Back to Point: Exploring Point-Language Models for Zero-Shot 3D Anomaly Detection

Supplementary Material

1. Implementation Details

Overall Architecture. Our framework is implemented upon the public **ULIP-2 PointBERT** backbone, which is pre-trained on large-scale 3D-language datasets such as Objaverse and ShapeNet with multi-modal supervision involving point clouds, images, and text, providing strong 3D-language alignment priors. We keep the PointBERT backbone as the 3D encoder and introduce two task-specific modules: the **Geometric Feature Creation Module (GFCM)** and the **Multi-Granularity Feature Embedding Module (MGFEM)**, which jointly enhance local geometric reasoning and cross-modal semantic fusion for zero-shot 3D anomaly detection.

Geometric Feature Creation Module (GFCM). The GFCM extracts learnable patch-level geometric descriptors to complement the backbone's semantic features. Given the original point cloud $\mathbf{P} \in \mathbb{R}^{B\times N\times 3}$ and local patch indices $\mathcal{I} \in \mathbb{Z}^{B\times G\times M}$, the module first gathers neighboring points per patch, forming $\mathbb{R}^{B\times G\times M\times 3}$ tensors. A PointNet-style encoder with three 1×1 convolutions (3→64→128→256) followed by BatchNorm and ReLU learns per-point geometry, after which a *max-pooling* operation aggregates features within each patch. Two fully connected layers (256→128→33) project the result to a 33-dimensional geometric descriptor, aligned with the FPFH dimension. These geometric embeddings are later used in a contrastive loss $\mathcal{L}_{geo}$ to align learned features with handcrafted geometry priors.

Multi-Granularity Feature Embedding Module (MGFEM). MFEM integrates three complementary feature sources: (i) multi-level semantic features from PointBERT, (ii) patch-level geometric embeddings (from FPFH or GFCM), and (iii) a global feature token $\mathbf{g}$. Each branch is first linearly mapped to a common dimension $D_{out} = 1280$ via two-layer MLPs (384→512→1280 for PointBERT and global, 33→512→1280 for geometry). For multi-layer PointBERT features, a learnable weight vector $\mathbf{w} \in \mathbb{R}^L$ (where $L = 4$ for layers $\{2, 5, 8, 11\}$) is normalized via $\mathrm{softmax}$ and used to compute a weighted sum:

$$\mathbf{F}_{\mathrm{P}} = \sum_{\ell=1}^{L} \alpha_\ell \, \phi_{\mathrm{P}}(\mathbf{F}^{(\ell)}), \quad \alpha_\ell = \mathrm{softmax}(w_\ell).$$

The fused semantic, geometric, and global features are concatenated and passed through a fusion projection MLP ($3D_{out} \to D_{out} \to D_{out}$), producing the final patch embeddings $\mathbf{E} \in \mathbb{R}^{B\times N\times D_{out}}$. These embeddings are aligned with textual representations to compute patch-level similarity maps for anomaly localization.

Parameter Grouping and Optimization. We adopt the **AdamW** optimizer (β_1=0.9, β_2=0.999) with a three-way parameter grouping strategy:

- **Prompt parameters** (`prompt_learner`) use a higher learning rate $\eta_{\text{prompt}} = \eta$ without weight decay.
- **Decay group:** all other linear/MLP/projection weights excluding `norm`, `bias`, `embed`, and `prompt` parameters use $\eta_{\text{proj}} = 0.5\eta$ with weight decay 1×10^{-4}.
- **No-decay group:** parameters containing `bias`, `norm`, `ln`, `embed`, or `prompt` use the same η_{proj} with zero weight decay.

This grouping stabilizes prompt adaptation while regularizing backbone weights. Global gradient clipping is applied at each iteration to prevent exploding gradients.

Learning Rate Schedule. The learning rate scheduler combines a 10% **linear warm-up** phase (start factor 0.1) followed by **cosine annealing** with a minimum learning rate of 1×10^{-6}:

$$\eta_t = \begin{cases} 0.1\eta + \frac{0.9\eta}{T_w} t, & t < T_w, \\ \frac{1}{2}\eta(1 + \cos(\pi(t - T_w)/(T - T_w))), & t \geq T_w. \end{cases}$$

Here, T and T_w denote total and warm-up iterations, respectively. This schedule ensures smooth convergence and stabilizes later-stage training.

Training Configuration. We train for 100 epochs with a batch size of 2. Default hyperparameters are: `num_points=2048`, `learning_rate=0.001`, `depth=9`, `n_ctx=8`, `t_n_ctx=4`. All experiments are conducted on a single NVIDIA RTX 4090 GPU. For reproducibility, we set the base random seed to 111 and repeat each run 5 times (`seed_base+run_idx`), averaging results over 10 total runs.

2. Additional Quantitative Results

Per-class results on Anomaly-ShapeNet. To provide a more comprehensive view of model behavior across differ-

Table 1. Comparison of AUROC results at the object and point levels (%) of various methods on Anomaly-ShapeNet.

Method	ashtray0	bag0	bottle0	bottle1	bottle3	bowl0	bowl1	bowl2	bowl3	bowl4	bowl5	bucket0	bucket1	cap0
BTF	57.8 / 51.2	41.0 / 43.0	59.7 / 55.1	51.0 / 49.1	56.8 / 72.0	56.4 / 52.4	26.4 / 46.4	52.5 / 42.6	38.5 / 68.5	66.4 / 56.3	41.7 / 51.7	61.7 / 61.7	32.1 / 68.6	66.8 / 52.4
M3DM	57.7 / 57.7	53.7 / 63.7	57.4 / 66.3	63.7 / 63.7	54.1 / 53.2	63.4 / 65.8	66.3 / 66.3	68.4 / 69.4	61.7 / 65.7	46.4 / 62.4	40.9 / 48.9	30.9 / 69.8	50.1 / 69.9	55.7 / 53.1
PatchCore$_{\text{FPFH}}$	58.7 / 59.7	57.1 / 57.4	60.4 / 65.4	66.7 / 68.7	57.2 / 51.2	50.4 / 52.4	63.9 / 53.1	61.5 / 62.5	53.7 / 32.7	49.4 / 72.0	55.8 / 35.8	46.9 / 45.9	55.1 / 57.1	58.0 / 47.2
PatchCore$_{\text{PMAE}}$	59.1 / 49.5	60.1 / 67.4	51.3 / 55.3	60.1 / 60.6	65.0 / 65.3	52.3 / 52.7	62.9 / 52.2	45.8 / 51.5	57.9 / 58.1	50.1 / 50.1	59.3 / 56.2	59.3 / 58.6	56.1 / 57.4	58.9 / 54.4
CPMF	35.3 / 61.5	64.3 / 65.5	52.0 / 52.1	48.2 / 57.1	40.5 / 43.5	78.3 / 74.5	63.9 / 48.8	62.5 / 63.5	65.8 / 64.1	68.3 / 68.3	68.5 / 68.4	48.2 / 48.6	60.1 / 60.1	60.1 / 56.2
Reg3D-AD	59.7 / 69.8	70.6 / 71.5	48.6 / 88.6	69.5 / 69.6	52.5 / 62.5	67.1 / 77.5	52.5 / 61.5	49.0 / 59.3	34.8 / 65.4	66.3 / 80.0	59.3 / 69.1	61.0 / 61.9	75.2 / 75.2	69.3 / 63.2
IMRNet	67.1 / 67.1	66.6 / 66.8	55.2 / 55.6	70.0 / 70.2	64.0 / 64.1	68.1 / 78.1	70.2 / 70.5	68.5 / 68.4	59.9 / 59.9	67.6 / 57.6	71.0 / 71.5	58.0 / 58.5	77.1 / 77.4	73.7 / 71.5
PO3AD	99.9 / 96.2	83.3 / 94.9	90.0 / 91.2	93.3 / 84.4	92.6 / 88.0	92.2 / 97.8	82.9 / 91.4	83.3 / 91.8	88.1 / 93.5	98.1 / 96.7	84.9 / 94.1	85.3 / 75.5	78.7 / 89.9	87.7 / 95.7
BTP (Ours)	75.2 / 85.6	53.8 / 94.6	75.8 / 95.7	62.7 / 87.0	79.2 / 95.6	92.1 / 94.0	63.8 / 80.3	76.1 / 90.1	59.7 / 80.6	44.5 / 81.3	78.6 / 88.8	61.9 / 83.5	79.1 / 94.0	57.3 / 94.1

Method	cap3	cap4	cap5	cup0	cup1	eraser0	headset0	headset1	helmet0	helmet1	helmet2	helmet3	jar0	micro.
BTF	52.7 / 68.7	46.8 / 46.9	37.3 / 37.3	40.3 / 63.2	52.1 / 56.1	52.5 / 63.7	37.8 / 57.8	51.5 / 47.5	55.3 / 50.4	34.9 / 44.9	60.2 / 60.5	52.6 / 70.0	42.0 / 42.3	56.3 / 58.3
M3DM	42.3 / 60.5	77.7 / 71.8	63.9 / 65.5	53.9 / 71.5	55.6 / 55.6	62.7 / 71.0	57.7 / 58.1	61.7 / 58.5	52.6 / 59.9	42.7 / 42.7	62.3 / 62.3	37.4 / 65.5	44.1 / 54.1	35.7 / 35.8
PatchCore$_{\text{FPFH}}$	45.3 / 65.3	75.7 / 59.5	79.0 / 79.5	60.0 / 65.5	58.6 / 59.6	65.7 / 81.0	58.3 / 58.3	63.7 / 46.4	54.6 / 54.8	48.4 / 48.9	42.5 / 45.5	40.4 / 73.7	47.2 / 47.8	38.8 / 48.8
PatchCore$_{\text{PMAE}}$	47.6 / 48.8	72.7 / 72.5	53.8 / 54.5	61.0 / 51.0	55.6 / 85.6	67.7 / 37.8	59.1 / 57.5	62.7 / 42.3	55.6 / 58.0	55.2 / 56.2	44.7 / 65.1	42.4 / 61.5	48.3 / 48.7	48.8 / 88.6
CPMF	55.1 / 55.1	55.3 / 55.3	69.7 / 55.1	49.7 / 49.7	49.9 / 50.9	68.9 / 68.9	64.3 / 69.9	45.8 / 45.8	55.5 / 55.5	58.9 / 54.2	46.2 / 51.5	52.0 / 52.0	61.0 / 61.1	50.9 / 54.5
Reg3D-AD	72.5 / 71.8	64.3 / 81.5	46.7 / 46.7	51.0 / 68.5	53.8 / 69.8	34.3 / 75.5	53.7 / 58.0	61.0 / 62.6	60.0 / 60.0	38.1 / 62.4	61.4 / 82.5	36.7 / 62.0	59.2 / 59.9	41.4 / 59.9
IMRNet	77.5 / 70.6	65.2 / 75.3	65.2 / 74.2	64.4 / 64.3	75.7 / 68.8	54.8 / 54.8	72.0 / 70.5	67.6 / 47.6	59.7 / 59.8	60.0 / 60.4	64.1 / 64.4	57.3 / 66.3	78.0 / 76.5	75.5 / 74.2
PO3AD	85.9 / 94.8	79.2 / 94.0	67.0 / 86.4	87.1 / 90.9	83.3 / 93.2	99.5 / 97.4	80.8 / 82.3	92.3 / 90.7	76.2 / 87.8	96.1 / 94.8	86.9 / 93.2	75.4 / 84.6	86.6 / 87.1	77.6 / 81.0
BTP (Ours)	60.5 / 82.8	65.6 / 92.6	62.7 / 95.8	71.5 / 91.5	51.8 / 80.4	85.1 / 95.1	60.1 / 83.5	57.2 / 80.4	50.4 / 88.1	48.6 / 71.9	64.3 / 85.6	51.8 / 80.0	77.9 / 87.6	93.4 / 94.0

Method	shelf0	tap0	tap1	vase0	vase1	vase2	vase3	vase4	vase5	vase7	vase8	vase9	Average
BTF	16.4 / 46.4	52.5 / 52.7	57.3 / 56.4	53.1 / 61.8	54.9 / 54.9	41.0 / 40.3	71.7 / 60.2	42.5 / 61.3	58.5 / 58.5	44.8 / 57.8	42.4 / 55.0	56.4 / 56.4	49.3 / 55.0
M3DM	56.4 / 55.4	75.4 / 65.4	73.9 / 71.2	42.3 / 60.8	42.7 / 60.2	73.7 / 73.7	43.9 / 65.8	47.6 / 65.5	31.7 / 64.2	65.7 / 51.7	66.3 / 55.1	66.0 / 66.3	55.2 / 61.6
PatchCore$_{\text{FPFH}}$	49.4 / 61.3	75.3 / 73.3	76.6 / 76.8	45.8 / 65.5	45.3 / 45.3	72.1 / 72.1	44.9 / 43.0	50.6 / 50.5	57.9 / 44.7	69.3 / 69.3	66.6 / 57.5	62.9 / 63.6	56.8 / 58.0
PatchCore$_{\text{PMAE}}$	52.3 / 54.3	45.8 / 85.8	53.8 / 65.1	44.7 / 67.7	55.2 / 55.1	74.1 / 74.2	46.0 / 46.5	51.6 / 52.3	57.9 / 57.2	65.0 / 65.1	66.2 / 36.4	62.9 / 42.3	56.2 / 57.7
CPMF	68.5 / 78.3	35.9 / 45.8	69.7 / 57.1	45.4 / 45.8	34.5 / 48.6	58.2 / 58.2	58.2 / 58.2	51.4 / 51.4	61.8 / 65.1	39.7 / 50.4	52.9 / 52.9	60.9 / 54.5	55.9 / 57.3
Reg3D-AD	68.8 / 68.8	67.6 / 58.9	64.1 / 74.1	53.3 / 54.8	70.2 / 60.2	60.5 / 40.5	65.0 / 51.1	50.0 / 75.5	52.0 / 62.4	46.2 / 88.1	62.0 / 81.1	59.4 / 69.4	57.2 / 66.8
IMRNet	69.6 / 60.5	73.6 / 68.1	90.0 / 69.9	78.8 / 53.5	72.9 / 68.5	75.2 / 61.4	74.2 / 40.1	63.0 / 52.4	75.7 / 68.2	77.1 / 59.3	72.1 / 63.5	71.8 / 69.1	74.9 / 65.0
PO3AD	57.3 / 66.3	74.5 / 78.3	68.1 / 69.2	85.8 / 95.5	74.2 / 88.2	95.2 / 97.8	82.1 / 88.4	67.5 / 90.2	85.2 / 93.7	96.6 / 98.2	73.9 / 95.0	83.0 / 95.2	83.9 / 89.8
BTP (Ours)	67.2 / 84.7	43.0 / 83.1	45.4 / 84.2	93.4 / 93.9	55.3 / 80.6	82.3 / 95.7	46.2 / 81.2	61.2 / 83.8	59.5 / 81.6	76.8 / 86.2	70.3 / 90.7	55.9 / 85.1	65.4 / 87.2

ent object categories, we report the detailed per-class AUROC results on the Anomaly-ShapeNet dataset in Tab. 1. Each cell presents the AUROC at both the object and point levels, where the former reflects global anomaly discrimination and the latter quantifies the accuracy of fine-grained localization.

Overall, our proposed BTP demonstrates consistently superior or comparable performance compared with representative baselines such as BTF, M3DM, PatchCore, CPMF, Reg3D-AD, and IMRNet. The improvements are particularly evident in categories with complex geometry or subtle local defects (*e.g., bottle, bowl, vase, helmet*), indicating that BTP can effectively capture local structural irregularities while maintaining strong overall recognition.

Furthermore, the stable per-class performance of BTP suggests better generalization to unseen shapes and surface variations, benefiting from its multi-granularity feature fusion and intrinsic 3D representation learning. These detailed results further confirm the robustness and practicality of BTP in diverse real-world 3D industrial inspection scenarios.

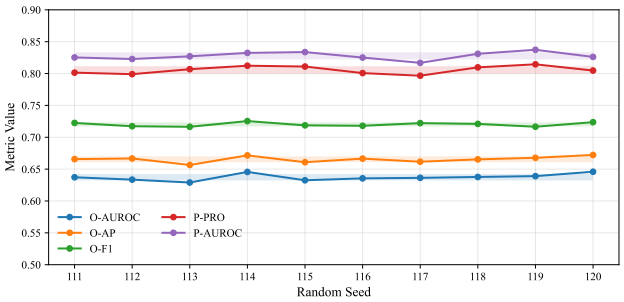
Stability Analysis Across Random Seeds. To evaluate the robustness of our framework with respect to random initialization on the Real3D-AD dataset, we conduct 10 independent runs for each training class (*chicken*, *gemstone*, and *car*), and report both object-level and point-level metrics. Fig. 1 visualizes the variation trends across different random seeds. The first three subfigures correspond to individual training classes, while the last one shows the averaged trend across all. All metrics exhibit small fluctuations ($<$2% for object-level and $<$5% for point-level results), demonstrating the strong stability and reproducibility of our method.

3. Additional Visualizations

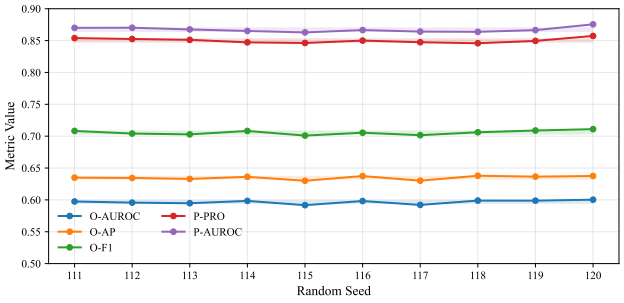
3.1. Additional Qualitative Results

As shown in Fig. 2, BTP exhibits distinct behaviors when trained on different source categories and tested on geometrically dissimilar target categories. The results clearly reveal that the intrinsic geometric nature of each object class plays a crucial role in determining anomaly localization quality.

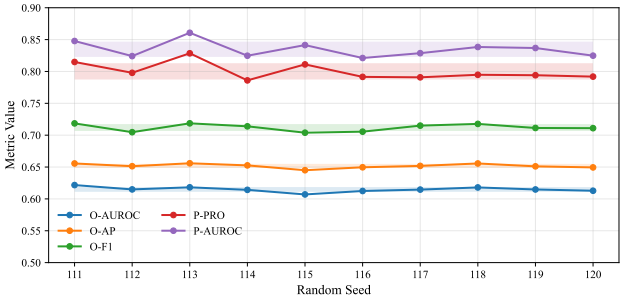
Specifically, some object classes (*e.g., cup, helmet, bottle*) exhibit smooth and rounded surfaces, where anomalies often appear as local protrusions or dents. When the model is trained on such round and compact geometries, it tends to learn a "smooth-surface prior," making it sensitive to any sharp or elongated structures in unseen categories. Consequently, when tested on sharper or more slender shapes (*e.g., tap, microphone, seahorse*), the model may incorrectly highlight normal edges or thin extensions as anomalous regions. Conversely, when trained on categories



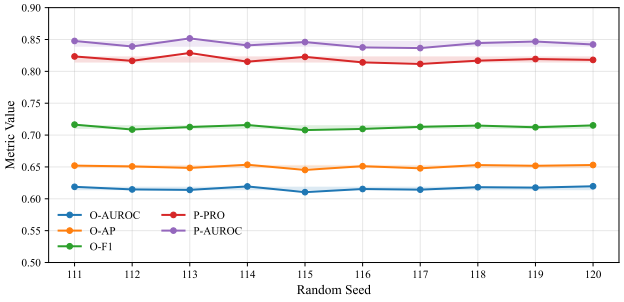
(a) chicken



(b) gemstone

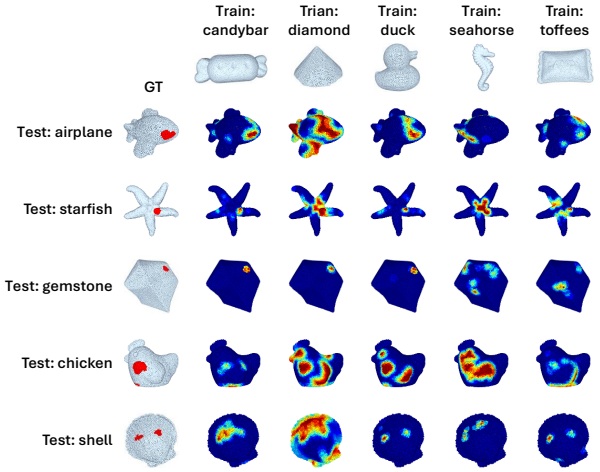


(c) car

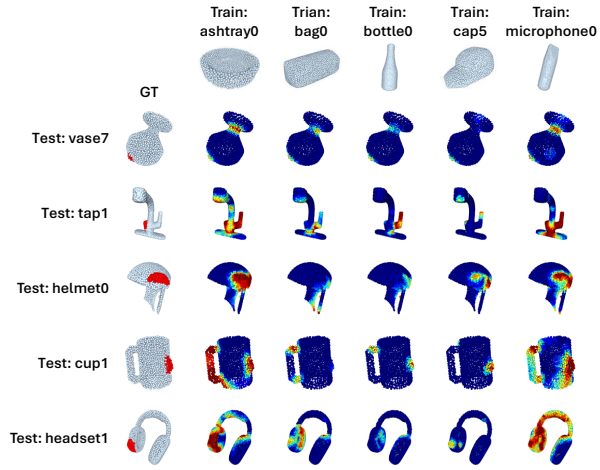


(d) Mean

Figure 1. Stability of performance across random seeds. Each curve shows the variation of object-level (O-AUROC, O-AP, O-F1) and point-level (P-PRO, P-AUROC) metrics over 10 runs. Subfigures (a)-(c) correspond to the three training classes, while (d) presents the averaged trend across all classes. The low variance indicates that the model converges consistently and is robust to random initialization.



(a) Cross-category visualization on Real3D-AD.



(b) Cross-category visualization on Anomaly-ShapeNet.

Figure 2. Cross-category anomaly localization visualization on Real3D-AD and Anomaly-ShapeNet. Each column corresponds to a training category, and each row shows testing categories unseen during training.

with complex or spiky geometries (*e.g., seahorse, starfish*), the model becomes more tolerant to high-curvature regions, leading to under-detection of subtle surface defects in smooth objects.

This phenomenon reflects a geometry-dependent bias that naturally emerges from pre-trained point-language alignment in the zero-shot setting. The model implicitly transfers shape priors from the training category to the testing one, and its generalization largely depends on the geometric compatibility between them. For instance, geometrically similar categories (*e.g., bottle* $\rightarrow$ *cup*) tend to yield more accurate and spatially consistent anomaly maps, while geometrically dissimilar pairs (*e.g., duck* $\rightarrow$ *gemstone*) show dispersed or misplaced activations.

These findings highlight both the strength and limitation

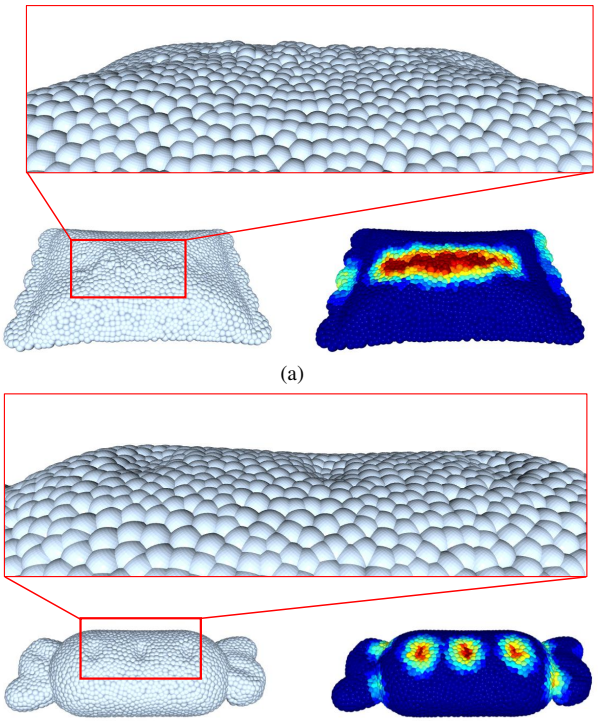
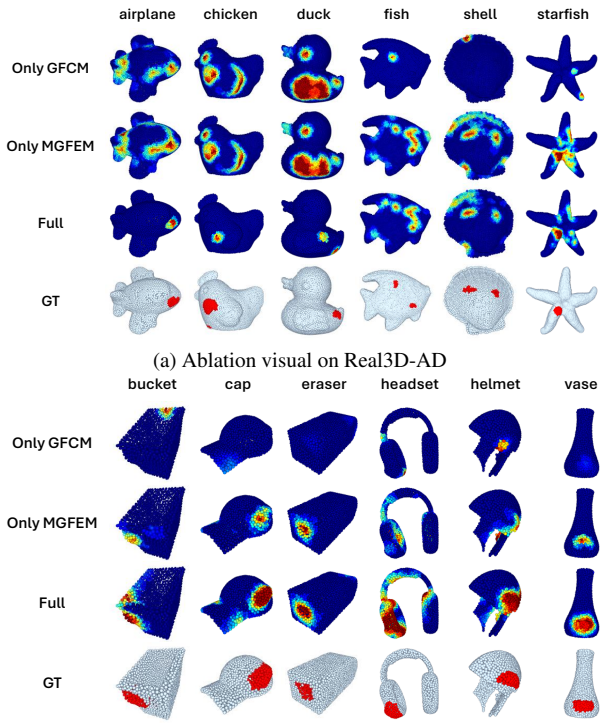


Figure 3. Failure case visualizations on unseen categories.



(a) Ablation visual on Real3D-AD

(b) Ablation visual on Anomaly-ShapeNet

Figure 4. Ablation visualization on Real3D-AD and Anomaly-ShapeNet. Each column corresponds to a testing category, and each row shows results with different module configurations.

of zero-shot 3D anomaly detection: although the model can generalize across categories without any anomalous samples, its geometric reasoning is still guided by the learned priors from the source class. To further enhance this capability, we envision incorporating few-shot guided adaptation, where a small number of normal and/or anomalous samples from new categories are used to refine the geometric embedding space. Such an approach could bridge the gap between purely zero-shot and fully supervised paradigms, allowing more reliable and category-adaptive anomaly localization in diverse 3D environments.

3.2. Failure Case Analysis

As shown in Fig. 3(a), the model mistakenly identifies the shallow decorative bumps on the surface of the toffee object as anomalous regions. This misclassification occurs because the network has learned "protrusion-like" cues associated with defects from other seen categories, which are incorrectly transferred when similar geometric patterns appear in normal objects. In contrast, Fig. 3(b) shows that the model highlights several concave surface areas of the candybar as anomalies, since "dent-like" features were previously correlated with defective regions during training. These cases reveal that the concept of anomaly learned in a zero-shot setting can still be influenced by category-specific priors, leading to a semantic boundary shift between normal and abnormal geometries when encountering unseen categories with distinct surface patterns. Improving the contextual adaptability of the learned normality representations may therefore help mitigate such false positives.

3.3. Ablation Visualizations

As shown in Fig. 4(a) and (b), we visualize anomaly localization results on Real3D-AD and Anomaly-ShapeNet, respectively, under three configurations: using only the Geometric Feature Creation Module (GFCM), using only the Multi-Granularity Feature Embedding Module (MGFEM), and using the full BTP model.

The GFCM encodes explicit geometric priors by aligning learned features with local descriptors, enhancing sensitivity to curvature variations and surface discontinuities. When used alone, it yields sharp but noisy activations that highlight local edges without semantic distinction. In contrast, the MGFEM aggregates multi-level point features to capture contextual semantics but tends to over-smooth fine geometric details, resulting in incomplete localization of subtle anomalies.

By jointly leveraging both modules, the BTP achieves a balance between geometric precision and semantic consistency, leading to compact, accurate, and reliable localiza-

tion across diverse object geometries.

4. Broader Impact

Our study explores the potential of zero-shot 3D anomaly detection through pre-trained point-language models (PLMs), which enables semantic understanding of 3D objects without requiring any anomalous data. In future research, we plan to extend our framework by incorporating RGB, allowing complementary 2D texture cues to enrich geometric representations and improve the recognition of subtle or texture-dependent defects. Moreover, we will investigate semi-zero-shot settings, where a small number of normal samples are provided during training. Such a setup is meaningful because certain types of anomalies are inherently defined relative to normal patterns; having limited normal references can help the model establish a stronger baseline of normality and thus enhance the detection of relative or context-dependent anomalies. We believe these directions will further improve the robustness, interpretability, and practical applicability of zero-shot 3D anomaly detection in real-world industrial environments.