	86.0	–	93.6
BTM (Ours)	**87.70±0.33**	**95.19±0.08**	88.33±0.19	**94.03±0.06**

Results. BTM leads the listed Real3D-AD baselines on both metrics, exceeding Simple3D's P-AUROC by 2.89 percentage points. It also leads point-level comparisons on Anomaly-ShapeNet, while its O-AUROC is below CASL. Thus, localization gains are consistent across these benchmarks, whereas object-level gains vary.

4.3 Ablation Studies

Table 2 compares Full BTM with three alternatives. Here M1 denotes Manifold-Aware Memory, M2 denotes Corrected Tangent-Normal Scoring, and M3 denotes Spatial-Context Gated Routing.

Full − M1: Euclidean Local Memory. We replace graph-geodesic memory construction with Euclidean local PCA patterns (Euclidean full-bundle memory). Neighborhoods are selected by direct Euclidean distance in FPFH feature space instead of shortest paths on the mutual-kNN graph. CTN scoring and spatial-context gated routing remain unchanged; only normal memory construction is replaced.

Full − M2: Raw Normal Scoring. We remove CTN's finite-support correction and use the raw normal residual $\|n_i\|_2$, without subtracting orthogonal thickness or adding tangential support excess. The graph-geodesic memory, selected references, and routing weights remain those of Full BTM; invalid patterns keep the Euclidean fallback.

Full − M3: Global Feature Routing. We remove spatial candidate constraints and gated context reranking, replacing them with global feature top-3 retrieval over all memory anchors. References are ranked and inversely weighted by FPFH feature distance, while the graph-geodesic memory and CTN scoring remain fixed.

For Full − M2 and Full − M3, registration, point mapping, and top-1% object aggregation stay those of Full BTM. All three ablated configurations (Full − M1, Full − M2, and Full − M3) use the same seeds $\{0, 1, 42\}$, with three runs per configuration on each benchmark.

Table 2: Three-module system comparison (AUROC, %; mean ± sample SD over available runs). Bold marks column bests.

Configuration	Real3D-AD		Anomaly-ShapeNet	
	O-AUROC ↑	P-AUROC ↑	O-AUROC ↑	P-AUROC ↑
Full − M1	81.78±0.89	90.25±0.05	**89.47±0.81**	90.54±0.15
Full − M2	87.27±0.19	94.66±0.09	87.41±0.48	91.38±0.11
Full − M3	79.90±0.29	91.22±0.03	88.06±0.42	91.79±0.07
Full BTM	**87.70±0.33**	**95.19±0.08**	88.33±0.19	**94.03±0.06**

Full BTM has higher P-AUROC than all three alternatives on both benchmarks. On Real3D-AD, its O-AUROC gain is 7.80 points over Full − M3 and approximately 0.42 points over Full − M2. On Anomaly-ShapeNet, Full − M1 has higher O-AUROC, so improved localization does not imply improved object discrimination.

Category-Level Statistics. After averaging runs within categories, paired category bootstrap and sign-flip tests assess effects with equal benchmark weights. Joint P-AUROC gains are $+4.217$, $+1.591$, and $+3.106$ points for M1–M3; all joint 95% intervals exclude zero and all three comparisons survive Holm correction (Appendix Table 7). However, M2's Real3D-AD interval includes zero; its joint gain is primarily driven by Anomaly-ShapeNet.

4.4 Mechanism Analysis

Local Normal Structure. We use leave-one-template-out evaluation over four normal sources per category on all 12 Real3D-AD categories. Compactness is measured by r_{95}, the rank needed to explain 95% of variance; fidelity is the held-out normal residual at matched ranks $\{2, 4, 8, 12, 16\}$. The two local methods share anchors and routing, and the K-means comparison matches the number of local models. As shown in Figure 3, the graph-geodesic local representation has a macro-average r_{95} of 6.97, a 48.1% reduction from 13.44 for Global PCA, and achieves lower held-out residuals at all five matched ranks. Compared with Euclidean Local PCA ($r_{95} = 7.50$), it provides similar held-out fidelity with a more compact representation, indicating that graph-geodesic neighborhoods describe normal variation with fewer local principal directions. Detailed settings are provided in Appendix A.5.1.

Reference Routing. With memory and CTN fixed, the seed-42 Real3D-AD comparison improves P-AUROC from 91.25% under global feature top-3 to 95.13% under full routing. Spatial feature top-3 without context reaches 95.16%: spatial candidate constraints retain the main gain. Context selectively refines low-similarity queries, adding a small O-AUROC gain (87.26% to 87.45%) without a P-AUROC gain. Across three shared gating thresholds, O-AUROC remains 0.28–0.31 points above the no-context control (Appendix A.6). This supports a modest supplementary role under the tested thresholds; these same-seed results do not establish stability across seeds. Fixed-pool source-identity diagnostics provide additional evidence of small reranking gains (Appendix A.5.2). Identity misses alone do not establish structural incompatibility.

Figure 4 illustrates reference selection and anomaly responses on three real-defect examples; visualization details and selection limitations are given in Appendix A.5.2.

Selective Effects of Finite-Support Corrections. We compare raw normal, corrected normal, and CTN scores with frozen memory and routing on Real3D-AD (12 categories; seeds $\{0, 1, 42\}$). Raw

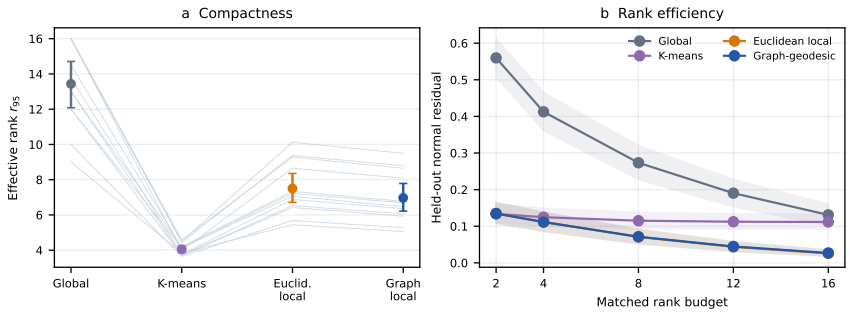


Figure 3: Local normal structure on Real3D-AD. (a) Category-level r_{95} values and the macro-average with its 95% CI. (b) Held-out residuals at matched ranks, averaged equally over categories, with 95% CIs. Rank and residual jointly characterize representation quality.

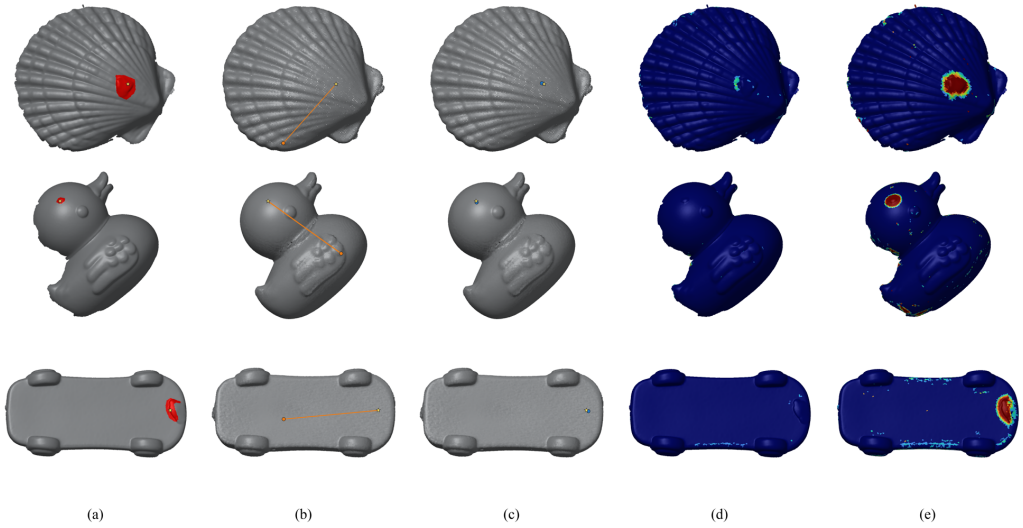


Figure 4: Routing examples with CTN scoring fixed: Shell, Duck, and Car (rows). (a) Real defect annotations in red. (b,c) Global feature-L2 and BTM top-1 reference selections for the same starred query point. (d,e) Corresponding weighted top-3 CTN heatmaps, with a shared color scale within each row.

normal scores aggregate $\|n_i\|_2$ with the routing weights; corrected normal scores instead aggregate $[\|n_i\|_2 - \tau_i]_+$, with the same Euclidean fallback for invalid patterns. Two-fold 95th-percentile thresholds use only normal scans from the opposite scan-family fold. Normal suppression measures alarms removed by normal correction at the fixed raw-normal threshold. Tangent recovery measures previously raw-unflagged points for which adding the tangent term crosses the CTN threshold while corrected normal alone does not. Table 3 shows selective suppression of existing normal alarms and selective recovery of previously unflagged anomalies: the differences are positive in all 12 categories and survive Holm correction across four tests. Net sensitivity change is not significant. This retrospective diagnostic uses normal test scans for threshold estimation and does not establish prospective calibration; full endpoint definitions, tests, and attribution are in Appendix A.5.3.

Table 3: Conditional event rates (%) and selectivity differences (points): normal minus anomaly for suppression; anomaly minus normal for recovery. Category-bootstrap 95% CIs and Holm-adjusted p-values across four tests; conditioning and full results are in Appendix A.5.3.

Comparison	Normal	Anomaly	Δ	95% CI	p_{Holm}
Normal suppression	39.534	7.626	$+31.908$	$[+19.502, +44.988]$	0.001953
Tangent recovery	1.045	13.446	$+12.401$	$[+8.611, +16.171]$	0.001953

4.5 Sensitivity and Robustness

Registration Residuals. Spatial routing tolerates small alignment residuals but loses advantage under larger rotations; both metrics remain above global retrieval throughout the tested range (Table 4). Only retrieval coordinates are perturbed, so this does not establish end-to-end scan robustness.

Table 4: Registration-residual sensitivity on Real3D-AD (seed 42; AUROC, %). Perturbations report mean $\pm$ sample SD across three directions or axes; v is voxel size, and memory and CTN scoring are fixed.

Setting	P-AUROC	O-AUROC
Default BTM	95.13	87.45
Global feature top-3	91.25	79.94
Translation $0.5v$	95.13 ± 0.01	87.69 ± 0.09
Translation $1v$	95.10 ± 0.01	87.66 ± 0.16
Translation $2v$	94.85 ± 0.02	87.14 ± 0.36
Rotation 1°	95.12 ± 0.01	87.63 ± 0.03
Rotation 3°	94.89 ± 0.02	87.24 ± 0.09
Rotation 5°	93.80 ± 0.02	84.47 ± 0.46

Neighborhood Scale. Detection is relatively stable around the default scale (Table 5), but coverage completion changes memory size, precluding a fixed-budget comparison. Both studies use seed-42 Real3D-AD; full protocols and gating results appear in Appendix A.6.

Table 5: Neighborhood-scale sensitivity on Real3D-AD (seed 42; AUROC and rates in %). Counts and rates are category averages; fallback rates first average over objects within each category.

$(K_{\text{graph}}, K_{\text{local}})$	P-AUROC	O-AUROC	Anchors (k)	Invalid (%)	Fallback (%)
$(8, 64)$	94.93	87.37	88.41	7.68	13.94
$(16, 32)$	95.28	87.47	95.13	4.55	4.71
$(16, 64)$ (default)	95.13	87.45	86.11	4.84	5.16
$(16, 128)$	94.87	87.65	77.95	5.13	6.02
$(32, 64)$	95.17	87.67	85.05	3.99	3.15

5 Conclusion

BTM separates normal pattern construction, reference routing, and support-based scoring for training-free 3D anomaly detection. It achieves $95.19 \pm 0.08\%$ and $94.03 \pm 0.06\%$ P-AUROC on Real3D-AD and Anomaly-ShapeNet, leading the reported point-level comparisons. Module comparisons and mechanism analyses support the roles of compatible references and local normal variation, with spatial constraints accounting for most observed routing gains. Performance remains dependent on registration and normal-data coverage; single-seed robustness analyses limit generalization claims. These results motivate further controlled evaluation of local support and reference selection.

AI Use Statement

AI assisted translation, language polishing, and coding, with personal revision of the text and code by the submitting author and full responsibility retained by the authors (Appendix B.1).

Ethics Statement

This work uses public or appropriately licensed point-cloud data for industrial inspection, whose deployment requires application-specific validation, human verification, and responsible data handling (Appendix B.2).

Reproducibility Statement

We document the method, implementation, datasets, metrics, seed coverage, and analysis protocols in the main text and appendices, with a reproducibility guide in Appendix B.3.

References

Yunkang Cao, Xiaohao Xu, and Weiming Shen. Complementary pseudo multimodal feature for point cloud anomaly detection. *Pattern Recognition*, 156:110761, 2024. doi: 10.1016/j.patcog.2024.110761.

Jiayi Cheng, Can Gao, Jie Zhou, Jiajun Wen, Tao Dai, and Jinbao Wang. MC3D-AD: A unified geometry-aware reconstruction model for multi-category 3D anomaly detection. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 837–845, 2025. doi: 10.24963/ijcai.2025/94.

Yuqi Cheng, Yihan Sun, Hui Zhang, Weiming Shen, and Yunkang Cao. Towards high-resolution 3D anomaly detection: A scalable dataset and real-time framework for subtle industrial defects. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(5):3327–3334, 2026. doi: 10.1609/aaai.v40i5.37328.

Yu-Min Chu, Chieh Liu, Ting-I Hsieh, Hwann-Tzong Chen, and Tyng-Luh Liu. Shape-guided dual-memory learning for 3D anomaly detection. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6185–6194, 2023.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: A patch distribution modeling framework for anomaly detection and localization. In *Pattern Recognition. ICPR International Workshops and Challenges*, volume 12664 of *Lecture Notes in Computer Science*, pp. 475–489. Springer, 2021. doi: 10.1007/978-3-030-68799-1_35.

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9737–9746, 2022.

Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. CFLOW-AD: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 98–107, 2022.

Eliahu Horwitz and Yedid Hoshen. Back to the feature: Classical 3D features are (almost) all you need for 3D anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2968–2977, 2023.

Xi Jiang, Jianlin Liu, Jinbao Wang, Qiang Nie, Kai Wu, Yong Liu, Chengjie Wang, and Feng Zheng. SoftPatch: Unsupervised anomaly detection with noisy data. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15433–15445, 2022. doi: 10.52202/068431-1123.

Xueyang Kang, Zizhao Li, Tian Lan, Dong Gong, Kourosh Khoshelham, and Liangliang Nan. Hierarchical point-patch fusion with adaptive patch codebook for 3D shape anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. Outlier detection in arbitrarily oriented subspaces. In *Proceedings of the 12th IEEE International Conference on Data Mining*, pp. 379–388, 2012. doi: 10.1109/ICDM.2012.21.

Wenqiao Li, Xiaohao Xu, Yao Gu, Bozhong Zheng, Shenghua Gao, and Yingna Wu. Towards scalable 3D anomaly detection and localization: A benchmark via 3D anomaly synthesis and a self-supervised learning network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22207–22216, 2024.

Hanzhe Liang, Guoyang Xie, Chengbin Hou, Bingshu Wang, Can Gao, and Jinbao Wang. Look inside for more: Internal spatial modality perception for 3D anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(5):5146–5154, 2025. doi: 10.1609/aaai.v39i5.32546.

Jiaqi Liu, Guoyang Xie, Ruitao Chen, Xinpeng Li, Jinbao Wang, Yong Liu, Chengjie Wang, and Feng Zheng. Real3D-AD: A dataset of point cloud anomaly detection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 30402–30415, 2023a. doi: 10.52202/075280-1324.

Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. SimpleNet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20402–20411, 2023b.

Wenxi Lv, Qinliang Su, and Wenchao Xu. One-for-All few-shot anomaly detection via instance-induced prompt learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2022.

Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (FPFH) for 3D registration. In *IEEE International Conference on Robotics and Automation*, pp. 3212–3217, 2009. doi: 10.1109/ROBOT.2009.5152473.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001. doi: 10.1162/089976601750264965.

Chengyu Tao, Xuanming Cao, and Juan Du. G2SF: Geometry-guided score fusion for multimodal industrial anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20551–20560, 2025.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. doi: 10.1126/science.290.5500.2319.

Michael E. Tipping and Christopher M. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, 11(2):443–482, 1999. doi: 10.1162/089976699300016728.

Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8032–8041, 2023.

Linchun Wu, Qin Zou, Yuanhao Yue, and Zhongyuan Wang. Geometry-aligned and anomaly-aware reconstruction for 3D anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14648–14657, 2026.

Guoyang Xie, Jinbao Wang, Jiaqi Liu, Yaochu Jin, and Feng Zheng. Pushing the limits of fewshot anomaly detection in industry vision: GraphCore. In *The Eleventh International Conference on Learning Representations*, 2023.

Chaoran Xu, Chengkan Lv, Qiyu Chen, Feng Zhang, and Zhengtao Zhang. MRAD: Zero-shot anomaly detection with memory-driven retrieval. In *The Fourteenth International Conference on Learning Representations*, 2026.

Jianan Ye, Weiguang Zhao, Xi Yang, Guangliang Cheng, and Kaizhu Huang. PO3AD: Predicting point offsets toward better 3D point cloud anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1353–1362, 2025.

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*, volume 35, 2022. doi: 10.52202/068431-0330.

Yuyang Yu, Zhengwei Chen, Xuemiao Xu, Lei Zhang, Haoxin Yang, Yongwei Nie, and Shengfeng He. Registration is a powerful rotation-invariance learner for 3D anomaly detection. In *Advances in Neural Information Processing Systems*, volume 38, pp. 5073–5099, 2025. doi: 10.52202/085713-0180.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRAEM: A discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8330–8339, 2021.

Yaohua Zha, Xue Yuerong, Chunlin Fan, Yuansong Wang, Tao Dai, Ke Chen, and Shu-Tao Xia. CASL: Curvature-augmented self-supervised learning for 3D anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(15):12340–12348, 2026. doi: 10.1609/aaai.v40i15.38226.

Zhenyue Zhang and Hongyuan Zha. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. *SIAM Journal on Scientific Computing*, 26(1):313–338, 2004. doi: 10.1137/S1064827502419154.

Baozhu Zhao, Xiaohan Zhang, Jingfeng Guo, and Qi Liu. PointCore: An efficient framework for unsupervised point cloud anomaly detection using joint local-global features. *Neural Networks*, 197:108446, 2026. doi: 10.1016/j.neunet.2025.108446.

Bozhong Zheng, Jinye Gan, Xiaohao Xu, Xintao Chen, Wenqiao Li, Xiaonan Huang, Na Ni, and Yingna Wu. Bridging 3D anomaly localization and repair via high-quality continuous geometric representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27063–27072, 2025.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations*, 2024a.

Zheyuan Zhou, Le Wang, Naiyu Fang, Zili Wang, Lemiao Qiu, and Shuyou Zhang. R3D-AD: Reconstruction via diffusion for 3D anomaly detection. In *Computer Vision – ECCV 2024*, pp. 91–107, 2024b.

A Supplementary Experiments and Reproducibility Details

This appendix provides additional related work, data and implementation details, evaluation and statistical protocols, detection results, and mechanism analyses of normal memory, reference routing, and CTN scoring. It also describes parameter sensitivity, method limitations, and limitations of the experimental sources.

A.1 Extended Related Work

2D Anomaly Detection. Anomaly detection in 2D images has advanced rapidly, particularly under normal-only target-domain training. Existing methods fall broadly into feature embedding and reconstruction paradigms. Feature embedding approaches construct a normal representation space and detect departures from it through knowledge distillation (Deng & Li, 2022), one-class discriminative learning (Liu et al., 2023b), probabilistic distribution modeling (Defard et al., 2021; Gudovskiy et al., 2022), memory-bank retrieval (Roth et al., 2022; Jiang et al., 2022), or vision–language alignment (Zhou et al., 2024a). Reconstruction approaches instead learn to recover normal images or features and identify anomalies from reconstruction residuals (Zavrtanik et al., 2021; You et al., 2022). These mechanisms have also been extended to limited-reference and cross-category generalization settings, including few-shot and zero-shot detection (Xie et al., 2023; Lv et al., 2025; Xu et al., 2026). When moving from structured images to 3D point clouds, however, normality modeling must further account for unordered samples, irregular density, spatial correspondence, and pose variation, motivating dedicated 3D anomaly detection methods.

Additional 3D Representations. Beyond the methods discussed in the main text, reconstruction work also extends to multi-category detection (Cheng et al., 2025). Feature-based work explores multimodal or dual memories (Wang et al., 2023; Cao et al., 2024; Chu et al., 2023), rotation-invariant representations (Yu et al., 2025), and hierarchical point-patch representations with adaptive codebooks (Kang et al., 2026).

Baseline Methods. We include feature-memory, reconstruction, and self-supervised baselines to compare BTM with alternative representations and anomaly-scoring strategies. BTF retrieves normal references using raw coordinates or handcrafted FPFH features (Horwitz & Hoshen, 2023); PatchCore uses a coreset memory with FPFH or learned PointMAE features in the reported 3D adaptations (Roth et al., 2022; Liu et al., 2023a). M3DM, originally a multimodal fusion method, is evaluated through its point-cloud features on these benchmarks (Wang et al., 2023; Liu et al., 2023a; Li et al., 2024). CPMF combines 3D descriptors with features from rendered 2D views of the same point cloud (Cao et al., 2024), while ISMP adds internal spatial representations (Liang et al., 2025). Reg3D-AD aligns test clouds with normal prototypes (Liu et al., 2023a); Simple3D combines multiscale geometric descriptors with local spatial aggregation (Cheng et al., 2026), providing a particularly relevant feature-based comparison.

IMRNet reconstructs masked geometry (Li et al., 2024); R3D-AD uses diffusion reconstruction (Zhou et al., 2024b); and PO3AD predicts point offsets that correct simulated defects (Ye et al., 2025). PASDF represents normal geometry with a continuous signed distance field (Zheng et al., 2025), while AARD uses geometry-aligned diffusion and anomaly-aware references for reconstruction (Wu et al., 2026). CASL learns point-cloud representations through curvature-guided reconstruction and subsequently fine-tunes for anomaly classification (Zha et al., 2026). Together, these comparisons place BTM's explicit local-support scoring alongside reference retrieval, geometry recovery, and learned representations. Results follow the published benchmark-specific configurations; we do not assume identical feature backbones or training procedures.

A.2 Implementation Details

Dataset Background. We evaluate our method on two pure point-cloud benchmarks. Real3D-AD comprises 1,254 objects across 12 categories, with each object containing between approximately 40,000 and several million points; only a small number of normal templates are provided during the training phase (Liu et al., 2023a). The original version of Anomaly-ShapeNet consists of 1,600 samples across 40 categories, with each sample containing approximately 8,000–30,000 points, covering six types of anomalies: protrusions, indentations, holes, fractures, bends, and cracks (Li et al., 2024).

In this paper, we utilize all 12 categories from the former and the original 40 categories from the latter, excluding the subsequent 10-category extension.

Data Loading and Alignment. The Real3D-AD and Anomaly-ShapeNet adapters use the label conventions `good`/lowercase `gt` and `positive`/uppercase `GT`, respectively. Training and normal test points are read from PCD files. Anomalous test points and their labels are read jointly from the same XYZ+label ground-truth table to preserve pointwise alignment. Evaluation covers all target categories of both benchmarks, with each anomalous test sample paired with its pointwise ground truth.

Training and Inference Details. The first normal template in sorted order serves as the reference for each category. Normal templates undergo registration and visible-surface augmentation before FPFH extraction; the standardizer is fitted only to normal training descriptors. Training features not covered by the initial graph-geodesic neighborhoods are appended as anchors. Local PCA and support statistics are then fitted, and feature, coordinate, and context indices are stored. Inference follows Sections 3.3–3.4, using CTN as the primary anomaly score. Scoring-mechanism comparisons compute different score components under the same routing.

Registration and Augmentation Details. The first normal template in sorted order is the category reference. Registration uses FPFH-RANSAC for coarse alignment and point-to-plane ICP for refinement, retaining at most two pose hypotheses. Deterministic lay-flat visible-surface views augment the normal training geometry.

Local Covariance, Bundle Validity, and Fallback. For neighborhood $\mathcal{G}_i$ of size q_i, the mean-centered covariance used to estimate the PCA directions is

$$\bar{\phi}_i = \frac{1}{q_i} \sum_{j \in \mathcal{G}_i} \phi_j, \qquad \Sigma_i = \frac{1}{\max(q_i - 1, 1)} \sum_{j \in \mathcal{G}_i} (\phi_j - \bar{\phi}_i)(\phi_j - \bar{\phi}_i)^\top. \tag{12}$$

Support is nevertheless measured relative to anchor a_i, using the projections in Eq. 3. Local PCA uses float64 covariance and the minimum rank explaining 95% of variance, subject to the rank bounds in Table 6 and the numerical effective rank. Patterns with too few neighbors, insufficient rank, nonfinite values, or excessive orthogonality error are marked invalid. Their test scores fall back to Euclidean query-to-anchor feature distance. For valid orthonormal bases, tangent and normal components satisfy the Pythagorean identity up to numerical error. This geometric property provides no probabilistic calibration guarantee.

Routing Implementation Details. For each retained pose, we retrieve the 64 spatially nearest anchors, merge the results, and retain 32 candidates by core FPFH distance. Both base and reranked routes select three references. Context radii are $\{5v, 10v, 20v\}$, with neighbor limits $\{128, 256, 512\}$, where v is the voxel size; the minimum neighbor count is listed in Table 6. Reranking requires valid contexts at every scale for both query and candidate. If the query context is invalid or too few candidates are valid, the base route and its feature-distance weights are retained. Successful reranking uses inverse fused-distance weights.

Computational Complexity. For N normal training features, the mutual graph stores $O(Nk_g)$ edges. A truncated Dijkstra search is performed for each final anchor, and local eigendecomposition operates on a fixed 33-dimensional covariance matrix. Inference cost is dominated by registration, FPFH extraction, spatial candidate retrieval, and context computation for low-similarity queries. The initial coreset ratio differs from the final memory ratio after coverage completion.

Table 6: Main parameters of the reported BTM runs. v denotes voxel size and ASN denotes Anomaly-ShapeNet. Category-specific similarity thresholds are stored in the corresponding run configurations.

Module	Parameter	Value
FPFH	Voxel size (Real3D / ASN)	0.265 / 0.01
FPFH	Normal / feature radius	$2.333v$ / $5.25v$
FPFH	Max. normal / feature neighbors	40 / 140
Initial memory	Coreset ratio / size cap	0.20 / 100,000
Graph	Mutual neighbors	16
Local PCA	Neighbors / explained variance	64 / 0.95
Local PCA	Min.–max. rank / precision	2–16 / float64
Routing	Spatial candidates per pose	64
Routing	Local pool / selected bundles	32 / 3
Context	Radius	$5v, 10v, 20v$
Context	Max. / min. neighbors	128,256,512 / 3
Fusion	Core / per-scale context weight	$1/2$ / $1/6$
Score	Tangential support / normal thickness	nearest-rank $Q_{.95}$ / mean
Random seeds	Seed set (usage specified in the text)	0, 1, 2, 3, 42

A.3 Evaluation Protocols and Baseline Sources

Statistical Methods and Units of Analysis. System comparisons use the same test objects and retain preprocessing steps that are not part of the intervention. For the three-module comparisons, each configuration is first averaged over its available runs within each category, using the seed coverage in Appendix A.8. Inference pairs these category means, with 10,000 paired category bootstrap resamples. Joint effects weight the two benchmark macro-effects equally. Two-sided stratified Monte Carlo sign-flip tests use 2,000,000 draws, with Holm correction across the three primary comparisons. Categories are the units of statistical inference. The seed-42 routing diagnostics are descriptive analyses outside this primary hypothesis family.

Sources of Baseline Results. The baselines in Table 1 are reported values from published papers and their official supplements. We distinguish results from the original method authors from benchmark evaluations in other papers. Literature reports do not imply that all methods were run by their original authors under our protocol.

For Real3D-AD, object-level BTF, PatchCore, and M3DM results come from Table 4 of the original Real3D-AD paper; point-level results come from the published evaluation in ISMP Table 1(b) (Liu et al., 2023a; Liang et al., 2025). Some of the latter values differ from the early benchmark tables and agree with AARD Table 2. Reg3D-AD uses its original Table 4 object-level result and the point-level value 70.5 from ISMP Table 1(b), compared with 70.0 in the original Real3D-AD Table 6. Object-level IMRNet, R3D-AD, and PO3AD values come from Tables 2, 1, and 3 of their respective papers. R3D-AD's point-level value 59.2 is for the complete model D in its Table 3 (Li et al., 2024; Zhou et al., 2024b; Ye et al., 2025). PASDF, Simple3D, and AARD results come from their original Table 1, Table 3, and Tables 1–2, respectively (Zheng et al., 2025; Cheng et al., 2026; Wu et al., 2026).

For Anomaly-ShapeNet, BTF, PatchCore, M3DM, CPMF, Reg3D-AD, and IMRNet use the 40-category results from Table 3 and official supplementary Table S1 of the benchmark paper (Li et al., 2024); the earlier methods were evaluated by the benchmark authors. R3D-AD's object-level result comes from its official supplementary Table 3; PO3AD uses its original Tables 1–2; Simple3D and CASL use their original Tables 3 and 2, respectively (Zhou et al., 2024b; Ye et al., 2025; Cheng et al., 2026; Zha et al., 2026). AARD's point-level value 93.6 comes from its official supplementary Table 1. Its main-paper Table 3 specifies 15 categories. We could not confirm that its object-level result uses the same 40-category macro-averaging protocol, so we exclude the object-level value 93.1. A dash indicates that no verifiable result with an applicable protocol is included, not that the method cannot produce the metric.

Discrepancies in Reported ISMP Results. For Real3D-AD, we use O-AUROC 76.7 reported in ISMP Table 1(a) and its text. However, the arithmetic mean of the 12 listed categories is approximately 75.7, matching the value reproduced in AARD. We do not interpret this internal reporting discrepancy as an experimental improvement. P-AUROC 83.6 agrees with the original paper. Anomaly-ShapeNet P-AUROC 69.1 is confirmed by ISMP Table 2, but that paper does not provide the corresponding 40-category object-level table. The object-level value 75.7 is taken from Simple3D Table 3 (Cheng et al., 2026); it is a secondary report, not a value verified in the original ISMP paper (Liang et al., 2025; Cheng et al., 2026).

A.4 Additional Results

Complete Category-Level Statistics. Table 7 provides the complete category-level statistics summarized in Section 4.3. Differences are the main experiment's Full BTM minus each alternative, in P-AUROC percentage points. I/E/D counts improved, equal, and degraded categories. Interval and testing procedures are given in Appendix A.3.

Table 7: Complete category-level statistics for the three modules, grouped by dataset. Effects pair category means after averaging each configuration's available runs. Joint effects weight the two benchmarks equally. The Real3D-AD interval for M2 includes zero.

Module	ΔP-AUROC	95% CI	I/E/D
Real3D-AD			
M1	$+4.935$	$[+2.251, +7.956]$	11/0/1
M2	$+0.530$	$[-0.158, +1.330]$	7/0/5
M3	$+3.971$	$[+1.820, +6.264]$	10/0/2
Anomaly-ShapeNet			
M1	$+3.498$	$[+1.771, +5.338]$	33/0/7
M2	$+2.651$	$[+0.566, +4.726]$	30/0/10
M3	$+2.241$	$[+0.356, +4.037]$	29/0/11
Joint effects			p_{Holm}
M1	$+4.217$	$[+2.607, +5.956]$	$< 1.50 \times 10^{-6}$
M2	$+1.591$	$[+0.469, +2.697]$	6.99×10^{-3}
M3	$+3.106$	$[+1.707, +4.520]$	4.16×10^{-4}

Results Across Random Seeds. Table 8 reports per-seed macro-averages on Real3D-AD. Seed coverage is detailed in Appendix A.8.

Table 8: Main results by seed on Real3D-AD (12-category macro-average, AUROC, %). Object-level results use seeds $\{0, 1, 42\}$, for which complete point scores are available; point-level results use all five seeds. A dash indicates that point scores needed to compute O-AUROC are unavailable. Each final-row entry summarizes the available runs in its column.

Seed	O-AUROC ↑	P-AUROC ↑
0	88.07	95.22
1	87.57	95.09
2	–	95.18
3	–	95.30
42	87.45	95.13
Mean ± SD	87.70 ± 0.33	95.19 ± 0.08

A.5 Mechanism Analyses

A.5.1 Local Normal Structure

Evaluation Protocol for Local Normal Structure. We examine local normal structure through representation compactness and held-out fidelity. Figure 3 compares Global PCA, K-means PCA with a matched model count, Euclidean Local PCA, and Graph-geodesic Local PCA. Compactness is measured by r_{95} and fidelity by held-out normal residuals at matched ranks $R \in \{2, 4, 8, 12, 16\}$. The two local methods share anchors and routing.

Global covariance includes both within-pattern variation $\mathbb{E}_k[\Sigma_k]$ and between-pattern variation $\mathrm{Cov}_k(\mu_k)$, whereas local PCA fits only within-neighborhood variation. We perform leave-one-template-out evaluation over four normal physical sources per category, aggregate folds first, and then compute macro-averages and bootstrap 95% CIs across 12 categories. For Chicken, when the strict registration threshold rejects every candidate, we use the best finite rigid candidate. The other categories use the fixed registration protocol.

Although K-means PCA achieves the lowest representation rank, its held-out residual remains relatively high, so compactness must be interpreted alongside fidelity. Graph-geodesic and Euclidean Local PCA have broadly similar held-out residuals, with differences depending on the matched rank.

A.5.2 Reference Routing

Real-Defect Reference-Selection Cases. Figure 4 compares global feature-L2 routing with full BTM routing on three Real3D-AD test samples: Shell (558_bulge), Duck (684_bulge), and Car (651_bulge_cut). Query features, registration, normal memory, and CTN scoring are held fixed. Column (a) shows the benchmark's real defect annotations; columns (b,c) display the top-1 reference selected by each method for the same marked anomalous query point. Columns (d,e) show CTN scores aggregated over the respective top-3 references with their routing weights. The displayed top-1 reference therefore illustrates one component of each weighted score, rather than the entire scoring computation.

Visualization Protocol. Each row uses the same camera, projection, lighting, and color mapping for both methods. Let U be the 99th percentile of the pooled original-point scores from both methods for that sample, and let $T = 0.25U$. Scores below T use a deep-blue base color; scores from T to U follow a blue-to-red scale, and scores above U saturate. These are display settings computed without anomaly labels, not detection thresholds; they do not change scores or AUROC. All original points are rendered, without score smoothing or interpolation.

Interpretation and Scope. In these selected examples, BTM retrieves spatially closer references and produces clearer responses over the annotated defects. The examples were selected retrospectively and do not estimate how often this behavior occurs across the benchmark. They compare complete routing procedures and do not isolate spatial filtering, gating, or context reranking. Real defect labels do not supply independent reference-identity ground truth, and a distant reference alone does not establish structural incompatibility. The separate identity experiment below uses known source correspondences and a fixed candidate pool.

Independent Routing Ground Truth. Reference-identity evaluation starts from normal templates with known point identities. We apply known rigid transformations, optionally adding coordinate noise, point dropout, or visibility occlusion. The inverse known transformation maps each retained query ID to its source ID. An identity hit requires the reference anchor's provenance to contain that exact point identity. Ground truth is not inferred from coordinate distance for other templates or anchors lacking identity records; an identity miss does not imply a semantic or structural error. Identity matches are determined before FPFH and context computation, independently of retrieval results and anomaly scores. We report top-1 and top-3 hit rates and mean reciprocal rank, together with category-level improvements and whether any high-similarity queries are reranked.

Reranking with a Fixed Candidate Pool. With the spatial candidate pool fixed, context reranking provides a small additional gain in source-identity recovery (Table 9). This diagnostic measures

the context contribution within a fixed pool, whereas M3 in Section 4.3 compares detection with complete routing procedures.

In Table 9, each query's spatial candidate pool remains fixed. Reference-row identities are stored before feature extraction; ground truth is independent of anomaly labels, final routes, and anomaly scores. Queries whose ground truth is absent from the pool count as misses in end-to-end metrics. The seed-42 identity evaluation covers 19 conditions: clean queries; $1/3/5^\circ$ rotations; $0.5/1/2$-voxel translations; Gaussian coordinate noise of $0.05/0.10/0.20$ voxels; $25/50/75\%$ point dropout; and $75/50/25\%$ visibility along $+z/-z$. Conditions are weighted equally within categories, then all 12 categories are weighted equally. This analysis selects the best finite registration pose without quality-threshold filtering. Its intervention point and metrics differ from the post-registration residual experiment in Section 4.5.

Table 9: Routing reranking with a fixed spatial candidate pool (Real3D-AD, seed 42). Metrics are averaged equally over conditions and categories and reported on their original proportion scale. Intervals use 10,000 paired category bootstrap resamples.

Metric	Base ranking	Reranked	Paired difference	95% CI
Top-1	0.28042	0.28122	$+0.00079$	$[0.00039, 0.00125]$
Top-3	0.36161	0.36243	$+0.00081$	$[0.00013, 0.00150]$
MRR	0.31566	0.31647	$+0.00081$	$[0.00029, 0.00135]$

Top-1 improves in 10/12 categories; Top-3 and MRR each improve in 9/12 categories. No high-similarity queries are reranked. Paired differences use unrounded statistics, not subtraction of the rounded values shown here.

Negative Controls for Routing. Random-reference, farthest-feature-reference, and spatial-only controls share BTM's finite-support memory and CTN scoring. Only the reference-selection rule changes, testing its influence on detection.

Table 10 reports equally weighted macro-averages over all 12 Real3D-AD categories for seed 42. BTM has higher point-level AUROC than registered-space-only routing N3 in every category. This control shows that routing affects detection scores, but provides neither independent routing ground truth nor uncertainty across seeds.

Table 10: Routing negative controls on Real3D-AD (AUROC, %): a single seed-42 run.

ID	Routing	O-AUROC ↑	P-AUROC ↑
N1	Random bundle	71.89	68.39
N2	Farthest-feature bundle	47.19	53.57
N3	Registered-space top-3 only	79.79	91.59

Routing Diagnostics with Real Labels. This seed-42 Real3D-AD diagnostic compares feature top-3 and safe-gated routing with the same normal memory, registration, FPFH, and CTN scoring. It covers 1,206 test samples across all 12 categories; the two strategies select different top-1 references at 66.48% of original-resolution points. Normal test scans in each category are sorted by a fixed hash. The first half determines a separate $Q_{.95}$ threshold for each strategy; the remaining normal scans and all anomalous scans are used to compute FP/FN. This measures detection changes under real labels. Reference-identity recovery is evaluated independently using the protocol in Appendix A.5.2.

Table 11: Supplementary routing diagnostics with real labels on Real3D-AD, seed 42. O/P-AUROC uses all test samples and the main-experiment reference values; object scores use the top-1% mean; FP/FN are macro-averaged category error rates (%) after an independent split of normal scans. This table does not measure independent reference-routing accuracy.

Routing	O-AUROC ↑	P-AUROC ↑	FP ↓	FN ↓
Global feature top-3	79.94	91.25	6.63	35.83
BTM routing	87.45	95.13	6.59	23.73

A.5.3 CTN Corrections and Finite Support

Cross-Fitted Analysis of CTN Corrections. This analysis covers all 12 Real3D-AD categories and seeds $\{0, 1, 42\}$, totaling 3,618 sets of test-scan scores. Scoring rules share a fixed model, registration, memory, support bounds, routes, and weights. Normal controls are points from normal scans; anomaly-labeled points and normal background in anomalous scans are summarized separately. Voxel mappings retain each voxel's positive and negative original-point counts rather than labeling mixed voxels entirely anomalous.

Family splits and calibration. The first numeric identifier in each filename defines a scan family within a category; complete, cut, and copy scans remain in the same fold. Families are sorted by a fixed SHA-256 hash of category, anomaly-presence indicator, and family identifier, then assigned alternately to two folds shared by all three seeds. The shell/551 family contains both good and bulge scans and remains together, with every scan retaining its original labels. Each fold uses only normal scans from the opposite fold to estimate separate weighted $Q_{.95}$ thresholds for s_N, s_C, s_{CTN}, using the inverse empirical CDF without interpolation. An alarm requires a score strictly above its threshold. No evaluation family contributes to its calibration set, and anomalous scans are never used to estimate thresholds. Calibration uses normal scans from the benchmark test set, making this a retrospective analysis rather than independently collected calibration validation or an end-to-end source-held-out evaluation.

Aggregation and inference. Original points of the relevant label are weighted equally within a scan; scans containing that label are weighted equally within a family; families are also weighted equally. Conditional rates divide event mass by conditioning mass within each category and seed. The three seeds are averaged within category, then categories receive equal weight. Consequently, a ratio of two macro-averaged population proportions need not equal the reported macro-averaged conditional rate. We use two-sided exact sign-flip tests over 12 categories (2^{12} sign assignments) and 10,000 paired category bootstrap resamples. Seeds are not treated as independent categories. The four endpoints in Table 12 form one Holm-correction family; only two appear in the main text. Marginal 95% CIs condition on the observed calibration data and folds and exclude uncertainty from collecting new calibration data. Distinct filename families do not guarantee independent physical objects. Tests assume that category-difference signs are exchangeable under the null.

Endpoint definitions. Let s_N, s_C, s_{CTN} denote aggregated raw-normal, corrected-normal, and CTN scores, and let $\theta_N, \theta_{\text{CTN}}$ be the raw and CTN thresholds. Normal suppression compares normal minus anomaly probabilities of $s_C \le \theta_N$ conditional on $s_N > \theta_N$. This fixed-threshold comparison isolates score reduction and is not an improvement in the recalibrated false-alarm rate. Tangent recovery compares anomaly minus normal probabilities of $s_C \le \theta_{\text{CTN}} < s_{\text{CTN}}$ conditional on $s_N \le \theta_N$, requiring the tangent increment to cause threshold crossing. Net sensitivity compares anomaly sensitivity under separately calibrated CTN and raw thresholds. Tangent contribution compares conditional means of $(s_{\text{CTN}} - s_C)/\theta_N$ among raw-unflagged points, with anomaly minus normal as the direction. It is related to tangent recovery, not an independent replication. These four mechanism tests form a hypothesis family separate from the three primary module comparisons.

Table 12: Complete CTN mechanism tests. Effects and intervals in the first three rows are in percentage points; the last row is in units of the raw threshold. Counts are positive/equal/negative categories. All four tests are retained and jointly Holm-corrected.

Endpoint	Effect	95% CI	p_{Holm}	+/0/−
Normal suppression	$+31.908$	$[+19.502, +44.988]$	0.001953	12/0/0
Tangent recovery	$+12.401$	$[+8.611, +16.171]$	0.001953	12/0/0
Net sensitivity	$+1.483$	$[-0.497, +3.965]$	0.357910	6/0/6
Tangent contribution	$+0.120$	$[+0.083, +0.163]$	0.001953	12/0/0

Table 13: Category-level CTN mechanism differences averaged over three seeds. The first three columns are in percentage points; the last is in units of the raw threshold. Columns report normal-suppression selectivity, tangent-recovery selectivity, net sensitivity, and tangent-contribution differences.

Category	Suppression	Recovery	Net sensitivity	Contribution
Airplane	+41.509	+14.945	+1.902	+0.151
Candybar	+56.479	+8.505	-0.884	+0.038
Car	+12.860	+10.667	-2.511	+0.077
Chicken	+16.443	+15.439	-0.435	+0.165
Diamond	+23.753	+20.819	-0.542	+0.284
Duck	+1.579	+1.875	-1.740	+0.046
Fish	+50.925	+17.323	+0.805	+0.122
Gemstone	+6.750	+3.382	+1.111	+0.082
Seahorse	+68.946	+19.142	+9.404	+0.190
Shell	+21.245	+22.591	+10.838	+0.175
Starfish	+19.272	+4.602	+0.034	+0.052
Toffees	+63.133	+9.526	-0.182	+0.060

Table 14: Cross-fitted alarm rates and event transitions (%), each normalized by all points of the corresponding label group. The first three columns use separately normal-calibrated thresholds. New alarm denotes raw-unflagged points flagged by CTN; Lost alarm is the reverse. Tangent crossing is the subset of new alarms caused by the tangent increment. New alarms are false positives for normal points and recovered misses for anomalous points.

Point group	Raw	Corrected	CTN	New alarm	Lost alarm	Tangent crossing
Normal scans	5.317	5.317	5.335	1.128	1.110	0.990
Anomaly labels	74.782	76.042	76.266	3.363	1.880	2.559
Anomalous-scan background	7.699	7.919	7.932	1.450	1.218	1.212

Support attribution and scope. Each valid route is assigned to one of four regions according to $\|t_i\| \le \rho_i$ and $\|n_i\| \le \tau_i$; Euclidean fallbacks are recorded separately. The normal reduction is $\sum_i \alpha_i(\|n_i\| - s_{C,i})$ and the tangent increment is $\sum_i \alpha_i(s_{\text{CTN},i} - s_{C,i})$. Both are zero for fallback routes. Contributions partitioned by support region sum to the corresponding aggregate score difference. To align scales, contributions are divided by the raw threshold for their category, seed, and fold, aggregated within label groups using the family weights above, and macro-averaged over categories and seeds. Shares in Figure 5 are ratios of these contribution masses, not means of category-level shares. Neither a global constant subtraction nor a shuffled-τ control is included, so general score shrinkage remains an alternative explanation for normal suppression. Contributions from thickness,

registration error, and feature noise are not separately identified. This component analysis covers all 12 Real3D-AD categories.

From support regions to score contributions. Figure 5 decomposes normal reductions and tangent increments using the actual routing weights. Among suppressed normal false alarms, 76.231% of the normal reduction comes from routes outside only the normal bound. Among recovered anomalies, 99.971% of the tangent increment comes from routes outside both bounds. Thus, the observed mechanism primarily subtracts local thickness offsets and strengthens normal responses that have not yet crossed the detection threshold, rather than mainly zeroing strictly in-support points or recovering tangent-only violations. Support boundaries differ from detection thresholds; region occupancy is neither a direct anomaly decision nor an out-of-sample coverage guarantee.

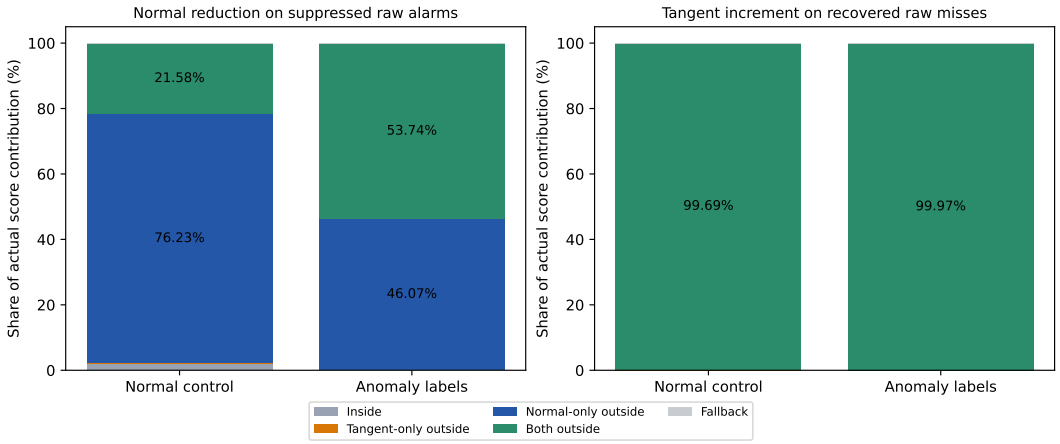


Figure 5: Support regions and realized score contributions. Left: each region's share of normal reduction among raw alarms suppressed by normal correction. Right: each region's share of tangent increments among raw-unflagged points whose threshold crossing is caused by that increment. Normal columns describe false alarms; anomaly columns describe lost detections or recovered misses. Shares are ratios of macro-averaged normalized contribution masses, not point counts. Fallback contributions are separate. Zero in-support tangent increments follow from the formula and are not independent statistical findings.

After accounting for newly missed anomalies, CTN's net anomaly-sensitivity gain over raw scoring is not statistically significant. Complete tests, category-level results, and alarm rates appear in Tables 12, 13, and 14, respectively.

A.6 Sensitivity, Robustness, and Context Reranking

This section supplements the registration-residual and graph-neighborhood protocols for Tables 4 and 5 in Section 4.5, and reports gating-threshold sensitivity and context-reranking controls.

Shared Evaluation Protocol. Experiments here and in Section 4.5 use all 12 Real3D-AD categories and model seed 42. AUROC is computed per category and then macro-averaged with equal category weights. P-AUROC uses original-resolution points; object scores average the largest $k = \max(1, \lceil 0.01N \rceil)$ scores for an object with N original points. Default BTM and global feature top-3 reference values follow the seed-42 main experiment; intervention rows report their respective evaluations. Registration-perturbation standard deviations reflect three deterministic directions or axes, not model-seed uncertainty. These sensitivity analyses are limited to Real3D-AD and the tested conditions and do not quantify uncertainty across model seeds.

Tolerance to Registration Residuals. We hold query features, context, normal memory, and mappings to original points fixed, perturbing only the registered coordinates used for spatial candidate retrieval. Translations and rotations are applied separately, with the same rigid transformation applied to all retained poses of each object. Rotations are centered at the centroid of the downsampled

points in the primary pose; translation magnitudes are expressed in units of the voxel size $v = 0.265$. The global feature top-3 baseline uses the same memory and CTN scoring and is unaffected by coordinate perturbations. Table 4 shows stable performance under small residuals. Larger rotations reduce the gains, but both point- and object-level AUROC remain above the global baseline across all tested perturbations, indicating that spatial constraints remain beneficial within this range. The error range at which this advantage disappears remains undetermined. These results concern routing tolerance to registration residuals and do not establish end-to-end robustness to scan perturbations.

Neighborhood-Scale Stability. We hold normal training features, the feature standardizer, and the initial coreset fixed and test five combinations of graph connectivity and local geodesic neighborhood size. For each setting, we rebuild graph neighborhoods, perform coverage completion, and recompute local PCA, support parameters, and anchor-dependent routing state. Table 5 reports detection performance, memory size, and invalid-pattern fallback statistics. Both detection metrics remain relatively stable around the default neighborhood scale. Coverage completion changes the final anchor count, while model validity and fallback rates also vary with these parameters. Thus, this comparison does not control final memory size and does not replace a controlled ablation of graph-geodesic versus Euclidean neighborhoods.

Gating and Context Experimental Setup. Normal memory, poses, query features, contexts, and spatial candidate pools are frozen; only category-specific gating thresholds are replaced by a shared $\theta \in \{0.925, 0.950, 0.975\}$. An additional control directly uses the base feature top-3 within the spatial pool, explicitly disabling context reranking rather than approximating it with an extreme threshold. Default BTM and global feature top-3 provide references. All conditions use the same memory and CTN scoring; complete results appear in Table 15. The no-context control retains spatial filtering, FPFH ranking within the pool, and feature-distance weights. It differs from N3 in Table 10, which selects and weights references directly by spatial distance.

Table 15: Gating thresholds and context-reranking controls on Real3D-AD (seed 42; AUROC, %). Conditions share normal memory, query features, and CTN scoring; gating variants and the no-context control also share spatial candidate pools. Results are 12-category macro-averages, with O-AUROC based on the top-1% mean of original point scores.

Setting	P-AUROC	O-AUROC
Default BTM	95.13	87.45
Global feature top-3	91.25	79.94
Shared $\theta = 0.925$	95.13	87.57
Shared $\theta = 0.950$	95.12	87.57
Shared $\theta = 0.975$	95.10	87.54
Spatial feature top-3 (no context)	95.16	87.26

Shared-Threshold Stability. Point- and object-level metrics under shared thresholds remain close to the default setting. Reranking rates increase from 15.62% to 40.03%, while O-AUROC remains 87.54–87.57%, exceeding the no-context control's 87.26% by 0.28–0.31 percentage points. This supports stability of a small object-level gain under the tested shared thresholds. These settings share the same seed and test samples and are not independent replications.

Contribution of Context Reranking. Context reranking selectively supplements low-similarity queries; high-similarity queries retain the base route. Gating uses feature-match similarity rather than anomaly labels. Selected queries may represent anomalies, normal variation, or unreliable matches and cannot be equated with anomalous points. Small aggregate metric changes are consistent with reranking only some queries, but the reranking rate alone neither establishes effectiveness nor explains every gain or loss. Top-1% object aggregation and point-level AUROC respond differently to score changes.

The observed evidence includes small object-level numerical gains over the no-context control for both default and shared-threshold routing. In the fixed-pool identity diagnostic, top-1 improves in 10 of 12 categories, with a positive 95% CI for the macro-average difference (Table 9). The former

supports a supplementary object-level benefit under the tested thresholds; the latter supports limited identity-recovery improvements. These are different objectives. P-AUROC does not improve; spatial candidate constraints account for most gains over global routing. Current evidence therefore supports a limited supplementary role for reranking. Selective reranking has not been directly compared with reranking every query.

A.7 Scope and Limitations

Reference selection can remain ambiguous when registration is inaccurate and the spatial candidate pool contains repeated local geometry. For connected components smaller than the target neighborhood size, bundle neighborhoods are limited by component size. Sparse boundaries can also destabilize normals and FPFH, while visible-surface augmentation does not guarantee coverage of all inspection views. Coverage completion adds uncovered training points as anchors; actual coreset storage savings must therefore be assessed using the completed memory.

A.8 Seed Coverage and Analysis Settings

Main Results and Seed Coverage. The Full BTM entries in both the main-results and system-comparison tables use seeds $\{0, 1, 2, 3, 42\}$ for Real3D-AD P-AUROC and $\{0, 1, 42\}$ for its O-AUROC. Full BTM uses $\{0, 1, 42\}$ for both metrics on Anomaly-ShapeNet. All three ablated configurations (Full $-$ M1, Full $-$ M2, and Full $-$ M3) use $\{0, 1, 42\}$ for both metrics on each benchmark. The category-level statistics in Table 7 use these same available runs for each configuration; pairing is across categories after averaging runs. Configurations share test samples, with each Real3D-AD run covering 12 categories and 1,206 test objects. Object scores use the top-1% mean in Section 3.4, followed by equally weighted category AUROC. Real3D-AD per-seed results are listed in Table 8.

Routing Diagnostics. Tables 10 and 11 use seed-42 Real3D-AD results, with metrics computed within categories and then equally macro-averaged. They are separate from the main multi-run system comparison. The real-label diagnostic covers 12 categories and 1,206 test samples with matched registration and query features. Real anomaly labels do not provide independent reference-identity ground truth.

CTN Analysis Settings. The CTN mechanism analysis compares scoring rules on the same test scans across three seeds, fixing support bounds, routes, and weights and aggregating through the original-point mapping. Scan-family folds are determined by category, family anomaly-presence indicator, and family identifier. Scans from a family remain together and retain their original labels. The four mechanism tests use Holm correction as detailed in Appendix A.5.3. This is a retrospective analysis, not a preregistered study.

B Detailed Statements

B.1 AI Use and Author Responsibility

We used AI assistants to translate and polish the manuscript and to assist with code writing, code checks, test development, and consistency checks between the method description and its implementation. The submitting author personally edited and revised the AI-assisted text and code, rather than merely reviewing or accepting the generated output. The authors take responsibility for the final manuscript and code, including the accuracy of references, equations, experimental results, and claims.

B.2 Ethical Considerations

This work studies industrial geometric inspection using public or appropriately licensed point-cloud data and does not require identity, biometric, or demographic attributes. Missed defects may allow faulty parts to pass inspection, while false alarms may cause unnecessary waste. AUROC therefore does not constitute a deployment guarantee. Deployment requires application-specific thresholds, scanner-domain-shift testing, human verification, and traceable failure handling. Data use must respect dataset licenses and the confidentiality of private scans.

B.3 Reproducibility Guide

Section 3 defines normal memory construction, reference routing, and CTN scoring; Section 4 describes datasets, metrics, and statistical methods. Appendix A.2 details data processing, registration, neighborhoods, and principal hyperparameters. Appendix A.8 specifies seed coverage and analysis settings. Data splits, threshold estimation, and statistical units for mechanism analyses are described in Appendix A.5.