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

Introduction

Background

Patch tokens compress intra-patch detail. A patch-aligned DCT branch recovers complementary texture cues; normal-only adaptation can collapse feature diversity.

Contributions

- ACMM injects patch-aligned spectral detail using spatial semantics.
- A-LoRA regularizes the adaptation subspace to preserve diversity.
- One model supports image-level detection and pixel-level localization.

Evaluation: MVTec-AD, VisA and MMAD.

Experiments

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

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**

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

Model Architecture

Frequency stream: first K = 128 zigzag coefficients.

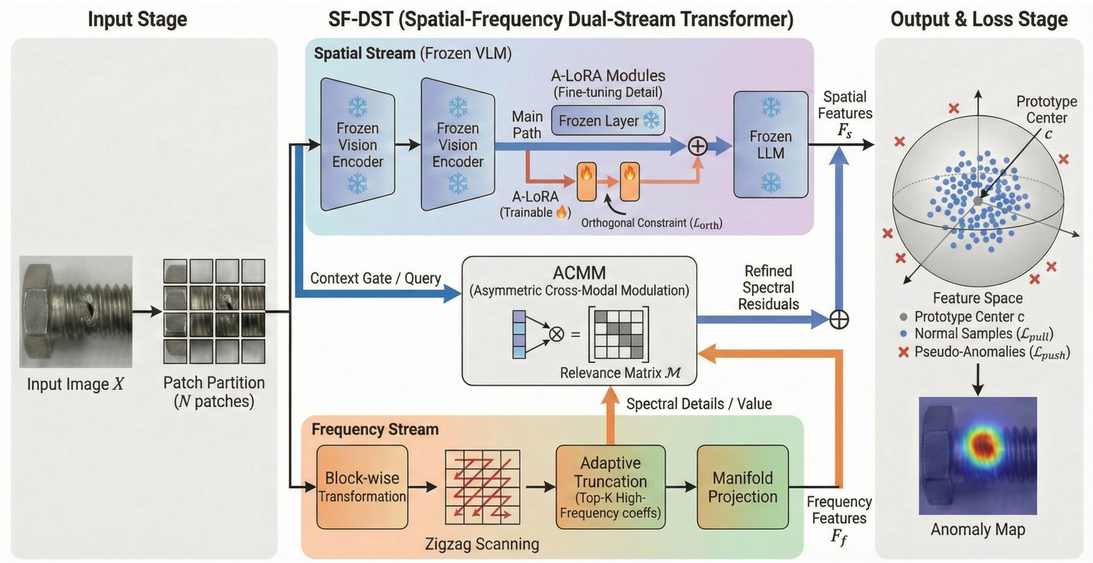


Figure 1: Overview of our SF-DST. We employ a dual-stream architecture to capture both spatial semantics and high-frequency details. The ACMM module fuses these streams by selectively injecting spectral residuals into the frozen backbone. For adaptation, A-LoRA applies orthogonal regularization ($\mathcal{L}_{orth}$)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

ACMM

Spatial tokens query spectral features. A zero-initialized gate starts with no spectral residual.

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

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

A-LoRA

Orthogonality discourages redundant directions in the adaptation subspace.

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

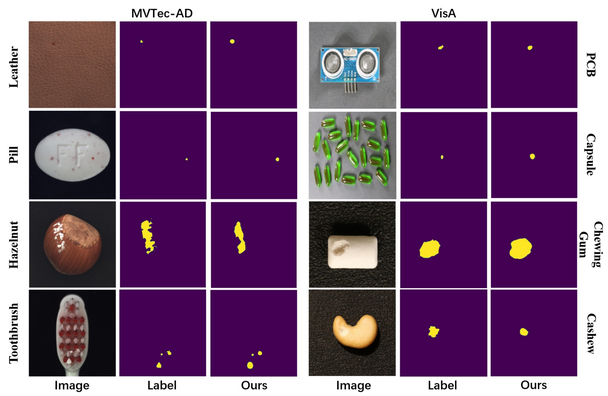
$$\mathcal{L}_{total} = \mathcal{L}_{pull}(x_{nor}) + \mathcal{L}_{push}(x_{anom}) + \lambda \cdot \mathcal{L}_{orth} \tag{9}$$

Momentum prototype update

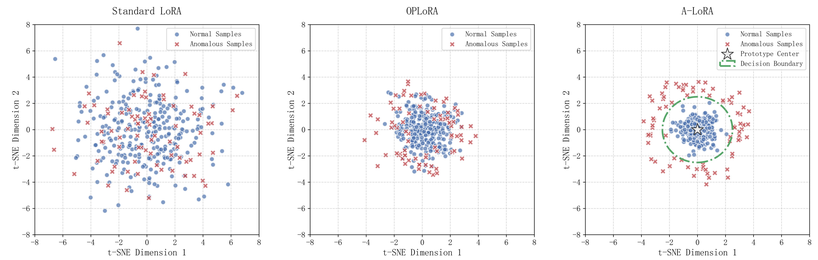
$$\mathbf{c}^{(t)} \leftarrow \mu \cdot \mathbf{c}^{(t-1)} + (1-\mu) \cdot \bar{\mathbf{z}}_{nor}^{(t)} \tag{8}$$

Normal-only training; CutPaste 1:1; r = 16, λ = 0.1.
Pull normal features toward the momentum prototype; push CutPaste features beyond margin ξ = 0.5.

Visualization



Examples: input, ground truth, prediction.



(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.

Frequency + ACMM: textures +3.79 pp; objects +1.38 pp (both A-LoRA).
VisA I-AUROC: +2.4 pp over protocol-matched HVQ-Trans.

* Corresponding author: songyh@xjtu.edu.cn

NSFC 62350083. No competing interests.