SF-DST: Adapting Vision-Language Models for Anomaly Detection via Asymmetric Modulation and A-LoRA

Zhipeng Wang **Xiaofei Li** **Dong Zhao** **Yonghong Song***
School of Software Engineering, Xi'an Jiaotong University, Xi'an, China
{wangzp2333, lxf0714, zhaodong}@stu.xjtu.edu.cn, songyh@xjtu.edu.cn

Abstract

Vision-Language Models (VLMs) show great potential for Industrial Anomaly Detection (IAD) but often overlook subtle high-frequency defects due to their semantic bias. Additionally, adapting them with only normal training images risks rank collapse of the low-rank adaptation subspace. To address these issues, we propose a unified framework termed SF-DST. First, we design **S**patial-**F**requency **D**ual-**S**tream **T**ransformer, which explicitly captures intra-patch details via a parallel, patch-aligned discrete cosine transform branch. It incorporates Asymmetric Cross-Modal Modulation (ACMM), in which spatial semantics act as a context gate that selectively injects spectral residuals through a zero-initialized gate, substantially mitigating the texture blindness of VLM encoders. Second, Anomaly-Aware LoRA (A-LoRA) enforces orthogonal regularization on the adaptation subspace while optimizing a hypersphere-based objective. Spectral diagnostics confirm that A-LoRA keeps the adaptation subspace well-conditioned: it triples the effective rank of the weight update over vanilla LoRA, doubles that of the learned feature covariance, and better preserves the pre-trained representation. Experiments on MVTec-AD, VisA and MMAD show that A-LoRA outperforms five PEFT alternatives including DoRA under an identical backbone and objective, and that SF-DST is competitive with strong specialist detectors, achieving the best VisA image-level AUROC among protocol-matched methods, while additionally providing language-based defect understanding that vision-only detectors cannot offer.

1 Introduction

Industrial Anomaly Detection (IAD) is essential for intelligent manufacturing, aiming to identify rare defects among a vast majority of normal samples [4, 44]. Traditional methods often lack scalability as they require training from scratch for each category [30]. Recently, Large Vision-Language Models (VLMs) have shown great potential for IAD due to their robust pre-trained representations [29, 43]. Methods like WinCLIP [15] and AnomalyGPT [9] leverage these general priors to enable few-shot or zero-shot detection. However, directly applying these generic models to industrial tasks faces significant architectural and optimization challenges.

The first challenge is the Semantic-Detail Paradox. Generic encoders prioritize semantic stability, effectively acting as low-pass filters that suppress high-frequency noise [7]. However, industrial anomalies (e.g., microscopic scratches) are often high-frequency signals. Consequently, standard VLMs suffer from texture blindness. The second challenge is rank collapse during adaptation. In the one-class data setting, no real anomaly is observed during training; even when pseudo-anomalies are synthesized from the normal images themselves, the supervision remains narrow and highly

*Corresponding author.

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

redundant. Under this regime, parameter-efficient tuning methods like LoRA [13] tend to degenerate into a redundant low-rank subspace [1]: we observe that the columns of the LoRA up-projection progressively align during training, and naive LoRA even *degrades* a strong frozen VLM backbone (Table 3).

To address these issues, we propose Spatial-Frequency Dual-Stream Transformer (SF-DST). Recognizing the importance of spectral cues, we add a parallel branch for spectral analysis. Specifically, we design Asymmetric Cross-Modal Modulation (ACMM). This mechanism uses the spatial semantics of the VLM as a context gate to selectively retrieve intra-patch high-frequency details from the Discrete Cosine Transform (DCT) domain. ACMM substantially mitigates the texture blindness of the backbone without disrupting its pre-trained semantic features.

For robust optimization, we introduce Anomaly-Aware LoRA (A-LoRA). Unlike standard fine-tuning, A-LoRA imposes Orthogonal Regularization on the adaptation matrices. This constraint keeps the low-rank subspace well-conditioned and counteracts feature collapse [2]. Additionally, we adopt the hypersphere objective of Deep SVDD [33], which pulls normal samples towards a compact center while pushing pseudo-anomalies synthesized from normal images [19] away, carving out a precise decision boundary in the feature space.

The main contributions of our work are summarized as follows:

- We propose Spatial-Frequency Dual-Stream Transformer (SF-DST), a framework tailored for adapting VLMs to industrial anomaly detection. Its Asymmetric Cross-Modal Modulation (ACMM) injects patch-aligned spectral residuals into the semantic stream; the gain is largest on texture categories (+3.8 I-AUROC), directly targeting the texture blindness of generic encoders.

- We introduce Anomaly-Aware LoRA (A-LoRA) for stable one-class adaptation. Spectral diagnostics show that vanilla LoRA collapses to an effectively near-rank-5 update at nominal rank 16, whereas A-LoRA triples the effective rank, reduces basis coherence by an order of magnitude, and better preserves the pre-trained representation.

- Under an identical backbone and objective, A-LoRA outperforms five PEFT alternatives including DoRA, PiSSA and AdaLoRA, with a stable ordering across random seeds. SF-DST is competitive with strong specialist detectors on MVTec-AD and VisA, achieving the best VisA image-level AUROC among protocol-matched methods and improving PRO over HVQ-Trans by 3.6/6.9 points, while also supporting MMAD-style multimodal defect understanding.

2 Related Works

2.1 Frequency-Domain Learning in Vision Transformers

While Vision Transformers (ViTs) have achieved remarkable success in global semantic understanding, they often exhibit a low-frequency bias, filtering out high-frequency texture details essential for identifying subtle anomalies [5]. In contrast to the spatial domain, the frequency domain naturally segregates structural semantics (low-frequency) from textural details and noise (high-frequency). Early deep frequency methods utilized Fast Fourier Transform (FFT) for global spectral analysis [28]. For instance, recent deepfake detection frameworks leverage FFT to capture GAN-induced spectral artifacts invisible in the RGB space [32]. However, global FFT lacks local spatial awareness. To address this, block-wise Discrete Cosine Transform (DCT) has emerged as a potent alternative, offering high energy compaction and real-valued coefficients compatible with standard attention mechanisms [11]. Several studies have integrated DCT into CNNs for channel attention or compression artifact reduction. In the context of ViTs, FreqCross [38] and FMSI [25] proposed dual-branch architectures fusing spatial and frequency features. However, these methods typically employ symmetric fusion or simple concatenation. Our method instead uses an asymmetric configuration in which spatial semantic tokens query the patch-aligned DCT tokens, and the retrieved spectral residual is injected through a zero-initialized gate, guiding the model towards semantically improbable high-frequency signals (e.g., scratches on a smooth surface) rather than generic texture.

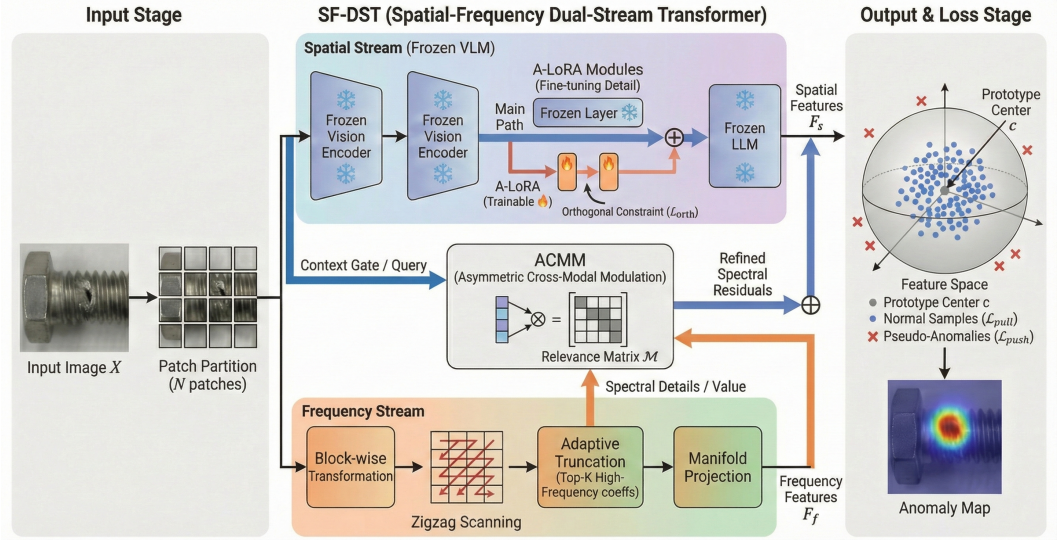


Figure 1: Overview of our SF-DST. We employ a dual-stream architecture to capture both spatial semantics and high-frequency details. The ACMM module fuses these streams by selectively injecting spectral residuals into the frozen backbone. For adaptation, A-LoRA applies orthogonal regularization ($\mathcal{L}_{orth}$) to prevent rank collapse, optimizing a hypersphere-based decision boundary.

2.2 Large Multimodal Models for Anomaly Detection

The paradigm of Industrial Anomaly Detection (IAD) is shifting from training specialist models from scratch to adapting foundational Large Multimodal Models (LMMs). Models like CLIP enabled zero-shot IAD via text-prompting, as demonstrated by WinCLIP [15] and AnomalyCLIP [42], which define normality through compositional prompt ensembles. More recently, AnomalyGPT [9] and Myriad [21] integrated Large Language Models (LLMs) with vision encoders to enable conversational anomaly localization. Despite their versatility, these generalist models suffer from the domain gap: their pre-training on web-scale natural images does not generalize well to the microscopic defects and specific lighting conditions of industrial settings [8]. In parallel, vision-only detectors built on DINOv2 features, such as AnomalyDINO [6] and Dinomaly [10], achieve strong anomaly scores and maps but provide no language interface for describing or reasoning about defects. Furthermore, full fine-tuning of LMMs is computationally prohibitive and induces substantial drift from the pre-trained representation (Sec. 4.4). Our approach retains the robust semantic backbone of InternVL3 while injecting domain-specific industrial priors through a lightweight, frequency-aware adapter, striking a balance between generalizability and precision.

2.3 Parameter-Efficient Fine-Tuning and Orthogonal Adaptation

To adapt large models with minimal overhead, Parameter-Efficient Fine-Tuning (PEFT) techniques like Low-Rank Adaptation (LoRA) [13] have become standard. LoRA freezes pre-trained weights and injects trainable low-rank matrices; subsequent variants improve its capacity and stability through adaptive rank allocation (AdaLoRA [41]), principal-subspace initialization (PiSSA [27]), or magnitude–direction decomposition (DoRA [22]). However, in the One-Class Anomaly Detection (OC-AD) setting, where only normal samples are available during training, the adapted features are prone to dimensional collapse, a phenomenon studied in self-supervised learning through covariance and spectral diagnostics [3, 17]. To mitigate this, recent works in continual learning have proposed orthogonal regularization. O-LoRA [36] and OPLoRA [37] enforce orthogonality between learnable subspaces and pre-trained subspaces to preserve prior knowledge. Building on this foundation, our A-LoRA introduces Anomaly-Aware Orthogonal Regularization. Unlike standard continual learning, our constraint acts on the adaptation basis itself and is specifically designed to prevent the normal-class manifold from collapsing into a trivial solution, ensuring that the feature space remains sufficiently high-dimensional to differentiate anomalies while adapting to the target distribution via prototype contrastive learning [23] and synthetic anomaly generation (e.g., CutPaste [19]).

3 Methodology

3.1 Spatial-Frequency Dual-Stream Transformer

The design inspiration for SF-DST stems from the dual characteristics of Image Forensics and the human visual system. When identifying objects, humans focus on both the overall shape of the object (spatial semantics) and are highly sensitive to sudden changes in local textures (frequency domain anomalies). Therefore, we designed a dual-stream architecture comprising spatial and frequency domains.

3.1.1 Spatial Stream

To extract robust high-level semantic representations from input images, the Spatial Stream leverages the pre-trained visual encoder of InternVL3-8B [43]. Given an input image $X \in \mathbb{R}^{B\times C\times H\times W}$, we first flatten it into a sequence of N fixed-size patches $X_p \in \mathbb{R}^{B\times N\times (P^2\cdot C)}$, where P denotes the patch size. These patches are initially processed through the Frozen Vision Encoder Φ_{ViT} to extract visual tokens. Crucially, to benefit from the advanced semantic alignment capabilities of Large Multimodal Models, these visual tokens are subsequently projected and fed into the Frozen LLM component. The final spatial features $F_s \in \mathbb{R}^{B\times N\times D}$ are derived from the output embeddings of the LLM layers, ensuring that the feature map maintains rich semantic context while preserving spatial correspondence, where D represents the hidden dimension. Crucially, distinct from standard classification paradigms that rely on the global [CLS] token, our approach strictly retains the sequence of patch tokens and discards the [CLS] token. This design choice is motivated by the requirement of industrial anomaly detection to preserve dense, localized spatial semantics, ensuring that the feature map $\mathbf{F}_s$ maintains a one-to-one spatial correspondence with the original image grid for subsequent pixel-aligned anomaly localization.

3.1.2 Frequency Stream

To explicitly capture fine-grained textural anomalies that are typically filtered out by the low-pass spectral bias of standard Vision Transformers, we design a Frequency Stream operating in the discrete cosine domain. This branch processes visual signals through a pipeline of block-wise transformation, spectral filtering, and manifold projection. Since the spatial stream emits one token per $P \times P$ patch, all structure interior to a patch is compressed into a single descriptor; the patch grid thus acts as the effective resolution limit of the semantic pathway. The frequency stream is deliberately aligned to the same grid, so that every non-DC coefficient it retains describes intra-patch structure above this limit, making the two streams complementary rather than redundant.

Block-wise Spectral Transformation. Unlike global frequency analyses which lose local spatial context, we adopt a patch-aligned strategy. Let the input image be partitioned into non-overlapping patches $\mathcal{P}_{i,j} \in \mathbb{R}^{P\times P\times C}$, corresponding to the spatial grid coordinates (i, j). For each channel c, we apply the 2D Discrete Cosine Transform (DCT-II) to decouple the spatial intensity into spectral coefficients. The frequency map $\mathcal{D}_c^{(i,j)}(u, v)$ is computed as:

$$\begin{aligned}\mathcal{D}_c^{(i,j)}(u,v) = \alpha_u\alpha_v \sum_{x=0}^{P-1}\sum_{y=0}^{P-1} \mathcal{P}_c^{(i,j)}(x,y) \\ \cdot \cos\left[\frac{\pi u(2x+1)}{2P}\right]\cos\left[\frac{\pi v(2y+1)}{2P}\right]\end{aligned} \tag{1}$$

where $u, v \in \{0, \ldots, P-1\}$ denote the horizontal and vertical frequency indices, and α serves as the orthogonal normalization factor. This transformation concentrates the signal energy into the upper-left corner (low-frequency) of the spectral block, while high-frequency textures and noise are distributed towards the bottom-right.

Energy Compaction and Adaptive Truncation. To translate the 2D spectral grid into a learnable sequence while preserving the frequency hierarchy, we employ a Zigzag Scanning operator $\mathcal{Z}(\cdot)$. This operator traverses the spectrum from the DC component (0,0) to the highest frequency $(P-1, P-1)$ along the anti-diagonals, flattening the matrix into a 1D vector ordered by spectral energy density. Critically, to mitigate the interference of extreme high-frequency sensor noise and reduce computational redundancy, we introduce a spectral truncation mechanism. We retain only the

top-K coefficients from the zigzag sequence, denoted as vector $\mathbf{z}_c^{(i,j)} \in \mathbb{R}^K$. Together with the residual injection of Sec. 3.2, this yields a two-sided, band-pass contribution of the frequency branch: truncation sets the upper cutoff by discarding the noise-dominated spectral tail, while the lower band is already carried by $\mathbf{F}_s$ (the DC term is, up to normalization, the patch mean retained by the tokenizer), so the branch has no incentive to re-inject it:

$$\mathbf{z}_c^{(i,j)} = \text{Trunc}_K\left(\mathcal{Z}\left(\mathcal{D}_c^{(i,j)}\right)\right) \tag{2}$$

Manifold Projection and Alignment. The raw spectral coefficients $\mathbf{z}^{(i,j)}$ from all channels are concatenated and projected into the high-dimensional latent space D via a learnable linear transformation W_f and bias b_f. Since the block-wise DCT is locally invariant and agnostic to the global position of the patch, we inject a learnable positional embedding $\mathbf{P}_{pos}$ to restore spatial awareness. The final output of the frequency branch, $\mathbf{F}_f \in \mathbb{R}^{B\times N\times D}$, is formulated as:

$$\mathbf{F}_f^{(i,j)} = \text{ReLU}\left(W_f \cdot \underset{c}{\text{Concat}}(\mathbf{z}_c^{(i,j)}) + b_f\right) + \mathbf{P}_{pos}^{(i,j)} \tag{3}$$

3.2 Asymmetric Cross-Modal Modulation

While the spatial stream provides a stable semantic anchor, the frequency stream offers a rich but noisy reservoir of high-frequency details. Direct fusion often leads to feature contamination, where irrelevant background noise from the spectral domain corrupts the semantic manifold. To address this, we propose Asymmetric Cross-Modal Modulation (ACMM). ACMM builds on the cross-attention operator, but formulates the fusion as a conditional feature modulation process with three deliberate design choices: (i) *asymmetry*, where only the spatial→frequency direction is instantiated, so spatial semantics act as a context gate; (ii) *residual injection* rather than replacement, so the spectral pathway can only perturb the semantic manifold; and (iii) a *zero-initialized gate*, so fusion starts as the identity on $\mathbf{F}_s$. Each choice is load-bearing: symmetric cross-attention over the same streams reaches only 97.2/96.8 versus 98.3/97.5 for ACMM (Table 7).

Let $\mathbf{F}_s \in \mathbb{R}^{N\times D}$ denote the semantic context from the spatial stream and $\mathbf{F}_f \in \mathbb{R}^{N\times D}$ be the candidate spectral details. We first project them into a shared interaction subspace to measure their local compatibility. We define a Relevance Matrix $\mathcal{M} \in \mathbb{R}^{N\times N}$, which quantifies the necessity of supplementing the i-th spatial location with the j-th spectral component:

$$\mathcal{M}_{i,j} = \frac{\exp\left((\mathbf{F}_s\mathbf{W}_{ctx})_i \cdot (\mathbf{F}_f\mathbf{W}_g)_j^\top/\tau\right)}{\sum_{k=1}^{N}\exp\left((\mathbf{F}_s\mathbf{W}_{ctx})_i \cdot (\mathbf{F}_f\mathbf{W}_g)_k^\top/\tau\right)} \tag{4}$$

Here, $\mathbf{W}_{ctx}$ transforms the spatial features into a context vector, while $\mathbf{W}_g$ extracts a gating signal from the frequency domain. τ is a temperature parameter controlling the sharpness of the selection. Rather than pooling the two streams symmetrically, $\mathcal{M}_{i,j}$ here functions as a semantic filter: it suppresses spectral noise that is semantically inconsistent with the current object part (e.g., texture variations in the background should be ignored). Based on this filtering mechanism, we synthesize the Refined Spectral Residual $\mathbf{R}_f$:

$$\mathbf{R}_f = \mathcal{M} \cdot (\mathbf{F}_f\mathbf{W}_{inj}) \tag{5}$$

where $\mathbf{W}_{inj}$ prepares the spectral features for injection. Finally, we fuse the information using a learnable adaptive scaling strategy. Rather than overwriting the spatial semantics, we treat the refined spectral information as a perturbative correction term added to the semantic backbone:

$$\mathbf{F}_{out} = \text{LayerNorm}\left(\mathbf{F}_s + \gamma \cdot \mathbf{R}_f\right) \tag{6}$$

The scalar γ is initialized to 0, ensuring the optimization trajectory starts from the robust pre-trained semantic manifold and progressively injects high-frequency texture cues only when they reduce the anomaly detection loss. This design effectively avoids the semantic degradation often caused by introducing heterogeneous frequency features.

3.3 Anomaly-Aware Low-Rank Adaptation

Standard full-parameter fine-tuning of large-scale VLMs (e.g., InternVL3) on small industrial datasets causes substantial drift from the pre-trained representation, and naive low-rank adaptation is unstable when only normal samples are available. To address this, we propose Anomaly-Aware Low-Rank Adaptation (A-LoRA), a mechanism that introduces a lightweight, orthogonally constrained adaptation subspace to fine-tune the feature manifold specifically for anomaly discrimination.

3.3.1 Orthogonally Constrained Subspace Projection

We freeze the pre-trained weights $W_0 \in \mathbb{R}^{d\times k}$ of the backbone and inject trainable low-rank matrices to approximate the weight updates $\Delta W = \mathbf{BA}$, where $\mathbf{B} \in \mathbb{R}^{d\times r}$ and $\mathbf{A} \in \mathbb{R}^{r\times k}$ ($r \ll d$). While standard LoRA is parameter-efficient, in a one-class setting the adaptation matrices suffer from rank collapse, where the columns of $\mathbf{B}$ progressively align into redundant directions during training (Sec. 4.4), failing to capture the diverse intra-class variance of normal samples. To enforce the learning of a distinct and informative basis for the adaptation subspace, we impose an Orthogonal Regularization explicitly on the up-projection matrix $\mathbf{B}$. This constraint ensures that the basis vectors of the learned subspace remain mutually orthogonal and normalized:

$$\mathcal{L}_{orth} = \left\|\mathbf{B}^\top\mathbf{B} - \mathbf{I}\right\|_F^2 \tag{7}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Minimizing $\mathcal{L}_{orth}$ keeps the adaptation subspace well-conditioned, so that all r directions of the low-rank bottleneck remain available for anomaly discrimination. It also acts as an anti-degeneracy device for the hypersphere objective below: a pull-only objective admits the trivial constant map, the *hypersphere collapse* of Deep SVDD [33], and constraining the adaptation basis counteracts this degeneracy on the adapter rather than on the network.

3.3.2 Hypersphere Metric Learning and Optimization

To strictly orient the adapted feature manifold towards anomaly detection, we employ the hypersphere objective of Deep SVDD [33]. We model the distribution of normal features as a compact hypersphere centered at a learnable prototype vector $\mathbf{c} \in \mathbb{R}^D$.

Prototype Estimation via Momentum. To avoid optimization instability caused by a randomly initialized center, we update the prototype $\mathbf{c}$ using a momentum-based moving average strategy. At iteration t, let $\bar{\mathbf{z}}_{nor}^{(t)}$ be the mean feature of the current mini-batch of normal samples. The center is updated as:

$$\mathbf{c}^{(t)} \leftarrow \mu \cdot \mathbf{c}^{(t-1)} + (1-\mu)\cdot \bar{\mathbf{z}}_{nor}^{(t)} \tag{8}$$

where μ is a momentum coefficient. This ensures that the center $\mathbf{c}$ stably tracks the global mean of the normal data distribution.

Overall Training Objective. The total optimization objective is designed to enforce Compactness for normal samples and Separability for anomalies. We generate synthetic pseudo-anomalies (x_{anom}) from normal samples (x_{nor}) using CutPaste [19]. They are synthesized online at a 1:1 ratio, one per normal image and re-sampled at every iteration, and are built solely from normal patches, so no real defect is observed at any point; this follows the standard one-class IAD protocol and keeps the pull and push terms on a common scale. The total loss function $\mathcal{L}_{total}$ is formulated as a weighted sum of three components:

$$\mathcal{L}_{total} = \mathcal{L}_{pull}(x_{nor}) + \mathcal{L}_{push}(x_{anom}) + \lambda \cdot \mathcal{L}_{orth} \tag{9}$$

- $\mathcal{L}_{pull}$: Minimizes the cosine distance between normal samples and the prototype $\mathbf{c}$ to condense the normal manifold:

$$\mathcal{L}_{pull} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{f(x_{nor}^{(i)})\cdot\mathbf{c}}{\|f(x_{nor}^{(i)})\|\|\mathbf{c}\|}\right) \tag{10}$$

- $\mathcal{L}_{push}$: Penalizes pseudo-anomalies that fall within a safety margin ξ of the center, effectively pushing them outside the normal boundary:

$$\mathcal{L}_{push} = \frac{1}{M}\sum_{j=1}^{M}\max\left(0, \xi - \left(1 - \frac{f(x_{anom}^{(j)})\cdot\mathbf{c}}{\|f(x_{anom}^{(j)})\|\|\mathbf{c}\|}\right)\right) \tag{11}$$

By jointly optimizing these terms, A-LoRA carves out a precise, spherical decision boundary in the high-dimensional space, where the orthogonal constraint $\mathcal{L}_{orth}$ ensures the geometric integrity of the adaptation, and the metric losses $\mathcal{L}_{pull/push}$ enforce semantic discriminability.

Table 1: Few-shot IAD result on MVTec-AD and VisA datasets. Each cell reports image-level (left) and pixel-level (right) AUROC (in %). The best and second-best results are respectively marked in bold and underlined.

Method	MVTec-AD			VisA		
	1-shot	2-shot	4-shot	1-shot	2-shot	4-shot
PatchCore	83.4, 92.0	86.3, 93.3	88.8, 94.3	79.9, 95.4	81.6, 96.1	85.3, 96.8
WinCLIP	93.1, 95.2	94.4, 96.0	95.2, 96.2	83.8, 96.4	84.6, 96.8	87.3, 97.2
PromptAD	86.3, 91.8	89.2, 92.2	90.6, 92.4	80.8, 96.3	84.3, 96.9	85.7, 97.2
KAG-prompt	<u>95.8</u>, <u>96.2</u>	**96.6**, <u>96.5</u>	<u>97.1</u>, 96.7	**91.6**, <u>97.0</u>	**92.7**, **97.4**	<u>93.3</u>, **97.7**
AnomalyGPT	94.1, 95.3	95.5, 95.6	96.3, 96.2	87.4, 96.2	<u>88.6</u>, 96.4	90.6, 96.7
IIPAD	94.2, **96.4**	95.7, **96.7**	96.1, **97.0**	85.4, 96.9	86.7, <u>97.2</u>	88.3, <u>97.4</u>
Ours	**96.2**, **96.4**	<u>96.5</u>, <u>96.5</u>	**97.4**, <u>96.9</u>	<u>91.4</u>, **97.1**	**92.7**, **97.4**	**93.4**, 97.3

4 Experiments

4.1 Experimental Setups

Datasets. We evaluate our proposed method on two widely recognized benchmarks for industrial anomaly detection: MVTec AD and VisA. Additionally, to comprehensively evaluate the performance of multimodal large models in anomaly detection, we also assessed our approach on the MMAD dataset. (1) MVTec-AD [4] consists of 15 categories, including 5 distinct texture types (e.g., carpet, grid) and 10 object types (e.g., bottle, transistor), containing a total of 5,354 distinct high-resolution images. It serves as the primary testbed for verifying the model's ability to handle both structural and textural anomalies. (2) VisA [44] includes 12 subsets with complex structures and multiple objects in single views, posing a greater challenge for anomaly localization and data-hungry adaptation. (3) MMAD [16] comprises 8,366 industrial images spanning 38 product categories and 244 defect types, generating 39,672 multiple-choice questions to establish the most comprehensive MLLM capability benchmark in the industrial sector.

Evaluation Metrics. For a comprehensive evaluation, we adopt widely used metrics following prior works. For anomaly detection, we report image-level AUROC (I-AUROC). For anomaly localization, we employ pixel-level AUROC (P-AUROC), and Area Under the Per-Region-Overlap Curve (PRO). As for MMAD dataset, we follow the evaluation metrics outlined in the original paper [16].

Implementation Details. We use the pre-trained InternVL3-8B as the backbone, i.e., the same variant reported as a baseline in Table 6, so that the comparison against the unadapted backbone is like-for-like. Input images are resized to 512×512. The patch size P is set to 16. We retain the top $K = 128$ spectral coefficients after Zigzag scanning. The rank of A-LoRA is set to $r = 16$, and the orthogonality weight λ is set to 0.1. The learning rate is set to 10^{-4}, and the margin ξ of the push loss is set to 0.5. Pseudo-anomalies are synthesized online at a 1:1 ratio (Sec. 3.3). Following the reporting practice of the compared baselines, benchmark tables report single runs; the central adaptation comparison is additionally repeated over three seeds (Table 4).

4.2 Experimental Results

We report results across few-shot IAD, unsupervised IAD, and MMAD. Extended analyses are presented in the Appendix: detailed MMAD scores and qualitative examples (Appendix C), fusion ablations, hyperparameter sensitivity, efficiency and robustness (Appendix D), and the full mechanism diagnostics (Appendix E).

Few-shot IAD Results. Few-shot IAD is challenging because it targets unseen products with only a handful of normal exemplars. We compare against PatchCore [30], WinCLIP [15], KAG-prompt [34], AnomalyGPT [9], PromptAD [20], and IIPAD [26]. As shown in Table 1, SF-DST achieves the best or second-best image-level and pixel-level AUROC in 11 of the 12 settings across 1/2/4-shot on MVTec-AD and VisA, confirming strong few-shot transfer.

Unsupervised IAD Results. We further compare with UniAD [40], DiAD [12], GLAD [39], and HVQ-Trans [24]. SF-DST reaches 98.3/97.5 (I-AUROC/P-AUROC) on MVTec-AD and 93.7/98.4

on VisA (Table 2). Among protocol-matched methods, it achieves the best VisA image-level AUROC ($+2.4$ over HVQ-Trans) and improves PRO over HVQ-Trans by 3.6 and 6.9 points on the two benchmarks, remaining within 1 point of GLAD on PRO. For context, we also list the published results of the recent DINOv2-based vision-only detectors AnomalyDINO [6] and Dinomaly [10], whose training and reference-set protocols differ from ours. These detectors produce anomaly scores and maps only, whereas SF-DST retains a language pathway and is directly evaluated on MMAD below, a task that pure detectors cannot address.

Table 2: Unsupervised IAD results on MVTec-AD and VisA datasets. The best and second-best results among protocol-matched methods are respectively marked in bold and underlined. †DINOv2-based vision-only detectors with different training/reference protocols (AnomalyDINO: full-shot, 672 resolution); published results are listed for context and excluded from ranking.

Benchmark	Method	Image-level	Pixel-level	
		AU-ROC	AU-ROC	PRO
MVTec-AD	WinCLIP	95.2	96.2	-
	UniAD	96.5	96.8	90.7
	DiAD	97.2	96.8	90.7
	GLAD	97.5	<u>97.3</u>	**95.2**
	HVQ-Trans	<u>98.0</u>	<u>97.3</u>	91.4
	Ours	**98.3**	**97.5**	<u>95.0</u>
	AnomalyDINO†	99.5	98.2	95.0
	Dinomaly†	99.6	98.4	94.8
VisA	WinCLIP	78.1	79.6	56.8
	UniAD	88.8	96.2	85.5
	DiAD	86.8	96.0	75.2
	GLAD	90.1	97.4	**94.1**
	HVQ-Trans	<u>91.3</u>	**98.5**	86.3
	Ours	**93.7**	<u>98.4</u>	<u>93.2</u>
	AnomalyDINO†	97.6	98.8	96.1
	Dinomaly†	98.7	98.7	94.5

MMAD Dataset Results. We compare against mainstream VLMs on MMAD [16] (Fig. 2; per-task scores in Appendix Table 6). SF-DST improves over its own InternVL3-8B backbone on six of the seven subtasks, with large gains in anomaly discrimination ($+4.6$), defect classification ($+7.8$) and object classification ($+9.8$). It achieves the best defect localization and object analysis scores among all compared models, including the much larger closed-source GPT-4o-mini [14] and Gemini-1.5-flash [35], and ranks second in anomaly discrimination and defect classification. The remaining gap to closed-source systems lies mainly in defect description and object classification.

Qualitative Comparison. Qualitative examples on MVTec-AD and VisA are shown in Appendix Fig. 3. The method handles both subtle texture anomalies and large structural defects, consistent with the texture-category gains of the frequency stream (Sec. 4.4).

4.3 Ablation Studies

Component ablation. Table 3 yields three observations. (1) The frequency stream and ACMM improve the frozen model from 94.9/94.0 to 96.3/96.1. (2) Naive LoRA *degrades* the strong VLM backbone, both without ($94.9 \rightarrow 94.5$) and with ($96.3 \rightarrow 95.5$) the frequency stream, which is precisely the failure mode A-LoRA targets. (3) A-LoRA reverses this degradation in both cases, and its benefit is amplified by the frequency stream: $+1.2$ I-AUROC over the frozen model without it versus $+2.0$ with it, reaching 98.3/97.5.

Comparison with PEFT methods. Table 4 compares A-LoRA with five PEFT methods. All adapters share the full SF-DST architecture, backbone, input resolution, training budget and the hypersphere + CutPaste objective, so differences isolate the adaptation parameterization rather than the objective. A-LoRA outperforms the strongest alternative, DoRA, by $+0.9$ I-AUROC and vanilla LoRA by $+2.8$, and the ordering is stable over three seeds with the lowest variance for A-LoRA. Notably, adapters with a better-conditioned update perform better: across the six adapters, I-AUROC follows the

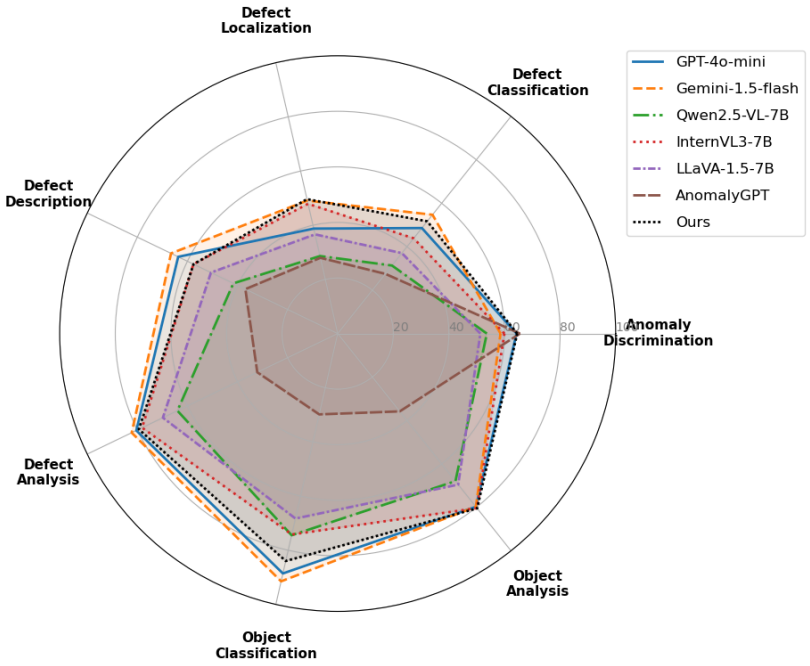


Figure 2: Radar chart comparing our proposed method with six representative VLMs on MMAD dataset.

Table 3: Ablation studies on MVTec-AD using AUROC. Abbreviations: SF-DST (Spatial-Frequency Dual-Stream Transformer), ACMM (Asymmetric Cross-Modal Modulation), OPLoRA (Orthogonal Projection LoRA), A-LoRA (Anomaly-Aware LoRA).

SF-DST	ACMM	Adaptation Strategy	Results
-	-	-	94.9 / 94.0
-	-	LoRA	94.5 / 93.9
-	-	**A-LoRA**	96.1 / 95.8
✓	-	None	95.9 / 95.3
✓	✓	None	96.3 / 96.1
✓	✓	LoRA	95.5 / 95.7
✓	✓	OPLoRA	96.0 / 95.9
✓	✓	**A-LoRA**	**98.3 / 97.5**

effective rank of ΔW almost monotonically (Spearman $\rho = 0.94$), which motivates the mechanism analysis below.

4.4 Mechanism Analysis

A-LoRA prevents rank collapse. We directly examine the geometry of the adaptation subspace and of the learned features (Table 5; full results in Appendix E). At nominal rank 16, the vanilla LoRA update is effectively near-rank-5, with a condition number σ_1/σ_{16} of 180 and a mean column coherence of 0.42. Collapse is a dynamic process: under vanilla LoRA the orthogonality residual $\|\mathbf{B}^\top\mathbf{B} - \mathbf{I}\|_F$ grows monotonically from 2.1 to 12.7 during training, whereas A-LoRA drives it down to 0.92 and holds it there. A-LoRA reduces both the condition number and the coherence by over an order of magnitude, and the effect propagates to the representation: following DirectCLR and VICReg [3, 17], the effective rank of the covariance of ℓ_2-normalized patch features roughly doubles ($18.6 \rightarrow 39.1$). A-LoRA also better preserves the pre-trained representation, with a linear CKA [18] of 0.91 to the frozen InternVL3 features on held-out natural images, versus 0.84 for LoRA and 0.61 for full fine-tuning.

Table 4: Comparison of adaptation strategies on MVTec-AD under an identical SF-DST architecture, backbone, resolution, training budget and objective. $\mathrm{erank}(\Delta W)$ is the effective rank [31] of the update at nominal rank $r = 16$, averaged over adapted layers. The last column reports mean $\pm$ std of I-AUROC over three seeds.

Adaptation	**I-AUROC**	**P-AUROC**	$\mathbf{erank}(\boldsymbol{\Delta W})\uparrow$	**I-AUROC (3 seeds)**
None (frozen)	96.3	96.1	-	-
LoRA [13]	95.5	95.7	4.8	95.4 ± 0.3
AdaLoRA [41]	96.8	96.4	9.5	-
PiSSA [27]	97.0	96.6	11.8	-
DoRA [22]	<u>97.4</u>	<u>97.0</u>	<u>12.2</u>	97.3 ± 0.3
OPLoRA [37]	96.0	95.9	10.9	-
A-LoRA (Ours)	**98.3**	**97.5**	**14.6**	$\mathbf{98.2 \pm 0.2}$

Table 5: Mechanism diagnostics on MVTec-AD ($r = 16$). Full FT: full fine-tuning. Coh.: mean $|\cos|$ between columns of $\mathbf{B}$. $\mathrm{erank}(\mathrm{Cov})$: effective rank of the covariance of ℓ_2-normalized patch features. CKA: linear CKA between adapted and frozen representations on held-out natural images. Higher is better for erank and CKA; lower is better for σ_1/σ_{16}, Coh. and $\|\mathbf{B}^\top\mathbf{B} - \mathbf{I}\|_F$.

Method	$\mathbf{erank}(\boldsymbol{\Delta W})$	$\boldsymbol{\sigma_1/\sigma_{16}}$	**Coh.**	$\|\mathbf{B}^\top\mathbf{B} - \mathbf{I}\|_F$	$\mathbf{erank(Cov)}$	**CKA**
Full FT	-	-	-	-	-	0.61
LoRA	4.8	180.0	0.42	12.7	18.6	0.84
OPLoRA	10.9	39.8	0.16	-	-	-
A-LoRA	**14.6**	**8.7**	**0.05**	**0.92**	**39.1**	**0.91**

The frequency stream recovers texture cues. Splitting MVTec-AD into its 5 texture and 10 object categories, adding the frequency stream and ACMM improves I-AUROC by $+3.79$ on textures ($94.72 \to 98.51$) but by $+1.38$ on objects ($96.81 \to 98.19$), i.e., the gain concentrates where defects are texture-like. A fixed-capacity band-stop ablation (Appendix E) further shows that masking the mid-band zigzag coefficients 32–63 causes the largest drop (-1.44 I-AUROC), and the truncation sweep exhibits an interior optimum at $K = 128$ with degradation on both sides, the signature of the band-pass behavior described in Sec. 3.1.

Limitations and Future Work. SF-DST builds on an 8B VLM, so its inference cost (307 ms/image, of which the dual-stream overhead accounts for only 7.7%) is higher than that of lightweight vision-only detectors; distilling the adapted model into a compact student is a natural next step. Extending the evaluation from public benchmarks to production lines with expert-calibrated operating points is left for future work.

5 Conclusion

In this work, we addressed the semantic-detail paradox and rank collapse that hinder VLM adaptation for industrial anomaly detection. We proposed SF-DST, a unified framework combining spectral modeling with orthogonally constrained adaptation. Asymmetric Cross-Modal Modulation (ACMM) injects patch-aligned spectral residuals via context-guided gating, yielding the largest gains on texture defects, while Anomaly-Aware LoRA (A-LoRA) keeps the adaptation subspace well-conditioned, tripling the effective rank of the update and outperforming strong PEFT alternatives such as DoRA. SF-DST is competitive with strong specialist detectors on MVTec-AD and VisA while additionally supporting language-based defect understanding on MMAD. This work underscores the value of integrating physical priors with geometric constraints, paving the way for future research in logical anomaly detection.

Acknowledgments and Disclosure of Funding

This work is supported by the National Natural Science Foundation of China under Grant 62350083. The authors declare no competing interests.

References

[1] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 7319–7328, 2021.

[2] Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep networks? *Advances in Neural Information Processing Systems*, 31, 2018.

[3] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

[4] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad – a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[5] ShouYu Chen, Liang Hu, JunTao Wang, Usman Naseem, Zhongyuan Lai, and Qi Zhang. Rf-dtr: A multi-stage dct token regression network for progressive rib fracture mask refinement. In James Kwok, editor, *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 9149–9157. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/1017. URL `https://doi.org/10.24963/ijcai.2025/1017`. AI4Tech: AI Enabling Technologies.

[6] Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. AnomalyDINO: Boosting patch-based few-shot anomaly detection with DINOv2. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1319–1329, 2025.

[7] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Tao Gong, Qi Chu, Bin Liu, Wei Zhou, and Nenghai Yu. Fe-clip: Frequency enhanced clip model for zero-shot anomaly detection and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21220–21230, 2025.

[9] 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.

[10] Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20405–20415, 2025.

[11] Amine Haboub, Hamza Baali, and Abdesselam Bouzerdoum. Dct-based channel attention for multivariate time series classification. *IEEE Open Journal of the Computer Society*, 2025.

[12] Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 8472–8480, 2024.

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[14] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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

[16] Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. MMAD: A comprehensive benchmark for multimodal large language models in industrial anomaly detection. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=JDiER86r8v`.

[17] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022.

[18] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529, 2019.

[19] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9664–9674, 2021.

[20] Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Promptad: Learning prompts with only normal samples for few-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16838–16848, 2024.

[21] Yuanze Li, Haolin Wang, Shihao Yuan, Ming Liu, Debin Zhao, Yiwen Guo, Chen Xu, Guangming Shi, and Wangmeng Zuo. Myriad: Large multimodal model by applying vision experts for industrial anomaly detection. *arXiv preprint arXiv:2310.19070*, 2023.

[22] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *International Conference on Machine Learning*, 2024.

[23] Haodong Lu, Dong Gong, Shuo Wang, Jason Xue, Lina Yao, and Kristen Moore. Learning with mixture of prototypes for out-of-distribution detection. *arXiv preprint arXiv:2402.02653*, 2024.

[24] Ruiying Lu, YuJie Wu, Long Tian, Dongsheng Wang, Bo Chen, Xiyang Liu, and Ruimin Hu. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. *Advances in Neural Information Processing Systems*, 36:8487–8500, 2023.

[25] Xinyu Luo and Yu Wang. Frequency-domain masking and spatial interaction for generalizable deepfake detection. *Electronics*, 14(7):1302, 2025.

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

[27] Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems*, 2024.

[28] Ji Qi, Xinchang Zhang, Dingqi Ye, Yongjian Ruan, Xin Guo, Shaowen Wang, and Haifeng Li. Sfnet: Fusion of spatial and frequency-domain features for remote sensing image forgery detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 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*, pages 14318–14328, 2022.

[31] Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference*, pages 606–610, 2007.

[32] Gona Rozhbayani, Amel Tuama, and Huda Hamza Abdulkhudhur. Fusing frequency and spatial transformers for robust detection of ai-generated images. *Ingenierie des Systemes d'Information*, 30(10):2719, 2025.

[33] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

[34] Fenfang Tao, Guo-Sen Xie, Fang Zhao, and Xiangbo Shu. Kernel-aware graph prompt learning for few-shot anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7347–7355, 2025.

[35] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[36] Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuan-Jing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10658–10671, 2023.

[37] Yifeng Xiong and Xiaohui Xie. Oplora: Orthogonal projection lora prevents catastrophic forgetting during parameter-efficient fine-tuning. *arXiv preprint arXiv:2510.13003*, 2025.

[38] Guang Yang. Freqcross: A multi-modal frequency-spatial fusion network for robust detection of stable diffusion 3.5 generated images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6577–6584, 2025.

[39] Hang Yao, Ming Liu, Zhicun Yin, Zifei Yan, Xiaopeng Hong, and Wangmeng Zuo. Glad: Towards better reconstruction with global and local adaptive diffusion models for unsupervised anomaly detection. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024.

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

[41] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023.

[42] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*, 2023.

[43] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[44] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 392–408, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20056-4.

A Algorithm for Training SF-DST

Algorithm 1 Training procedure for SF-DST

Input: Training dataset $\mathcal{D}_{train}$ (normal samples), Frozen VLM backbone (Encoder Φ_{ViT}, LLM Φ_{LLM}), Batch size B, Learning rate η, Momentum μ, Margin ξ, Hyper-parameters λ, K.
Initialization:
  Initialize trainable parameters Θ: ACMM weights (W_{ctx}, W_g, W_{inj}), Frequency projection (W_f), A-LORA matrices (A, B).
  Initialize prototype center c using the mean of the first batch features.
repeat
  Sample a batch of normal images $x_{nor} \sim \mathcal{D}_{train}$.
  Generate pseudo-anomalies x_{anom} via CutPaste(x_{nor}).
  Construct training batch $X = \{x_{nor}, x_{anom}\}$.
  Extract spatial features: $F_s = \Phi_{LLM}(\Phi_{ViT}(X))$.
  Compute spectral coefficients: $\mathcal{D} = \text{DCT}(X)$.
  Adaptive Truncation: $z = \text{Trunc}_K(\text{Zigzag}(\mathcal{D}))$.
  Extract frequency features: $F_f = \text{ReLU}(W_f \cdot z + b_f) + P_{pos}$.
  Compute Relevance Matrix $\mathcal{M}$ via Eq. (4).
  Synthesize refined residuals: $R_f = \mathcal{M} \cdot (F_f W_{inj})$.
  Fuse features: $F_{out} = \text{LayerNorm}(F_s + \gamma \cdot R_f)$.
  Apply A-LoRA adaptation to get final embeddings Z.
  Calculate compactness loss $\mathcal{L}_{pull}$ for x_{nor} (Eq. 9).
  Calculate margin loss $\mathcal{L}_{push}$ for x_{anom} (Eq. 10).
  Calculate orthogonal regularization $\mathcal{L}_{orth} = \|B^T B - I\|_F^2$.
  Total objective: $\mathcal{L}_{total} = \mathcal{L}_{pull} + \mathcal{L}_{push} + \lambda \mathcal{L}_{orth}$.
  Update trainable parameters $\Theta \leftarrow \Theta - \eta \nabla_\Theta \mathcal{L}_{total}$.
  Update prototype center: $c^{(t)} \leftarrow \mu c^{(t-1)} + (1-\mu)\bar{Z}_{nor}^{(t)}$.
until convergence or max iterations reached.
Output: Optimized parameters Θ^*, Prototype center c^*.

B Algorithm for Inference SF-DST

Algorithm 2 Inference Stage with SF-DST

Input: Test image x_{test}, Trained parameters Θ^*, Prototype center c^*, Frozen Backbone.
Output: Anomaly Score Map S_{map}, Image-level Score S_{img}.
Procedure:
  Get spatial features F_s from frozen VLM.
  Get frequency features F_f via DCT and projection.
  Fuse via ACMM: $F_{out} = \text{LayerNorm}(F_s + \gamma \cdot (\mathcal{M} \cdot F_f W_{inj}))$.
  Adapt features via A-LoRA: $Z = \text{Forward}(F_{out}; \Theta^*)$.
  Compute pixel-wise cosine distance to center c^*:
    $S_{map}^{(i)} = 1 - \frac{Z^{(i)} \cdot c^*}{\|Z^{(i)}\| \|c^*\|}, \quad \forall i \in \{1, \ldots, N\}$
  Compute image-level score (e.g., max of map):
    $S_{img} = \max(S_{map})$.
Return: S_{map}, S_{img}.

C Additional Experimental Results

C.1 Detailed MMAD Table

Table 6: Detailed quantitative results of each method on MMAD dataset. Abbreviations: Anomaly Discrimination (AD), Defect Classification (DC), Defect Localization (DL), Defect Description (DD), Defect Analysis (DA), Object Classification (OC), Object Analysis (OA). The best and second-best results are respectively marked in bold and underlined.

Model	Anomaly	Defect				Object	
	Discr. (AD)	Class. (DC)	Loc. (DL)	Desc. (DD)	Anal. (DA)	Class. (OC)	Anal. (OA)
AnomalyGPT	**65.57**	27.49	27.97	36.86	32.11	29.84	35.82
LLaVA-1.5-7B	51.33	37.04	36.62	50.60	69.79	68.29	69.53
Qwen2.5-VL-7B	53.65	31.33	28.62	41.66	63.99	74.46	67.94
InternVL3-8B	59.97	43.85	47.91	57.60	78.10	74.18	<u>80.37</u>
GPT-4o-mini	64.33	48.58	38.75	<u>63.68</u>	<u>80.40</u>	<u>88.56</u>	79.74
Gemini-1.5-flash	58.58	**54.70**	<u>49.10</u>	**66.53**	**82.24**	**91.47**	79.71
Ours	<u>64.57</u>	<u>51.68</u>	**49.62**	57.56	79.40	83.97	**80.46**

C.2 Qualitative Comparison

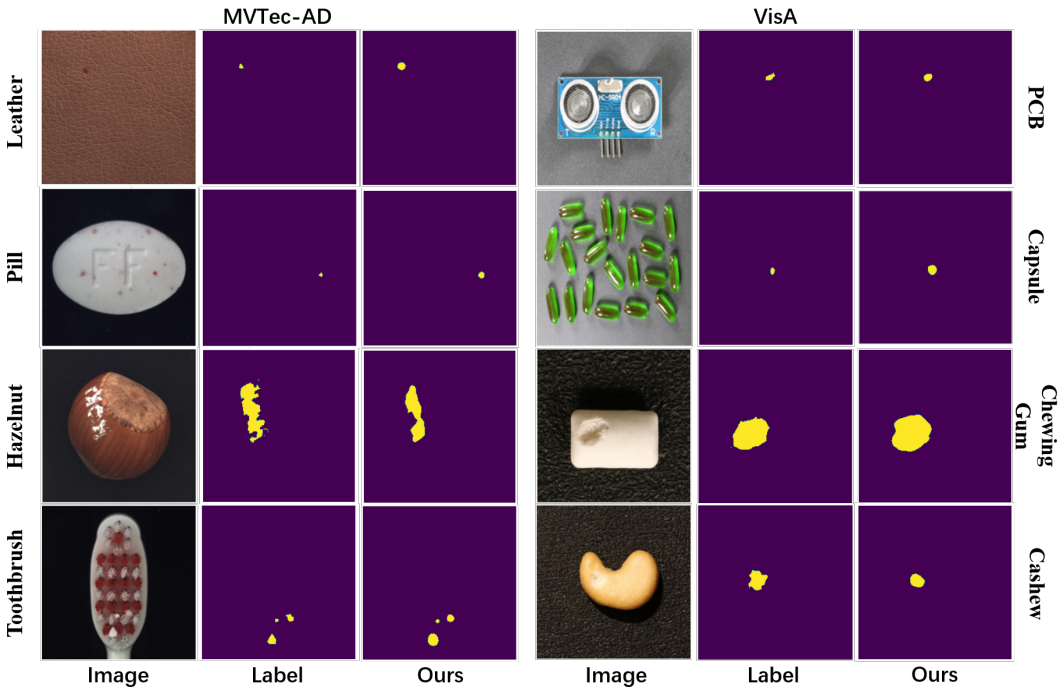


Figure 3: Qualitative results of our method on MVTec-AD and VisA datasets.

D Additional Analyses

D.1 ACMM Fusion Baseline Comparison

Table 7: Ablation on fusion strategies evaluated on MVTec-AD. Symmetric cross-attention is evaluated at a parameter count matched to ACMM.

Fusion Strategy	**I-AUROC (%)**	**P-AUROC (%)**
Simple Addition	96.5	96.0
Concatenation	97.0	96.2
Symmetric Cross-Attention	97.2	96.8
ACMM (Ours)	**98.3**	**97.5**

D.2 Sensitivity Analyses

Table 8: Sensitivity analysis for A-LoRA rank size (r) with fixed orthogonal weight $\lambda = 0.1$.

Rank Size (r)	**I-AUROC (%)**	**P-AUROC (%)**
$r = 8$	97.9	97.1
$r = 16$ **(Ours)**	**98.3**	**97.5**
$r = 32$	98.2	97.2
$r = 64$	98.0	96.8

Table 9: Sensitivity analysis for orthogonal weight (λ) with fixed rank size $r = 16$.

Orthogonal Weight (λ)	**I-AUROC (%)**	**P-AUROC (%)**
$\lambda = 0.01$	96.5	95.8
$\lambda = 0.05$	97.8	96.9
$\lambda = 0.1$ **(Ours)**	**98.3**	**97.5**
$\lambda = 0.2$	98.0	97.2
$\lambda = 0.5$	97.4	96.5

Table 10: Sensitivity analysis for spectral truncation coefficient (K).

Truncation Coefficient (K)	**I-AUROC (%)**	**P-AUROC (%)**
$K = 32$	95.2	94.6
$K = 64$	97.4	96.2
$K = 128$ **(Ours)**	**98.3**	**97.5**
$K = 256$	97.6	96.9

Table 11: Sensitivity analysis for push loss margin (ξ).

Margin (ξ)	**I-AUROC (%)**	**P-AUROC (%)**
$\xi = 0.1$	96.8	96.1
$\xi = 0.3$	97.7	96.8
$\xi = 0.5$ **(Ours)**	**98.3**	**97.5**
$\xi = 0.7$	98.1	97.3
$\xi = 0.9$	97.3	96.4

D.3 Efficiency and Robustness Under Domain Shift

Table 12: Robustness evaluation on VisA under simulated domain shifts, compared with the VLM-based WinCLIP [15].

Scenario	Method	I-AUROC	P-AUROC
Standard VisA	WinCLIP	78.1	79.6
	SF-DST (Ours)	**93.7**	**98.4**
+ Gaussian Noise	WinCLIP	65.4	68.2
	SF-DST (Ours)	**91.5**	**92.2**
+ Brightness Variation	WinCLIP	68.1	72.5
	SF-DST (Ours)	**92.1**	**94.6**

D.4 Inference Time Comparison

Table 13: Inference time and throughput on MVTec-AD test set, with a per-module breakdown of the dual-stream overhead (22.1 ms, 7.7% over the frozen backbone).

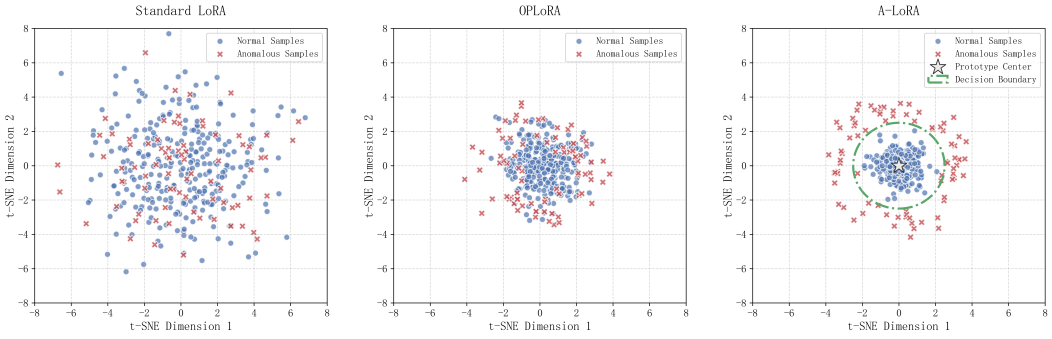
Model	Inference Time (ms/img) ↓	Throughput (FPS) ↑
Frozen VLM Backbone (InternVL3-8B)	284.9	3.5
+ DCT frequency stream	+ 10.7	-
+ ACMM	+ 8.2	-
+ Hypersphere scoring	+ 3.2	-
SF-DST (Ours)	**307.0**	**3.2**

D.5 Metric Completeness (AP and F1-max)

Table 14: Unsupervised IAD results with AP and F1-max in addition to AUROC/PRO.

Benchmark	Method	Image-level	Pixel-level		AP	F1-max
		AU-ROC	AU-ROC	PRO		
MVTec-AD	Ours	98.3	97.5	95.0	75.4	77.2
VisA	Ours	93.7	98.4	93.2	71.3	74.9

D.6 Feature-Space Visualization Across Adaptation Strategies



(a) Standard LoRA (b) OPLoRA (c) A-LoRA

Figure 4: t-SNE visualization of the learned feature distributions of normal and anomalous samples under different adaptation strategies.

E Mechanism Diagnostics and Further Analyses

Definitions. For each adapted layer, $\Delta W = \mathbf{BA}$ with $\mathbf{B} \in \mathbb{R}^{d\times r}$, $\mathbf{A} \in \mathbb{R}^{r\times k}$ and $r = 16$. The effective rank [31] of a matrix M is $\mathrm{erank}(M) = \exp(-\sum_i p_i \log p_i)$ with $p_i = \sigma_i(M)/\sum_j \sigma_j(M)$, averaged over adapted layers. Feature statistics are computed on the covariance of ℓ_2-normalized patch features on the MVTec-AD test set, following DirectCLR and VICReg [3, 17].

E.1 Subspace Dynamics and Feature Collapse

Table 15 shows that rank collapse is a dynamic process rather than a property of initialization: under vanilla LoRA the orthogonality residual grows monotonically as the columns of $\mathbf{B}$ progressively align, whereas A-LoRA drives it down within the first epochs and holds it. Table 16 shows that the effect propagates to the representation, with A-LoRA roughly doubling every dimensionality statistic of the learned features.

Table 15: Orthogonality residual $\|\mathbf{B}^\top\mathbf{B} - \mathbf{I}\|_F$ during training on MVTec-AD.

Epoch	1	5	10	20	final
Vanilla LoRA	2.1	5.4	8.6	11.9	12.7
A-LoRA	1.8	1.2	0.96	0.91	**0.92**

Table 16: Dimensional collapse of the learned patch features on the MVTec-AD test set.

Method	$\mathrm{erank}(\mathrm{Cov})\uparrow$	**#Eigenvalues for 95% variance↑**	**Mean per-dim. std↑**
Vanilla LoRA	18.6	21	0.041
A-LoRA	**39.1**	**44**	**0.079**

E.2 Preservation of the Pre-trained Representation

Table 17 measures the drift of the adapted representation from the frozen InternVL3-8B representation via linear CKA [18] and mean cosine similarity on held-out natural images, together with the MMAD general-object subtasks as a retention probe. A-LoRA drifts the least and retains the highest object-level understanding.

Table 17: Representational drift with respect to the frozen backbone.

Method	**CKA↑**	**Mean cosine↑**	**MMAD OC↑**	**MMAD OA↑**
Full fine-tuning	0.61	0.69	76.4	78.1
Vanilla LoRA	0.84	0.89	79.3	79.8
A-LoRA	**0.91**	**0.95**	**83.97**	**80.46**

E.3 Seed Stability

Table 18 repeats the central adaptation comparison over three seeds. The ordering LoRA $<$ DoRA $<$ A-LoRA is preserved in every case, and A-LoRA exhibits the smallest variance on both metrics.

Table 18: Mean $\pm$ std over three seeds on MVTec-AD.

Method	**I-AUROC**	**P-AUROC**
Vanilla LoRA	95.42 ± 0.34	95.63 ± 0.27
DoRA	97.28 ± 0.29	96.91 ± 0.22
A-LoRA	$\mathbf{98.16 \pm 0.21}$	$\mathbf{97.39 \pm 0.17}$

E.4 Frequency-Band Analysis

With $K = 128$ fixed, we divide the retained zigzag coefficients into four contiguous bands and mask one band at a time, without changing the architecture or parameter count (Table 19). The mid band (32–63) carries the most defect-relevant information, while removing either edge of the retained range has a smaller effect. Together with the truncation sweep (Table 10), where $K = 32$ discards defect-bearing intra-patch structure and $K = 256$ (the full P^2 spectrum, i.e., no truncation) readmits the noise floor, this supports the band-pass behavior of the frequency branch. Table 20 confirms that the branch primarily contributes patch-internal texture information.

Table 19: Fixed-capacity band-stop ablation on MVTec-AD ($K = 128$).

Masked zigzag band	**I-AUROC**	**P-AUROC**	**ΔI-AUROC**
None (full SF-DST)	98.3	97.5	-
0–31	97.88	97.09	-0.42
32–63	96.86	96.27	-1.44
64–95	97.43	96.74	-0.87
96–127	97.76	97.02	-0.54

Table 20: I-AUROC on the 5 texture and 10 object categories of MVTec-AD, without and with the frequency stream and ACMM (both with A-LoRA).

Group	**w/o frequency stream**	**Full SF-DST**	**Gain**
Texture categories	94.72	98.51	**+3.79**
Object categories	96.81	98.19	+1.38

E.5 Deployment-Oriented Evaluation

Beyond threshold-free AUROC, Table 21 reports operating-point and stability statistics. The threshold is calibrated on a held-out normal-only split, so no test labels are used to select it, unlike the oracle-thresholded F1-max in Table 14. Per-category I-AUROC remains above 91% on every MVTec-AD category and above 87% on every VisA category.

Table 21: Deployment-oriented statistics. FPR@95%TPR and precision/recall are computed on the MVTec-AD test set (1,258 anomalous and 467 normal images, i.e., an anomaly prevalence of 72.9%); precision depends on this prevalence.

Metric	**Result**
MVTec-AD FPR@95%TPR	8.13%
MVTec-AD precision / recall at the normal-only calibrated threshold	90.87 / 94.36
MVTec-AD per-category I-AUROC (min / max / std)	91.43 / 99.91 / 2.24
VisA per-category I-AUROC (min / max / std)	87.96 / 97.38 / 2.71

E.6 Comparison with DINOv2-based Detectors

Dinomaly [10] and AnomalyDINO [6] report strong image-level scores with a lightweight DINOv2 encoder (Table 2). SF-DST is complementary to these detectors: its pixel-level AUROC and PRO are close to theirs on MVTec-AD (97.5 / 95.0 versus 98.2–98.4 / 94.8–95.0), and, because it retains the language pathway of the VLM, it can additionally answer questions about defect type, location, cause and object attributes (Table 6), which these detectors cannot address at all.

F Broader Impacts

SF-DST targets automated visual inspection in manufacturing. Detecting subtle surface defects earlier and more reliably can reduce product recalls, material waste and the safety risks of defective

components reaching end users, and the language interface lowers the barrier for inspectors to understand and act on detected defects. As with any automated inspection system, missed defects or false alarms could have downstream costs if the system is deployed without human oversight; we therefore recommend using SF-DST as a decision-support tool with operating thresholds validated on the target production line. Increased automation may also shift the role of human inspectors, which deployment should take into account. The method is trained only on public industrial datasets and involves no personal data.

G Licenses of Existing Assets

All assets used in this work are publicly available and are credited to their creators. MVTec-AD [4] is released under the CC BY-NC-SA 4.0 license, which permits non-commercial research use. VisA [44] is released under the CC BY 4.0 license. MMAD [16] is provided for research use; its images are drawn from public datasets (MVTec-AD, MVTec LOCO AD, VisA, GoodsAD and DS-MVTec) and remain subject to their respective licenses. The InternVL3 backbone [43] is released under the MIT License, and its Qwen2.5 language component is licensed under Apache-2.0. Baseline results are taken from or reproduced according to the cited publications. All assets are used for non-commercial research purposes in accordance with these terms.

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 main claims made in 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: The paper has already discussed the limitations of the work performed by the authors.

   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: The paper does not contain formal theorems or proofs; its claims are supported empirically.

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: The paper fully discloses 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)

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: All datasets used are public benchmarks. We will publicly release the code and configuration files for every reported number upon publication.

   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: The paper specifies all the training and test details necessary to understand the results.

   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: The central adaptation comparison is reported as mean $\pm$ std over three seeds (Table 4 and Appendix Table 18); following the reporting practice of the compared baselines, the remaining benchmark tables report single runs.

   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: Inference time and its per-module breakdown are reported in Appendix Table 13.

   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: The research conducted in the paper conform, in every respect, with 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: Positive impacts (more reliable industrial inspection, reduced waste and safety risks) and potential negative impacts (costs of missed defects or false alarms without human oversight, changes to inspection jobs) are discussed in Appendix F.

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: The 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: All datasets and models used are cited, and their licenses and terms of use (CC BY-NC-SA 4.0 for MVTec-AD, CC BY 4.0 for VisA, research use for MMAD, MIT/Apache-2.0 for InternVL3) are stated in Appendix G.

    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 will release the complete code when the paper is accepted, and the code will be well documented.

    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: [Yes]

    Justification: The spatial stream uses the output embeddings of the frozen LLM component of InternVL3-8B as features (Sec. 3.1). The LLM is a standard, publicly available component; only the lightweight modules described in Sec. 3.3 and Sec. 3.2 are trained.

    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.