Align3D-AD: Cross-Modal Feature Alignment and Dual-Prompt Learning for Zero-shot 3D Anomaly Detection

Letian Bai[1], **Xuanming Cao**[1], **Juan Du**[1,2]*, **Chengyu Tao**[3]
[1] Smart Manufacturing Thrust, The Hong Kong University of Science and Technology (Guangzhou)
[2] The Hong Kong University of Science and Technology
[3] College of Mechanical and Vehicle Engineering, Hunan University
lbai799@connect.hkust-gz.edu.cn, xcao743@connect.hkust-gz.edu.cn
juandu@ust.hk, taochengyu@hnu.edu.cn

Abstract

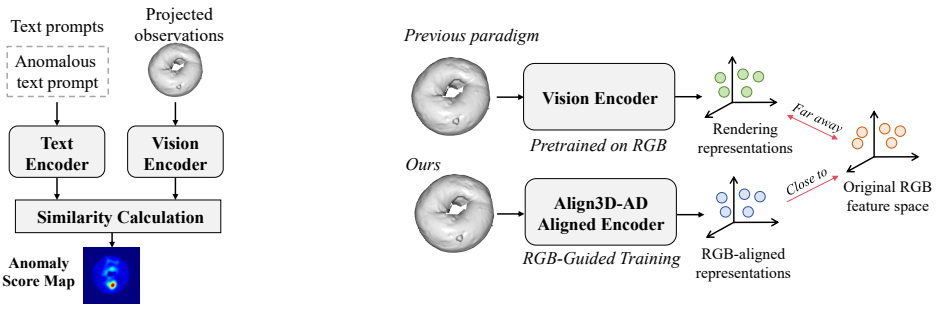
Zero-shot 3D anomaly detection aims to identify anomalies without access to training data from target categories. However, existing methods mainly rely on projecting 3D observations into multi-view representations that primarily capture geometric cues rather than realistic visual semantics and process them with vision encoders pretrained on RGB data, leading to a significant domain gap between the encoder and the projected representations. To address this issue, we propose Align3D-AD, a unified two-stage framework that leverages the RGB modality from auxiliary categories as cross-modal guidance for zero-shot 3D anomaly detection. First, we introduce a cross-modal feature alignment paradigm that maps rendering features into the RGB semantic space. Unlike prior works that implicitly rely on pretrained encoders, our method enables direct semantic transfer from RGB observations. A semantic consistency reweighting strategy is further introduced to refine feature alignment by reweighting local regions according to holistic semantic consistency. Second, we propose a modality-aware prompt learning framework with dual-prompt contrastive alignment. By assigning independent prompts to RGB-aligned and rendering features, our method captures complementary semantics across modalities, while the contrastive alignment further enhances prompt representations to improve discriminability. Extensive experiments on MVTec3D-AD, Eyecandies, and Real3D-AD demonstrate that Align3D-AD outperforms existing zero-shot methods overall under both one-vs-rest and cross-dataset settings, highlighting its generalization capability and robustness. The source code is available at https://github.com/Protein111/Align3D-AD.

1 Introduction

Surface anomaly detection plays a crucial role in industrial quality control [32, 14, 33, 6, 2, 12]. Compared to traditional 2D image inspection, which is highly susceptible to lighting conditions and lacks 3D geometric details, automated detection based on 3D point clouds has garnered increasing attention [7]. Due to the extreme scarcity of anomalies in industrial scenarios, most prevailing methods fall into the category of unsupervised approaches, learning the feature distribution of normal samples primarily through memory banks [30, 19, 35] or teacher-student architectures [1, 31]. However, these unsupervised methods still rely heavily on a large amount of normal data from the target category. When faced with practical industrial constraints such as privacy protection or the absence of new product prototypes [37, 38, 17, 16], training data for the target category is often completely inaccessible, prompting the emergence of zero-shot 3D anomaly detection tasks. Specifically, the

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Zero-shot 3D anomaly detection paradigm.

(b) Implicit domain gap reduction with cross-modal alignment.

Figure 1: Motivation of Align3D-AD. (a) An overview of the existing framework for zero-shot 3D anomaly detection. (b) A comparison between the previous paradigm and the proposed method. By incorporating RGB guidance during training, the proposed model yields RGB-aligned representations, thereby effectively mitigating the domain gap inherent in the vision encoder.

paradigm seeks to train a general model using auxiliary data, thereby obviating the requirement for training samples from the target category [13].

Although zero-shot anomaly detection has been extensively investigated in the 2D domain [20, 37, 9, 28], research on 3D remains relatively limited, with only a few recent attempts. For example, PointAD [38] integrates the hybrid representation learning of points and pixels to comprehend structural anomalies, while GS-CLIP [13] dynamically generates geometry-aware text prompts embedded with 3D priors. Fundamentally, this prevailing paradigm relies on the projection of 3D point clouds into multi-view 2D observations [8], aiming to transfer the strong generalization capability of the vision-language model (CLIP) [29] to the 3D domain (as illustrated in Figure 1a). However, the strong semantic understanding capability of CLIP is established in the RGB feature space characterized by rich appearance and texture cues, whereas multi-view renderings generated from 3D point clouds are highly abstract and contain exclusively geometric information, thereby leading to a critical domain gap in this paradigm. This cross-modal inconsistency makes it difficult for the model to extract deep discriminative features when provided solely with geometric inputs.

To bridge this gap, we propose Align3D-AD, a unified two-stage framework that leverages the RGB modality as cross-modal guidance during training. The core motivation is to utilize the RGB modality as the cross-modal guidance during the training stage (as shown in Figure 1b). Specifically, the first stage introduces the feature aligner to project rendering features into the RGB feature space, further refined by a semantic consistency reweighting strategy activated in the later phase of training. In the second stage, we propose a modality-aware prompt learning mechanism that assigns independent prompts to both aligned and rendering features, thereby capturing complementary anomaly characteristics and semantics. A dual-prompt contrastive alignment is further introduced to improve the consistency and discrimination capability of prompt representations.

The main contributions of this paper are summarized as follows:

- We propose Align3D-AD, a unified two-stage framework for zero-shot 3D anomaly detection. By introducing the RGB modality as cross-modal guidance during training, the framework learns RGB-aligned semantic representations for rendering features, enabling the model to exploit rich visual semantics even when only 3D observations are available.
- We design a simple yet effective semantic consistency reweighting strategy. This mechanism adaptively assigns local alignment weights by aggregating cross-modal semantic consistency across patches, emphasizing semantically consistent regions and enabling more robust and effective feature alignment.
- We propose a modality-aware prompt learning scheme with dual-prompt contrastive alignment. By assigning independent text prompts to both RGB-aligned and rendering branches, the framework captures modality-specific semantics and further improves the effectiveness of prompt representation learning through contrastive alignment.

- Extensive experiments on MVTec3D-AD [3], Eyecandies [4], and Real3D-AD [24] across diverse evaluation settings demonstrate that Align3D-AD outperforms existing zero-shot methods overall, with only rendering observations required during inference.

The remainder of this paper is organized as follows. Section 2 reviews the literature on unsupervised 3D anomaly detection and zero-shot anomaly detection in both 2D and 3D scenarios. Section 3 introduces the proposed Align3D-AD framework in detail. Section 4 provides the experimental setup and evaluation results. Finally, Section 5 concludes the paper, followed by discussions on limitations and broader impacts.

2 Related Work

2.1 Unsupervised 3D Anomaly Detection

Unsupervised 3D anomaly detection predominantly learns the feature distribution of normal samples through feature-based or reconstruction-based paradigms [7]. Feature-based methods evaluate anomalies by measuring representational deviations from established normal samples [14]. Within this category, memory bank approaches [19, 35, 8, 11] store nominal features extracted from 3D descriptors or multi-view renderings to compute the feature distance of test samples. Concurrently, knowledge distillation frameworks [1, 31] utilize teacher-student architectures to quantify representational discrepancies. In contrast, reconstruction-based approaches [22, 36] identify anomalies by evaluating reconstruction errors. Nevertheless, all the unsupervised paradigms strictly necessitate extensive collections of normal training data, exposing a critical limitation that zero-shot methodologies aim to resolve.

2.2 Zero-shot 2D Anomaly Detection

To address the inaccessibility of normal samples, zero-shot anomaly detection has been extensively explored using the vision-language model (CLIP) [29]. WinCLIP [20] pioneers this by leveraging the native image-text alignment of CLIP for tuning-free anomaly detection. To overcome the limitations of hand-crafted prompts, AnomalyCLIP [37] learns object-agnostic text templates to capture generic anomaly semantics across domains, while AdaCLIP [9] introduces a hybrid learnable prompt mechanism to adaptively adjust the feature space of CLIP for varying inputs. To tackle the unpredictability of anomalies, Bayes-PFL [28] models the posterior distribution of prompts via a Bayesian framework, significantly enhancing model robustness compared to deterministic approaches. Furthermore, AA-CLIP [25] constructs decoupled text anchors to guide the alignment of patch-level visual features, effectively improving the inter-class discriminability between normal and anomalous representations. More recently, AnomalyVFM [15] leverages synthetic data generation and parameter-efficient adaptation to transform pretrained vision foundation models into effective 2D zero-shot anomaly detectors.

2.3 Zero-shot 3D Anomaly Detection

Zero-shot 3D anomaly detection methods also leverage CLIP, wherein 3D data are projected and rendered into images [8] to satisfy the input requirements of CLIP. PointAD [38] projects 3D point clouds into multi-view 2D renderings and integrates hybrid representation learning of points and pixels, thereby transferring the strong generalization capability of CLIP to zero-shot 3D anomaly detection. Subsequently, MVP-PCLIP [10] converts 3D point clouds into multi-view depth maps and introduces learnable visual and adaptive text prompting to fine-tune the vision-language model, thereby enhancing detection performance. Furthermore, GS-CLIP [13] dynamically generates geometry-aware text prompts embedded with 3D geometric priors and synergistically fuses multi-view representations of rendering images and depth maps, enabling the model to accurately identify complex 3D geometric anomalies. However, these methods overlook the domain gap inherent in the rendering modality, which significantly degrades the representational capacity of visual features. In contrast, the proposed Align3D-AD framework effectively bridges this domain gap and achieves superior detection results.

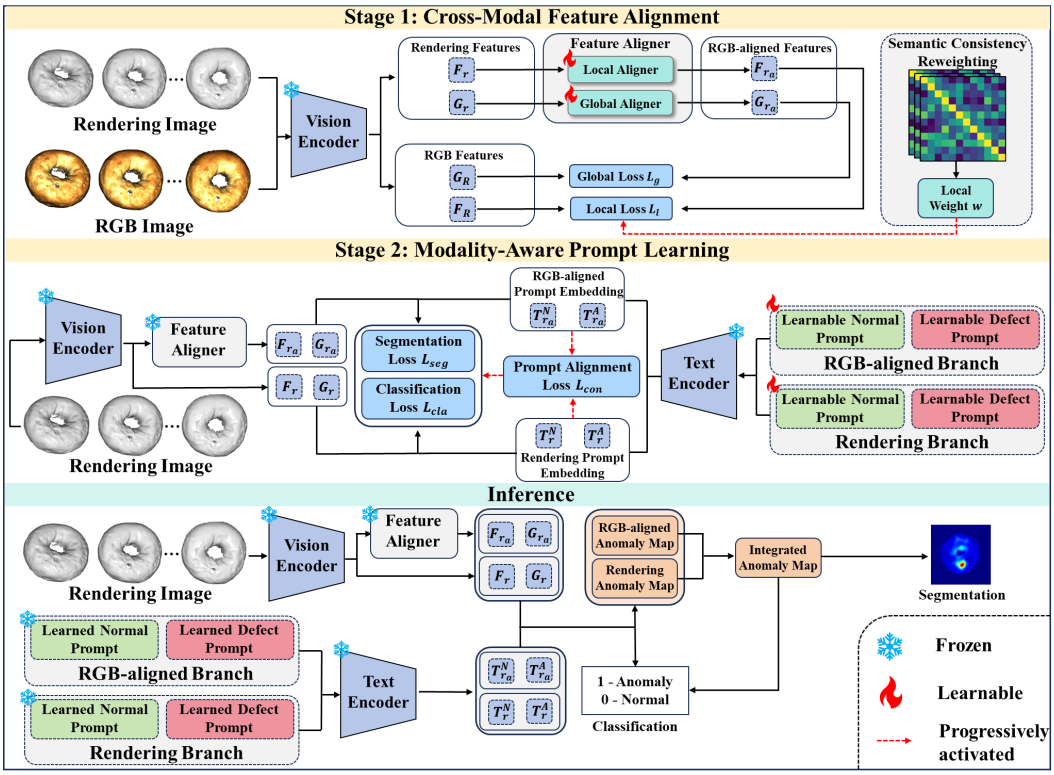


Figure 2: Overall framework of Align3D-AD. Our model consists of two stages: Cross-Modal Feature Alignment and Modality-Aware Prompt Learning. Stage 1 aligns rendering and RGB features in the RGB feature space, further refined by the semantic consistency reweighting strategy. Stage 2 performs modality-aware prompt learning for both RGB-aligned and rendering branches, with the dual-prompt contrastive alignment activated in the later phase to enhance prompt discriminability. During inference, anomaly predictions from both branches are integrated to produce the final anomaly detection results.

3 Align3D-AD

3.1 Overview

We propose Align3D-AD, a unified two-stage framework for zero-shot 3D anomaly detection that bridges the domain gap between rendering and RGB modalities and enhances anomaly discrimination through modality-aware prompt learning. An overview of the framework is shown in Figure 2. In the first stage, we perform cross-modal feature alignment to project rendering features into the RGB feature space through the learnable feature aligner. The semantic consistency reweighting strategy further improves the alignment by reweighting rendering patches based on the holistic semantic consistency across modalities. In the second stage, we perform modality-aware prompt learning for both RGB-aligned and rendering branches. To further improve the representational capability of modality-aware prompts, the dual-prompt contrastive alignment is incorporated in the later phase. By combining cross-modal alignment with structured prompt learning, Align3D-AD leverages complementary geometric and semantic cues to learn feature representations for zero-shot 3D anomaly detection.

3.2 Feature Extraction

Unlike existing zero-shot 3D anomaly detection methods that rely solely on projected observations derived from 3D point clouds, we incorporate RGB images during the training stage to provide complementary semantic cues. Specifically, given a 3D point cloud D_{3d} with P_n points, we render it from V viewpoints, obtaining multi-view rendering images $\{I_i^r\}_{i=1}^V$ and corresponding RGB images $\{I_i^R\}_{i=1}^V$. Here, $I_i^m \in \mathbb{R}^{H\times W}$ denotes the observation of modality $m \in \{R, r\}$ at the i-th view. A shared vision encoder extracts features from each view independently, producing the global feature

$G_i^m \in \mathbb{R}^d$ and local features $F_i^m = [f_{i,1}^m, f_{i,2}^m, \ldots, f_{i,N}^m]^\top \in \mathbb{R}^{N\times d}$, where $f_{i,n}^m \in \mathbb{R}^d$ corresponds to the n-th patch, N denotes the number of patches, and d is the feature dimension.

3.3 Cross-modal Feature Alignment

Feature alignment. Most large-scale vision encoders are trained on RGB images, resulting in a domain gap between the rendering and RGB modalities. To bridge this gap, we project rendering features into the RGB semantic space through the learnable feature aligner, enabling better learned semantic representations and reducing cross-modal discrepancies.

Formally, the rendering global feature is projected as $G_i^{r_a} = \mathcal{A}_g(G_i^r)$, and the local features are projected as $F_i^{r_a} = \mathcal{A}_l(F_i^r)$, where r_a denotes the RGB-aligned representation, i.e., rendering features projected into the RGB feature space, and $\mathcal{A}_g(\cdot)$ and $\mathcal{A}_l(\cdot)$ denote learnable MLPs. To enforce cross-modal consistency, we align the projected features with their RGB counterparts using cosine similarity. The global alignment loss $\mathcal{L}_g$ and the local alignment loss $\mathcal{L}_l$ are defined as:

$$\mathcal{L}_g = \frac{1}{V}\sum_{i=1}^{V}(1 - \langle G_i^{r_a}, G_i^R\rangle), \quad \mathcal{L}_l = \frac{1}{V}\sum_{i=1}^{V}\frac{1}{N}\sum_{n=1}^{N}(1 - \langle f_{i,n}^{r_a}, f_{i,n}^R\rangle), \tag{1}$$

where $\langle\cdot,\cdot\rangle$ denotes cosine similarity. The overall alignment loss is defined as $\mathcal{L}_{\text{align}} = \mathcal{L}_g + \mathcal{L}_l$, which enforces global and local feature consistency simultaneously.

Semantic consistency reweighting. Although feature alignment enforces consistency in the RGB feature space, it relies on local matching and treats all patches equally, which may be suboptimal in the presence of cross-modal discrepancies. This motivates us to move beyond local correspondence and incorporate holistic semantic information. To this end, we propose a simple yet effective reweighting strategy that adaptively modulates the contribution of each rendering patch based on the holistic semantic consistency with the RGB modality. To ensure stability, the reweighting strategy is applied only to the local alignment loss $\mathcal{L}_l$ and is activated in the later phase of cross-modal feature alignment.

Specifically, we compute the semantic consistency matrix $C_i = F_i^r(F_i^R)^\top \in \mathbb{R}^{N\times N}$ for each view i, where $(C_i)_{n,j}$ measures the similarity between the n-th rendering patch and the j-th RGB patch. For each rendering patch $f_{i,n}^r$, we aggregate its consistency with all RGB patches as $c_{i,n} = \frac{1}{N}\sum_{j=1}^{N}(C_i)_{n,j}$. Based on this score, we compute the local weight $w_{i,n} = 1 + \lambda \cdot \sigma(c_{i,n})$, where $\sigma(\cdot)$ is the sigmoid function and λ controls the reweighting strength. The reweighted local alignment loss is defined as:

$$\mathcal{L}_l^w = \frac{1}{V}\sum_{i=1}^{V}\frac{1}{N}\sum_{n=1}^{N} w_{i,n}(1 - \langle f_{i,n}^{r_a}, f_{i,n}^R\rangle). \tag{2}$$

The overall reweighted alignment loss in this phase is defined as $\mathcal{L}_{\text{align}}^w = \mathcal{L}_g + \mathcal{L}_l^w$. Consequently, rendering patches exhibiting higher holistic semantic consistency with the RGB modality receive larger alignment weights. By leveraging aggregated cross-modal semantic interactions, the proposed strategy adaptively emphasizes such regions during feature alignment, leading to more robust and discriminative representations.

3.4 Modality-aware Prompt Learning

Building upon the aligned features, we aim to learn anomaly semantics in a zero-shot manner. Accordingly, we propose a modality-aware prompt learning strategy that models anomaly representations for both RGB-aligned and rendering branches in separate semantic spaces.

Dual-prompt learning. Following PointAD [38], we construct learnable text prompts to model normal and anomalous semantics. Nevertheless, due to the discrepancy between RGB-aligned and rendering features, a shared prompt space is insufficient to capture modality-specific characteristics. We therefore introduce modality-specific prompts t_m^N and t_m^A for $m \in \{r_a, r\}$, corresponding to normal and anomalous states. These prompts are encoded into text embeddings T_m^N and T_m^A via a text encoder. Based on text embeddings, the anomaly prediction for the i-th view is computed as

$$\hat{y}_i^m = \frac{\exp(\langle G_i^m, T_m^A\rangle/\tau)}{\exp(\langle G_i^m, T_m^N\rangle/\tau) + \exp(\langle G_i^m, T_m^A\rangle/\tau)},$$

where τ denotes the temperature coefficient. The final prediction $\hat{y}^m$ is obtained by averaging the predictions across all views. For anomaly segmentation, we compute 2D anomaly score maps as

$$M_{i,A}^m = \frac{\exp(\langle F_i^m, T_m^A \rangle)}{\exp(\langle F_i^m, T_m^N \rangle) + \exp(\langle F_i^m, T_m^A \rangle)}.$$

These maps are then back-projected to 3D space as $M_A^m = \frac{1}{V} \sum_{i=1}^{V} R_i^{-1}(\mathrm{Up}(M_{i,A}^m)) \odot H_i$, where $\mathrm{Up}(\cdot)$ denotes bilinear interpolation, $R_i^{-1}(\cdot)$ denotes the inverse projection from the i-th view to 3D space, $\odot$ denotes element-wise multiplication, and H_i is the visibility mask for the corresponding view. The same procedure is applied to obtain $M_{i,N}^m$ and M_N^m for the normal state.

The supervision consists of image-level labels $y \in \{0, 1\}$, point-level annotations $Y \in \{0, 1\}^{P_n}$, and 2D masks $Y_i \in \{0, 1\}^{H \times W}$. We jointly optimize classification and segmentation objectives for each modality m, denoted as $\mathcal{L}_{cls}^m$ and $\mathcal{L}_{seg}^m$, respectively. For image-level anomaly classification, we define the classification loss based on Cross-entropy (CE) as

$$\mathcal{L}_{cls}^m = \frac{1}{V} \sum_{i=1}^{V} \mathrm{CE}(\max\{Y_i\}, \hat{y}_i^m) + \mathrm{CE}(y, \hat{y}^m).$$

For anomaly segmentation, we supervise both 3D point-level predictions and 2D pixel-level score maps, and the segmentation objectives at the 3D and 2D levels are defined using the Dice loss [27] and the Focal loss [23] as

$$\mathcal{L}_{seg}^{m,3D} = \mathrm{Dice}(M_A^m, Y) + \mathrm{Dice}(M_N^m, I - Y),$$

$$\mathcal{L}_{seg}^{m,2D} = \frac{1}{V} \sum_{i=1}^{V} \mathrm{Dice}(M_{i,A}^m, Y_i) + \mathrm{Dice}(M_{i,N}^m, I - Y_i) + \mathrm{Focal}(M_{i,N}^m \oplus M_{i,A}^m, Y_i),$$

where I is an all-one matrix and $\oplus$ denotes channel-wise concatenation. Therefore, the overall segmentation loss is $\mathcal{L}_{seg}^m = \mathcal{L}_{seg}^{m,3D} + \mathcal{L}_{seg}^{m,2D}$, and the final objective is given by $\mathcal{L} = \sum_{m \in \{r_a, r\}} \left(\mathcal{L}_{cls}^m + \mathcal{L}_{seg}^m \right)$.

Dual-prompt contrastive alignment. Although modality-specific prompts are optimized independently for each branch, they still exhibit limited semantic expressiveness. To address this limitation, we introduce a dual-prompt contrastive alignment that enhances prompt discriminability in the semantic space while preserving cross-modal consistency.

Concretely, for each modality $m \in \{r_a, r\}$, the prompts are optimized while treating the complementary modality $\bar{m}$ as a fixed reference. The contrastive alignment loss for modality m is defined as:

$$\begin{aligned} \mathcal{L}_{con}^m = \frac{1}{2} \Bigg[& - \log \frac{\exp(\langle T_m^N, T_{\bar{m}}^N \rangle)}{\exp(\langle T_m^N, T_{\bar{m}}^N \rangle) + \exp(\langle T_m^N, T_m^A \rangle)} \\ & - \log \frac{\exp(\langle T_m^A, T_{\bar{m}}^A \rangle)}{\exp(\langle T_m^A, T_{\bar{m}}^A \rangle) + \exp(\langle T_m^A, T_m^N \rangle)} \Bigg], \end{aligned} \tag{3}$$

The overall objective in this phase is defined as $\mathcal{L} = \sum_{m \in \{r_a, r\}} \left(\mathcal{L}_{cls}^m + \mathcal{L}_{seg}^m + \lambda_{con} \mathcal{L}_{con}^m \right)$. To ensure stable optimization, this regularization is activated in the later phase of modality-aware prompt learning, after the prompts have captured coarse normal and anomalous semantics.

3.5 Training and Inference

Training. We adopt a two-stage training scheme with progressively refined objectives. In the first stage, the model is initially optimized using the alignment loss $\mathcal{L}_{align}$ and subsequently refined with the reweighted alignment loss $\mathcal{L}_{align}^w$ when the reweighting strategy is activated. In the second stage, we jointly optimize classification and segmentation objectives $\mathcal{L} = \sum_{m \in \{r_a, r\}} \left(\mathcal{L}_{cls}^m + \mathcal{L}_{seg}^m \right)$ and further incorporate the contrastive loss $\mathcal{L}_{con}$ in the later phase of modality-aware prompt learning to refine text prompt representations.

Inference. Following Section 3.4, 3D anomaly score maps M_A^m are computed for both modalities, yielding $M_A^{r_a}, M_A^r \in \mathbb{R}^{P_n}$ and corresponding image-level predictions $\hat{y}^{r_a}$ and $\hat{y}^r$. The final 3D

anomaly score map and image-level prediction are obtained via linear fusion $M_A = \alpha M_A^{r_a} + (1-\alpha)M_A^{r}$, $\hat{y} = \alpha\hat{y}^{r_a} + (1-\alpha)\hat{y}^{r} + \max(M_A)$, where $\alpha \in [0,1]$ controls the contribution of each feature representation, and $\max(M_A)$ denotes the maximum anomaly score in M_A.

4 Experiment

4.1 Experiment Setup

Datasets and evaluation settings. We evaluate our method under two settings: one-vs-rest and cross-dataset. In the one-vs-rest setting, experiments are conducted on MVTec3D-AD [3] and Eyecandies [4], where one object class is used as auxiliary training data and the remaining classes are used for evaluation. In the cross-dataset setting, the model is trained on a single category from MVTec3D-AD [3] and evaluated on Eyecandies [4] and Real3D-AD [24] to assess cross-dataset generalization. All experiments are repeated three times, and the average results are reported. More details on dataset descriptions, evaluation settings, and visualizations of the rendered multi-view observations are provided in Appendix A.

Evaluation metrics. We adopt standard metrics to evaluate both classification and segmentation performance. For classification, we report the Area Under the Receiver Operating Characteristic curve (AUROC) and Average Precision (AP), denoted as O-R and O-A, respectively. For segmentation, we report AUROC and Per-Region Overlap (PRO), denoted as P-R and P-P, respectively.

Implementation details. We adopt CLIP ViT-L/14@336px as the visual encoder. We generate 9 views for each object, where each view consists of both rendering and RGB observations. Our model is trained in two stages. In the first stage, we perform feature alignment for 250 epochs. The feature aligner consists of two separate MLPs, each with three layers of dimension 768 and GELU activation [18]. The semantic consistency reweighting strategy is activated in the final 50 epochs. In the second stage, we conduct modality-aware prompt learning for 15 epochs. The dual-prompt contrastive alignment is introduced in the final 5 epochs. We use the Adam optimizer with a learning rate of 1×10^{-3} for both stages. All experiments are conducted on NVIDIA RTX 4090 GPUs. More details are provided in Appendix B.

Table 1: One-vs-rest results on MVTec3D-AD and Eyecandies. Best results in **bold**. * indicates results reproduced by ourselves.

Model	MVTec3D-AD				Eyecandies			
	O-R	O-A	P-R	P-P	O-R	O-A	P-R	P-P
CLIP [29]	61.2	85.8	80.3	54.4	66.7	69.2	81.2	37.9
AA-CLIP [25]	74.2	88.5	87.1	61.6	66.5	68.0	85.9	39.3
MVP-PCLIP [10]	81.3	92.7	94.6	83.6	69.3	72.7	90.8	67.8
GS-CLIP* [13]	78.7	92.9	94.6	80.6	68.3	73.4	92.3	72.0
PointAD [38]	82.0	94.2	95.5	84.4	69.1	73.8	92.1	71.3
ZUMA [26]	81.9	-	95.6	**87.0**	67.1	-	92.5	73.6
Align3D-AD	**83.0**	**94.3**	**95.9**	86.5	**71.3**	**76.6**	**92.6**	**74.2**

Table 2: Cross-dataset results on Eyecandies and Real3D-AD. Best results in **bold**. * indicates results reproduced by ourselves.

Model	Eyecandies				Real3D-AD		
	O-R	O-A	P-R	P-P	O-R	O-A	P-R
AA-CLIP [25]	63.6	65.3	82.6	39.8	70.2	71.7	55.2
MVP-PCLIP [10]	66.7	70.7	88.3	66.0	74.9	75.6	73.9
GS-CLIP* [13]	70.1	73.9	91.4	70.4	74.8	77.5	72.6
PointAD [38]	69.5	74.3	91.8	71.4	75.9	77.9	71.6
Align3D-AD	**70.7**	**75.3**	**92.0**	**72.6**	**76.1**	**79.3**	**76.7**

Table 3: Ablation study of key components. * denotes RGB-aligned feature representation only, and † denotes rendering feature representation only.

Component				Metric			
DpL	FA	SCR	DpCA	O-R	O-A	P-R	P-P
✔	✗	✗	✗	81.9	94.0	95.3	83.7
✔	✔*	✗	✗	79.8	93.0	95.6	84.2
✔	✔†	✗	✗	82.0	94.0	95.2	83.4
✔	✔	✗	✗	82.4	94.0	95.7	85.6
✔	✔	✔	✗	82.6	94.1	95.6	85.1
✔	✔	✗	✔	82.6	94.2	**95.9**	86.3
✔	✔	✔	✔	**83.0**	**94.3**	**95.9**	**86.5**

Table 4: Ablation study of different feature representations under the one-vs-rest setting.

Modality	MVTec3D-AD				Eyecandies			
	O-R	O-A	P-R	P-P	O-R	O-A	P-R	P-P
RGB-aligned	80.1	93.2	95.6	84.3	69.5	73.2	92.5	73.5
Rendering	82.5	94.2	95.6	84.7	70.6	76.1	92.3	73.3
Align3D-AD	**83.0**	**94.3**	**95.9**	**86.5**	**71.3**	**76.6**	**92.6**	**74.2**

4.2 Main Results

One-vs-rest results. As shown in Table 1, our method outperforms 3D-based approaches overall. Compared with PointAD, which solely relies on rendering images, our method achieves improvements

of +1.0% in O-R, +0.1% in O-A, +0.4% in P-R, and +2.1% in P-P on MVTec3D-AD, while also yielding notable gains on Eyecandies. Compared with ZUMA, although our performance on MVTec3D-AD is slightly lower than ZUMA for the P-P metric, the overall performance remains superior, indicating stronger anomaly discrimination at both object and point levels. The performance gains can be attributed to three key components. First, the introduction of RGB-aligned features effectively bridges the domain gap between rendering and RGB modalities. Second, the proposed dual-prompt learning strategy enables complementary modeling of anomaly semantics across modalities. Third, the contrastive alignment further facilitates effective modality-aware prompt learning by enforcing structured relationships between normal and anomalous representations. Together, these designs result in more accurate anomaly detection.

Qualitative analysis. Figure 3 presents qualitative comparisons among rendering, RGB-aligned, and integrated predictions, highlighting their complementary roles in anomaly detection. As shown in the first row, rendering features effectively capture geometric and structural information, enabling the detection of anomalies that are less distinguishable in the RGB-aligned space. In contrast, RGB-aligned features demonstrate a stronger advantage by leveraging richer semantic cues in other cases. The second row shows that RGB-aligned features can also achieve more accurate localization, enabling more precise delineation of anomaly boundaries and reducing excessive erroneous segmentation, whereas rendering predictions are less discriminative. This variation arises because rendering representations effectively encode geometric structures but have limited semantic expressiveness. Therefore, although rendering features exhibit advantages for certain samples, incorporating RGB-aligned features yields superior overall performance and produces more accurate anomaly predictions than using rendering-only feature representations.

Cross-dataset results. We further evaluate the zero-shot generalization capability of our model under the cross-dataset setting. As shown in Table 2, our method consistently outperforms existing approaches on both Eyecandies and Real3D-AD, demonstrating strong transferability across datasets with varying scene distributions. Methods relying solely on 3D information exhibit suboptimal generalization when encountering unseen semantics and appearance variations. In contrast, our method integrates semantic information aligned with RGB observations and geometric information gained from rendering representations, resulting in more robust anomaly detection across domains. Notably, even though Real3D-AD does not provide RGB data, our model still achieves strong performance, highlighting the importance of leveraging RGB-aligned representations to compensate for the limited semantic information.

4.3 Ablation Study

Ablation of key components. To evaluate the effectiveness of each component, we conduct ablation studies on MVTec3D-AD by progressively enabling different modules, including dual-prompt learning (DpL), feature alignment (FA), semantic consistency reweighting (SCR), and dual-prompt contrastive alignment (DpCA), with results shown in Table 3. Under the setting with only DpL, the performance is relatively limited across all metrics. After introducing FA, overall performance is improved, with the integration of RGB-aligned and rendering features yielding more significant gains than using either feature representation alone. Building upon the integrated setting, SCR further enhances feature representations by emphasizing holistic semantic consistency, while DpCA improves the discriminability of prompt representations. With all components enabled, the model achieves the best performance, validating the effectiveness of our design.

Feature representation analysis. To evaluate the effectiveness of feature alignment, we compare three configurations: using RGB-aligned representations, rendering representations, and integrated representations on MVTec3D-AD and Eyecandies. As shown in Table 4, relying on a single representation source leads to suboptimal performance. Compared to rendering features, RGB-aligned features achieve slightly lower performance while providing complementary semantic information. Nevertheless, incorporating RGB-aligned features consistently improves performance over using rendering features alone across all metrics on both datasets. These results demonstrate that RGB-aligned and rendering features are complementary, highlighting the effectiveness of the proposed cross-modal feature alignment strategy in bridging the domain gap. We provide more analysis in Appendix C and Appendix E to further evaluate the effectiveness of RGB-aligned representations and illustrate how the semantic consistency reweighting strategy improves the cross-modal feature alignment for learning more effective RGB-aligned feature representations.

Impact of optimization epochs for SCR and DpCA. Table 5 and Table 6 present the impact of different optimization epochs for SCR and DpCA, respectively. As shown in these tables, the performance varies with the number of optimization epochs for both SCR and DpCA. For SCR, performance is suboptimal with a small number of optimization epochs, reaches its peak at 50 epochs, and then slightly degrades with additional optimization epochs, as it tends to overemphasize semantic information and suppress geometric cues, leading to excessive alignment. Similarly, for DpCA, performance peaks at 5 epochs and gradually declines with further optimization, which can be attributed to the prompt space becoming overly constrained under prolonged training. These results suggest that an appropriate number of optimization epochs is critical, as insufficient training limits effectiveness while excessive optimization may harm generalization. Nevertheless, the performance remains relatively stable under different optimization epochs, indicating the robustness of our model.

Table 5: Impact of SCR optimization epochs.

Optimization Epochs	O-R	O-A	P-R	P-P
10	82.4	94.1	95.7	85.6
30	82.6	94.1	95.5	85.0
50	**83.0**	**94.3**	**95.9**	**86.5**
100	82.5	94.1	95.6	85.3
200	82.4	94.1	95.4	84.6
250	82.4	94.1	95.8	86.0

Table 6: Impact of DpCA optimization epochs.

Optimization Epochs	O-R	O-A	P-R	P-P
1	82.6	94.2	95.7	85.7
3	82.6	94.1	95.7	85.7
5	**83.0**	**94.3**	**95.9**	**86.5**
8	82.7	94.2	95.7	85.9
10	82.5	94.1	95.7	85.8
15	82.2	94.0	95.3	84.1

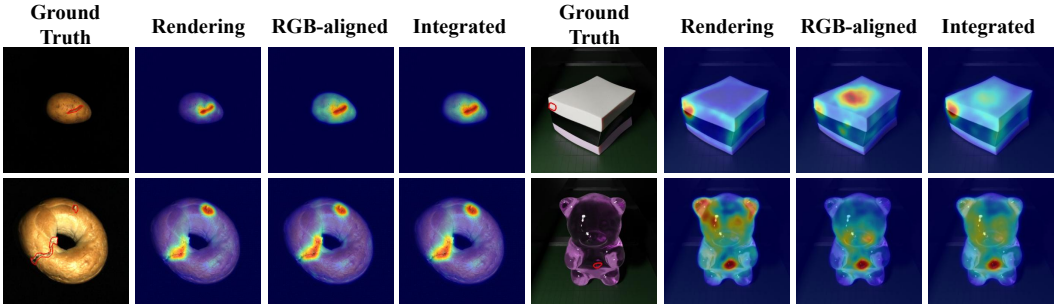


Figure 3: Qualitative visualization on MVTec3D-AD and Eyecandies.

Impact of different CLIP backbones. We further investigate the impact of different CLIP backbone sizes on MVTec3D-AD. As shown in Table 7, Align3D-AD achieves the best performance with CLIP ViT-L/14@336px, while the performance gradually decreases when using smaller backbones or lower input resolutions. This is expected since smaller backbones generally provide less discriminative visual-semantic representations. We further compare Align3D-AD with existing zero-shot 3D anomaly detection methods using the same CLIP ViT-B/32 backbone. As shown in Table 8, Align3D-AD consistently outperforms PointAD and GS-CLIP across all evaluation metrics, demonstrating the effectiveness of our cross-modal feature alignment and modality-aware prompt learning.

Table 7: Impact of different CLIP backbones on MVTec3D-AD.

Backbone	O-R	O-A	P-R	P-P
ViT-L/14@336px	**83.0**	**94.3**	**95.9**	**86.5**
ViT-L/14	77.9	92.3	95.0	82.7
ViT-B/16	74.6	90.7	94.5	81.2
ViT-B/32	72.2	90.1	92.3	73.4

Table 8: Comparison with existing methods using CLIP ViT-B/32 on MVTec3D-AD.

Model	O-R	O-A	P-R	P-P
PointAD	71.8	90.0	91.5	67.9
GS-CLIP	70.4	89.1	91.9	69.8
Align3D-AD	**72.2**	**90.1**	**92.3**	**73.4**

Generalization to other vision-language backbones. We further replace CLIP with SigLIP2-L/16 [34] and TIPSv2-L/14 [5] while keeping the training protocols unchanged. As shown in Tables 9 and 10, stronger vision-language backbones do not consistently improve zero-shot 3D anomaly detection performance. Specifically, stronger RGB-pretrained semantic representations do not necessarily transfer better to rendering observations, which exhibit a substantial domain gap from natural RGB images. Nevertheless, stronger backbones can still benefit fine-grained localization. For example, TIPSv2-L/14 improves the P-R of Align3D-AD from 95.9 to 97.1. This observation is also

consistent with recent findings in SubspaceAD [21], indicating that increased backbone capacity does not necessarily translate into improved anomaly detection performance.

Table 9: Results with SigLIP2-L/16 backbone on MVTec3D-AD.

Model	O-R	O-A	P-R	P-P
PointAD	70.1	88.9	95.7	85.0
GS-CLIP	69.2	88.1	95.7	84.5
Align3D-AD	**73.7**	**90.5**	**95.8**	**85.5**

Table 10: Results with TIPSv2-L/14 backbone on MVTec3D-AD.

Model	O-R	O-A	P-R	P-P
PointAD	80.1	93.2	96.7	85.3
GS-CLIP	77.5	92.1	96.6	84.8
Align3D-AD	**80.8**	**93.5**	**97.1**	**86.0**

Complexity analysis. We investigate the computational efficiency of mainstream methods, including PointAD and GS-CLIP, in terms of inference time per image, frames per second (FPS), and GPU memory usage, with results reported in Table 11. Compared with PointAD, which relies on rendering images, our method exhibits slightly higher inference time per image and GPU memory usage, while exhibiting a slightly lower FPS. Compared with GS-CLIP, which utilizes point clouds, rendering images, and depth maps, our method achieves consistently better efficiency. Overall, our method obtains the best performance, achieving a more favorable trade-off between computational cost and accuracy. More details are provided in Appendix K.

Table 11: Comparison of inference time per image, frame per second (FPS), and GPU memory usage on MVTec3D-AD.

Model	Time (s)	FPS	Memory (MB)
GS-CLIP	0.62	1.62	4978
PointAD	0.18	5.68	2460
Ours	0.19	5.30	2791

5 Conclusion

This paper studies the challenging problem of zero-shot 3D anomaly detection. We propose Align3D-AD, a unified framework that bridges rendering and RGB modalities to learn transferable anomaly representations. By introducing cross-modal feature alignment, the model transfers semantic knowledge from RGB observations to rendering features. A semantic consistency reweighting strategy is further employed to enhance cross-modal feature alignment by leveraging holistic semantic consistency. Furthermore, a modality-aware prompt learning scheme with dual-prompt contrastive alignment is developed to capture modality-specific anomaly semantics and enhance the discriminability of prompt representations. Extensive experiments on multiple benchmarks demonstrate that our method achieves superior performance under both one-vs-rest and cross-dataset settings, highlighting its strong generalization capability and effectiveness.

Limitations. In practice, introducing RGB guidance increases model complexity due to the additional modality and modality-specific prompt learners. Moreover, the reliance on multi-view observations makes the model sensitive to rendering quality, which may affect performance.

Broader impact. This work takes an initial step toward bridging the domain gap in zero-shot 3D anomaly detection by explicitly incorporating RGB information into the training stage. We hope this perspective can encourage future research to move beyond purely 3D information and focus more on cross-modal understanding for 3D tasks.

Acknowledgments and Disclosures

Funding. This work was supported in part by the National Natural Science Foundation of China under Grant No. 72371219, and Guangdong Project under Grant No. 2024TQ08A432. (Corresponding author: Juan Du).

Competing Interests. The authors declare that they have no financial or non-financial competing interests that could be perceived as influencing the research presented in this paper.

References

[1] 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.

[2] 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 (CVPR)*, pages 4183–4192, 2020.

[3] 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 (VISIGRAPP 2022) - Volume 5: VISAPP*, pages 202–213. INSTICC, SciTePress, 2022. ISBN 978-989-758-555-5. doi: 10.5220/0010865000003124.

[4] 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.

[5] Bingyi Cao, Koert Chen, Kevis-Kokitsi Maninis, Kaifeng Chen, Arjun Karpur, Ye Xia, Sahil Dua, Tanmaya Dabral, Guangxing Han, Bohyung Han, Joshua Ainslie, Alex Bewley, Mithun Jacob, René Wagner, Washington Ramos, Krzysztof Choromanski, Mojtaba Seyedhosseini, Howard Zhou, and Andre Araujo. TIPSv2: Advancing vision-language pretraining with enhanced patch-text alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 29325–29335, 2026.

[6] Tri Cao, Jiawen Zhu, and Guansong Pang. Anomaly detection under distribution shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6511–6523, 2023.

[7] Xuanming Cao, Chengyu Tao, Yifeng Cheng, and Juan Du. IAENet: An importance-aware ensemble model for 3D point cloud-based anomaly detection. *Information Fusion*, 130:104097, 2026. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2025.104097.

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

[9] Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. AdaCLIP: Adapting CLIP with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, pages 55–72. Springer, 2024.

[10] Yuqi Cheng, Yunkang Cao, Guoyang Xie, Zhichao Lu, and Weiming Shen. Toward zero-shot point cloud anomaly detection: a multiview projection framework. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 56(3):1747–1760, 2026. doi: 10.1109/TSMC.2025.3648581.

[11] 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*, pages 6185–6194, 2023.

[12] Niv Cohen and Yedid Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020.

[13] Zehao Deng, An Liu, and Yan Wang. GS-CLIP: Zero-shot 3D anomaly detection by geometry-aware prompt and synergistic view representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 35587–35596, June 2026.

[14] Juan Du, Chengyu Tao, Xuanming Cao, and Fugee Tsung. 3D vision-based anomaly detection in manufacturing: A survey. *Frontiers of Engineering Management*, 12(2):343–360, 2025.

[15] Matic Fučka, Vitjan Zavrtanik, and Danijel Skočaj. AnomalyVFM – transforming vision foundation models into zero-shot anomaly detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 35555–35566, 2026.

[16] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Hao Li, Ming Tang, and Jinqiao Wang. FiLo: Zero-shot anomaly detection by fine-grained description and high-quality localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2041–2049, 2024.

[17] 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*, volume 38, pages 1932–1940, 2024.

[18] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

[19] 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 (CVPR) Workshops*, pages 2968–2977, June 2023.

[20] 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 (CVPR)*, pages 19606–19616, 2023.

[21] Chris Lendering, Eren Akdag, and Egor Bondarev. SubspaceAD: Training-free few-shot anomaly detection via subspace modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28557–28566, 2026.

[22] 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 (CVPR)*, pages 22207–22216, 2024.

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[24] 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. *Advances in Neural Information Processing Systems*, 36:30402–30415, 2023.

[25] Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S Kevin Zhou. AA-CLIP: Enhancing zero-shot anomaly detection via anomaly-aware CLIP. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4744–4754, 2025.

[26] Yunfeng Ma, Min Liu, Shuai Jiang, Jingyu Zhou, Yuan Bian, Xueping Wang, and Yaonan Wang. ZUMA: Training-free zero-shot unified multimodal anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(6):6601–6614, 2026. doi: 10.1109/TPAMI.2026.3658856.

[27] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.

[28] Zhen Qu, Xian Tao, Xinyi Gong, Shichen Qu, Qiyu Chen, Zhengtao Zhang, Xingang Wang, and Guiguang Ding. Bayesian prompt flow learning for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 30398–30408, 2025.

[29] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[30] 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 (CVPR)*, pages 14318–14328, 2022.

[31] 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 2592–2602, 2023.

[32] Chengyu Tao and Juan Du. PointSGRADE: Sparse learning with graph representation for anomaly detection by using unstructured 3D point cloud data. *IISE Transactions*, 57(2):131–144, 2025.

[33] 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*, pages 20551–20560, 2025.

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

[35] 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 (CVPR)*, pages 8032–8041, 2023.

[36] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Cheating depth: Enhancing 3D surface anomaly detection via depth simulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2164–2172, 2024.

[37] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 49705–49737, 2024.

[38] 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. *Advances in Neural Information Processing Systems*, 37:84866–84896, 2024.

A Dataset

Dataset choice and description. We conduct experiments on three publicly available 3D anomaly detection datasets: MVTec3D-AD, Eyecandies, and Real3D-AD. MVTec3D-AD contains high-resolution 3D scans from 10 object categories with diverse semantics and provides both point-level annotations and corresponding RGB images. Eyecandies is a synthetic dataset with 10 categories, also offering paired RGB observations. Real3D-AD is a real-world industrial dataset covering 12 categories, containing both normal and defective instances but without RGB information.

Since our framework leverages RGB information during training, we use MVTec3D-AD and Eyecandies for training and evaluation under the one-vs-rest setting. In addition, we consider a more challenging cross-dataset setting, where auxiliary data from one dataset is used to evaluate all categories in another dataset. Specifically, the model is trained on MVTec3D-AD and evaluated on Eyecandies and Real3D-AD. Notably, despite the absence of RGB data in Real3D-AD, our model still generalizes effectively, as the cross-modal alignment learned during training enables rendering features to produce RGB-aligned representations, thereby compensating for the limited semantic expressiveness of 3D data.

Evaluation settings. Under the one-vs-rest setting, we report the average performance by alternately selecting different object categories as the auxiliary training set, including *Carrot*, *Cookie*, and *Dowel* for MVTec3D-AD, and *Confetto*, *Licorice Sandwich*, and *Peppermint Candy* for Eyecandies, with the average performance reported.

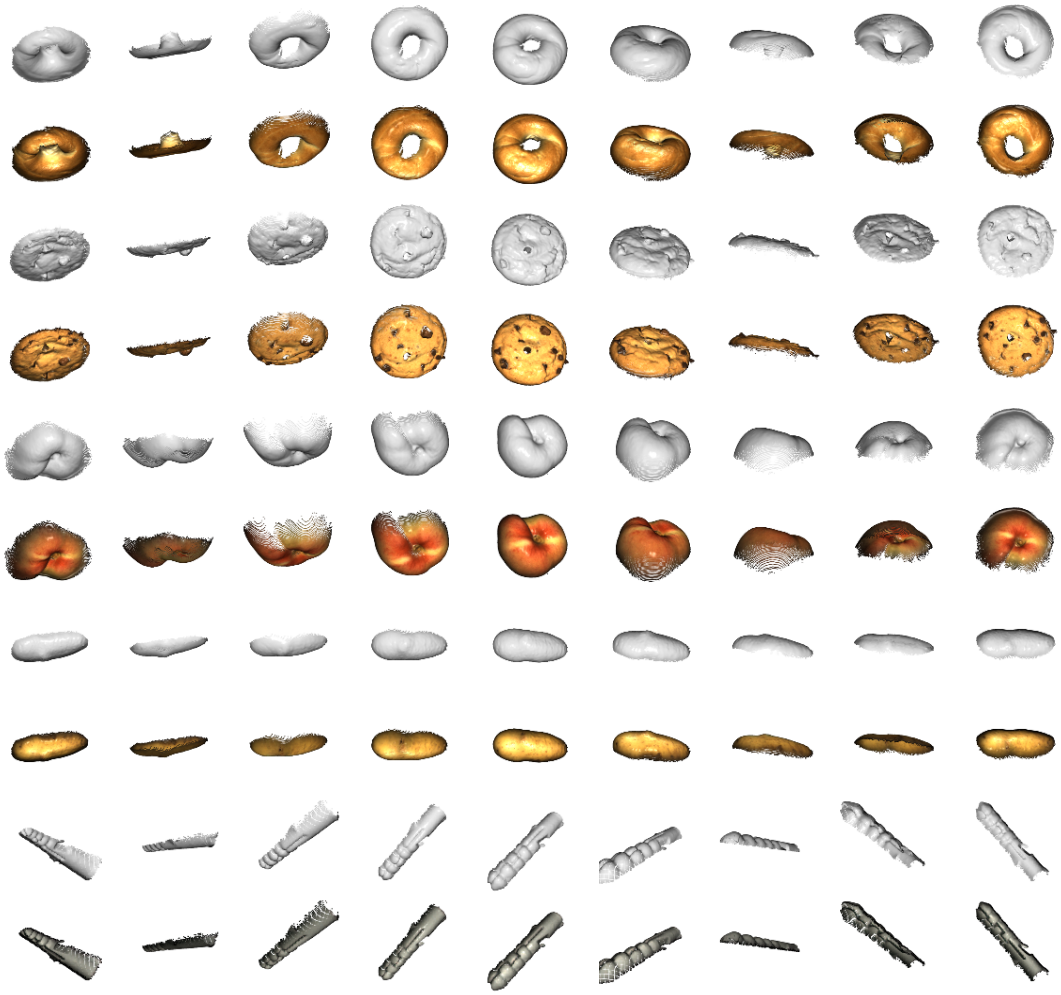


Figure 4: Visualization of the rendered multi-view observations on MVTec3D-AD, including both rendering and corresponding RGB images from different views.

For the cross-dataset setting, the model is trained on the *Carrot*, *Cookie*, and *Dowel* categories from MVTec3D-AD and directly evaluated on all categories of Eyecandies and Real3D-AD, with the average performance reported.

Visualization of the rendered multi-view datasets. We present visualizations of the rendered observations used in our framework to illustrate the characteristics of the multi-view observations, as shown in Figure 4 for MVTec3D-AD and Figure 5 for Eyecandies. For each 3D point cloud sample, rendering images along with corresponding RGB observations are generated from diverse viewpoints. From these visualizations, it can be observed that rendering images primarily capture geometric structures and shape information, whereas RGB observations provide richer semantic and appearance cues. By leveraging this complementary information, Align3D-AD constructs RGB-aligned representations to bridge the domain gap between rendering and RGB modalities, leading to more discriminative and robust anomaly detection.

However, due to the inherent characteristics of the collected point cloud data, the multi-view observations obtained through the rendering process may exhibit missing regions, blank areas, and noise artifacts. These imperfections introduce additional challenges for both cross-modal feature alignment and zero-shot 3D anomaly detection, as establishing reliable correspondences across modalities becomes more difficult under incomplete and noisy observations. Further analysis is provided in Appendix J.

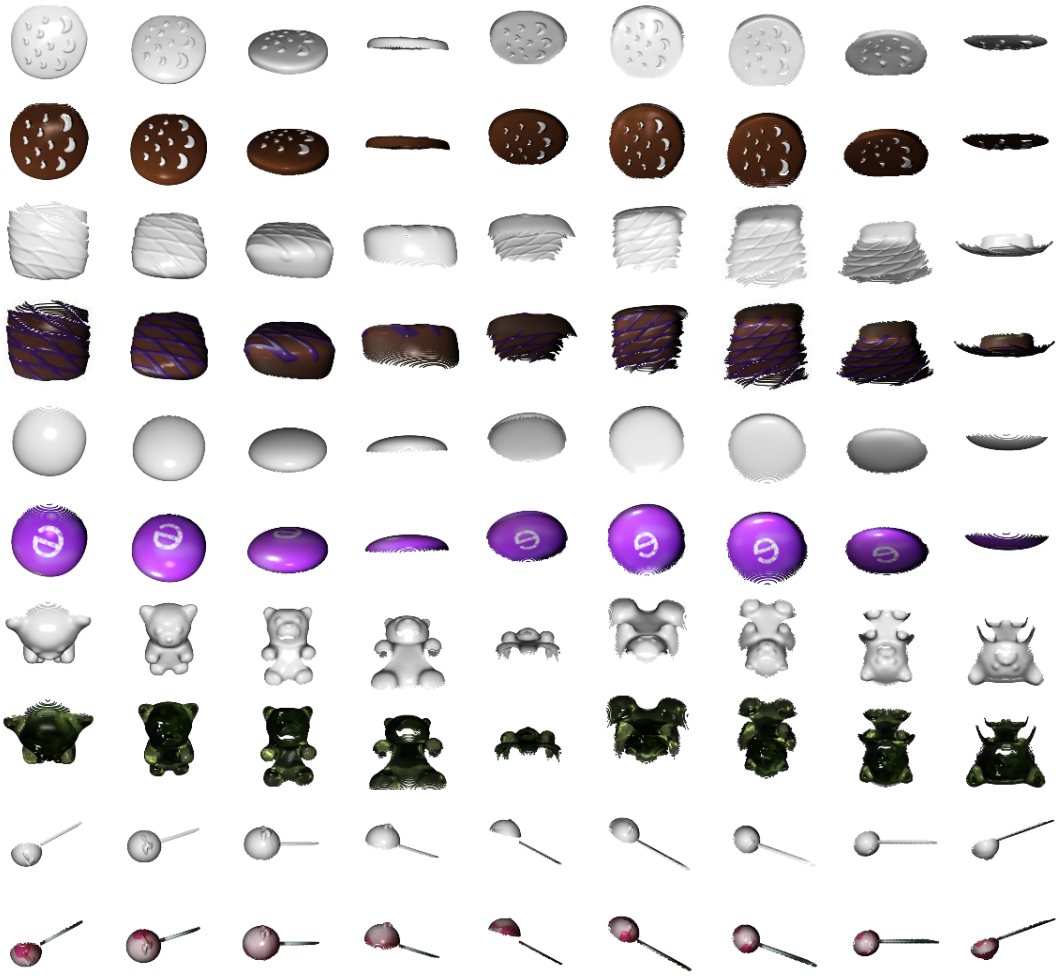


Figure 5: Visualization of the rendered multi-view observations on Eyecandies, including both rendering and corresponding RGB images from different views.

B Implementation Details

We adopt CLIP ViT-L/14@336px as the visual encoder with pre-trained weights and keep it frozen throughout training. For each 3D object, we use the Open3D library to generate 9 views by rotating point clouds around the X-axis with angles $\{-\frac{4\pi}{5}, -\frac{3\pi}{5}, -\frac{2\pi}{5}, -\frac{\pi}{5}, 0, \frac{\pi}{5}, \frac{2\pi}{5}, \frac{3\pi}{5}, \frac{4\pi}{5}\}$ for all categories. Correspondingly, the 3D point-level annotations are projected to obtain 2D masks for each view, following the standard rendering and projection process as PointAD [38]. All images are resized to 336×336.

In the first stage, we train the feature aligner for 250 epochs to project rendering features into the RGB feature space. The feature aligner consists of two independent MLPs for global and local features, each comprising three fully connected layers with a hidden dimension of 768 and GELU activation. To improve efficiency, image features are pre-computed and cached before training. In the final 50 epochs, the semantic consistency reweighting strategy is introduced to refine feature local alignment. In the second stage, we optimize modality-specific prompt learners for 15 epochs while keeping the feature aligner frozen. In the final 5 epochs, the dual-prompt contrastive alignment is incorporated to enhance prompt discriminability while preserving cross-modal consistency.

The modality fusion coefficient α during inference is set to 0.5, the reweighting parameter λ in SCR is set to 1, the contrastive alignment loss weight λ_{con} in DpCA is set to 0.05, and the temperature coefficient τ is set to 0.07 for most categories. It is worth noting that these parameters may require fine-tuning for specific categories to balance distributions across modalities and ensure stable convergence in zero-shot 3D anomaly detection. All learnable components are optimized using the Adam optimizer

Table 12: Comparison of RGB-aligned and rendering feature representations under different auxiliary training categories on MVTec3D-AD under the one-vs-rest setting.

Category	RGB-aligned				Rendering			
	O-R	O-A	P-R	P-P	O-R	O-A	P-R	P-P
Carrot	79.8	93.0	96.0	85.8	79.6	93.4	95.6	84.5
Cookie	77.0	92.4	96.9	87.0	82.2	94.1	96.4	85.3
Dowel	81.6	93.2	94.3	80.7	84.2	94.4	95.3	84.8

Table 13: Comparison of RGB-aligned and rendering feature representations under different auxiliary training categories on Eyecandies under the one-vs-rest setting.

Category	RGB-aligned				Rendering			
	O-R	O-A	P-R	P-P	O-R	O-A	P-R	P-P
Confetto	67.6	70.5	93.0	74.6	69.3	74.2	92.3	73.7
Licorice Sandwich	66.5	71.5	92.0	73.1	69.3	75.4	91.6	71.1
Peppermint Candy	71.1	73.6	92.1	71.8	70.9	75.9	92.6	74.3

with a learning rate of 1×10^{-3}. During the training stage, the batch size is set to 4. All experiments are implemented using PyTorch 2.7.1 and executed on NVIDIA RTX 4090 GPUs.

Note that since RGB information is only introduced in stage 1 to facilitate cross-modal feature alignment and reduce the domain gap between RGB and rendering modalities, while stage 2 and inference rely solely on rendering images, our method remains consistent with the zero-shot 3D anomaly detection setting.

C Analysis of RGB-aligned Feature Representations

In this section, we investigate the role of RGB-aligned feature representations, with results under different auxiliary training categories reported in Table 12 for MVTec3D-AD and Table 13 for Eyecandies. The leftmost column corresponds to different auxiliary training categories used under the one-vs-rest setting. According to Table 12 and Table 13, although RGB-aligned features are slightly inferior to rendering features, they still achieve competitive overall performance, indicating the ability to capture rich semantic information complementary to geometric cues.

We further analyze the segmentation performance. As reported in Table 4 in Section 4.3, RGB-aligned features achieve comparable performance to rendering features on MVTec3D-AD, while outperforming rendering features on Eyecandies. Moreover, as shown in Table 12 and Table 13, for both the P-R and P-P metrics under different auxiliary training categories, RGB-aligned features outperform rendering features in most auxiliary training settings, including *Carrot* and *Cookie* on MVTec3D-AD, as well as *Confetto* and *Licorice Sandwich* on Eyecandies. These results suggest that RGB-aligned representations provide complementary semantic cues that are beneficial for anomaly localization. Although the overall performance of RGB-aligned features is slightly inferior to rendering features, this gap is primarily driven by specific auxiliary training categories, namely *Dowel* on MVTec3D-AD and *Peppermint Candy* on Eyecandies.

These observations further demonstrate that, even without RGB inputs during inference, the learned feature aligner enables the model to extract semantically enriched representations from rendering observations, validating the effectiveness of the proposed cross-modal feature alignment.

We further evaluate the complementary effects of RGB-aligned and rendering feature representations under the cross-dataset setting. As shown in Table 14, on both Eyecandies and Real3D-AD, the integrated model consistently achieves better overall performance than either single branch. In particular, although the rendering branch generally achieves stronger performance individually, incorporating the RGB-aligned branch further improves the results across the evaluation metrics. These results demonstrate that the RGB-aligned and rendering representations provide complementary information and that their integration remains effective under cross-dataset generalization.

Table 14: Ablation study of different feature representations under the cross-dataset setting.

Modality	Eyecandies				Real3D-AD		
	O-R	O-A	P-R	P-P	O-R	O-A	P-R
RGB-aligned	69.1	73.7	91.6	71.7	71.1	74.4	75.5
Rendering	69.6	74.6	**92.0**	72.1	76.0	79.1	76.4
Align3D-AD	**70.7**	**75.3**	**92.0**	**72.6**	**76.1**	**79.3**	**76.7**

D Effectiveness of Paired RGB Semantic Guidance

To further verify whether the performance improvement originates from effective RGB semantic transfer rather than the additional branch structure or model parameters, we conduct a control experiment on MVTec3D-AD while keeping the entire network architecture and training configuration unchanged. Specifically, we replace the paired RGB observations with mismatched RGB observations during cross-modal feature alignment. As shown in Table 15, using mismatched RGB guidance leads to clear performance degradation across all metrics compared with paired RGB guidance. In particular, O-R and P-P decrease from 83.0 and 86.5 to 78.5 and 82.3, respectively. These results demonstrate that the performance gains mainly arise from meaningful semantic transfer from paired RGB observations, rather than merely from introducing additional parameters or an additional branch.

Table 15: Comparison between paired and mismatched RGB guidance on MVTec3D-AD.

RGB Guidance	O-R	O-A	P-R	P-P
Mismatch	78.5	92.5	95.0	82.3
Paired	**83.0**	**94.3**	**95.9**	**86.5**

E Analysis of Semantic Consistency Reweighting

To better understand the behavior of the semantic consistency reweighting strategy, we visualize the local weight by overlaying it onto the rendering images, where higher weights are represented in red and lower weights are represented in green. As shown in Figure 6, the learned weights exhibit spatial variations across different regions of the rendering observations. This spatial pattern demonstrates that the proposed strategy adaptively modulates local alignment weights according to cross-modal semantic consistency.

This phenomenon highlights the adaptive nature of the proposed reweighting strategy. By aggregating cross-modal semantic interactions, the strategy assigns larger alignment weights to rendering patches exhibiting higher holistic semantic consistency with the RGB modality.

With the incorporation of the reweighting strategy, the optimization process adaptively adjusts the contribution of individual patches to the local alignment objective, encouraging more effective semantic transfer from RGB observations. By leveraging globally aggregated semantic consistency, the proposed strategy facilitates robust cross-modal feature alignment and improves the quality of RGB-aligned representations.

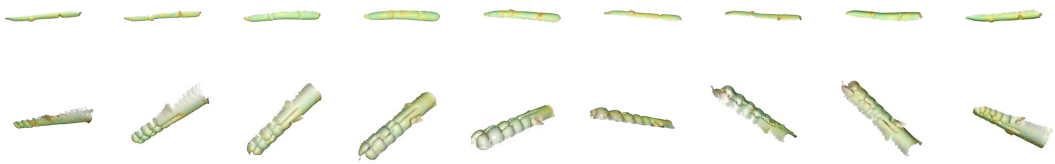


Figure 6: Visualization of semantic consistency weights on rendering images. Higher weights are represented in red, whereas lower weights are represented in green.

F Impact of the Number of Views

We investigate the impact of the number of views on model performance, with the results shown in Figure 7. As illustrated, increasing the number of views leads to consistent performance improvements across all metrics, as more views provide richer information for both rendering and RGB modalities, enabling the model to capture more detailed geometric and semantic cues. However, when the number of views increases from 9 to 11, a slight performance drop is observed. This demonstrates that incorporating excessive views may introduce redundant and noisy information, which can negatively affect model performance.

G Hyperparameter Sensitivity Analysis

We further investigate the sensitivity of Align3D-AD to the main hyperparameters, including the inference fusion weight α and the contrastive loss weight λ_{con}, on MVTec3D-AD, with results shown in Tables 16 and 17.

For the fusion weight α, the best performance is achieved at $\alpha = 0.5$, suggesting that a balanced fusion of RGB-aligned and rendering features is more effective. This indicates that the two branches provide complementary information: RGB-aligned features enhance semantic understanding, while rendering features preserve geometric cues. Excessively relying on either modality may weaken this complementarity.

For λ_{con}, the results remain stable across different values, demonstrating the robustness of our method to the contrastive loss weight. With $\lambda_{con} = 0.05$ as the default, this indicates that the dual-prompt contrastive alignment consistently provides beneficial regularization without requiring precise hyperparameter tuning.

Table 16: Sensitivity analysis of the inference fusion weight α on MVTec3D-AD.

	O-R	O-A	P-R	P-P
$\alpha = 0$	82.5	94.2	95.6	84.7
$\alpha = 0.2$	82.8	**94.3**	95.8	86.0
$\alpha = 0.5$	**83.0**	**94.3**	**95.9**	**86.5**
$\alpha = 0.8$	81.7	93.8	95.8	86.1
$\alpha = 1$	80.1	93.2	95.6	84.3

Table 17: Sensitivity analysis of the contrastive loss weight λ_{con} on MVTec3D-AD.

	O-R	O-A	P-R	P-P
$\lambda_{\text{con}} = 0.01$	82.7	94.2	95.6	85.2
$\lambda_{\text{con}} = 0.05$	**83.0**	**94.3**	**95.9**	**86.5**
$\lambda_{\text{con}} = 0.1$	82.7	94.2	95.6	85.5
$\lambda_{\text{con}} = 1$	82.9	**94.3**	95.8	86.1
$\lambda_{\text{con}} = 10$	82.2	94.1	95.6	85.2

H Impact of Prompt Length

We study the impact of prompt length on model performance, with the results shown in Figure 8. As the length of learnable prompts increases, the model achieves improved performance, indicating that longer prompts help capture richer normal and anomalous semantics. However, when the length further increases (e.g., from 12 to 16), a performance drop is observed. This confirms that excessively long prompts may introduce redundant information, while overly short prompts are insufficient to model anomaly semantics. An appropriate prompt length (i.e., 12) achieves the best overall performance.

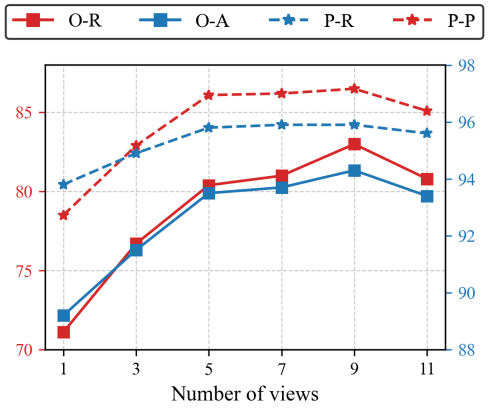


Figure 7: Impact of number of views on anomaly detection performance.

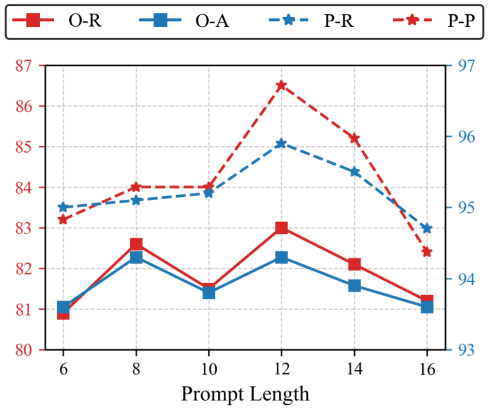


Figure 8: Impact of prompt length on anomaly detection performance.

I Detailed Results

For the one-vs-rest setting, we provide detailed per-category results on MVTec3D-AD in Table 18 and Eyecandies in Table 19, including all evaluation metrics and their mean values to offer a more comprehensive analysis of the model performance. For the cross-dataset setting, we also provide detailed per-category results on Eyecandies in Table 20 and Real3D-AD in Table 21, including all evaluation metrics and their mean values.

Table 18: Per-category results on MVTec3D-AD under the one-vs-rest setting.

Metric	Bagel	Cable gland	Carrot	Cookie	Dowel	Foam	Peach	Potato	Rope	Tire	Mean
O-R	98.8	63.7	96.5	94.4	73.2	61.3	90.9	99.0	92.6	59.8	83.0
O-A	99.7	89.3	99.3	98.5	92.2	86.7	97.4	99.8	97.1	83.2	94.3
P-R	98.7	93.1	99.4	89.0	95.8	89.7	99.2	99.8	99.4	95.3	95.9
P-P	97.4	73.1	97.8	79.4	85.6	63.7	97.8	99.3	94.0	77.3	86.5

Table 19: Per-category results on Eyecandies under the one-vs-rest setting.

Metric	Candy Cane	Chocolate Cookie	Chocolate Praline	Confetto	Gummy Bear	Hazelnut Truffle	Licorice Sandwich	Lollipop	Marshmallow	Peppermint Candy	Mean
O-R	54.1	52.0	71.3	78.1	72.0	66.2	79.1	81.8	77.8	80.4	71.3
O-A	53.9	60.2	81.5	85.3	79.0	70.1	86.0	78.9	84.3	86.6	76.6
P-R	97.9	92.3	90.0	94.2	90.2	87.9	93.3	97.6	87.9	95.1	92.6
P-P	87.1	72.9	66.6	72.4	73.0	62.0	74.1	84.1	64.6	84.8	74.2

Table 20: Per-category results on Eyecandies under the cross-dataset setting.

Metric	Candy Cane	Chocolate Cookie	Chocolate Praline	Confetto	Gummy Bear	Hazelnut Truffle	Licorice Sandwich	Lollipop	Marshmallow	Peppermint Candy	Mean
O-R	53.2	49.1	75.9	80.3	72.1	63.2	79.4	75.6	80.3	77.8	70.7
O-A	55.7	53.8	84.3	86.4	78.6	68.9	85.1	70.6	84.9	84.4	75.3
P-R	97.8	91.5	91.2	92.9	89.2	86.9	93.3	98.0	87.4	91.6	92.0
P-P	86.8	71.0	70.0	69.0	69.5	61.5	72.7	87.6	62.8	74.8	72.6

Table 21: Per-category results on Real3D-AD under the cross-dataset setting.

Metric	Airplane	Car	Candybar	Chicken	Diamond	Duck	Fish	Gemstone	Seahorse	Shell	Starfish	Toffees	Mean
O-R	50.7	71.4	86.5	60.5	100.0	47.4	82.5	84.6	79.5	87.0	84.2	78.5	76.1
O-A	58.1	74.7	89.9	66.3	100.0	54.8	85.8	84.3	84.0	85.0	87.2	82.0	79.3
P-R	63.2	69.9	83.4	76.0	93.9	49.5	82.6	84.8	80.7	78.3	80.0	77.5	76.7

J Failure Cases

As illustrated in Figure 9, we observe that the characteristics of datasets can lead to incomplete or degraded rendering results in both RGB and rendering modalities. Since our framework relies on accurate cross-modal feature alignment, high-quality rendered observations are essential for learning

consistent and accurate feature representations. Therefore, when the rendering quality is insufficient, these defects can adversely affect the model in the following aspects. First, the incomplete structures may introduce misleading and inaccurate signals, resulting in incorrect anomaly predictions. Second, the feature aligner fails to establish reliable correspondence between modalities, making it difficult to learn effective RGB-aligned features. As a result, the overall detection performance degrades on such challenging samples.

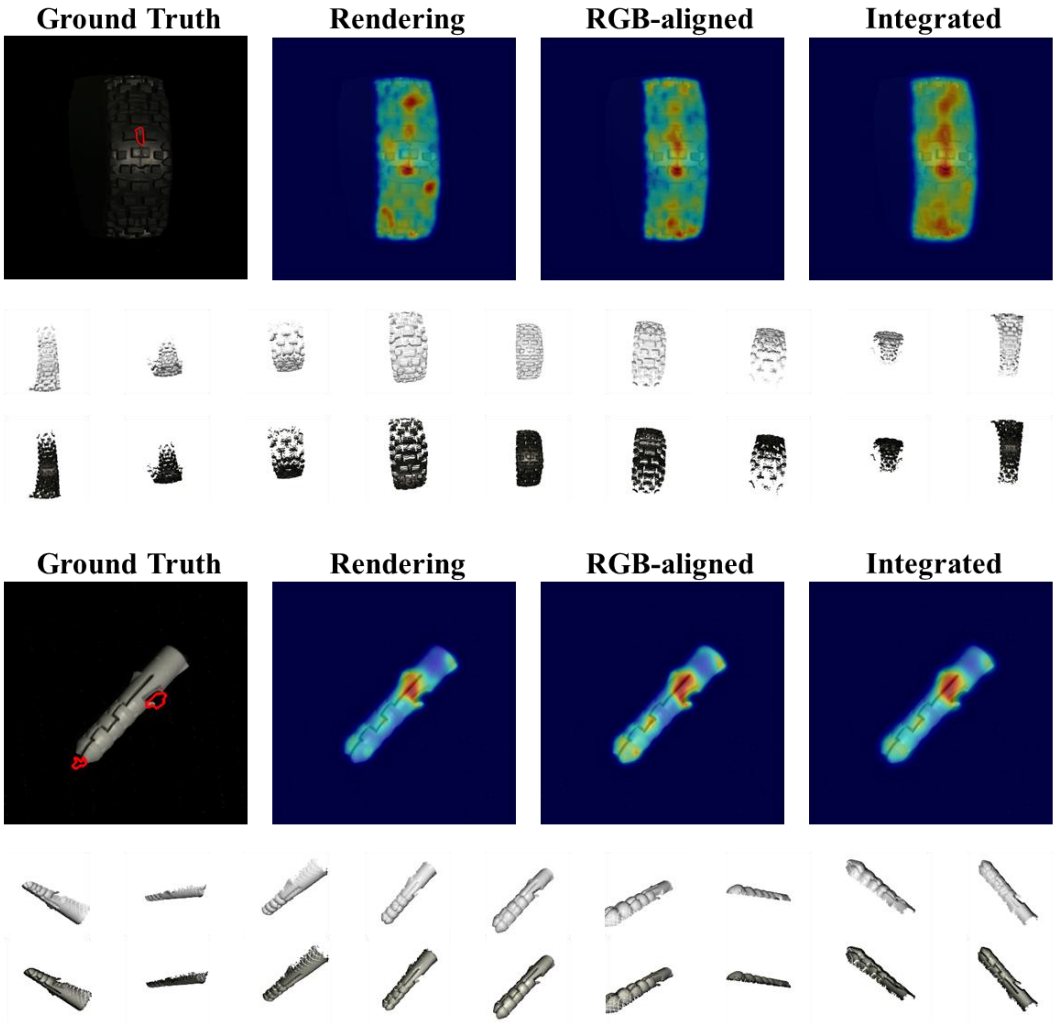


Figure 9: Failure cases of Align3D-AD on MVTec3D-AD.

K Implementation Details for Complexity Analysis

We compare the computational efficiency of our method with PointAD [38] and GS-CLIP [13] on MVTec3D-AD. Since PointAD and GS-CLIP report complexity analysis results on NVIDIA RTX 3090 GPUs in the original papers, we re-evaluate these methods on our hardware using the same 3D modality setting for a fair comparison. Specifically, all experiments are conducted on an NVIDIA RTX 4090 GPU, and the GPU is kept free during evaluation to avoid interference from other processes. The batch size during inference is set to 1 for all methods. The results reported in Table 11 reflect the efficiency of all methods under the same experimental conditions. As shown in Table 11, our method exhibits slightly higher computational costs compared with PointAD, which relies solely on rendering images. In contrast, when compared with GS-CLIP, which leverages multiple modalities including point clouds, rendering images, and depth maps, our method demonstrates consistently

better efficiency. Generally, our approach achieves a more favorable balance between computational cost and accuracy, highlighting the practical effectiveness of our model.

L Additional Visualization Results

We also provide more zero-shot visualization results of our model on MVTec3D-AD in Figure 10 and Eyecandies in Figure 11. From the additional visualization results, we observe that both modalities are informative while exhibiting subtle differences across samples. Although rendering representations effectively capture geometric and structural information, they are limited in expressing semantic cues, which may lead to suboptimal predictions in scenarios with ambiguous or subtle anomalies. In contrast, RGB-aligned representations provide richer semantic information, improving the recognition of semantically ambiguous defects that are difficult to distinguish from geometric cues alone. By integrating both representations, the model leverages the complementary information and achieves superior overall performance compared to the rendering-only feature representation, resulting in more reliable anomaly detection across all metrics.

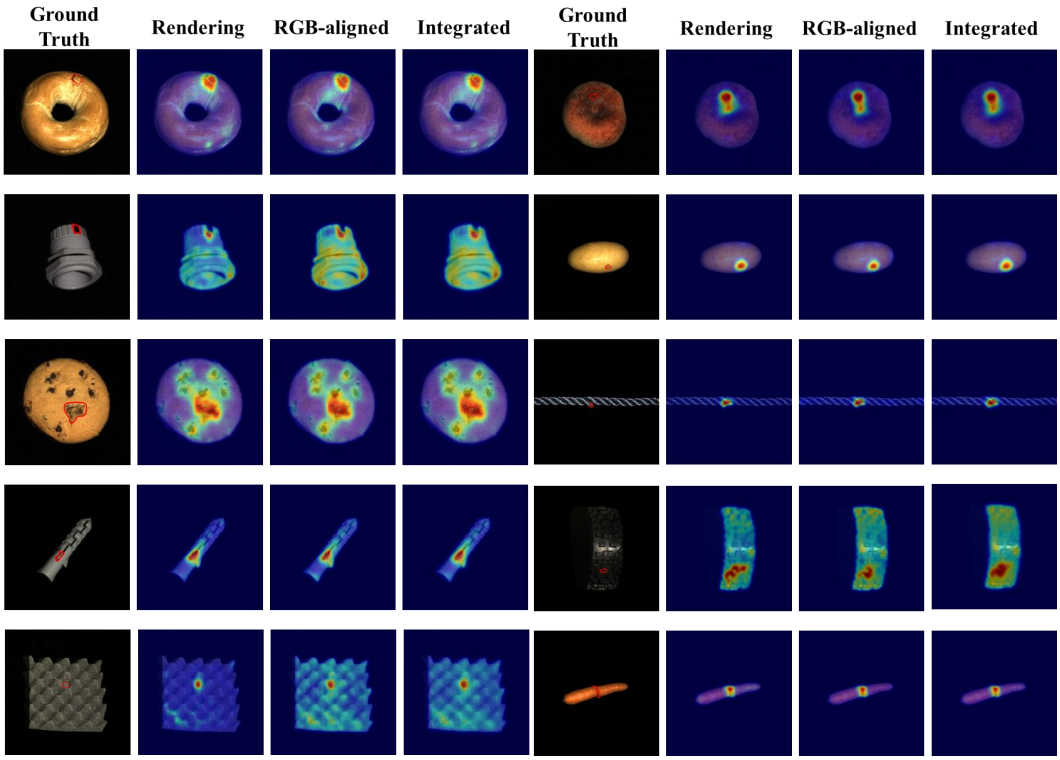


Figure 10: Visualization of anomaly score maps on MVTec3D-AD.

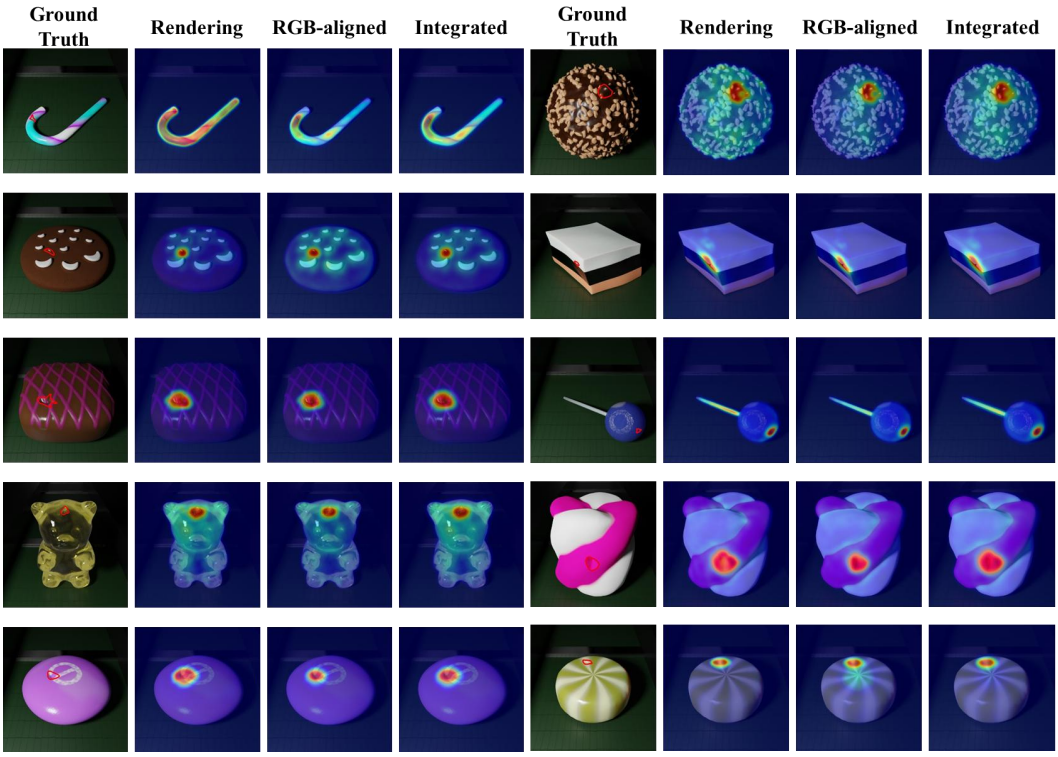


Figure 11: Visualization of anomaly score maps on Eyecandies.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in the conclusion section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide a detailed illustration of our proposed algorithm in our main paper, Appendix A, Appendix B, and Appendix K.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The source code is publicly available at the GitHub repository linked in the abstract.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We report the experimental setting and details in our main paper, Appendix B and Appendix K.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the average results across multiple runs and multiple datasets in our main paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the complexity analysis in our main paper and Appendix K.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our paper obeys the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss broader impacts in the conclusion section.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators and original owners of assets (e.g., code, data, models), used in the paper are properly credited, and the license and terms of use are explicitly mentioned and properly respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have released new assets proposed in our paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.