GRASP: Enhancing Zero-Shot Multimodal Anomaly Detection via Geometric Refinement and Semantic Prompting

Zhongbin Sun[1,2] *, **Yuze Cui**[1,2], **Yong Zhou**[1,2]
[1]School of Computer Science and Technology / School of Artificial Intelligence, China University of Mining and Technology, China
[2]Mine Digitization Engineering Research Center of the Ministry of Education, China
{zhongbin, ts24170003a31ld, yzhou}@cumt.edu.cn

Abstract

Zero-shot multimodal anomaly detection is critical for identifying structural defects that are often invisible to traditional RGB images, particularly in scenarios lacking target domain samples. However, existing methodologies face two significant impediments: the prohibitive computational overhead caused by relying on multi-view rendering for depth processing, and the inability of static text prompts to adapt to fine-grained local anomalies. To address these challenges, a unified zero-shot multimodal anomaly detection framework GRASP is proposed. Firstly, a Frequency Domain Enhancement module is introduced to replace costly rendering with spectral transformations, directly synthesizing high-fidelity depth images for efficient utilization. Secondly, a Prompt-Conditioned Variational module is designed to bridge the semantic gap by grounding global textual descriptions into local visual nuances. Finally, a Dual Cross-Injection Alignment module is proposed to enable robust feature fusion for enhancing anomaly classification performance, while a Pyramid Anomaly Map Recalibration module further refines anomaly localization across multiple scales. Extensive experiments on MVTec 3D-AD and Eyecandies demonstrate that GRASP establishes a new state-of-the-art, yielding substantial improvements of 3.0 points in I-AUROC and 2.7 points in AUPRO compared to the SOTA method.

1 Introduction

Multimodal anomaly detection plays a pivotal role in industrial quality control, leveraging complementary geometric information to identify structural defects invisible in conventional RGB images [Bergmann *et al.*, 2020; Pang *et al.*, 2022; Xie *et al.*, 2023; Gu *et al.*, 2024]. While essential, the application of such systems is often hindered by the difficulty of collecting sufficient normal training samples in dynamic industrial environments. Such scarcity has motivated a shift toward zero-shot anomaly detection (ZSAD) [Ming *et al.*, 2022; Jeong *et al.*, 2023; Zhou *et al.*, 2024a], empowered by Vision-Language Models (VLMs) like CLIP [Radford *et al.*, 2021]. By aligning visual concepts with textual descriptions, these models offer strong generalization capabilities across diverse tasks [Zhou *et al.*, 2022; Rao *et al.*, 2022; Khattak *et al.*, 2023; Kirillov *et al.*, 2023].

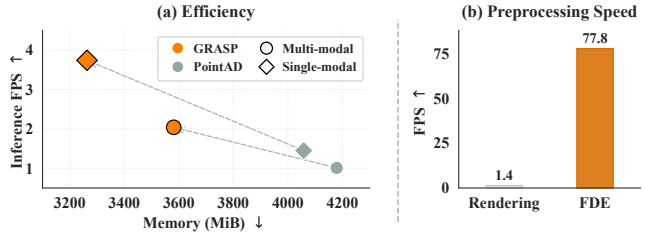

Figure 1: (a) Comparison of inference performance between GRASP and PointAD under unimodal setting and multimodal setting. (b) Comparison of the proposed FDE module with traditional rendering-based preprocessing.

While the integration of 3D geometry with zero-shot capabilities shows promise, critical impediments regarding efficiency and feature interaction remain. Firstly, processing 3D data is computationally expensive. Projection-based pipelines are typically employed by state-of-the-art methods [Zhou *et al.*, 2024b; Chen *et al.*, 2023; Bergmann and Sattlegger, 2023], rendering 3D point clouds into 2D images. This process incurs prohibitive overhead, creating a bottleneck for real-time deployment. Secondly, a misalignment exists between global semantic prompts and local anomalies. Standard text prompts describe holistic objects, whereas industrial defects are typically fine-grained and localized. Furthermore, most existing methods primarily rely on external semantic matching to identify anomalies, while overlooking the fact that anomalies are defined not only by semantic inconsistency but also by statistical deviations from their local context.

To systematically resolve these challenges, a unified framework, **GRASP** (**G**eometric **R**efinement **A**nd **S**emantic **P**rompting), is proposed to enhance zero-shot multimodal anomaly detection while reducing computational cost. As demonstrated in Figure 1(a), compared with PointAD, GRASP consistently achieves superior inference speed and significantly reduced computational cost. In GRASP, a Fre-

*Corresponding Author

quency Domain Enhancement (**FDE**) module is firstly introduced to bypass the rendering bottleneck, utilizing spectral transformations to directly synthesize depth images. As shown in Figure 1(b), compared with the rendering operation, preprocessing efficiency is improved by over $50\times$ with FDE. To bridge the semantic gap, a Prompt-Conditioned Variational (**PCV**) module is incorporated for feature refinement, alongside a Dual Cross-Injection Alignment (**DCIA**) module for bidirectional multimodal fusion. Finally, a Pyramid Anomaly Map Recalibration (**PAMR**) module is utilized to highlight anomalies through multi-scale statistical consistency. Experimental results on MVTec 3D-AD [Bergmann *et al.*, 2019] and Eyecandies [Bonfiglioli *et al.*, 2022] indicate that GRASP consistently surpasses existing state-of-the-art methods across all four metrics. Notably, it yields substantial improvements of 3.0 points in image-level AUROC and 2.7 points in pixel-level AUPRO relative to the previous best method MCL-AD [Li *et al.*, 2025]. Furthermore, ablation study is conducted to demonstrate the effectiveness of each proposed module, and we also provide the universality analysis for the proposed FDE module.

The main contributions are summarized as follows:

- A novel ZSAD framework GRASP is proposed to utilize mathematical transformations in the spectral domain to replace costly 3D rendering.
- The PCV and PAMR modules are introduced to bridge the gap between semantics and local defects through semantic alignment and internal statistical coherence.
- The DCIA module is designed for multimodal fusion, which effectively integrates the geometric feature from depth map and the semantic feature from RGB image.
- Extensive experiments demonstrate that GRASP consistently achieves SOTA performance across different metrics in both multimodal and unimodal settings.

2 Related Work

2.1 Zero-shot Anomaly Detection

Traditional 2D anomaly detection typically operates in an unsupervised manner on benchmarks like MVTec AD [Bergmann *et al.*, 2019], relying on representation-based paradigms such as teacher–student networks [Deng and Li, 2022], normalizing flows [Rudolph *et al.*, 2021; Gudovskiy *et al.*, 2022], and memory banks [Roth *et al.*, 2022], or reconstruction-based approaches involving autoencoders [Zavrtanik *et al.*, 2021; You *et al.*, 2022] and diffusion models [Zavrtanik *et al.*, 2022]. While some methods introduce synthetic defects for supervised learning [Li *et al.*, 2021; Liu *et al.*, 2023], their dependency on target-domain training data limits flexibility. To address this, Zero-Shot Anomaly Detection (ZSAD) has emerged, leveraging Vision-Language Models like CLIP [Radford *et al.*, 2021] to align visual and textual features without training samples. Recent works have effectively adapted CLIP for this task: WinCLIP [Jeong *et al.*, 2023] utilizes handcrafted prompts with multi-scale aggregation; AnomalyCLIP [Zhou *et al.*, 2024a] introduces learnable prompt tuning; and AdaCLIP [Sun *et al.*, 2023] and PromptAD [Tian *et al.*, 2023] further enhance generalization via category-aware semantics and negative prompting, respectively. However, these methods remain largely confined to the 2D domain, neglecting geometric anomalies.

2.2 Multimodal Fusion

To transcend the limitations of RGB-only detection, multimodal anomaly detection integrates 3D depth or point cloud data [Bergmann *et al.*, 2022; Bonfiglioli *et al.*, 2022] to capture structural defects. Early fusion strategies adapted 2D architectures to depth inputs, such as AST [Rudolph *et al.*, 2022] for background removal and 3D-ST [Bergmann and Sattlegger, 2023] for local geometric description, while subsequent memory-based methods like M3DM [Wang *et al.*, 2023] and reconstruction pipelines like EasyNet [Chen *et al.*, 2023] focused on hybrid feature fusion. More recently, projection-based methods such as CPMF [Cao *et al.*, 2024], PointAD [Zhou *et al.*, 2024b] and MCL-AD [Li *et al.*, 2025], convert 3D data into multi-view images to leverage 2D pre-trained feature extractors. Despite their effectiveness, these approaches often suffer from high computational overhead due to multi-view rendering or rely on extensive target-domain training. This motivates our proposed lightweight framework, which directly synergizes RGB and depth features for efficient zero-shot detection without costly rendering processes.

3 Method

3.1 Overview

The overall architecture of the proposed method GRASP is illustrated in Figure 2. Specifically, given paired RGB images and raw depth maps as inputs, the depth stream is firstly processed by the Frequency Domain Enhancement (FDE) module, where spectral transformations are applied to reconstruct high-fidelity representations. Subsequently, hierarchical visual features are extracted from both the RGB and enhanced depth streams via a frozen image encoder. In the textual branch, learnable prompts are projected into global textual context representations by a frozen encoder, whereas for capturing diverse semantic nuances, the Prompt-Conditioned Variational (PCV) module is introduced to map learnable prompts into local textual context representations. Finally, the Dual Cross-Injection Alignment (DCIA) module is designed to unify the multimodal information flows, and the final anomaly score is obtained by combining its output with the global textual context representations. Moreover, the fused features are spatially refined by the Pyramid Anomaly Map Recalibration (PAMR) module to generate the precise pixel-level anomaly map.

3.2 Frequency Domain Enhancement Module

As illustrated in Figure 3, the proposed FDE module is designed to convert raw depth inputs into representations with enhanced structural fidelity. Given depth maps $x \in \mathbb{R}^{H\times W}$, the processing pipeline comprises two sequential stages. The first stage executes structure-aware local enhancement via gradient-based modulation and Local Contrast Normalization (LCN), while the second stage performs frequency domain stripe attenuation. Through this two-stage refinement, the

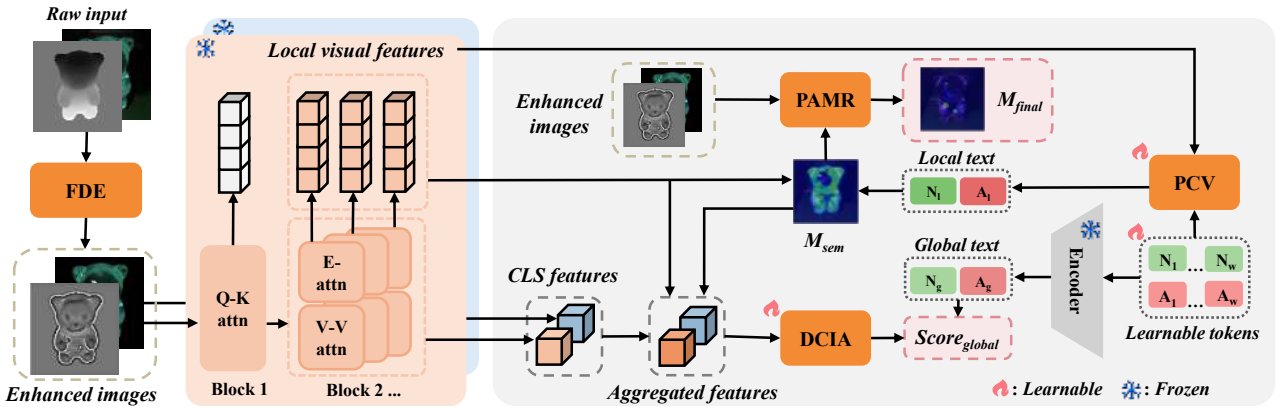

Figure 2: Framework of GRASP.

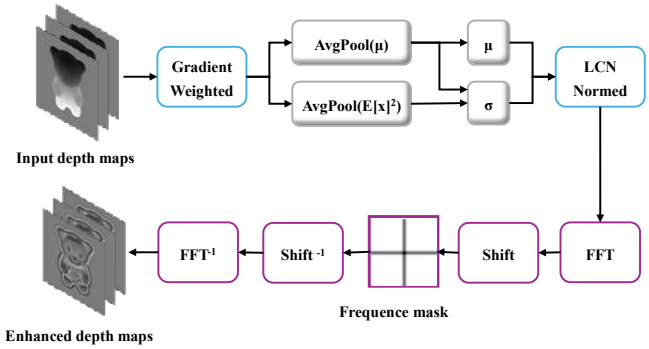

Figure 3: Frequency Domain Enhancement (FDE) Module.

module effectively preserves depth variations corresponding to physically meaningful geometric structures.

Structure-Aware Local Enhancement

Local geometric variations are first characterized by computing the gradient magnitude along the horizontal and vertical directions, derived as:

$$G(x) = \sqrt{\mathrm{Grad}_h(x)^2 + \mathrm{Grad}_v(x)^2 + \varepsilon} \tag{1}$$

, where $\mathrm{Grad}_d(x)$ means computing gradient along the d direction and a small constant ε is introduced to ensure numerical stability. This gradient magnitude serves as a robust indicator of geometric transition strength, effectively capturing high-frequency structural details such as edges and boundaries.

Building upon the derived gradient magnitude, a nonlinear modulation mechanism is formulated to self-adaptively amplify structural discontinuities. Specifically, a modulation weight map x_w is computed as:

$$x_w = 1 + \beta \cdot \tanh\big(\gamma \cdot G(x)\big) \tag{2}$$

, where $\beta > 0$ controls the enhancement amplitude and $\gamma > 0$ regulates sensitivity to gradient variations. Subsequently, the gradient-aware depth representation x_{grad} is generated by imposing this weight onto the raw input:

$$x_{\mathrm{grad}} = \mathrm{Norm}(x \odot x_w) \tag{3}$$

, where $\odot$ denotes element-wise multiplication and $\mathrm{Norm}(\cdot)$ represents per-sample min–max normalization. This operation effectively highlights structural boundaries while constraining the dynamic range for numerical stability.

While the gradient modulation effectively highlights transitions, the resulting representation remains susceptible to global depth bias, particularly where large-scale object geometries overshadow subtle defect signatures. To mitigate this, the Local Contrast Normalization (LCN) mechanism is introduced to enforce statistical consistency within local neighborhoods. Formally, for each coordinate (i, j) in x_{grad}, we define a $k \times k$ sliding window and compute the local statistics efficiently using average pooling operations:

$$\begin{aligned} \mu(i,j) &= \mathrm{AvgPool}_k(x_{\mathrm{grad}}), \\ \sigma(i,j) &= \sqrt{\mathrm{AvgPool}_k(x_{\mathrm{grad}}^2) - \mu(i,j)^2 + \varepsilon}. \end{aligned} \tag{4}$$

Based on these statistics, the locally normalized depth map x_{lcn} is derived via a nonlinear mapping:

$$x_{\mathrm{lcn}} = \mathrm{Norm}\left(\tanh\left(\alpha \cdot \frac{x_{\mathrm{grad}} - \mu}{\sigma + \varepsilon}\right)\right) \tag{5}$$

, where α amplifies local contrast. By standardizing intensity distributions within local windows, this operation effectively decouples fine-grained structural details from low-frequency global variations, thereby enhancing the prominence of potential anomalies.

Frequency Domain Stripe Attenuation

Applying local normalization to continuous geometric transitions inevitably induces periodic stripe-like artifacts, which can be misclassified as anomalies. To mitigate the stripe-like artifacts exacerbated by the preceding stage, the Frequency Domain Stripe Attenuation sub-module is introduced, leveraging the Fourier Transform to isolate and suppress the spectral signatures of these periodic disturbances. To be specific, the enhanced depth map is first transformed into the frequency domain:

$$F(u, v) = \mathrm{Shift}\big(\mathrm{FFT}(x_{\mathrm{lcn}})\big) \tag{6}$$

, where $\mathrm{FFT}(\cdot)$ denotes the Fast Fourier Transform and $\mathrm{Shift}(\cdot)$ moves the zero-frequency component to the center of

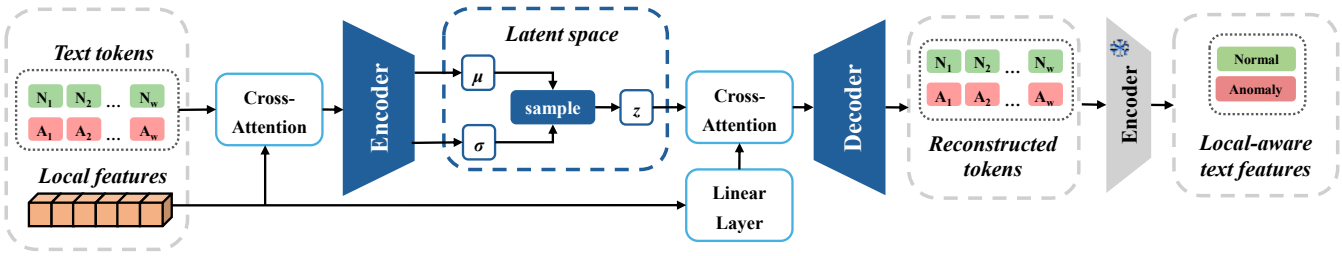

Figure 4: Prompt-Conditioned Variational (PCV) Module.

the spectrum, with (u, v) denoting the frequency coordinates relative to the spectral center (u_c, v_c). The stripe patterns are modeled as narrow-band spectral components concentrated along the principle axes. Their suppression is achieved using a separable Gaussian attenuation mask:

$$M(u,v) = \left(1 - \eta \cdot e^{-\frac{(v-v_c)^2}{2(Wr)^2}}\right) \cdot \left(1 - \eta \cdot e^{-\frac{(u-u_c)^2}{2(Hr)^2}}\right) \quad (7)$$

, where r controls the effective bandwidth ratio relative to image dimensions (H, W) and $\eta \in (0, 1)$ determines the attenuation strength. Finally, the clean depth map is reconstructed via the inverse transformation:

$$\tilde{x} = \mathrm{FFT}^{-1}\left(\mathrm{Shift}^{-1}(F \odot M)\right) \quad (8)$$

, where FFT^{-1} and Shift^{-1} denote the inverse operations of their respective forward functions. This frequency domain filtering effectively dampens the energy of enhancement-induced stripes while preserving the low- and mid-frequency components that encode genuine geometric structures.

The proposed FDE module resolves the intrinsic conflict between detail enhancement and noise amplification through a cascade of efficient analytical operations. Unlike complex rendering-based pre-processing or heavy reconstruction networks, our approach relies solely on fast matrix and frequency transformations, thereby offering superior computational efficiency. This lightweight design ensures that the resulting multi-channel representation $\tilde{x}$ enriches the feature space with global, local, and denoised geometric cues, while introducing negligible latency overhead compared to conventional pseudo-color rendering pipelines.

3.3 Prompt-Conditioned Variational Module

To effectively align enhanced geometric cues with high-level semantics, the limitations of static textual spaces which often struggle to capture fine-grained industrial structural patterns must be addressed. Therefore, the PCV module illustrated in Figure 4 is introduced to generate visually grounded prompt representations. Formally, utilizing a CLIP-based encoder augmented with E-attn [Salehi *et al.*, 2025], spatially dense visual features $V \in \mathbb{R}^{N \times d}$ are extracted. These features serve to condition the base prompt embedding $T \in \mathbb{R}^{t \times d}$, where N and t denote the number of visual patches and prompt tokens, respectively.

Latent Space Encoding

To embed visual context into textual representations, a cross-modal attention mechanism is employed where text embeddings T serve as queries and visual features V as keys/values. This dynamically aligns semantic tokens with visually relevant regions:

$$H = T + \mathrm{Attn}(T, V, V). \quad (9)$$

The aggregated representation H captures the prompt distribution explicitly supported by visual evidence, accounting for spatial heterogeneity in anomaly localization.

Subsequently, the encoder maps H to a variational posterior $[\mu, \log \sigma^2] = \mathrm{Encoder}(H)$. To ensure differentiability, the latent variable Z is sampled via the reparameterization trick:

$$Z = \mu + \sigma \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \quad (10)$$

Crucially, this formulation acts as a controlled information bottleneck. Unlike rigid feature concatenation, the variational projection filters out task-irrelevant noise, ensuring the final representation remains robust while strictly grounded in visual evidence.

Variational Prompt Generation

To mitigate information loss within the bottleneck, the decoder is augmented with a secondary cross-attention mechanism that re-injects visual cues. The latent variable is refined via

$$\widehat{Z} = Z + \mathrm{Attn}(Z,\ \mathrm{Linear}(V),\ \mathrm{Linear}(V)) \quad (11)$$

, and subsequently converted into the enhanced token-wise prompt representation $T' = \mathrm{Decoder}(\widehat{Z})$. The training objective balances semantic adaptivity with linguistic stability. A reconstruction loss employing a hinge-based formulation is introduced to control the semantic shift. This allows visually induced modifications only within a tolerance margin m:

$$\mathcal{L}_{\mathrm{rec}} = \max(0, 1 - m - \langle T', T \rangle). \quad (12)$$

This mechanism disentangles directional alignment from intensity consistency, preventing uncontrolled semantic drift. Combined with the standard KL divergence regularization, the total objective is defined as:

$$\mathcal{L}_{pcv} = \mathcal{L}_{\mathrm{rec}} + \mathcal{L}_{\mathrm{KL}}. \quad (13)$$

3.4 Dual Cross-Injection Alignment Module

Existing multimodal anomaly detection methodologies often rely on naive fusion strategies that fail to account for the inherent modal gap, leading to feature redundancy or the suppression of subtle anomalous cues. To address these limitations, the DCIA module is proposed. To be specific, DCIA first employs a semantic-guided aggregation strategy

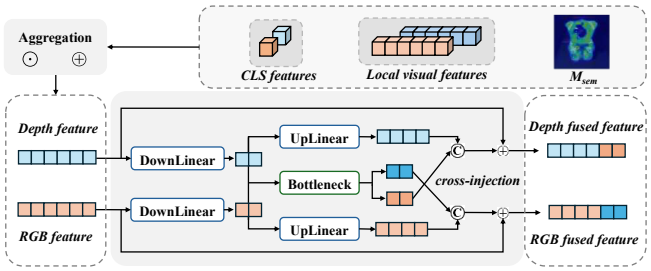

Figure 5: Dual Cross-Injection Alignment (DCIA) Module.

to refine image-level representations. Subsequently, a symmetric cross-injection mechanism is implemented to bridge the inherent modal gap and facilitate bidirectional information flow, ensuring the fused features are both semantically aligned and structurally enriched. The outputs of DCIA are combined with the global textual context representations (shown in Figure 2), and their cosine similarity is computed to obtain the final anomaly score.

Semantic-Guided Feature Aggregation

To extract discriminative image-level representations while preserving fine-grained details, enhanced prompt tokens T' is utilized to guide the aggregation of visual features. Specifically, normal and anomaly text anchors t_n, t_a are derived from T'_n, T'_a by CLIP text encoder. Subsequently, a preliminary semantic anomaly map M_{sem} is computed via cosine similarity:

$$M_{sem}(i,j) = \text{softmax}(\langle V(i,j) \cdot t_n \rangle, \langle V(i,j) \cdot t_a \rangle). \quad (14)$$

Rather than employing standard global average pooling, M_{sem} is utilized as a spatial attention mask to re-calibrate the aggregation process. The aggregated features f for each modality are formulated as:

$$f = v + \frac{\sum_{i,j}^{h,w} (M_{sem}(i,j) \odot V(i,j))}{\sum_{i,j}^{h,w} M_{sem}(i,j)} \quad (15)$$

, where v and V denote the `cls` token and local feature maps, respectively. This ensures that the global representations f_r (RGB) and f_d (Depth) are specifically enriched with localized evidence.

Shared Latent Projection

With f_r and f_d obtained, the DCIA module addresses the modality-specific noise by projecting features into a compact latent manifold via linear layer W:

$$z_r = \phi(W_r f_r), \qquad z_d = \phi(W_d f_d) \quad (16)$$

, where $\phi(\cdot)$ denotes ReLU activation. This compression forces the model to discard high-frequency noise while retaining essential semantic density. To further bridge the modal gap, a shared transformation U maps these latent vectors into a common semantic subspace:

$$u_r = U z_r, \qquad u_d = U z_d. \quad (17)$$

By sharing U, the network is constrained to align semantically corresponding structures from both modalities into similar regions of the feature space, establishing u_r and u_d as modality-agnostic priors.

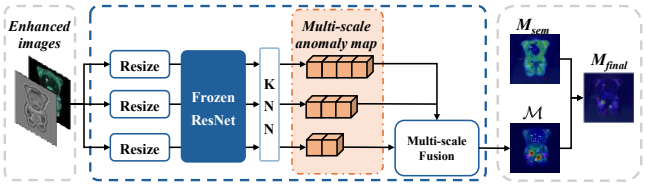

Figure 6: Pyramid Anomaly Map Recalibration (PAMR) Module.

Symmetric Cross-Injection

The core innovation of DCIA lies in its complementary reconstruction phase. Rather than independently recovering features, a symmetric cross-injection strategy is implemented where each modality utilizes the semantic prior of another modality to guide its reconstruction. The updates are synthesized as:

$$\Delta_r = \text{Concat}(W'_r z_r, u_d), \quad \Delta_d = \text{Concat}(W'_d z_d, u_r). \quad (18)$$

Here, u_d acts as a geometric guide for the RGB stream, while u_r provides texture context for the depth stream. This forces the network to "fill in" the information bottleneck by mining complementary cues from the paired modality. Finally, to preserve the original signal fidelity and prevent catastrophic forgetting, these cross-enhanced updates are integrated via residual connections:

$$\widetilde{f}_r = f_r + \Delta_r, \qquad \widetilde{f}_d = f_d + \Delta_d. \quad (19)$$

This formulation ensures that multimodal fusion acts as a constructive refinement, yielding features that are both semantically aligned and structurally enriched.

Based on DCIA, the final image-level anomaly score can be calculated as:

$$Score_{global} = \sum_{i}^{m} \frac{(s_i - 0.5)^2}{\sum_j^m (s_j - 0.5)^2} \cdot s_i \quad (20)$$

, where s_m is obtained by calculating the cosine similarity between the fused features $\widetilde{f}_m$ of the m modal and the global text. This weighting mechanism adaptively highlights discriminative modalities while suppressing uncertain ones by calculating the squared deviation of s_i from 0.5, thereby enhancing the robustness of multimodal fusion. To jointly optimize global anomaly scoring and pixel-wise localization, a hybrid objective function is introduced. During training, the Binary Cross-Entropy Loss ($\mathcal{L}_{\text{global}}$) is employed as the sample-level objective to optimize the model for accurate anomaly probability estimation in an end-to-end manner. For the segmentation branch, the Focal Loss and Dice Loss ($\mathcal{L}_{\text{local}}$) is utilized on M_{sem}. The total training objective is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{global}} + \mathcal{L}_{\text{local}} + \mathcal{L}_{pcv} \quad (21)$$

3.5 Pyramid Anomaly Map Recalibration Module

While the semantic pipeline established by the PCV provides a robust vision-language alignment, exclusive reliance on external textual prototypes remains vulnerable to semantic hallucinations. This pathology occurs when complex but normal

Category	Method	Multimodal				Unimodal			
		P-AUROC	P-AUPRO	I-AUROC	I-AP	P-AUROC	P-AUPRO	I-AUROC	I-AP
CLIP-based Methods	PointCLIP [Zhu *et al.*, 2023]	46.3	-	48.5	50.9	43.9	-	45.2	48.0
	AnomalyCLIP [Zhou *et al.*, 2024a]	79.6	45.4	65.7	68.1	76.6	38.7	50.9	49.5
	FAPrompt [Zhu *et al.*, 2025]	90.7	70.2	63.8	65.2	76.4	39.3	55.4	53.1
	Crane [Salehi *et al.*, 2025]	94.8	82.9	65.7	66.9	82.7	45.2	57.9	56.6
	PointAD-CoOp [Zhou *et al.*, 2024b]	94.4	80.3	76.3	78.9	91.8	70.5	69.1	73.8
	PointAD [Zhou *et al.*, 2024b]	94.0	80.7	78.6	80.8	91.8	71.4	69.5	74.3
	MCL-AD [Li *et al.*, 2025]	95.8	85.4	81.1	83.2	-	-	-	-
Few-shot Methods	M3DM [Wang *et al.*, 2023]	94.8	79.5	64.4	64.1	91.0	68.4	59.5	58.0
	Shape-Guided [Chu *et al.*, 2023]	90.3	66.9	53.6	56.5	85.8	56.3	53.2	51.2
	CFM [Costanzino *et al.*, 2024]	95.2	80.5	66.3	69.7	-	-	-	-
	CIF [Lin *et al.*, 2026]	93.6	75.4	75.0	-	88.7	59.0	59.2	-
	GRASP	**97.4**	**88.1**	**84.1**	**85.3**	**92.1**	**72.5**	**73.0**	**75.8**

Table 1: Comparison with SOTA methods under multimodal and unimodal settings. Best results are highlighted in bold red.

geometric textures are misidentified as defects due to the inherent ambiguity of zero-shot descriptions. To resolve this, the PAMR module is introduced as a rigorous internal verification layer that validates semantic predictions through structural statistical consistency, as shown in Figure 6.

Recognizing that industrial defects manifest across diverse scales, a pyramid strategy is employed to process images at multiple resolutions, extracting hierarchical deep representations F_s. Unlike the semantic branch which is constrained by linguistic definitions, this process is entirely data-driven. For each patch feature $F_{p,s}$ at scale s, its "isolation" in the feature space is quantified by computing the average cosine distance to its k-nearest neighbors within the same image:

$$\mathcal{M}_{p,s} = \frac{1}{k} \sum_{F_{q,s} \in \Omega_k(F_{p,s})} \left(1 - \frac{F_{p,s} \cdot F_{q,s}}{|F_{p,s}| \cdot |F_{q,s}|}\right) \tag{22}$$

, where $\Omega_k(F_{p,s})$ denotes the set of k most similar features in the current map. This metric effectively captures local incoherence: true anomalies naturally emerge as statistical outliers relative to the dominant normal patterns, resulting in a scale-invariant recalibration map $\mathcal{M}$.

The final decision is synthesized by fusing the external semantic prediction M_{sem} with the internal statistical evidence $\mathcal{M}$ through an ensemble strategy:

$$M_{\text{final}} = \lambda \cdot M_{\text{sem}} + (1 - \lambda) \cdot \sqrt{M_{\text{sem}} \odot \mathcal{M}} \tag{23}$$

, where λ is a balancing factor. The core innovation of this formulation lies in the geometric term $\sqrt{M_{\text{sem}} \odot \mathcal{M}}$, acting as a strict consistency gate. In contrast to standard arithmetic summation, which allows high-confidence semantic errors to persist, the geometric term necessitates simultaneous high activation from both semantic and structural sources. If a region is semantically ambiguous but statistically coherent with its surroundings, the geometric term forces the final score towards zero. This dual-evidence mechanism ensures that detections are not only semantically grounded but also statistically validated, effectively suppressing false positives and enhancing the reliability of zero-shot inspection.

4 Experiments

Datasets. GRASP is trained on the dataset MVTec 3D-AD and evaluated on the dataset Eyecandies, each containing 10 object categories with corresponding RGB and depth information, to simulate scenarios of unseen categories. MVTec 3D-AD captures real-world industrial anomalies with structured scenes, while Eyecandies features more diverse and colorful object types, providing a challenging testbed for zero-shot generalization.

Metrics. Following the evaluation protocol in PointAD [Zhou *et al.*, 2024b], two sets of standard metrics are employed: I-AUROC and I-AP for global image-level anomaly detection, and P-AUROC and P-AUPRO for local pixel-level anomaly segmentation. Since GRASP utilizes depth maps rather than raw point clouds, all metrics adhere to the RGB-based evaluation protocol to ensure methodological consistency.

Baselines. To comprehensively evaluate the effectiveness of the proposed GRASP, extensive comparisons are conducted against a wide range of state-of-the-art methods. Specifically, for zero-shot evaluation, representative CLIP-based approaches are selected, including PointCLIP [Zhu *et al.*, 2023], AnomalyCLIP [Zhou *et al.*, 2024a], FAPrompt [Zhu *et al.*, 2025], Crane [Salehi *et al.*, 2025], PointAD [Zhou *et al.*, 2024b], and MCL-AD [Li *et al.*, 2025]. Furthermore, to establish rigorous performance boundaries, prominent methods M3DM [Wang *et al.*, 2023], Shape-Guided [Chu *et al.*, 2023], CFM [Costanzino *et al.*, 2024] and CIF [Lin *et al.*, 2026] are included under few-shot setting as references. Notably, these few-shot models are evaluated under a challenging 2-shot setting to serve as a competitive baseline. All comparisons are performed across both multimodal and unimodal benchmarks to verify the versatility of the framework.

Implementation Details. The pre-trained CLIP model from `open_clip` is adopted as both the visual and text encoder, with all encoder weights frozen. Following AnomalyCLIP [Zhou *et al.*, 2024a], the number of learnable tokens in object-agnostic prompt templates is set to $w = 12$. Both RGB and depth inputs are resized to 336×336 before being fed into the visual encoder. During training, DCIA is trained for 20 epochs, while the remaining components are trained for 6 epochs. All experiments are conducted on a single NVIDIA RTX 2080Ti GPU using PyTorch 2.0.0.

Modules				Metrics			
$\mathcal{M}_{\text{FDE}}$	$\mathcal{M}_{\text{PCV}}$	$\mathcal{M}_{\text{PAMR}}$	$\mathcal{M}_{\text{DCIA}}$	P-AUROC	P-AUPRO	I-AUROC	I-AP
Baseline				94.8 / 82.7	82.9 / 45.2	65.7 / 56.6	66.9 / 57.9
✓				95.4 / 89.3	82.7 / 59.2		
✓	✓			97.3 / 92.0	86.8 / 70.5	82.3 / 73.0	83.5 / 75.8
✓	✓	✓		97.4 / 92.1	88.1 / 72.5		
✓	✓		✓	97.3 / 92.0	86.8 / 70.5	84.1 / 73.0	85.3 / 75.8
✓	✓	✓	✓	**97.4 / 92.1**	**88.1 / 72.5**	**84.1 / 73.0**	**85.3 / 75.8**

Table 2: Ablation study of the GRASP framework on MVTec 3D-AD and Eyecandies. Results are reported as Multimodal/Unimodal, with the best results highlighted in bold red.

Method	Setting	P-AUROC	P-AUPRO	I-AUROC	I-AP
AnomalyCLIP	w/o	90.9 / 76.6	71.4 / 38.7	57.4 / 49.5	57.9 / 50.9
	w	93.3 / 87.0	76.7 / 60.3	63.8 / 61.6	65.3 / 65.0
FAPrompt	w/o	90.7 / 76.4	70.2 / 39.3	63.8 / 53.9	65.5 / 55.4
	w	92.9 / 89.1	75.6 / 65.3	69.6 / 69.8	71.6 / 71.3

Table 3: Evaluation of universality for the FDE module under different settings. Results are presented as Multimodal/Unimodal.

4.1 Main Results

As reported in Table 1, GRASP establishes a new state-of-the-art across all evaluated metrics in the multimodal setting, achieving a peak I-AUROC of 84.1% and a P-AUPRO of 88.1%. These results represent a significant advancement over the previous leading method, MCL-AD, with substantial improvements of 3.0 points in image-level detection and 2.7 points in pixel-level localization. Furthermore, in the unimodal scenario of depth maps, GRASP continues to demonstrate exceptional robustness even when color or texture information is missing. To be specific, GRASP achieves an I-AUROC of 73.0% and a P-AUPRO of 72.5%, widening the performance gap against the strong baseline PointAD by 3.5 and 1.1 points, respectively.

These consistent improvements across diverse settings validate the synergistic effectiveness of the proposed modules. At the foundational level, FDE establishes a high-fidelity geometric baseline by bypassing traditional rendering artifacts, ensuring the visual encoder receives denoised and structurally accurate features. This high-quality input allows the PCV module to effectively address the limitations of static prompts by capturing diverse semantic nuances, which significantly enhances the model's generalization capabilities across unseen industrial categories. To integrate these heterogeneous signals, the DCIA mechanism facilitates robust bidirectional interaction between RGB and geometric streams, effectively resolving semantic misalignment through symmetric cross-modal alignment. Finally, the PAMR module serves as a critical internal verification gate that utilizes multi-scale statistical consistency to suppress semantic hallucinations, ensuring that anomaly predictions are not only semantically grounded but also statistically distinct from the dominant normal patterns of the object.

4.2 Ablation Study

The contribution of each individual component is systematically evaluated in the ablation study, as presented in Table 2. Starting from the baseline, the proposed modules are progressively incorporated, revealing a synergistic evolution in performance. Specifically, with the integration of FDE, a substantial improvement is observed, particularly in the I-AUROC and I-AP metrics, which establishes a robust foundation for feature extraction. Subsequently, the pixel-level localization capability is significantly enhanced by the PCV module, yielding sharp increases in P-AUPRO and P-AUROC. This localization precision is then further refined by the addition of the PAMR module. Furthermore, an alternative ablation setting excluding PAMR highlights the independent impact of the DCIA module, which elevates image-level anomaly detection metrics by precisely aligning the geometric cues from the depth stream with the semantic prompts from the RGB stream. Ultimately, the integration of all four modules demonstrates that the proposed components are not merely additive but highly complementary, collectively forming a unified system for high-precision anomaly detection.

4.3 FDE Universality Analysis

Given that low-fidelity geometric inputs act as a universal bottleneck in multimodal anomaly detection, the FDE module is utilized to rectify data and improve ZSAD performance. To demonstrate its architecture-agnostic universality, FDE is integrated into the AnomalyCLIP and FAPrompt frameworks as a plug-and-play booster. As shown in Table 3, transformative improvements are achieved across all baselines, such as significantly increasing the unimodal P-AUPRO of AnomalyCLIP from 38.7% to 60.3% and the multimodal I-AUROC of FAPrompt from 63.8% to 69.6%. These results substantiate that the FDE module effectively captures subtle structural features, establishing it as a universal enhancement for zero-shot anomaly detection.

5 Conclusion

In this paper, a unified and robust framework GRASP is proposed to overcome the critical challenges of computational inefficiency and semantic gap in zero-shot multimodal anomaly detection. By transitioning from heavy 3D rendering to efficient spectral transformations via the FDE module, a preprocessing speedup of over 50x is achieved while maintaining superior structural fidelity. To mitigate the semantic gap between textual prompts and local visual details, the PCV module enables visually grounded prompt refinement, thereby enhancing the generalization capability of the framework. Furthermore, the DCIA module strengthens multimodal fusion by dynamically integrating bidirectional cross-modal information flows, ensuring robust and consistent feature alignment across modalities. Moreover, the PAMR module improves anomaly localization reliability by suppressing semantic hallucinations through multi-scale internal statistical consistency validation. Extensive experiments on the MVTec 3D-AD and Eyecandies datasets demonstrate that GRASP establishes a new state-of-the-art, providing a resilient and high-precision solution for real-world industrial inspection tasks.

Acknowledgements

This research is supported by the Fundamental Research Funds for the Central Universities under Grant No. 2021QN1075.

References

[Bergmann and Sattlegger, 2023] Paul Bergmann and David Sattlegger. Anomaly detection in 3D point clouds using deep geometric descriptors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2613–2623, 2023.

[Bergmann *et al.*, 2019] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9592–9600, 2019.

[Bergmann *et al.*, 2020] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4183–4192, 2020.

[Bergmann *et al.*, 2022] Paul Bergmann, Xin Jin, David Sattlegger, and Carsten Steger. The MVTec 3D-AD dataset for unsupervised 3D anomaly detection and localization. In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pages 202–213, 2022.

[Bonfiglioli *et al.*, 2022] Luca Bonfiglioli, Marco Toschi, Davide Silvestri, Nicola Fioraio, and Daniele De Gregorio. The Eyecandies dataset for unsupervised multimodal anomaly detection and localization. In *Proceedings of the Asian Conference on Computer Vision*, pages 3586–3602, 2022.

[Cao *et al.*, 2024] Yunkang Cao, Xiaohao Xu, and Weiming Shen. Complementary pseudo multimodal feature for point cloud anomaly detection. *Pattern Recognition*, 156:110761, 2024.

[Chen *et al.*, 2023] Ruitao Chen, Guoyang Xie, Jiaqi Liu, Jinbao Wang, Ziqi Luo, Jinfan Wang, and Feng Zheng. EasyNet: An easy network for 3D industrial anomaly detection. In *Proceedings of the ACM International Conference on Multimedia*, pages 7038–7046, 2023.

[Chu *et al.*, 2023] 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 International Conference on Machine Learning*, pages 6185–6194, 2023.

[Costanzino *et al.*, 2024] Alex Costanzino, Pierluigi Zama Ramirez, Giuseppe Lisanti, and Luigi Di Stefano. Multimodal industrial anomaly detection by crossmodal feature mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. CVPR.

[Deng and Li, 2022] 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*, pages 9727–9736, 2022.

[Gu *et al.*, 2024] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. AnomalyGPT: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1932–1940, 2024.

[Gudovskiy *et al.*, 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*, pages 98–107, 2022.

[Jeong *et al.*, 2023] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. WinCLIP: Zero-/Few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023.

[Khattak *et al.*, 2023] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3992–4003, 2023.

[Li *et al.*, 2021] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. CutPaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021.

[Li *et al.*, 2025] Gang Li, Tianjiao Chen, Mingle Zhou, Min Li, Delong Han, and Jin Wan. MCL-AD: Multimodal collaboration learning for zero-shot 3D anomaly detection. *CoRR*, abs/2509.10282, 2025.

[Lin *et al.*, 2026] Yuxuan Lin, Hanjing Yan, Xuan Tong, Yang Chang, Huanzhen Wang, Ziheng Zhou, Shuyong Gao, Yan Wang, and Wenqiang Zhang. Commonality in few: Few-shot multimodal anomaly detection via hypergraph-enhanced memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

[Liu *et al.*, 2023] 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*, pages 20402–20411, 2023.

[Ming *et al.*, 2022] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 35, pages 35087–35102, 2022.

[Pang *et al.*, 2022] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(2):1–38, 2022.

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

[Rao *et al.*, 2022] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. DenseCLIP: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022.

[Roth *et al.*, 2022] 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*, pages 14318–14328, 2022.

[Rudolph *et al.*, 2021] Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1907–1916, 2021.

[Rudolph *et al.*, 2022] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Asymmetric student-teacher networks for industrial anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2591–2601, 2022.

[Salehi *et al.*, 2025] Alireza Salehi, Mohammadreza Salehi, Reshad Hosseini, Cees G. M. Snoek, Makoto Yamada, and Mohammad Sabokrou. Crane: Context-guided prompt learning and attention refinement for zero-shot anomaly detections. *CoRR*, abs/2504.11055, 2025.

[Sun *et al.*, 2023] Yiyou Sun, Zhenmei Shi, Yingyu Liang, and Yixuan Li. When and how does known class help discover unknown ones? provable understandings through spectral analysis. In *Proceedings of the International Conference on Machine Learning*, volume 202, pages 33014–33043, 2023.

[Tian *et al.*, 2023] Yu Tian, Fengbei Liu, Guansong Pang, Yuanhong Chen, Yuyuan Liu, Johan W Verjans, Rajvinder Singh, and Gustavo Carneiro. Self-supervised pseudo multi-class pre-training for unsupervised anomaly detection and segmentation in medical images. *Medical Image Analysis*, page 102930, 2023.

[Wang *et al.*, 2023] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. Multimodal industrial anomaly detection via hybrid fusion. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 8032–8041, 2023.

[Xie *et al.*, 2023] Guoyang Xie, Jinbao Wang, Jiaqi Liu, Yaochu Jin, and Feng Zheng. Pushing the limits of few-shot anomaly detection in industry vision: Graphcore. In *Proceedings of the International Conference on Learning Representations*, 2023.

[You *et al.*, 2022] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 35, pages 4571–4584, 2022.

[Zavrtanik *et al.*, 2021] 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*, pages 8330–8339, 2021.

[Zavrtanik *et al.*, 2022] Vitjan Zavrtanik, Matej Kristan, and Danijel Skocaj. DSR - A dual subspace re-projection network for surface anomaly detection. In *Proceedings of the European Conference on Computer Vision*, volume 13691, pages 539–554, 2022.

[Zhou *et al.*, 2022] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[Zhou *et al.*, 2024a] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In *Proceedings of the International Conference on Learning Representations*, 2024.

[Zhou *et al.*, 2024b] Qihang Zhou, Jiangtao Yan, Shibo He, Wenchao Meng, and Jiming Chen. PointAD: Comprehending 3D anomalies from points and pixels for zero-shot 3D anomaly detection. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 37, pages 84866–84896, 2024.

[Zhu *et al.*, 2023] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. PointCLIP v2: Prompting CLIP and GPT for powerful 3D open-world learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2639–2650, 2023.

[Zhu *et al.*, 2025] Jiawen Zhu, Yew-Soon Ong, Chunhua Shen, and Guansong Pang. Fine-grained abnormality prompt learning for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22241–22251, 2025.