CROSSCLIP: BIDIRECTIONAL CROSS-MODAL INTERACTION WITH INSTANCE-ADAPTIVE FREQUENCY MODULATION FOR ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Universal anomaly detection aims to identify anomalies in unseen object categories and domains without target-specific training data. Recent methods leverage pretrained vision-language models, such as CLIP, to align visual and textual representations of normality and abnormality. However, they typically adapt the two modalities independently or rely on unidirectional interaction, limiting the use of complementary cross-modal information. To address this limitation, we propose **CrossCLIP**, which enables bidirectional cross-modal interaction through three adapters: a Vision-Conditioned Textual Adapter (VCTA), a Text-Conditioned Spectral Adapter (TCSA), and a Textual Reference Adapter (TRA). VCTA incorporates image information into text prompts to generate instance-specific normality and abnormality anchors, which in turn guide TCSA to perform instance-adaptive frequency modulation for capturing transferable anomalous patterns across diverse objects and domains. TRA further generates a text-grounded normality reference to complement normal-image-based comparison in zero- and few-shot settings. Extensive experiments on diverse industrial and medical datasets demonstrate that **CrossCLIP** consistently outperforms the state-of-the-art universal anomaly detection methods.

1 INTRODUCTION

Anomaly detection (AD) in images is a fundamental task across a wide range of real-world applications, including industrial defect inspection and medical lesion identification. Early AD frameworks (Fernando et al., 2021; Defard et al., 2021; He et al., 2024) are typically trained on domain-specific datasets, achieving strong performance within their target distribution but struggling to generalize to unseen categories or domains. Recently, the emergence of pretrained vision-language models (VLMs), such as CLIP (Radford et al., 2021), has demonstrated strong zero-shot recognition capabilities across diverse vision tasks, opening new opportunities for universal visual anomaly detection by enabling the identification of anomalies in novel or unseen object categories under zero- and few-shot settings.

Leveraging CLIP's zero-shot generalization capability, WinCLIP (Jeong et al., 2023) and APRIL-GAN (Chen et al., 2023) match patch-level features with handcrafted descriptions of normal and abnormal states for universal anomaly detection. However, VLMs are primarily trained to align visual representations with the semantic categories of foreground objects, rather than explicitly modeling normality and abnormality. Consequently, their ability to recognize visual anomalies is limited, resulting in suboptimal anomaly detection performance. Inspired by the Parameter-Efficient Fine-Tuning (PEFT) paradigm (Zhou et al., 2022b; Gao et al., 2024), which can adapt CLIP while preserving its generalization capability, subsequent works (Zhou et al., 2024; Cao et al., 2024; Ma et al., 2025; Gao et al., 2026) adapt CLIP for anomaly detection.

Despite these advances, existing approaches typically adapt visual features through visual adapters and textual features through prompt learning, either independently or with limited visual context. Such independent adaptation or unidirectional cross-modal interaction limits the model's ability to fully exploit the complementary information between visual and textual representations, as illustrated in Fig. 1. Moreover, recent studies Lin et al. (2023); Bi et al. (2024) have shown that deep

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

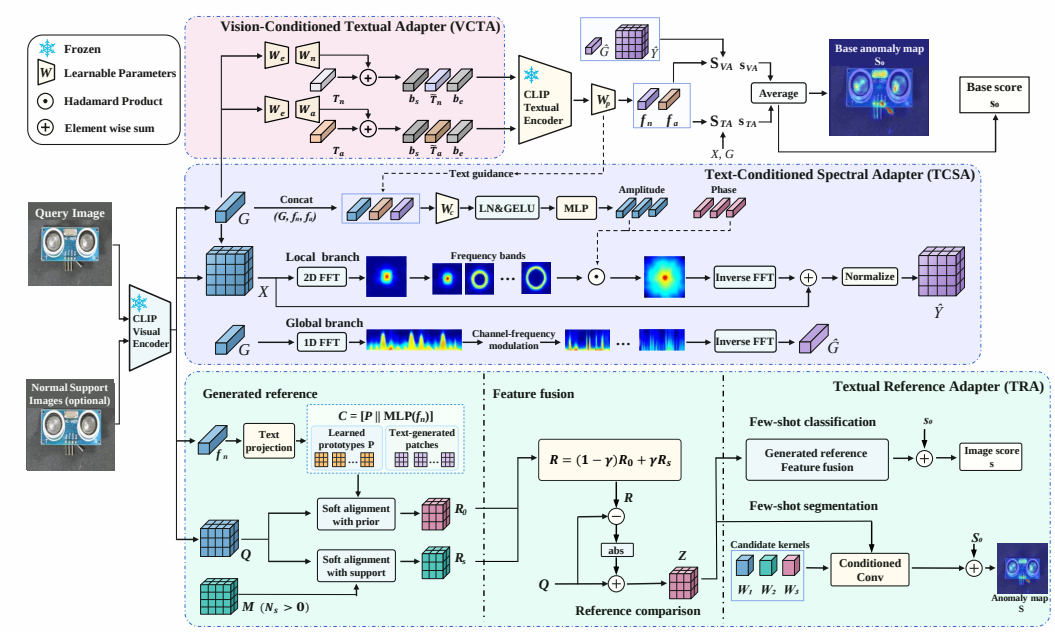

Figure 2: CrossCLIP overview. In CrossCLIP, the two modalities guide each other during adaptation: VCTA and TCSA exchange each other's modality information and TRA compares the query with textual-generated and normal reference.

models exhibit varying sensitivities to different frequency components during learning, suggesting that frequency-aware feature adaptation may improve the generalization of learned representations.

To address these challenges, we propose **CrossCLIP**, enables bidirectional cross-modal interaction throughout adaptation and alignment via three adapters while preserving frozen CLIP backbone: a Vision-Conditioned Textual Adapter (VCTA), a Text-Conditioned Spectral Adapter (TCSA), and a Textual Reference Adapter (TRA). As illustrated in Fig. 2, VCTA injects global image information into textual prompts to generate instance-specific normality and abnormality anchors. Meanwhile, TCSA performs instance-adaptive frequency modulation conditioned jointly on visual features and textual anchors, which selectively emphasizes informative frequency components to capture transferable discriminative features associated with anomalous patterns across diverse object categories and domains. Finally, TRA generates a text-grounded normality reference to complement normal-image-based comparison in both zero- and few-shot settings,overcoming the limitation of relying solely on normal support images and enabling more effective exploitation of cross-modal information. Our main contributions are summarized as follows:

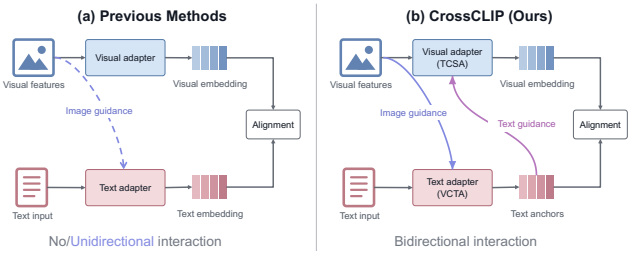

Figure 1: Comparison of previous methods and our method. They either without interaction or only interact with encoders.

- We propose **CrossCLIP**, a novel anomaly detection framework that facilitates bidirectional cross-modal interaction to learn generalizable visual and textual representations.
- We introduce an instance-adaptive frequency modulation mechanism conditioned jointly on visual features and textual anchors to capture transferable features associated with anomalous patterns across diverse object categories.
- We propose a textual reference adapter that generates a text-grounded normality reference to complement comparison with normal support images in both zero- and few-shot settings.

2 Related Work

CLIP-based anomaly detection. Leveraging CLIP's zero-shot generalization capability (Radford et al., 2021), a series of methods perform universal anomaly detection without target-specific training data. WinCLIP (Jeong et al., 2023) matches patch-level features with handcrafted descriptions of normal and abnormal states, while APRIL-GAN (Chen et al., 2023) introduces linear projections to map patch features into the joint vision-language embedding space. Following the prompt- and adapter-based PEFT paradigm (Zhou et al., 2022a; Gao et al., 2024), subsequent methods adapt CLIP to align visual and textual representations of normality and abnormality with learnable prompts or adapters. AnomalyCLIP (Zhou et al., 2024) introduces learnable prompts into CLIP encoders, while AdaCLIP (Cao et al., 2024) dynamically injects visual context from query images into textual prompt embeddings to enhance vision-language alignment. AA-CLIP (Ma et al., 2025) further orthogonalizes textual representations of normality and abnormality. PromptAD (Li et al., 2024) constructs explicit anomaly prompts from normal samples, and FiLo (Gu et al., 2024) leverages LLM-generated anomaly descriptions. AdaptCLIP (Gao et al., 2026) alternately optimizes textual and visual adapters with a prompt-query adapter for anomaly detection. Beyond PEFT-based adaptation, MRAD (Xu et al., 2026) explores retrieval-based anomaly detection by incorporating a memory bank to provide additional visual references. Despite these advances, existing methods mainly rely on independent adaptation or unidirectional cross-modal interaction, limiting their ability to fully exploit complementary information between visual and textual representations.

Frequency Learning. Frequency-domain representations have proven effective across diverse vision tasks, such as global token mixing (Rao et al., 2021), channel attention (Qin et al., 2021) and other recognition tasks (Chen et al., 2024). In domain adaptation and generalization, frequency information offers a useful prior for separating domain-invariant semantics from domain-specific styles (Yang & Soatto, 2020; Xu et al., 2021). FE-CLIP (Gong et al., 2025) utilizes fixed Fourier operation towards anomaly detection. NightAdapter (Bi et al., 2025) adapts illumination-insensitive and illumination-sensitive bands for nighttime segmentation. DFF (Lin et al., 2023) learns instance-adaptive attention masks on FFT-transformed features to emphasize transferable frequency components. Inspired by these works, we separate the visual feature spectrum into non-overlapping bands and learn instance-adaptive masks guided jointly by visual and textual cues. This selectively emphasizes anomaly-related frequency components and improves generalization across object categories and domains.

3 Methodology

3.1 Overview

Problem formulation. We train on an auxiliary dataset with pixel- and image-level annotations and evaluate on categories and domains unseen during training. Denote D as the dimension of the joint image-text embedding space and $H' \times W'$ is the patch grid. At inference, a query image X^q is accompanied by $N_s \geq 0$ normal support images, where $N_s = 0$ corresponds to the zero-shot setting. So a single trained model serves every N_s without re-training or re-calibration.

Cross-modal guided adaptation. The frozen CLIP visual encoder maps X^q to patch features $\mathbf{X} \in \mathbb{R}^{H' \times W' \times D}$ and a global feature $\mathbf{G} \in \mathbb{R}^D$. Instead of adapting each modality independently and aligning them only in the final score, every adapter of CrossCLIP is applied under a condition derived from the other modality of the current input:

$$\hat{\mathbf{u}}_m = \mathcal{A}_m\big(\mathbf{u}_m;\, \theta_m,\, \mathbf{c}_m\big), \qquad \mathbf{c}_{\mathrm{V}} = \mathbf{G}, \quad \mathbf{c}_{\mathrm{T}} = [\mathbf{G} \,\|\, \boldsymbol{f}_n \,\|\, \boldsymbol{f}_a], \quad \mathbf{c}_{\mathrm{S}} = \{\mathbf{G}, \boldsymbol{f}_n\}, \tag{1}$$

where $\mathbf{u}_m$ and θ_m denote the features and parameters of adapter $m \in \{\mathrm{V}, \mathrm{T}, \mathrm{S}\}$ (VCTA, TCSA and TRA, respectively), and $\boldsymbol{f}_n, \boldsymbol{f}_a \in \mathbb{R}^D$ are the normal and abnormal anchors produced by the Vision-Conditioned Textual Adapter (VCTA, Sec. 3.2). Visual evidence thus first shapes the textual anchors, which in turn steer the spectral adaptation of the visual features in the Text-Conditioned Spectral Adapter (TCSA, Sec. 3.3) and the comparison with normal references in the Textual Reference Adapter (TRA, Sec. 3.4). Since $\mathbf{c}_m$ is computed from the current input, the effective transformation varies across test images while θ_m remains fixed.

Training. VCTA and TCSA are optimized jointly. Given the ground-truth mask $\mathbf{Y}$ and image label y, each branch is supervised by a focal loss (Lin et al., 2017) and a Dice loss (Milletari et al., 2016) on the anomaly map and a binary cross-entropy (BCE) loss on the image-level score, following (Zhou et al., 2024; Gao et al., 2026).TRA is trained separately with the same objective applied to each layer, together with the regime-consistency loss of Eq. 17.

Scoring & prediction. All anomaly maps are read out by a shared operator $\mathcal{S}(\mathbf{U}) = \mathrm{Softmax}\big([\mathbf{U}\boldsymbol{f}_n^\top \,\|\, \mathbf{U}\boldsymbol{f}_a^\top]/\tau\big)$, whose abnormal channel is taken as the anomaly map (or score), with τ a temperature. Applying $\mathcal{S}$ to the frozen patch features $\mathbf{X}$ and global feature $\mathbf{G}$ gives the textual map $\mathbf{S}_{\text{TA}}$ and score s_{TA}, and applying it to the TCSA-adapted features gives $\mathbf{S}_{\text{VA}}$ and the image-level score s_{VA}. The prediction of the full model is

$$\mathbf{S} = \tfrac{1}{2}\big(\mathbf{S}_{\text{TA}} + \mathbf{S}_{\text{VA}}\big) + \frac{1}{N_l}\sum_{\ell} \mathbf{S}_{\text{TRA}}^{(\ell)}, \qquad s = \tfrac{1}{2}\big(s_{\text{TA}} + s_{\text{VA}}\big) + \frac{1}{N_l}\sum_{\ell} s_{\text{TRA}}^{(\ell)}, \tag{2}$$

where $\mathbf{S}_{\text{TRA}}^{(\ell)}$ and $s_{\text{TRA}}^{(\ell)}$ are the layer-ℓ outputs of TRA (Sec. 3.4), and N_l is the number of encoder layers used. The same expression is evaluated at every $N_s \geq 0$: TRA is always present, and the support images enter only through its mixing weight γ (Sec. 3.4), which is zero when $N_s = 0$. The output therefore has the same composition and scale in both regimes, which is what allows one trained model to be deployed at any N_s without re-calibration.

3.2 Vision-Conditioned Textual Adapter

VCTA learns generic representations of normality and abnormality and conditions them on the query image, so that the semantic features used by all downstream components already reflect the object under inspection.

Learnable context tokens. We maintain two sets of learnable context tokens $\mathbf{T}_n, \mathbf{T}_a \in \mathbb{R}^{L\times D_t}$ for the normal and abnormal states, where L is the number of context tokens and D_t is the embedding dimension of the CLIP text encoder.

Visual conditioning. A lightweight meta-encoder generates per-image residuals from the global visual feature $\mathbf{G} \in \mathbb{R}^D$:

$$\boldsymbol{\eta} = \mathrm{GELU}(\mathbf{W}_e\mathbf{G}), \quad \boldsymbol{\delta}_n = \mathbf{W}_n\boldsymbol{\eta}, \quad \boldsymbol{\delta}_a = \mathbf{W}_a\boldsymbol{\eta}, \tag{3}$$

where $\mathbf{W}_e \in \mathbb{R}^{D_m\times D}$ projects $\mathbf{G}$ to a bottleneck $\boldsymbol{\eta} \in \mathbb{R}^{D_m}$ ($D_m \ll D$), and $\mathbf{W}_n, \mathbf{W}_a$ are independent heads producing the residuals $\boldsymbol{\delta}_n, \boldsymbol{\delta}_a$. The residuals are added to the static tokens through learnable scalar gates:

$$\tilde{\mathbf{T}}_n = \mathbf{T}_n + \tanh(\alpha_n)\cdot\boldsymbol{\delta}_n, \qquad \tilde{\mathbf{T}}_a = \mathbf{T}_a + \tanh(\alpha_a)\cdot\boldsymbol{\delta}_a, \tag{4}$$

where α_n, α_a are initialized to zero, so that training starts from purely object-agnostic prompts and the visual conditioning is introduced gradually.

Prompt assembly and anchor extraction. The conditioned vectors are assembled with the frozen boundary embeddings $\boldsymbol{b}_s, \boldsymbol{b}_e$:

$$\mathbf{p}_n = [\boldsymbol{b}_s,\ \tilde{\mathbf{T}}_n,\ \boldsymbol{b}_e], \qquad \mathbf{p}_a = [\boldsymbol{b}_s,\ \tilde{\mathbf{T}}_a,\ \boldsymbol{b}_e], \qquad \mathbf{p}_n, \mathbf{p}_a \in \mathbb{R}^{(L+2)\times D_t}. \tag{5}$$

Feeding $\mathbf{p}_n$ and $\mathbf{p}_a$ into the frozen CLIP text encoder $\mathcal{T}(\cdot)$ and projecting the end-of-text token with the frozen projection $\mathbf{W}_p \in \mathbb{R}^{D_t\times D}$ yields the normalized anchors

$$\boldsymbol{f}_n = \frac{\mathcal{T}(\mathbf{p}_n)_{\text{eot}}\,\mathbf{W}_p}{\|\mathcal{T}(\mathbf{p}_n)_{\text{eot}}\,\mathbf{W}_p\|}, \qquad \boldsymbol{f}_a = \frac{\mathcal{T}(\mathbf{p}_a)_{\text{eot}}\,\mathbf{W}_p}{\|\mathcal{T}(\mathbf{p}_a)_{\text{eot}}\,\mathbf{W}_p\|}, \qquad \boldsymbol{f}_n, \boldsymbol{f}_a \in \mathbb{R}^D. \tag{6}$$

The textual anomaly map and score are then $\mathbf{S}_{\text{TA}} = \mathcal{S}(\mathbf{X})$ and $s_{\text{TA}} = \mathcal{S}(\mathbf{G})$. Beyond scoring, $\boldsymbol{f}_n$ and $\boldsymbol{f}_a$ serve as the cross-modal condition of TCSA and TRA.

3.3 Text-Conditioned Spectral Adapter

TCSA adapts the visual features in the frequency domain, with the modulation predicted for each input from the query image and the anchors produced by VCTA.

Frequency decomposition. TCSA operates on two branches. For the local branch, given the patch features $\mathbf{X} \in \mathbb{R}^{H' \times W' \times D}$, its frequency domain feature $\mathbf{F}$ can be obtained by 2D DFT:

$$\mathbf{F}(u,v) = \sum_{x=0}^{H'-1} \sum_{y=0}^{W'-1} \mathbf{X}(x,y)\, e^{-j2\pi\left(\frac{ux}{H'} + \frac{vy}{W'}\right)}, \tag{7}$$

with the modulated features recovered by the corresponding inverse transform:

$$\hat{\mathbf{X}}(x,y) = \frac{1}{H' \times W'} \sum_{u=0}^{H'-1} \sum_{v=0}^{W'-1} \hat{\mathbf{F}}(u,v)\, e^{j2\pi\left(\frac{ux}{H'} + \frac{vy}{W'}\right)}. \tag{8}$$

The global branch is similar but applies a 1D DFT to $\mathbf{G}$, giving the spectrum $\mathbf{H} = \mathcal{F}_1(\mathbf{G})$. Detailed process is presented in Appendix **??**. The transforms are applied per channel in the local branch and per image in the global branch. For the local branch, we partition the 2D frequency plane into K concentric annular bands by the normalized radial distance

$$\omega(u,v) = \frac{\sqrt{u^2 + v^2}}{\max_{(u',v')} \sqrt{u'^2 + v'^2}}, \tag{9}$$

assigning bin (u,v) to band $k = \mathrm{clip}(\lfloor \omega(u,v) \cdot K \rfloor, 0, K-1)$. This yields K non-overlapping masks $\{\mathbf{m}_k\}$ covering the plane, with band spectra $\mathbf{F}^{(k)} = \mathbf{F} \odot \mathbf{m}_k$, from low-frequency structure at $k{=}0$ to high-frequency detail at $k{=}K{-}1$.

Cross-modal dynamic modulation. A condition encoder fuses the visual information with the semantic features,

$$\boldsymbol{\kappa} = \mathrm{GELU}(\mathrm{LN}(\mathbf{W}_c\,[\mathbf{G} \,\|\, \boldsymbol{f}_n \,\|\, \boldsymbol{f}_a])), \tag{10}$$

where $\mathbf{W}_c$ is a learnable bias-free projection that fuses the three unit-norm vectors into a condition vector. Since $\boldsymbol{f}_n$ and $\boldsymbol{f}_a$ are themselves conditioned on $\mathbf{G}$ (Eq. 4), $\boldsymbol{\kappa}$ summarizes both what the image contains and what normality and abnormality mean for this image. From $\boldsymbol{\kappa}$, independent MLPs modulate amplitude and phase for every band,

$$\mathbf{A}_i = \mathrm{Softplus}\big(\mathrm{MLP}_i^A(\boldsymbol{\kappa})\big), \quad \boldsymbol{\Phi}_i = \pi \tanh\big(\mathrm{MLP}_i^\Phi(\boldsymbol{\kappa})\big), \quad i \in \{0,\ldots,K{-}1\} \cup \{g\}, \tag{11}$$

where $i = g$ denotes the global branch, and the corresponding coefficients are modulated as:

$$\hat{\mathbf{F}}^{(k)} = \mathbf{F}^{(k)} \odot \mathbf{A}_k \odot e^{j\boldsymbol{\Phi}_k}, \qquad \hat{\mathbf{H}} = \mathbf{H} \odot \mathbf{A}_g \odot e^{j\boldsymbol{\Phi}_g}. \tag{12}$$

and $\hat{\mathbf{F}} = \sum_k \hat{\mathbf{F}}^{(k)}$ is fed to Eq. 8. The amplitude amplifies the bands in which anomalous evidence is prominent and attenuates those dominated by normal content, whereas the phase adjusts the spatial arrangement encoded in the band, to which structural deviations such as misplaced or missing parts are particularly sensitive.

Reconstruction and scoring. The modulated local spectrum is mapped back and combined with the input through a residual connection,

$$\hat{\mathbf{Y}} = \mathbf{X} + \hat{\mathbf{X}}, \tag{13}$$

and the global feature is likewise recovered as $\hat{\mathbf{G}} = \mathcal{F}_1^{-1}(\hat{\mathbf{H}})$. Here $\mathbf{S}_{\mathrm{VA}} = \mathcal{S}(\hat{\mathbf{Y}})$ and $s_{\mathrm{VA}} = \mathcal{S}(\hat{\mathbf{G}})$. The anchors therefore play two roles: they *steer* the modulation through $\boldsymbol{\kappa}$ and they *evaluate* its outcome through $\mathcal{S}$, so that the visual adaptation is driven toward exactly the distinction the textual branch is designed to express.

3.4 Textual Reference Adapter

Few-shot AD typically matches query patches against a memory of normal supports and decodes the retrieval residual into an anomaly map. TRA makes the reference set itself elastic by generating a virtual reference from textual features and blending it with real normal references.

Text-Generated Reference. Let $\mathbf{Q}$ denote the query patch features at CLIP encoder layer ℓ, and let $\mathbf{M}$ be the concatenated patch features of the N_s normal support images, with $N_s = 0$ in the zero-shot regime. Independently of N_s, we obtain a text-generated reference $\mathbf{C}$ and the full reference set of TRA by:

$$\mathbf{C} = \big[\, \mathbf{P} \,\|\, \mathrm{MLP}(\boldsymbol{f}_n) \,\big], \quad \mathbf{B} = \mathbf{C} \cup \mathbf{M}, \tag{14}$$

where $\mathbf{P}$ is a set of layer-specific learnable visual prototypes. The MLP is a two-layer bottleneck whose output is reshaped into V pseudo-patch tokens and is shared across layers. The generated reference is thus grounded both in the visual statistics of normality observed during training and in the semantics carried by the textual branch, which constitutes the cross-modal coupling of TRA.

Soft alignment. A query patch is normally compared with its nearest reference, which commits the comparison to a single token and leaves the resulting distance dependent on how many references are available (Gao et al., 2026; Jeong et al., 2023). We instead compare against the whole reference set using a soft relaxation of nearest-neighbour retrieval. For a query set $\mathbf{Q} = \{\mathbf{q}_i\}$ and a reference set $\mathbf{B} = \{\mathbf{b}_j\}$, a temperature-scaled softmax is applied over the query-reference similarities to obtain a_{ij}, giving each query $\mathbf{q_i}$ a distribution over $\mathbf{B}$. We denote this soft alignment by $(\mathbf{d}, \mathbf{R}) = \mathrm{Align}_{\tau'}(\mathbf{Q}, \mathbf{C})$, where $\mathbf{d} = [d_i]$ and $\mathbf{R} = [\mathbf{r}_i]$ with:

$$d_i = \sum_j a_{ij} \big(1 - \cos(\mathbf{q}_i, \mathbf{b}_j)\big), \qquad \mathbf{r}_i = \sum_j a_{ij}\, \mathbf{b}_j. \tag{15}$$

Here d_i is the soft minimum of the cosine distance to $\mathbf{C}$, and $\mathbf{R} = [\mathbf{r}_i]$ collects the aggregated references, where each $\mathbf{r}_i$ acts as a soft nearest neighbour of $\mathbf{q}_i$ in $\mathbf{B}$.

Adaptive reference fusion. The two sources are matched separately, $(\mathbf{d}_0, \mathbf{R}_0) = \mathrm{Align}_{\tau_0}(\mathbf{Q}^{(\ell)}, \mathbf{C}^{(\ell)})$ for the generated reference and $(\mathbf{d}_s, \mathbf{R}_s) = \mathrm{Align}_{\tau_s}(\mathbf{Q}^{(\ell)}, \mathbf{M}^{(\ell)})$ for the real reference, with separate temperatures τ_0 and τ_s.

How much each source should be trusted depends on the image at hand, so the mixing weight $\gamma \in [0, 1]$ is predicted per instance from the query global feature and the mean distance of each stream on each layer ℓ, and is forced to zero whenever no support image is given.

$$\mathbf{d} = (1 - \gamma)\, \mathbf{d}_0 + \gamma\, \mathbf{d}_s, \qquad \mathbf{R} = (1 - \gamma)\, \mathbf{R}_0 + \gamma\, \mathbf{R}_s. \tag{16}$$

The zero-shot regime is thus simply the $\gamma = 0$ case of Eq. 16, in which the reference set falls back to the prior alone, and one set of parameters serves any $N_s \geq 0$.

Difference and decoding. Following AdaptCLIP (Gao et al., 2026), we use a similar fusion strategy. For layer ℓ, each query patch is fused with its aggregated reference by adding the elementwise absolute residual between the two, which gives the fused representation $\mathbf{Z}^{(\ell)}$. For segmentation, $\mathbf{Z}$ is upsampled by a convolution conditioned on the global query feature and the reference summary to obtain the anomaly map $\mathbf{S}^{(\ell)}_{\text{TRA}}$. For classification, we follow AdaptCLIP to obtain the image-level score $s^{(\ell)}_{\text{TRA}}$ (classification and conditioned convolution details in Appendix A.5.2).

Regime-consistency objective. Sharing one formulation across regimes does not by itself make the two predictions agree. We therefore evaluate TRA twice on each training sample, with and without the support pool, and distill the former into the latter:

$$\mathcal{L}_{\text{cons}} = \sum_{\ell} \mathrm{KL}\Big(\mathrm{sg}\big[\mathbf{S}^{(\ell)}_{\text{few}}\big] \;\Big\|\; \mathbf{S}^{(\ell)}_{\text{zero}} \Big), \tag{17}$$

where $\mathrm{sg}[\cdot]$ is the stop-gradient operator. This supervises the textual normal prior with the stronger support-conditioned signal and keeps the two regimes on a common output scale, so that a single trained model can be deployed at any $N_s \geq 0$ without re-calibration.

4 Experiment

4.1 Experiment setups

Datasets. We evaluate our method on 11 benchmarks. For industrial anomaly detection, we use MVTec AD (Bergmann et al., 2021), VisA (Zou et al., 2022), BTAD (Mishra et al., 2021), DTD-

Table 1: Image-level AUROC/AP of zero-shot AD methods in Industrial and Medical domains.

Domain	Dataset	CLIP† OpenCLIP	APRIL-GAN† CVPRw 2023	WinCLIP CVPR 2023	AnomalyCLIP ICLR 2024	AdaCLIP ECCV 2024	AA-CLIP† CVPR 2025	AdaptCLIP AAAI 2026	CrossCLIP
Industrial	MVTec	70.4/86.0	86.2/93.6	90.4/95.6	91.6/96.4	90.7/95.6	89.4/94.9	**93.5/96.7**	91.9/96.6
	VisA	54.5/62.5	77.5/80.8	75.5/78.7	82.0/85.3	81.7/84.3	77.9/81.5	84.8/87.6	**87.6/90.3**
	BTAD	50.7/64.8	73.8/69.7	68.2/70.9	88.3/88.2	89.9/**95.5**	91.8/82.6	91.0/92.2	**92.4**/92.6
	DTD	72.9/86.8	85.6/94.0	95.1/97.7	93.9/97.2	92.7/96.4	94.2/98.0	96.0/98.4	**97.6/99.1**
	MPDD	62.6/69.7	76.8/82.7	61.5/69.2	77.5/82.5	64.0/70.8	56.5/66.3	73.6/74.5	**79.5/86.9**
	Avg	62.2/74.0	80.0/84.2	78.1/82.4	86.7/89.9	83.8/88.5	82.0/84.7	87.8/89.9	**89.8/93.1**
Medical	Br35H	59.7/55.8	93.1/92.9	80.6/82.1	94.2/94.2	96.7/96.7	91.1/92.4	94.8/95.1	**97.2/97.2**
	Covid	71.2/74.7	31.9/41.0	74.7/**76.3**	77.7/55.5	69.4/43.6	66.2/65.3	86.5/54.4	**89.7**/74.3
	Avg	65.5/65.3	62.5/67.0	77.7/79.2	86.0/74.9	83.1/70.2	78.7/78.9	90.7/74.8	**93.5/85.8**

Table 2: Pixel-level AUROC/AP of zero-shot AD methods in Industrial and Medical domains.

Domain	Dataset	CLIP† OpenCLIP	APRIL-GAN† CVPRw 2023	WinCLIP CVPR 2023	AnomalyCLIP ICLR 2024	AdaCLIP ECCV 2024	AA-CLIP† CVPR 2025	AdaptCLIP AAAI 2026	CrossCLIP
Industrial	MVTec	45.7/4.2	88.4/39.3	82.3/18.2	91.1/34.5	88.3/39.1	91.8/45.4	90.9/38.3	**93.5/46.3**
	VisA	42.1/1.2	94.2/26.4	73.2/5.40	95.5/21.3	95.7/**31.0**	94.8/26.1	95.7/26.1	**95.8**/30.0
	BTAD	35.2/1.9	91.0/36.3	72.7/12.9	94.2/45.5	91.6/42.9	95.5/38.0	93.8/41.8	**96.2/61.8**
	DTD	30.0/1.6	94.3/67.0	79.5/9.80	97.9/62.6	98.3/75.2	97.2/62.9	97.7/68.7	**98.4/76.3**
	MPDD	56.0/1.7	94.8/28.2	71.2/14.1	96.5/28.9	95.5/25.9	95.9/22.7	95.9/25.3	**97.1/35.2**
	Avg	41.8/2.1	92.5/39.4	75.8/12.1	95.0/38.6	93.9/42.8	95.0/39.0	94.8/40.0	**96.2/49.9**
Medical	Kvasir	54.5/18.0	76.4/36.2	71.9/29.1	79.0/39.6	79.0/37.1	83.8/45.8	82.1/45.3	**92.8/77.1**
	Endo	56.8/17.4	83.0/43.3	71.8/25.2	84.2/46.6	83.5/41.3	87.9/51.6	86.5/52.0	**89.7/71.8**
	ClinicDB	62.2/13.5	81.9/32.2	74.6/22.4	85.0/41.8	80.9/26.0	86.3/41.9	86.0/45.0	**92.8/69.7**
	ColonDB	61.5/10.5	78.6/22.7	63.7/12.3	82.9/34.3	81.6/23.7	81.1/24.0	84.5/34.4	**87.9/44.9**
	Avg	58.8/14.9	80.0/33.6	70.5/22.3	82.8/40.6	81.3/32.0	84.8/40.8	84.8/44.2	**90.8/65.9**

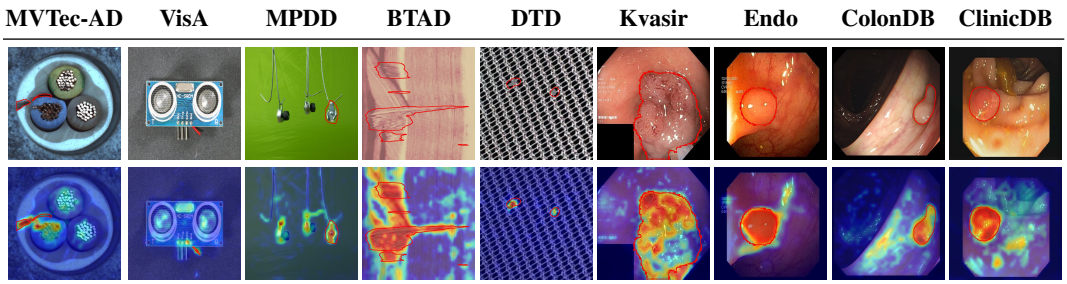

Figure 3: Qualitative results of CrossCLIP. Top: ground truth. Bottom: predicted heatmaps.

Synthetic (Aota et al., 2023), and MPDD (Jezek et al., 2021). For image-level medical anomaly detection, we use Br35H (Hamada, 2020) and Covid-19 (Peng et al., 2020) (chest X-ray) as they lack pixel-wise annotation. For pixel-level medical anomaly detection, we use Kvasir-Seg (Jha et al., 2020), Endo (Borgli et al., 2020), Colon-DB (Bernal et al., 2012) and Clinic-DB (Bernal et al., 2015) as evaluation datasets.

Evaluation Metrics. Following the previous work (Gao et al., 2026; Ma et al., 2025), we report image- and pixel-level Area Under the ROC Curve (AUROC) and Average Precision (AP) for anomaly classification and segmentation. Higher values indicate better performance. Results are averaged across all datasets within each domain to assess overall generalization. Best and second-best results are highlighted in **bold** and underline, respectively. Results with † are re-implemented with the same training and testing protocol.

Implementation details. We adopt the publicly available CLIP model ViT-L/14@336px trained by OpenAI (Radford et al., 2021), and images are resized to 518×518. We train our adapters on MVTec-AD (Bergmann et al., 2021) and test on other datasets; for evaluation of MVTec-AD, VisA (Zou et al., 2022) is utilized for training. We use Adam as the optimizer with a learning rate of 0.01 and a batch size of 8. We trained our model for 15 epochs. For a fair comparison, all competing methods share the same backbone, input resolution and patch-level feature extraction except for WinCLIP, whose default backbone is ViT-B-16+240. We trained our model on a single A800 GPU. FLOPs, parameter costs, FPS and memory usage are reported in the appendix.

Table 3: Image-level AUROC and pixel-level AP of few-shot AD methods (1-shot, 2-shot, 4-shot) on industrial datasets. Image-level AP and pixel-level AUROC are provided in the appendix.

Shot	Dataset	Image-level AUROC				Pixel-level AP			
		WinCLIP+ CVPR 2023	AnomalyCLIP+ ICLR 2024	AdaptCLIP AAAI 2026	**CrossCLIP** ours	WinCLIP+ CVPR 2023	AnomalyCLIP+ ICLR 2024	AdaptCLIP AAAI 2026	**CrossCLIP** ours
1-Shot	MVTec	93.6±0.4	95.2±0.2	94.5±0.5	**96.0±0.1**	38.3±0.8	40.8±0.1	53.7±0.9	**58.2±1.0**
	VisA	80.0±2.4	86.1±0.7	90.5±1.2	**91.5±0.3**	15.8±0.2	24.8±0.9	**38.9±0.3**	37.7±0.5
	BTAD	84.4±1.5	88.5±0.8	93.4±0.0	**94.7±0.7**	41.3±2.6	41.3±1.1	60.6±1.0	**62.7±0.3**
	DTD	97.9±0.2	98.0±0.2	98.0±0.0	**98.3±0.1**	47.8±0.9	67.4±0.4	76.9±0.1	**78.8±0.4**
	MPDD	69.3±2.9	83.4±2.6	83.8±2.2	**85.7±1.2**	29.8±2.0	34.2±0.8	33.5±2.5	**39.4±0.4**
	Avg	85.0±1.5	90.2±0.9	92.0±0.8	**93.2±0.5**	34.6±1.3	41.7±0.7	52.7±1.0	**55.4±0.5**
2-Shot	MVTec	94.5±1.0	95.4±0.1	95.7±0.6	**96.7±0.2**	39.5±0.6	41.5±0.1	55.1±0.5	**59.1±1.0**
	VisA	82.7±1.0	87.8±0.5	92.2±0.8	**93.1±0.3**	17.2±0.8	26.2±0.7	**40.7±0.6**	38.9±0.3
	BTAD	85.8±1.8	89.2±1.1	93.4±0.2	**94.7±0.6**	42.8±1.3	41.9±0.6	61.0±0.6	**62.4±0.2**
	DTD	98.1±0.2	98.2±0.1	98.3±0.0	**98.5±0.1**	48.2±0.9	68.1±0.2	77.4±0.2	**78.8±0.3**
	MPDD	69.3±2.3	83.4±1.5	84.4±0.7	**87.1±0.5**	30.7±1.1	35.3±1.1	35.0±0.7	**38.9±0.2**
	Avg	86.1±1.3	90.8±0.7	92.8±0.5	**94.0±0.3**	35.7±0.9	42.6±0.5	53.8±0.5	**55.6±0.4**
4-Shot	MVTec	95.3±0.1	96.1±0.1	96.6±0.3	**97.4±0.2**	41.2±0.9	42.4±0.0	57.2±0.8	**59.9±1.0**
	VisA	84.3±0.6	88.8±0.5	93.1±0.2	**93.1±0.3**	18.1±1.3	27.5±1.1	**41.8±0.6**	40.0±0.3
	BTAD	87.8±0.8	90.5±1.2	93.3±0.3	**95.0±0.5**	44.0±0.4	45.8±3.0	62.3±0.3	**63.5±0.3**
	DTD	98.2±0.0	98.4±0.1	98.5±0.1	**98.6±0.2**	49.3±0.1	68.5±0.2	78.2±0.2	**79.4±0.4**
	MPDD	71.2±1.6	86.3±1.8	86.8±1.1	**89.4±0.4**	32.0±0.2	36.8±1.0	37.4±1.1	**39.5±0.2**
	Avg	87.4±0.6	92.0±0.7	93.7±0.4	**94.7±0.3**	36.9±0.6	44.2±1.1	55.4±0.6	**56.5±0.4**

Table 4: Pixel-level AUROC and pixel-level AP of few-shot AD methods (1-shot, 2-shot, 4-shot) on medical datasets. Kvasir and Endo provide only pixel-level annotations.

Shot	Dataset	Pixel-level AUROC				Pixel-level AP			
		WinCLIP+ CVPR 2023	AnomalyCLIP+ ICLR 2024	AdaptCLIP AAAI 2026	**CrossCLIP** ours	WinCLIP+ CVPR 2023	AnomalyCLIP+ ICLR 2024	AdaptCLIP AAAI 2026	**CrossCLIP** ours
1-Shot	Kvasir	73.1±2.6	81.3±0.4	83.3±0.5	**94.6±0.5**	27.6±2.9	46.9±3.9	49.2±4.7	**75.5±1.9**
	Endo	72.3±0.3	84.8±1.5	86.5±1.2	**95.0±0.4**	23.6±0.1	47.8±4.9	52.4±4.7	**79.9±1.6**
	Avg	72.7±1.5	83.1±1.0	84.9±0.9	**94.8±0.5**	25.6±1.5	47.4±4.4	50.8±4.7	**77.7±1.8**
2-Shot	Kvasir	74.6±0.5	81.7±1.0	83.5±0.8	**94.6±0.4**	29.1±0.2	47.3±2.9	49.0±4.1	**75.8±1.9**
	Endo	75.5±1.6	85.1±1.4	86.6±1.1	**95.1±0.3**	27.6±2.3	49.6±4.8	53.1±4.2	**81.5±1.2**
	Avg	75.1±1.1	83.4±1.2	85.1±1.0	**94.8±0.4**	28.4±1.3	48.5±3.9	51.1±4.2	**78.6±1.6**
4-Shot	Kvasir	75.0±0.8	81.3±0.4	83.0±0.5	**94.3±0.5**	29.6±0.8	45.9±1.5	47.5±2.7	**75.0±1.9**
	Endo	75.3±0.3	84.5±0.9	85.8±0.7	**94.7±0.3**	27.7±0.5	49.2±3.4	52.2±3.1	**81.2±1.1**
	Avg	75.2±0.6	82.9±0.7	84.4±0.6	**94.5±0.4**	28.7±0.7	47.6±2.5	49.9±2.9	**78.1±1.5**

4.2 Main Results

We compare CrossCLIP with state-of-the-art zero-shot methods, including CLIP (Radford et al., 2021), APRIL-GAN (Chen et al., 2023), WinCLIP (Jeong et al., 2023), AnomalyCLIP (Zhou et al., 2024), AdaCLIP (Cao et al., 2024), AA-CLIP (Ma et al., 2025), and AdaptCLIP (Gao et al., 2026), and with few-shot methods, including WinCLIP+ (Jeong et al., 2023), AnomalyCLIP+ (Zhou et al., 2024; Gao et al., 2026), which augments AnomalyCLIP with WinCLIP+-style patch-level association, and AdaptCLIP (Gao et al., 2026).

Zero-shot performance. As shown in Tables 1 and 2, CrossCLIP attains the best average image- and pixel-level performance across the five industrial benchmarks, improving over the strongest prior method AdaptCLIP (Gao et al., 2026) by $+2.0/+3.2$ image-level and $+1.4/+9.9$ pixel-level AUROC/AP. The gain is concentrated in localization and is largest on BTAD and MPDD. Without any medical training data, the same model transfers to the medical domain and exceeds the best prior average by $+2.8/+6.6$ image-level and $+6.0/+21.7$ pixel-level. This may be because the corresponding lesions are relatively large and their information is mostly distributed in the low-frequency band, making them easy to detect in the frequency domain by simply amplifying low-frequency features.

Few-shot performance. As shown in Table 3, CrossCLIP achieves the best average image-level AUROC and pixel-level AP under all three shot settings on the five industrial benchmarks. The pixel-level gains over AdaptCLIP (Gao et al., 2026) are consistent across settings, which we attribute to the text-generated reference set and the instance-adaptive segmentation head of TRA. AdaptCLIP retains an advantage in pixel-level AP on VisA. On medical data (Table 4), the margin widens further, e.g., $+26.9$ average pixel-level AP in the 1-shot setting.

Table 5: Ablation on the adapter components.

VCTA	TCSA	TRA	Image-level AUROC/AP				Pixel-level AUROC/AP			
			MPDD	VisA	BTAD	DTD	MPDD	VisA	BTAD	DTD
✗	✗	✗	51.2/62.6	47.9/58.2	51.3/63.9	43.4/72.4	64.9/3.0	68.1/1.4	75.1/12.0	55.1/1.7
✗	✗	✓	66.2/74.8	79.4/81.4	86.8/80.9	88.5/95.5	96.7/25.1	93.5/14.8	95.6/40.8	95.4/62.4
✓	✗	✗	72.6/75.5	81.2/84.2	88.4/87.6	**97.9**/98.9	96.1/22.1	95.2/22.1	93.4/40.7	98.3/72.6
✗	✓	✗	77.3/83.7	78.6/82.9	81.8/89.9	97.8/**99.1**	96.4/26.1	94.4/22.8	90.8/40.6	**98.5**/75.6
✓	✓	✗	75.0/81.9	86.9/89.8	90.9/92.1	97.1/98.9	96.8/31.9	95.8/28.3	95.3/57.2	98.4/75.9
✗	✓	✓	75.3/78.9	80.4/84.0	83.0/85.6	94.8/98.1	93.6/21.1	95.6/27.0	85.8/27.9	97.8/72.0
✓	✗	✓	69.4/75.4	73.5/79.0	86.0/92.2	96.0/98.3	79.5/12.1	91.7/19.4	83.4/33.6	94.2/58.8
✓	✓	✓	**79.5/86.9**	**87.6/90.3**	**92.4/92.6**	97.6/**99.1**	**97.1/35.2**	**95.8/30.0**	**96.2/61.8**	98.4/**76.3**

Table 6: Ablation on conditioned instance-adpation and frequency-domain modulation in TCSA. The first row is a static spatial adapter (AdaptCLIP).

Conditioned	Frequency	Image-level AUROC/AP				Pixel-level AUROC/AP			
		VisA	DTD	MPDD	BTAD	VisA	DTD	MPDD	BTAD
✗	✗	84.8/87.6	96.0/98.4	73.6/74.6	91.0/92.2	94.8/26.1	97.2/62.9	95.9/22.7	93.8/41.8
✓	✗	84.6/87.3	95.7/98.4	**80.8**/81.7	90.0/90.7	95.3/24.8	97.0/69.1	96.5/29.7	93.7/43.1
✗	✓	86.7/89.4	97.1/98.7	75.0/79.2	91.8/**93.1**	95.7/26.8	**98.4/76.4**	96.9/31.5	96.1/58.5
✓	✓	**87.6/90.3**	**97.6/99.1**	79.5/**86.9**	**92.4**/92.6	**95.8/30.0**	**98.4**/76.3	**97.1/35.2**	**96.2/61.8**

Qualitative results. Fig. 3 visualizes representative zero-shot predictions of CrossCLIP on industrial and medical benchmarks. The anomaly maps closely follow the ground truth across diverse objects, textures, and imaging modalities, including thin, low-contrast defects that occupy only a few patches. Visualizations of frequency domain feature are provided in appendix A.11.

4.3 Ablation study

Effectiveness of the adapters. Table 5 evaluates VCTA, TCSA, and TRA individually and jointly. Each adapter alone improves substantially over frozen CLIP, yet the full model is best by a clear margin, especially in pixel-level AP (e.g., $+21.1$ on BTAD over the best single adapter). The joint gain exceeds what either adapter provides alone because, in CrossCLIP, the two adapters do not merely contribute independent scores: the anchors of VCTA steer the modulation of TCSA.

Dynamic and frequency-domain modulation. Tab. 6 disentangles the two design choices of TCSA, starting from a static spatial adapter. Instance-adaptation mainly improves image-level detection (e.g., $+7.2$ AUROC on MPDD), while frequency-domain modulation mainly benefits pixel-level localization, in line with the analysis in Sec. 3.3 that band-wise modulation raises the contrast of spatially subtle deviations. The dynamic-only variant is ahead on MPDD image-level AUROC, and the frequency-only variant is marginally ahead on BTAD image-level AP and DTD pixel-level AP. The frequency decomposition provides a compact set of controls that are effective for exposing anomalies, and dynamic prediction sets these controls differently for each input.

5 Conclusion

In this paper, we present CrossCLIP, a universal anomaly detection framework in which visual and textual representations guide each other during adaptation, covering the zero-shot and few-shot settings. CrossCLIP contains three components: a VCTA module that grounds the prototypes of normality and abnormality in the query image, a TCSA module that uses these prototypes to modulate visual features band by band in the frequency domain, and a TRA module that grounds the comparison with normal references in textual normality. Experiments on 11 industrial and medical benchmarks show that CrossCLIP achieves the best average image- and pixel level performance in both zero- and few-shot settings, with the largest gains in localization.

AI Use Statement

Large language models (LLMs) were used in the preparation of this paper to check grammar, improve wording and readability, format LaTeX tables, check code syntax, and assist with related-work

searches and literature review. Assistance with experimental code was limited to syntax checking. LLMs were not involved in research ideation, the design of the proposed method, the execution of experiments, the analysis of results, or the writing of substantive technical content. All suggestions provided by LLMs were carefully reviewed and, where appropriate, revised by the authors, who take full responsibility for the entire content of this paper, including its correctness and originality.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure that the results in this paper are reproducible. The architecture of CrossCLIP is described in Sec. 3.1, with the Vision-Conditioned Textual Adapter, the Text-Conditioned Spectral Adapter and the Textual Reference Adapter detailed in Sec. 3.2, Sec. 3.3 and Sec. 3.4, respectively. The complete frequency-domain formulation and the training objectives are given in Appendix A.5. All implementation details, including the frozen backbone (OpenAI CLIP ViT-L/14@336px), input resolution (518×518), optimizer, learning rate, batch size, number of epochs and hardware, are listed in Sec. 4.1, and the default hyperparameters of TCSA ($K = 6$, *reduction*$= 8$) are justified by the sensitivity analysis in the appendix. All 11 benchmarks used in this work are publicly available; their statistics and the cross-dataset training/testing protocol (training on MVTec AD and testing on the remaining datasets, and training on VisA when evaluating on MVTec AD) are described in Sec. 4.1 and Appendix A.7. The sources of all baseline results and the protocol used for the re-implemented methods are documented in the Competing Methods section of the appendix. Few-shot results are reported as mean$\pm$std over three runs, and computational costs are reported in the appendix. To facilitate reproduction, we will publicly release the code and trained model weights upon acceptance.

REFERENCES

Toshimichi Aota, Lloyd Teh Tzer Tong, and Takayuki Okatani. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5564–5572, January 2023.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4):1038–1059, 2021.

Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012.

Jorge Bernal, Francisco J Sánchez, Gloria Fernández-Esparrach, David Gil, Carmen Rodríguez, and Fernando Vilarino. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015.

Qi Bi, Jingjun Yi, Hao Zheng, Haolan Zhan, Yawen Huang, Wei Ji, Yuexiang Li, and Yefeng Zheng. Learning frequency-adapted vision foundation model for domain generalized semantic segmentation. *Advances in Neural Information Processing Systems*, 37:94047–94072, 2024.

Qi Bi, Jingjun Yi, Huimin Huang, Hao Zheng, Haolan Zhan, Yawen Huang, Yuexiang Li, Xian Wu, and Yefeng Zheng. Nightadapter: Learning a frequency adapter for generalizable night-time scene segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23838–23849, 2025. doi: 10.1109/CVPR52734.2025.02220.

Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Oda Gjestang, Javier Garcia, Thomas de Lange, Dag Johansen, Håvard D Johansen, et al. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1): 1–12, 2020.

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*, pp. 55–72. Springer, 2024.

Linwei Chen, Lin Gu, and Ying Fu. When semantic segmentation meets frequency aliasing. In *International Conference on Learning Representations (ICLR)*, 2024.

Xuhai Chen, Yue Han, and Jiangning Zhang. A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2. *arXiv preprint arXiv:2305.17382*, 2023. CVPR 2023 VAND Workshop Challenge.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International conference on pattern recognition*, pp. 475–489. Springer, 2021.

Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for medical anomaly detection–a survey. *ACM computing surveys (CSUR)*, 54(7): 1–37, 2021.

Bin-Bin Gao, Yue Zhou, Jiangtao Yan, Yuezhi Cai, Weixi Zhang, Meng Wang, Jun Liu, Yong Liu, Lei Wang, and Chengjie Wang. Adaptclip: Adapting clip for universal visual anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4095–4103, 2026.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International journal of computer vision*, 132(2):581–595, 2024.

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*, pp. 21220–21230, 2025.

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*, pp. 2041–2049, 2024.

Ahmed Hamada. Br35H :Brain Tumor Detection 2020. `https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection`, 2020.

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, pp. 8472–8480, 2024.

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*, pp. 19606–19616, 2023.

Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops*, pp. 66–71. IEEE, 2021.

Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Proceedings, Part II*, pp. 451–462. Springer, 2020.

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*, pp. 16838–16848, 2024.

Shiqi Lin, Zhizheng Zhang, Zhipeng Huang, Yan Lu, Cuiling Lan, Peng Chu, Quanzeng You, Jiang Wang, Zicheng Liu, Amey Parulkar, et al. Deep frequency filtering for domain generalization. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11797–11807. IEEE, 2023.

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 (ICCV)*, pp. 2980–2988, 2017.

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*, pp. 4744–4754, 2025.

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

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. VT-ADL: A vision transformer network for image anomaly detection and localization. In *2021 IEEE International Symposium on Industrial Electronics*, 2021.

Yifan Peng, Yu-Xing Tang, Sungwon Lee, Yingying Zhu, Ronald M Summers, and Zhiyong Lu. Covid-19-ct-cxr: a freely accessible and weakly labeled chest x-ray and ct image collection on covid-19 from the biomedical literature. *arXiv preprint arXiv:2006.06177*, 2020.

Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In *Advances in Neural Information Processing Systems*, 2021.

Chaoran Xu, Chengkan Lv, Qiyu Chen, Feng Zhang, and Zhengtao Zhang. Mrad: Zero-shot anomaly detection with memory-driven retrieval. In *International Conference on Learning Representations*, 2026.

Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022b.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *International Conference on Learning Representations (ICLR)*, 2024.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pp. 390–407. Springer, 2022.

Table 7: Image-level AP and pixel-level AUROC of few-shot AD methods (1-shot, 2-shot, 4-shot) on industrial datasets.

Shot	Dataset	Image-level AP				Pixel-level AUROC			
		WinCLIP+ CVPR 2023	AnomalyCLIP+ ICLR 2024	AdaptCLIP AAAI 2026	**CrossCLIP** ours	WinCLIP+ CVPR 2023	AnomalyCLIP+ ICLR 2024	AdaptCLIP AAAI 2026	**CrossCLIP** ours
1-Shot	MVTec	96.8±0.2	97.2±0.1	97.5±0.1	**98.0±0.1**	93.4±0.2	92.8±0.0	94.3±0.1	**95.7±0.2**
	VisA	81.7±1.5	87.7±1.1	92.3±0.9	**93.0±0.3**	94.7±0.1	96.4±0.1	96.8±0.0	**97.0±0.0**
	BTAD	80.5±3.3	74.2±1.5	**95.8±0.9**	94.1±0.9	95.6±0.2	95.3±0.3	96.6±0.2	**96.9±0.1**
	DTD	99.0±0.1	**99.2±0.2**	99.1±0.0	**99.2±0.1**	96.5±0.1	97.6±0.1	97.4±0.0	**98.3±0.1**
	MPDD	73.5±2.2	85.6±2.8	83.1±3.7	**87.2±1.2**	94.7±1.3	97.4±0.2	97.4±0.2	**97.7±0.1**
	Avg	86.3±1.5	88.8±1.1	93.6±1.1	**94.3±0.5**	95.0±0.4	95.9±0.1	96.5±0.1	**97.1±0.1**
2-Shot	MVTec	97.3±0.5	97.3±0.1	97.9±0.2	**98.4±0.1**	93.8±0.1	92.9±0.1	94.5±0.0	**95.9±0.3**
	VisA	84.0±0.7	89.1±0.7	93.6±0.6	**94.2±0.2**	95.1±0.1	96.7±0.1	97.1±0.0	**97.3±0.1**
	BTAD	82.5±3.2	75.4±1.5	**95.9±0.1**	94.2±0.7	95.7±0.1	95.5±0.2	96.7±0.1	**96.9±0.1**
	DTD	99.1±0.1	**99.3±0.1**	99.2±0.0	**99.3±0.1**	96.6±0.1	97.7±0.1	97.6±0.0	**98.3±0.1**
	MPDD	73.6±2.0	85.4±2.4	84.7±2.1	**87.9±0.6**	94.2±0.2	97.5±0.1	97.7±0.0	**97.8±0.1**
	Avg	87.3±1.3	89.3±1.0	94.3±0.6	**94.8±0.3**	95.1±0.1	96.1±0.1	96.7±0.0	**97.2±0.1**
4-Shot	MVTec	97.7±0.0	97.8±0.0	98.4±0.2	**98.8±0.1**	94.2±0.2	93.2±0.1	94.8±0.1	**96.1±0.3**
	VisA	85.5±0.9	90.1±0.7	**94.3±0.2**	94.2±0.3	95.1±0.2	96.9±0.1	97.3±0.0	**97.4±0.1**
	BTAD	88.1±1.4	77.5±3.1	**96.4±0.1**	94.5±0.8	95.9±0.1	95.7±0.1	96.8±0.0	**97.0±0.1**
	DTD	99.2±0.1	**99.4±0.1**	99.3±0.0	99.3±0.1	96.8±0.1	97.8±0.0	97.8±0.1	**98.4±0.1**
	MPDD	75.3±0.1	88.0±2.2	87.7±2.4	**90.2±0.6**	94.7±0.3	97.8±0.1	97.9±0.1	**98.0±0.1**
	Avg	89.2±0.5	90.6±1.2	95.2±0.6	**95.4±0.4**	95.3±0.2	96.3±0.1	96.9±0.1	**97.4±0.1**

Table 8: Computational overhead of CLIP-based anomaly detection methods. Parameters are reported as frozen + learnable (k-shot rows include the normal-sample memory bank). GFLOPs and peak memory are measured per image at batch size 1 in fp32 on a single NVIDIA A800-80GB GPU.

Method	k	CLIP Model	Input Size	#Params (M)	Learn. (M)	GFLOPs†	Peak Mem. (GiB)
APRIL-GAN	0	ViT-L/14@336px	518×518	431.91	3.15	1022.24	1.911
WinCLIP	0	ViT-B-16+240	240×240	208.38	0.00	673.45	0.866
AnomalyCLIP	0	ViT-L/14@336px	518×518	434.31	5.56	1222.89	1.998
AdaCLIP	0	ViT-L/14@336px	518×518	439.42	10.67	2678.62	1.936
AA-CLIP‡	0	ViT-L/14@336px	518×518	441.34	12.58	1041.61	1.940
AdaptCLIP	0	ViT-L/14@336px	518×518	429.37	0.61	1223.69	1.999
CrossCLIP (ours)	0	ViT-L/14@336px	518×518	433.00	5.06	1269.40	2.010
AdaptCLIP	1	ViT-L/14@336px	518×518	430.52	1.76	1249.59	2.193
CrossCLIP (ours)	1	ViT-L/14@336px	518×518	433.80	5.06	1286.70	2.020
AdaptCLIP	2	ViT-L/14@336px	518×518	430.52	1.76	1261.10	2.383
CrossCLIP (ours)	2	ViT-L/14@336px	518×518	433.80	5.06	1304.00	2.050
AdaptCLIP	4	ViT-L/14@336px	518×518	430.52	1.76	1284.13	2.759
CrossCLIP (ours)	4	ViT-L/14@336px	518×518	433.80	5.06	1338.50	2.070

A Technical Appendices and Supplementary Material

A.1 Additional Experiment results

In addition to the image-level AUROC and pixel-level AP reported in the main paper, Table 7 provides image-level AP and pixel-level AUROC under the 1-, 2-, and 4-shot settings on five industrial datasets.

A.2 Computation Cost Analysis

Computation Analysis As shown in Table 8, CrossCLIP introduces only 5.06M learnable parameters, comparable to existing CLIP-based baselines, while its GFLOPs and peak memory remain within a similar range across all shot settings. The additional FFT-based operations account for a negligible fraction of the total compute (0.16 GFLOPs), indicating that the proposed cross-modal adaptation adds minimal overhead over the frozen CLIP backbone.

Effect of the CLIP backbone. Tab. 9 compares ViT-B-16+240 and ViT-L/14@336px under the zero- and one-shot settings. ViT-L/14@336px consistently performs better, with the largest gap on MPDD, suggesting that stronger semantic capacity is particularly important for datasets with large viewpoint variation. We adopt ViT-L/14@336px as the default backbone.

Table 9: Experiments on different pre-trained CLIP models. Results are reported as AUROC/AP (%).

Shot	CLIP Model	Image-level AUROC/AP					Pixel-level AUROC/AP				
		MVTec	VisA	BTAD	MPDD	**Avg**	MVTec	VisA	BTAD	MPDD	**Avg**
0-Shot	ViT-B-16+240	87.2/94.3	72.4/77.7	82.6/82.7	63.7/70.8	76.5/81.4	91.9/38.2	93.7/17.3	92.4/38.2	94.3/22.9	93.1/29.2
	ViT-L/14@336px	**91.9/96.6**	**87.6/90.3**	**92.4/92.6**	**79.5/86.9**	**87.9/91.6**	**93.5/46.3**	**95.8/30.0**	**96.2/61.8**	**97.1/35.2**	**95.7/43.3**
1-Shot	ViT-B-16+240	92.2/96.9	83.6/86.7	88.6/92.9	70.5/78.3	83.7/88.7	95.2/50.3	95.9/30.0	95.4/54.1	96.5/33.8	95.8/42.1
	ViT-L/14@336px	**96.0/98.0**	**91.5/93.0**	**94.7/94.1**	**85.7/87.2**	**92.0/93.1**	**95.7/58.2**	**97.0/37.7**	**96.9/62.7**	**97.7/39.4**	**96.8/49.5**

Table 10: Ablation on cross-modal guidance in TCSA. "Independent" removes the textual anchors from the condition of TCSA, so that the visual adapter is adapted without textual guidance; "Guided" is the full CrossCLIP. Results (%) are reported on industrial benchmarks. Best results are in **bold**.

Dataset	Image-level AUROC		Image-level AP		Pixel-level AUROC		Pixel-level AP	
	Independent	Guided	Independent	Guided	Independent	Guided	Independent	Guided
MPDD	75.9	**79.5**	80.0	**86.9**	96.7	**97.1**	31.2	**35.2**
VisA	86.0	**87.6**	87.6	**90.3**	95.4	**95.8**	27.6	**30.0**
BTAD	**92.4**	**92.4**	**93.3**	92.6	95.3	**96.2**	53.4	**61.8**
DTD	97.1	**97.6**	98.9	**99.1**	98.2	**98.4**	**77.2**	76.3
Avg.	87.9	**89.3**	90.0	**92.2**	96.4	**96.9**	47.4	**50.8**

A.3 Additional Ablation studies

Independent vs. cross-modal guided adaptation. To isolate the contribution of textual guidance, Tab. 10 compares TCSA with and without the textual anchors in its condition. Guidance improves the average image-level AUROC/AP by $+1.4$/$+2.2$ and the average pixel-level AUROC/AP by $+0.5$/$+3.4$. The largest gains appear on BTAD ($+8.4$ pixel-level AP) and MPDD ($+6.9$ image-level AP), where defects are small relative to heterogeneous normal parts and knowing what counts as abnormal for the current image is most informative. The independent variant remains competitive on the texture-dominated DTD, where it is slightly better in pixel-level AP, and on BTAD image-level metrics. This is expected, since the anomalous bands of homogeneous textures can largely be identified from visual statistics alone, leaving less room for semantic guidance.

Hyperparameter sensitivity. We ablate two key hyperparameters of TCSA, including the channel reduction ratio (*reduction*) of the frequency MLP and the number of dynamic frequency bands (K). As shown in Table 11, Group I fixes $K = 6$ and varies *reduction*, where *reduction*$= 8$ achieves the best or second-best performance across most metrics with only 2.4M parameters, suggesting that overly large channel capacity may lead to overfitting. Group II fixes *reduction*$= 8$ and varies K, where performance improves up to $K = 6$ but saturates or slightly degrades at $K = 8$, indicating redundancy when too many frequency bands are used. Based on these results, we adopt *reduction*$=$ 8 and $K = 6$ as the default configuration, achieving a favorable balance between accuracy and model complexity.

Frequency modulation strategy. Tab. 12 compares modulating only the phase, only the amplitude, or both. Amplitude rescales the energy of each band and thereby controls which frequency components are amplified, whereas phase encodes spatial arrangement and is sensitive to structural deviations such as contours and part placement. The two are complementary: joint modulation is best on 12 of the 16 metrics and gives the highest averages, with phase-only modulation slightly ahead on BTAD image-level AUROC/AP and DTD image-level AUROC and pixel-level AP, and we adopt it by default.

A.4 Frequency Domain Analysis

We perform frequency domain analysis towards all datasets.



Figure 4: Spectral separability between normal and anomalous image regions across industrial and medical datasets. Stars indicate the observed peaks.

Table 11: Ablation study of CrossCLIP on key hyperparameters. The best and second-best results are highlighted in **bold** and underline, respectively.

Group	Config	Params	Image-level AUROC/AP				Pixel-level AUROC/AP			
			VisA	DTD	MPDD	BTAD	VisA	DTD	MPDD	BTAD
Group I ($K = 6$)	*reduction*= 2	12.4M	80.4/82.9	**98.2/99.2**	75.3/77.6	89.6/93.2	94.4/21.6	**98.5**/76.4	96.6/25.9	93.8/44.4
	reduction= 4	5.4M	82.3/84.8	97.6/99.0	73.4/79.1	90.7/**95.2**	95.4/24.4	98.2/74.6	96.8/28.9	94.9/48.5
	reduction= 8	2.4M	**87.6/90.3**	97.6/99.1	**79.5/86.9**	**92.4**/92.6	**95.8/30.0**	98.4/76.3	**97.1/35.2**	**96.2/61.8**
	reduction= 16	1.2M	83.0/85.3	98.1/99.1	77.9/85.9	90.1/90.6	**95.8**/27.8	**98.5/77.1**	96.9/32.6	94.3/52.6
Group II (*reduction* = 8)	$K = 2$	1.1M	86.6/89.2	**97.6/99.1**	69.9/75.7	92.4/**93.5**	**95.9**/28.2	98.4/75.5	96.6/25.9	**96.2**/60.3
	$K = 4$	1.8M	84.3/88.7	96.9/98.9	71.2/77.3	**93.2**/93.4	95.7/29.5	**98.5**/77.6	96.9/27.1	95.7/60.4
	$K = 6$	2.4M	**87.6/90.3**	**97.6/99.1**	**79.5/86.9**	92.4/92.6	95.8/**30.0**	98.4/76.3	**97.1/35.2**	**96.2/61.8**
	$K = 8$	3.1M	87.2/**90.3**	97.1/98.9	77.7/84.7	88.5/89.0	95.8/29.0	98.4/**77.9**	96.8/30.8	95.4/57.3

Table 12: Ablation study on the frequency-domain MLP modulation strategy in CrossCLIP. "Phase Only" modulates only the phase spectrum; "Amplitude Only" modulates only the amplitude spectrum; "Phase + Amplitude" modulates both simultaneously. The best results are highlighted in **bold**.

Phase	Amplitude	Image-level AUROC/AP				Pixel-level AUROC/AP			
		VisA	DTD	MPDD	BTAD	VisA	DTD	MPDD	BTAD
✓	✗	86.2/88.2	**97.7**/99.0	75.5/81.5	**92.5/93.0**	95.2/27.3	**98.4/77.2**	96.8/31.5	95.7/56.2
✗	✓	84.5/87.5	96.7/98.8	78.2/83.7	88.4/88.5	95.6/29.1	**98.4**/74.5	97.0/32.6	95.3/54.2
✓	✓	**87.6/90.3**	97.6/**99.1**	**79.5/86.9**	92.4/92.6	**95.8/30.0**	**98.4**/76.3	**97.1/35.2**	**96.2/61.8**

Same-image pairing. Each anomalous window is paired with a normal window cropped from the same image. Pairing within one image reduces the differences in illumination, acquisition conditions and content that would otherwise arise when windows from different images are compared.

Spectrum computation. Let $\mathbf{J} \in \mathbb{R}^{w\times w\times 3}$ be an image window with $w = 64$, and let $\mathbf{J}_t$ denote its t-th color channel. Each channel is transformed independently by an orthonormal 2D real DFT, i.e. the transform of Eq. 7 applied to $\mathbf{J}_t$ in place of $\mathbf{X}$ and scaled by $1/w$:

$$\Psi_t(u, v) = \mathcal{F}_2\{\mathbf{J}_t\}(u, v), \qquad t \in \{1, 2, 3\}, \tag{18}$$

The squared magnitudes are averaged over the three RGB channels to obtain the 2D power spectrum:

$$\mathcal{P}_{\mathbf{J}}(u, v) = \frac{1}{3}\sum_{t=1}^{3}\big|\Psi_t(u, v)\big|^2, \tag{19}$$

where (u, v) indexes the discrete Fourier grid and $\mathcal{P}_{\mathbf{J}}(u, v)$ is the channel-averaged power at that frequency. No spatial mean is subtracted, no window function (e.g. Hann) is applied, and no model features or learned parameters are involved. The one-sided grid returned by FFT, $u \in \{0, \dots, w-1\}$ and $v \in \{0, \dots, w/2\}$, is used directly, without re-weighting the conjugate-symmetric frequencies.

Let $\xi_1(u)$ and $\xi_2(v)$ be the vertical and horizontal spatial frequencies of grid point (u, v), i.e. $\xi_1(u) = u/w$ for $u < w/2$ and $(u - w)/w$ otherwise, and $\xi_2(v) = v/w$. The radial frequency is

$$\nu(u, v) = \sqrt{\xi_1(u)^2 + \xi_2(v)^2}, \tag{20}$$

measured in cycles per native image pixel. Unlike the normalized radial distance $\omega(u, v)$ of Sec. 3.3, which is defined on the grid of patch features, $\nu(u, v)$ is defined on raw pixels and is not normalized. Its maximum is

$$\nu_{\max} = \sqrt{0.5^2 + 0.5^2} = \sqrt{\tfrac{1}{2}} \approx 0.7071. \tag{21}$$

The Nyquist frequency along each axis is 0.5, whereas the 2D radial distance can exceed 0.5.

Radial binning. The radial range is divided into $N_\beta = 48$ equal-width bins with edges

$$\zeta_\beta = \frac{\beta}{N_\beta}\nu_{\max}, \qquad \beta = 0, 1, \dots, N_\beta, \tag{22}$$

where the offset 10^{-6} ensures that the outermost frequency points are included. The frequency points falling into the β-th bin are

$$\Omega_\beta = \big\{(u, v) : \zeta_{\beta-1} \le \nu(u, v) < \zeta_\beta\big\}, \qquad \beta = 1, \dots, N_\beta. \tag{23}$$

We first compute the mean power within each bin,

$$\bar{\mathcal{P}}_\beta(\mathbf{J}) = \frac{1}{|\Omega_\beta|} \sum_{(u,v)\in\Omega_\beta} \mathcal{P}_{\mathbf{J}}(u,v), \tag{24}$$

and then normalize across all bins,

$$\psi_\beta(\mathbf{J}) = \frac{\bar{\mathcal{P}}_\beta(\mathbf{J})}{\sum_{\beta'=1}^{N_\beta} \bar{\mathcal{P}}_{\beta'}(\mathbf{J})}, \tag{25}$$

where $|\Omega_\beta|$ is the number of discrete frequency points in bin β and $\psi_\beta(\mathbf{J})$ is the normalized value used in the paired comparison below. Because the power is averaged over the points of each bin before normalization, $\psi_\beta(\mathbf{J})$ is a normalized radially averaged power rather than the fraction of total energy contained in each annulus, and it differs in this respect from the band energies E_k^{n} and E_k^{Δ} of Appendix **??**. The normalization attenuates the effect of a uniform multiplicative brightness scaling, but it does not remove all illumination effects. For plotting, each bin is placed at the mean radius of the discrete frequency points it actually contains,

$$\bar{\nu}_\beta = \frac{1}{|\Omega_\beta|} \sum_{(u,v)\in\Omega_\beta} \nu(u,v). \tag{26}$$

Paired spectral separability. Consider a category with N_p valid image pairs, and let $\mathbf{J}_h^{\mathrm{a}}$ and $\mathbf{J}_h^{\mathrm{n}}$ denote the anomalous and normal windows of the h-th pair. Their difference in normalized spectrum at bin β is:

$$\Delta\psi_{h,\beta} = \psi_\beta\big(\mathbf{J}_h^{\mathrm{a}}\big) - \psi_\beta\big(\mathbf{J}_h^{\mathrm{n}}\big), \qquad h = 1, \ldots, N_p. \tag{27}$$

We define a sign function with tie correction,

$$\chi(z) = \begin{cases} 1, & z > 0, \\ \frac{1}{2}, & z = 0, \\ 0, & z < 0, \end{cases} \tag{28}$$

and compute the proportion of positive differences, with ties counted as one half:

$$\hat{\mu}_\beta = \frac{1}{N_p} \sum_{h=1}^{N_p} \chi\big(\Delta\psi_{h,\beta}\big) = \frac{1}{N_p} \sum_{h=1}^{N_p} \Big[\mathbf{1}\big(\Delta\psi_{h,\beta} > 0\big) + \tfrac{1}{2}\,\mathbf{1}\big(\Delta\psi_{h,\beta} = 0\big)\Big], \tag{29}$$

where $\mathbf{1}(\cdot)$ is the indicator function, equal to 1 if its condition holds and 0 otherwise. The paired spectral separability of bin β is then

$$\mathrm{Sep}_\beta = \big|2\hat{\mu}_\beta - 1\big|, \qquad 0 \le \mathrm{Sep}_\beta \le 1, \tag{30}$$

which is plotted as the blue curve in Fig. **??**. Sep_β therefore measures the directional consistency of the paired spectral differences. The absolute value is taken so that bands in which the anomalous region is consistently stronger and bands in which it is consistently weaker than the normal region are both captured.

A.5 Additional Method Details

A.5.1 Training Objectives

VCTA and TCSA are optimized jointly. Both anomaly maps are supervised by a combination of focal and Dice losses against the ground-truth mask, and the image-level logit by a binary cross-entropy loss; the losses of the textual and the spectral branch are summed with equal weight. TRA is trained separately under the same combination of objectives, together with the regime-consistency term of Eq. 17. Optimization hyperparameters are listed in Sec. 4.1.

A.5.2 TEXTUAL REFERENCE ADAPTER

We give the full form of the two decoding heads summarized in Sec. 3.4. The segmentation head maintains M candidate transposed-convolution kernels $\{\mathbf{W}_k\}_{k=1}^{M}$ and mixes them with coefficients predicted from the query global feature $\mathbf{G}_q$ and the reference summary $\mathbf{G}_r$:

$$\boldsymbol{\lambda} = \mathrm{Softmax}\big(\mathrm{MLP}([\mathbf{G}_q \,\|\, \mathbf{G}_r])\big) \in \Delta^{M-1}, \qquad \mathbf{W}^* = \sum_{k=1}^{M} \lambda_k \, \mathbf{W}_k. \tag{31}$$

The mixed kernel $\mathbf{W}^*$ is applied as a transposed convolution that upsamples $\mathbf{Z}^{(\ell)}$ to the input resolution $I \times I$, and its two-channel output is normalized by a softmax over the channel dimension:

$$\mathbf{S}_{\mathrm{s}}^{(\ell)} = \mathrm{Softmax}\big(\mathrm{ConvT}(\mathbf{Z}^{(\ell)}; \, \mathbf{W}^*)\big). \tag{32}$$

The classification head pools $\mathbf{Z}^{(\ell)}$ by averaging its global mean $\overline{\mathbf{Z}}^{(\ell)}$ with the mean of its top-10 activated patches, and maps the result to an image-level logit:

$$\bar{\mathbf{z}}^{(\ell)} = \frac{1}{2}\Big(\overline{\mathbf{Z}}^{(\ell)} + \text{TopK-mean}_{10}(\mathbf{Z}^{(\ell)})\Big), \qquad \hat{\mathbf{y}}^{(\ell)} = \mathrm{MLP}(\bar{\mathbf{z}}^{(\ell)}). \tag{33}$$

Unlike the segmentation head, the classification head uses static weights, since its pooling is by design insensitive to instance-level variation.

A.6 LIMITATIONS.

Although CrossCLIP achieves strong zero- and few-shot performance, several limitations remain. Firstly, the strong transfer to medical images may partly come from the fact that polyps and tumors are relatively large and dominated by low-frequency content. Performance on small, subtle lesions or on other modalities such as CT and MRI has not been verified. Secondly, performance remains limited on objects with large viewpoint variation or strong intra-class appearance changes, e.g., Bracket Brown on MPDD, and CrossCLIP does not surpass AdaptCLIP in few-shot pixel-level AP on VisA, where objects appear in multiple instances with complex layouts. Future work will explore learnable band partitions and training with weaker or synthetic supervision.

A.7 DATASETS DETAIL

We provide detailed descriptions of the 11 benchmarks used in our experiments, spanning industrial anomaly detection and medical anomaly detection domains.

MVTec AD (Bergmann et al., 2021) is a comprehensive benchmark for unsupervised anomaly detection and localization in industrial inspection. It contains 15 categories of objects and textures, comprising 5,354 high-resolution images in total. Training images are anomaly-free, while the test set includes both normal samples and samples with various types of defects, each accompanied by pixel-level ground-truth annotations.

VisA (Zou et al., 2022) (Visual Anomaly) is a large-scale industrial anomaly detection dataset containing 12 categories of objects, including PCB boards, capsules, and various other manufactured components. It consists of 10,821 images with 9,621 normal and 1,200 anomalous images, with pixel-level annotations available for all anomalous samples.

BTAD (Mishra et al., 2021) (BeanTech Anomaly Detection) is an industrial dataset capturing real-world scenarios from a manufacturing pipeline. It consists of 3 product categories with a total of 2,830 images. Both image-level labels and pixel-level segmentation masks are provided for anomalous test images.

DTD-Synthetic (Aota et al., 2023) is a synthetically generated anomaly detection dataset built upon the Describable Textures Dataset (DTD). Anomalies are procedurally injected into texture images, yielding a diverse set of defect patterns with precise pixel-level ground truth, which facilitates controlled evaluation of texture-based anomaly detection methods.

MPDD (Jezek et al., 2021) (Metal Parts Defect Detection) focuses on detecting surface defects on metal parts. It contains 6 categories of metal components with 1,346 training images and 1,282 test images. Pixel-level annotations are available for test samples exhibiting defects.

Table 13: Summary of datasets used for evaluation, spanning industrial and medical anomaly detection domains.

Domain	Dataset	Modality	Category	# Classes	Pose-Agnostic	Anomaly Annotations		Train	Test	
						Image-Level	Pixel-Level	# Normal	# Normal	# Anomaly
Industrial	MVTec	Photography	Obj & Texture	15	✗	✓	✓	3,629	467	1,258
	VisA	Photography	Obj	12	✗	✓	✓	8,659	962	1,200
	BTAD	Photography	Obj & Texture	3	✗	✓	✓	1,799	451	290
	DTD	Photography	Texture	12	✗	✓	✓	1,200	357	947
	MPDD	Photography	Obj	6	✓	✓	✓	888	176	282
Medical	Br35H	Radiology (MRI)	Brain	1	✓	✓	✗	0	1,500	1,500
	Covid	Radiology (X-ray)	Chest	1	✓	✓	✗	0	1,341	219
	Kvasir	Endoscopy	Gastrointestinal tract	1	✓	✗	✓	0	0	1,000
	Endo	Endoscopy	Gastrointestinal tract	1	✓	✗	✓	0	0	200
	ColonDB	Endoscopy	Gastrointestinal tract	1	✓	✗	✓	0	0	380
	ClinicDB	Endoscopy	Gastrointestinal tract	1	✓	✗	✓	0	0	612

Br35H (Hamada, 2020) is a brain MRI dataset for image-level binary classification of brain tumors. It contains 3,000 brain MRI scans, split into tumor-positive and tumor-negative samples. As pixel-wise annotations are not provided, this dataset is used exclusively for image-level anomaly detection evaluation.

Covid-19 (Peng et al., 2020) is a chest X-ray dataset designed for binary classification of COVID-19 infection. It comprises radiographic images from COVID-19-positive patients and healthy controls. Similar to Br35H, pixel-level annotations are absent, restricting evaluation to the image-level setting.

Kvasir-Seg (Jha et al., 2020) is an endoscopic dataset containing 1,000 gastrointestinal polyp images along with corresponding pixel-level segmentation masks. The images were collected and annotated by expert gastroenterologists, making it a standard benchmark for polyp segmentation and anomaly localization.

Endo (Borgli et al., 2020) is an endoscopy dataset providing colonoscopy images with associated pixel-wise ground truth annotations for lesion regions. It serves as a challenging benchmark due to the subtle appearance of lesions against complex mucosal backgrounds.

Colon-DB (Bernal et al., 2012) is a colonoscopy dataset comprising images of colonic polyps with pixel-level annotation masks. It is widely used for evaluating polyp detection and segmentation algorithms, featuring significant variation in polyp size, shape, and appearance.

Clinic-DB (Bernal et al., 2015) (CVC-ClinicDB) is a database of frames extracted from colonoscopy videos, containing 612 images with ground-truth binary masks indicating polyp regions. It is a commonly adopted benchmark for medical image segmentation and pixel-level anomaly detection in clinical settings.

A.8 Detailed Per-category datasets performances

We present per-category performances for each datasets. For both image-level and pixel-level classification and segmentation, Area Under the ROC Curve (AUROC), Average Precision (AP) and maximum F1 score (F1-max) are used as evaluation metrics.

A.9 Competing Methods

Result Sources. For fair comparison, the results of WinCLIP, AnomalyCLIP, AdaCLIP, and AdaptCLIP reported in Tables 1–3 are taken directly from AdaptCLIP (Gao et al., 2026) under the same evaluation protocol. CLIP, APRIL-GAN, and AA-CLIP are re-implemented under our training and testing protocol (marked with † in the tables), as no results under this protocol are publicly available. AnomalyCLIP+ follows the few-shot implementation described in (Gao et al., 2026). Below we briefly summarize each competing method.

WinCLIP (Jeong et al., 2023) pioneered zero-shot anomaly detection by leveraging CLIP, a vision-language foundation model. It crafts binary textual prompts and adopts a sliding multi-scale window strategy to enable fine-grained anomaly localization. A notable limitation is its substantial computational overhead and memory consumption, which restricts its applicability to high-resolution inputs or large backbone models. Since no official codebase has been released, our reported numerical results of WinCLIP are reproduced by (Gao et al., 2026) using an unofficial PyTorch implementation.

Table 14: Per-category few-shot (1-shot, 2-shot, 4-shot) performance of **CrossCLIP** on the MVTec AD dataset. Few-shot results are reported as mean±std over three runs. All metrics are reported in %.

Shot	Category	Image-level			Pixel-level		
		AUROC	AP	F_1-max	AUROC	AP	F_1-max
1-Shot	Bottle	98.40±0.10	99.47±0.06	97.60±0.80	95.17±0.29	73.23±0.75	68.33±0.76
	Cable	98.27±0.68	98.83±0.47	95.03±0.83	97.27±0.23	71.83±0.99	68.30±0.70
	Capsule	96.80±0.46	99.30±0.10	96.27±0.61	97.23±0.25	42.83±1.06	43.63±0.64
	Carpet	100.00±0.00	99.90±0.00	100.00±0.00	99.37±0.06	72.87±1.27	70.00±1.30
	Grid	98.97±0.23	99.47±0.06	97.10±0.44	95.57±1.63	41.07±1.81	43.97±1.17
	Hazelnut	99.97±0.06	99.70±0.00	99.53±0.40	98.23±0.21	69.87±0.92	66.10±0.62
	Leather	100.00±0.00	99.90±0.00	100.00±0.00	99.60±0.00	62.77±1.22	61.80±1.15
	Metal Nut	99.17±0.45	99.70±0.10	98.53±0.64	81.30±2.91	44.17±4.15	42.10±4.72
	Pill	94.03±0.32	98.17±0.15	96.20±0.62	92.37±0.67	44.20±1.13	43.97±1.25
	Screw	67.30±0.69	84.03±0.75	86.73±0.06	97.40±0.36	13.07±1.79	18.27±2.36
	Tile	99.27±0.21	99.60±0.00	97.63±0.06	97.80±0.17	79.47±0.85	73.73±1.14
	Toothbrush	99.70±0.30	99.47±0.06	98.90±0.95	99.17±0.12	73.43±2.15	72.93±1.55
	Transistor	97.07±0.93	95.23±1.07	89.73±2.83	90.73±1.17	51.57±3.26	49.03±3.15
	Wood	99.70±0.10	99.73±0.06	98.93±0.46	98.07±0.40	79.23±0.67	72.47±0.75
	Zipper	91.73±0.21	97.83±0.06	91.93±0.38	96.50±0.26	54.07±1.36	53.57±1.04
	Average	**96.00±0.10**	**98.03±0.06**	**96.30±0.17**	**95.70±0.20**	**58.23±0.98**	**56.53±1.08**
2-Shot	Bottle	98.60±0.17	99.50±0.00	97.07±0.46	95.33±0.32	73.53±0.80	69.13±0.83
	Cable	98.53±0.55	99.00±0.36	96.33±0.64	97.40±0.17	72.07±0.93	68.40±0.70
	Capsule	97.50±0.46	99.40±0.10	96.87±0.45	97.43±0.25	44.33±1.20	44.30±0.79
	Carpet	100.00±0.00	99.90±0.00	100.00±0.00	99.33±0.06	72.67±1.27	69.83±1.24
	Grid	99.20±0.20	99.53±0.06	97.40±0.00	96.23±1.25	40.00±1.77	44.23±1.37
	Hazelnut	99.97±0.06	99.70±0.00	99.53±0.40	98.30±0.17	69.87±0.84	66.20±0.60
	Leather	100.00±0.00	99.90±0.00	100.00±0.00	99.60±0.00	62.40±1.21	61.43±1.20
	Metal Nut	99.73±0.31	99.87±0.06	99.27±0.64	81.57±2.61	44.77±3.87	42.97±4.38
	Pill	93.70±0.56	97.83±0.06	96.40±0.61	92.50±0.70	44.40±1.22	44.23±1.40
	Screw	74.77±0.21	89.00±0.44	88.03±0.06	97.97±0.15	24.23±1.91	27.70±2.00
	Tile	99.40±0.17	99.63±0.06	97.47±0.32	97.83±0.12	79.57±0.67	73.73±1.11
	Toothbrush	99.90±0.17	99.50±0.00	99.47±0.92	99.27±0.06	73.10±1.73	73.67±0.74
	Transistor	97.17±1.01	95.23±1.24	90.73±2.14	90.67±1.19	51.67±3.18	48.97±3.01
	Wood	99.70±0.10	99.73±0.06	98.93±0.46	98.07±0.40	79.17±0.72	72.20±0.78
	Zipper	92.07±0.31	97.90±0.10	92.77±0.23	96.67±0.23	54.77±1.44	53.90±1.06
	Average	**96.67±0.15**	**98.37±0.06**	**96.67±0.15**	**95.87±0.25**	**59.10±0.95**	**57.40±0.95**
4-Shot	Bottle	98.67±0.15	99.50±0.00	97.07±0.46	95.50±0.26	74.10±0.80	69.80±0.89
	Cable	99.13±0.35	99.37±0.15	97.27±0.55	97.57±0.15	72.57±0.84	68.70±0.61
	Capsule	97.43±0.80	99.37±0.15	97.17±0.46	97.63±0.25	45.47±1.35	44.80±0.85
	Carpet	100.00±0.00	99.90±0.00	100.00±0.00	99.27±0.06	72.37±1.33	69.23±1.42
	Grid	99.10±0.17	99.50±0.00	97.40±0.00	96.07±1.12	40.57±1.40	45.00±1.04
	Hazelnut	99.97±0.06	99.70±0.00	99.30±0.00	98.20±0.17	68.80±0.79	65.27±0.57
	Leather	100.00±0.00	99.90±0.00	100.00±0.00	99.60±0.00	62.47±1.06	61.40±1.05
	Metal Nut	99.90±0.10	99.90±0.00	99.47±0.55	82.70±2.59	47.87±4.67	45.70±4.91
	Pill	94.47±0.35	98.30±0.17	96.70±0.17	93.13±0.74	46.37±1.42	45.83±1.56
	Screw	81.77±0.51	92.83±0.49	88.57±0.75	98.13±0.15	26.90±2.10	31.73±1.21
	Tile	99.37±0.15	99.63±0.06	97.30±0.35	97.87±0.06	79.43±0.72	73.70±1.15
	Toothbrush	99.90±0.17	99.50±0.00	99.47±0.92	99.23±0.06	73.20±1.57	73.53±0.67
	Transistor	97.90±0.69	96.10±0.78	92.17±2.46	91.07±1.18	53.57±3.18	50.57±3.02
	Wood	99.70±0.10	99.73±0.06	98.93±0.46	98.10±0.35	79.13±0.61	72.13±0.81
	Zipper	93.17±0.49	98.20±0.10	92.63±0.12	96.87±0.23	56.20±1.65	54.93±1.25
	Average	**97.37±0.15**	**98.77±0.06**	**96.90±0.17**	**96.07±0.25**	**59.93±0.98**	**58.17±1.18**

WinCLIP+ (Jeong et al., 2023) extends WinCLIP by incorporating visual-guided predictions alongside language-guided ones to boost both anomaly classification and segmentation performance. The language branch follows the same design as WinCLIP. The visual branch constructs a memory bank by storing multi-scale features extracted from a small set of normal reference images, and assigns anomaly scores based on the distance between each query feature and its nearest neighbor in the bank. The two branches are combined by averaging their respective anomaly scores.

Table 15: Per-category zero-shot (0-shot) and few-shot (1-shot, 2-shot, 4-shot) performance of **CrossCLIP** on the VisA dataset. Few-shot results are reported as mean±std over three runs. All metrics are reported in %.

Shot	Category	Image-level			Pixel-level		
		AUROC	AP	F_1-max	AUROC	AP	F_1-max
0-Shot	Candle	81.9	86.3	76.4	98.7	25.7	37.7
	Capsules	92.4	95.8	88.2	96.1	31.6	41.3
	Cashew	91.7	96.1	89.0	94.8	27.1	30.5
	Chewing Gum	96.7	98.7	95.5	99.6	83.2	76.9
	Fryum	90.0	95.4	88.3	94.5	22.7	29.6
	Macaroni1	89.3	91.8	82.2	99.5	27.0	34.7
	Macaroni2	83.0	85.0	77.0	99.3	20.5	30.9
	PCB1	86.2	88.9	79.2	92.7	12.2	18.0
	PCB2	83.8	84.5	77.1	93.6	21.4	30.6
	PCB3	71.0	73.1	71.1	90.0	16.3	23.2
	PCB4	89.2	90.4	82.8	92.7	16.0	20.2
	Pipe Fryum	95.7	97.8	93.6	98.6	56.2	59.9
	Average	**87.6**	**90.3**	**83.4**	**95.8**	**30.0**	**36.1**
1-Shot	Candle	93.87±0.72	94.47±0.57	86.27±0.55	98.97±0.06	26.70±0.44	41.90±0.20
	Capsules	95.83±0.93	97.37±0.59	93.30±1.47	96.37±0.32	24.00±1.22	34.47±1.27
	Cashew	96.00±0.26	98.23±0.06	92.73±0.38	97.63±0.78	53.90±3.00	57.63±2.57
	Chewing Gum	98.23±0.12	99.23±0.06	95.90±0.26	99.57±0.06	77.67±0.45	74.37±0.25
	Fryum	97.63±0.55	99.07±0.15	97.00±0.00	96.23±0.12	36.30±1.35	42.77±1.39
	Macaroni1	91.97±0.46	93.13±0.32	84.57±1.11	99.23±0.06	19.30±1.56	27.63±0.68
	Macaroni2	72.60±1.15	78.37±0.60	69.07±1.10	98.30±0.10	12.90±0.26	24.67±0.42
	PCB1	91.10±1.15	89.83±1.20	84.30±0.70	97.97±0.21	52.47±2.37	55.87±2.06
	PCB2	82.23±1.33	84.97±0.59	75.07±1.26	94.23±0.40	18.23±0.64	29.00±0.75
	PCB3	80.80±1.59	83.20±1.21	74.40±1.06	90.40±0.89	29.10±1.65	36.07±1.12
	PCB4	98.30±0.70	97.93±0.91	94.97±0.46	96.10±0.26	29.77±2.67	33.33±2.42
	Pipe Fryum	99.77±0.12	99.73±0.06	98.50±0.50	99.33±0.15	71.77±1.63	69.17±1.55
	Average	**91.53±0.31**	**92.97±0.25**	**87.20±0.26**	**97.00±0.00**	**37.70±0.46**	**43.90±0.62**
2-Shot	Candle	94.27±0.74	94.67±0.55	87.73±0.68	99.00±0.00	25.80±0.52	40.87±0.35
	Capsules	96.97±1.01	98.13±0.57	93.63±1.42	96.93±0.29	29.27±1.12	37.40±1.30
	Cashew	97.33±0.25	98.73±0.06	93.73±0.12	97.83±0.81	57.47±3.21	60.47±2.96
	Chewing Gum	98.50±0.10	99.40±0.00	96.73±0.23	99.60±0.00	78.00±0.40	74.43±0.25
	Fryum	97.43±0.55	98.97±0.15	96.00±0.50	96.33±0.12	37.30±1.44	43.80±1.54
	Macaroni1	91.53±0.67	92.87±0.47	84.27±1.30	99.23±0.06	18.77±1.77	27.17±0.60
	Macaroni2	77.87±1.27	82.70±0.69	71.83±1.33	98.87±0.06	15.00±0.26	26.80±0.10
	PCB1	93.57±1.27	92.37±1.27	86.47±1.04	98.13±0.15	56.87±2.01	59.67±1.86
	PCB2	87.63±0.98	89.50±0.79	81.17±0.67	94.97±0.45	19.40±0.70	30.57±0.84
	PCB3	83.17±1.50	85.13±1.59	76.40±0.46	90.63±0.74	32.07±1.75	38.33±1.14
	PCB4	98.80±0.56	98.50±0.70	95.47±1.10	96.60±0.26	33.43±2.39	36.40±2.21
	Pipe Fryum	99.80±0.10	99.77±0.06	98.50±0.50	99.13±0.15	63.53±2.41	64.37±2.21
	Average	**93.07±0.25**	**94.20±0.20**	**88.47±0.25**	**97.27±0.06**	**38.93±0.32**	**45.03±0.50**
4-Shot	Candle	95.83±0.49	95.80±0.36	90.27±1.02	99.07±0.06	25.17±0.46	40.33±0.35
	Capsules	96.83±0.96	98.03±0.57	93.47±0.85	97.17±0.23	29.80±1.25	37.97±1.37
	Cashew	97.07±0.35	98.63±0.06	93.43±0.38	97.90±0.95	59.57±3.72	62.00±2.81
	Chewing Gum	98.40±0.17	99.37±0.06	96.33±0.21	99.60±0.00	78.23±0.45	74.43±0.31
	Fryum	98.00±0.61	99.13±0.21	96.50±0.50	96.43±0.12	38.37±1.36	44.60±1.28
	Macaroni1	92.23±0.68	93.17±0.46	84.03±0.93	99.30±0.00	18.80±1.74	27.43±0.67
	Macaroni2	77.00±1.04	82.43±0.76	71.30±1.08	98.90±0.10	15.70±0.26	27.40±0.10
	PCB1	94.40±1.35	93.07±1.31	88.83±2.52	98.17±0.21	59.60±1.93	61.50±1.35
	PCB2	87.03±1.76	88.67±1.27	80.33±2.60	95.53±0.45	20.90±0.70	32.50±1.39
	PCB3	81.60±1.04	83.47±0.91	74.87±1.65	91.03±0.57	34.27±1.90	40.73±1.12
	PCB4	99.10±0.46	98.83±0.50	96.00±1.32	96.73±0.32	34.43±2.29	37.03±2.14
	Pipe Fryum	99.80±0.17	99.73±0.06	98.67±0.76	99.17±0.12	64.97±2.11	65.83±1.88
	Average	**93.10±0.30**	**94.20±0.30**	**88.67±0.61**	**97.43±0.06**	**40.00±0.30**	**46.00±0.40**

APRIL-GAN (Chen et al., 2023) builds upon WinCLIP's CLIP-based framework and introduces additional linear projection layers at each encoder stage to map patch-level visual features into the joint text-image embedding space, enabling more accurate patch-level anomaly scoring. For zero-

Table 16: Per-category zero-shot (0-shot) and few-shot (1-shot, 2-shot, 4-shot) performance of **CrossCLIP** on the BTAD dataset. Few-shot results are reported as mean±std over three runs. All metrics are reported in %.

Shot	Category	Image-level			Pixel-level		
		AUROC	AP	F_1-max	AUROC	AP	F_1-max
0-Shot	01	97.7	99.1	96.0	95.4	60.5	61.6
	02	84.5	97.6	93.2	96.0	71.3	66.2
	03	94.9	80.9	76.5	97.1	53.7	55.0
	Average	**92.4**	**92.6**	**88.6**	**96.2**	**61.8**	**60.9**
1-Shot	01	96.53±0.55	98.80±0.10	96.10±0.61	96.17±0.06	60.80±1.08	60.90±1.05
	02	90.73±1.40	98.60±0.20	93.87±0.47	96.80±0.20	72.40±0.61	66.17±0.86
	03	96.70±1.01	84.97±2.59	77.97±1.50	97.57±0.25	55.03±0.15	55.33±0.06
	Average	**94.67±0.71**	**94.13±0.87**	**89.33±0.31**	**96.87±0.12**	**62.73±0.25**	**60.80±0.10**
2-Shot	01	96.93±0.25	98.93±0.06	96.13±0.59	96.33±0.06	61.10±1.11	61.10±1.05
	02	90.30±1.35	98.47±0.21	93.93±0.51	96.83±0.15	72.07±0.64	66.10±0.92
	03	96.97±0.86	85.20±2.19	78.13±1.76	97.57±0.25	53.93±0.12	54.57±0.15
	Average	**94.70±0.56**	**94.20±0.72**	**89.47±0.35**	**96.90±0.10**	**62.37±0.15**	**60.60±0.00**
4-Shot	01	97.20±0.26	98.97±0.06	96.47±0.58	96.40±0.00	61.20±1.21	61.20±1.05
	02	90.47±1.36	98.53±0.25	94.07±0.55	96.90±0.10	72.07±0.72	66.07±0.93
	03	97.43±0.91	86.07±2.57	78.07±3.02	97.63±0.25	57.27±0.25	57.03±0.12
	Average	**95.03±0.50**	**94.50±0.82**	**89.53±0.75**	**96.97±0.12**	**63.53±0.25**	**61.40±0.10**

shot detection, anomaly maps are obtained by comparing the projected patch features against text embeddings derived from state and template ensembles. For few-shot detection, multi-scale features of normal reference images are stored in per-stage memory banks, and anomaly scores are computed via cosine similarity between query and reference features; the zero-shot and few-shot maps are then summed to produce the final prediction. Notably, the linear layers are trained only on the source domain and require no additional fine-tuning with reference images at test time.

AnomalyCLIP (Zhou et al., 2024) introduces learnable, object-agnostic text prompts that encode universal notions of normality and abnormality, decoupled from the semantic content of specific foreground objects. This allows the model to generalize across diverse categories without category-specific tuning. Training requires an auxiliary dataset containing both normal and anomalous samples from a source domain. We adopt the officially released model weights for evaluating both anomaly classification and pixel-level segmentation.

AnomalyCLIP+ serves as one of our few-shot baselines, proposed in (Gao et al., 2026) by augmenting AnomalyCLIP (Zhou et al., 2024) with a patch-level memory association mechanism analogous to that in WinCLIP+. The final anomaly score is obtained by fusing the zero-shot prediction from the learnable text prompts with the few-shot prediction derived from normal image features stored in the memory bank.

AdaCLIP (Cao et al., 2024) enhances vision-language alignment by dynamically injecting visual context from query images into textual prompt embeddings, promoting richer cross-modal interaction. In contrast to AnomalyCLIP, AdaCLIP is trained on a broader collection of auxiliary data covering both industrial and medical domains, which may introduce domain bias and complicate fair cross-domain comparison. To ensure a controlled evaluation, we retrain AdaCLIP following the same training protocol as AnomalyCLIP and report results uniformly across all datasets for zero-shot anomaly detection.

AdaptCLIP (Gao et al., 2026) proposes a universal anomaly detection framework built on two key insights. First, visual and textual adapters should be optimized *alternately* rather than jointly: the visual adapter learns adaptive patch and global tokens with fixed textual embeddings, while the textual adapter learns two-class prompt embeddings with fixed visual tokens, avoiding overfitting on limited source data. Second, few-shot comparative learning should incorporate both contextual and aligned residual features: for each query patch, its nearest neighbor is retrieved from the normal support pool, and the absolute residual is fused with the original query token to preserve semantic

Table 17: Per-category zero-shot (0-shot) and few-shot (1-shot, 2-shot, 4-shot) performance of **CrossCLIP** on the DTD-Synthetic dataset. Few-shot results are reported as mean±std over three runs. All metrics are reported in %.

Shot	Category	Image-level			Pixel-level		
		AUROC	AP	F_1-max	AUROC	AP	F_1-max
0-Shot	Blotchy 099	99.4	99.7	98.8	99.7	84.1	76.7
	Fibrous 183	99.0	99.7	98.1	99.3	81.4	74.4
	Marbled 078	98.4	99.6	98.1	99.4	80.6	74.1
	Matted 069	91.7	97.9	92.0	98.2	66.2	62.3
	Mesh 114	91.7	96.8	90.8	96.6	70.6	68.1
	Perforated 037	98.2	99.5	96.9	96.4	73.6	69.2
	Stratified 154	99.8	99.8	99.4	99.6	85.3	76.8
	Woven 001	99.2	99.6	98.6	99.8	84.6	77.5
	Woven 068	97.3	98.6	96.1	99.1	65.5	59.8
	Woven 104	100.0	99.9	100.0	98.2	76.7	71.6
	Woven 125	100.0	99.9	100.0	99.8	87.4	78.8
	Woven 127	96.2	97.6	94.2	94.7	59.4	61.1
	Average	**97.6**	**99.1**	**96.9**	**98.4**	**76.3**	**70.9**
1-Shot	Blotchy 099	99.97±0.06	99.90±0.00	99.80±0.35	99.80±0.00	92.57±0.42	86.17±0.67
	Fibrous 183	100.00±0.00	99.90±0.00	100.00±0.00	99.63±0.06	91.43±0.40	84.67±0.45
	Marbled 078	100.00±0.00	99.90±0.00	100.00±0.00	99.43±0.06	87.63±0.47	82.13±0.64
	Matted 069	99.57±0.15	99.77±0.06	99.40±0.00	99.70±0.00	89.60±0.46	82.73±0.55
	Mesh 114	93.00±0.26	97.27±0.15	90.43±0.87	96.23±0.51	64.80±1.65	62.83±1.83
	Perforated 037	97.20±0.36	99.27±0.06	95.60±0.35	94.53±0.12	69.77±0.38	67.00±0.20
	Stratified 154	99.53±0.23	99.80±0.00	98.53±0.75	99.60±0.00	85.73±0.91	76.30±1.06
	Woven 001	100.00±0.00	99.80±0.00	100.00±0.00	99.70±0.00	78.33±0.49	73.90±0.30
	Woven 068	95.20±0.52	97.37±0.32	91.73±0.40	98.97±0.06	64.60±0.26	59.27±0.15
	Woven 104	99.33±0.38	99.73±0.12	98.10±0.70	97.03±0.67	71.57±0.91	67.57±0.47
	Woven 125	100.00±0.00	99.90±0.00	100.00±0.00	99.77±0.06	87.83±0.60	80.17±0.59
	Woven 127	95.60±0.50	97.17±0.35	94.30±1.05	95.17±0.15	60.93±1.02	63.77±0.47
	Average	**98.27±0.06**	**99.17±0.06**	**97.30±0.10**	**98.30±0.10**	**78.77±0.35**	**73.90±0.50**
2-Shot	Blotchy 099	100.00±0.00	99.90±0.00	100.00±0.00	99.83±0.06	92.63±0.40	86.27±0.67
	Fibrous 183	100.00±0.00	99.90±0.00	100.00±0.00	99.70±0.00	91.80±0.36	84.93±0.45
	Marbled 078	100.00±0.00	99.90±0.00	100.00±0.00	99.43±0.06	87.63±0.47	81.93±0.64
	Matted 069	99.63±0.12	99.77±0.06	99.40±0.00	99.73±0.06	89.73±0.45	82.70±0.60
	Mesh 114	93.37±0.21	97.43±0.12	90.93±0.81	96.30±0.53	64.60±1.65	62.50±1.78
	Perforated 037	97.53±0.38	99.30±0.10	96.80±0.35	94.83±0.12	70.73±0.42	67.30±0.20
	Stratified 154	99.90±0.10	99.90±0.00	99.60±0.35	99.60±0.00	85.67±1.03	76.30±1.23
	Woven 001	100.00±0.00	99.80±0.00	100.00±0.00	99.80±0.00	77.43±0.76	73.80±0.26
	Woven 068	96.13±0.55	98.00±0.35	93.37±1.46	98.97±0.06	65.97±0.29	60.70±0.26
	Woven 104	99.53±0.21	99.80±0.00	98.53±0.38	96.90±0.61	70.80±0.95	66.97±0.47
	Woven 125	100.00±0.00	99.90±0.00	100.00±0.00	99.77±0.06	87.63±0.61	79.97±0.61
	Woven 127	96.13±0.55	97.47±0.32	95.63±0.81	95.27±0.15	61.53±0.91	63.77±0.25
	Average	**98.53±0.12**	**99.27±0.06**	**97.87±0.23**	**98.33±0.06**	**78.83±0.31**	**73.93±0.45**
4-Shot	Blotchy 099	100.00±0.00	99.90±0.00	100.00±0.00	99.90±0.00	92.70±0.46	86.27±0.67
	Fibrous 183	100.00±0.00	99.90±0.00	100.00±0.00	99.70±0.00	91.67±0.38	84.87±0.40
	Marbled 078	100.00±0.00	99.90±0.00	100.00±0.00	99.47±0.06	87.73±0.47	81.80±0.70
	Matted 069	99.60±0.10	99.80±0.00	99.17±0.40	99.73±0.06	89.63±0.45	82.57±0.55
	Mesh 114	93.07±0.42	97.30±0.17	91.23±1.60	96.20±0.53	64.80±1.56	63.13±1.57
	Perforated 037	97.73±0.29	99.37±0.06	97.00±0.00	95.20±0.10	71.83±0.42	68.07±0.25
	Stratified 154	99.87±0.06	99.87±0.06	99.40±0.00	99.60±0.00	85.60±0.79	76.33±0.93
	Woven 001	100.00±0.00	99.80±0.00	100.00±0.00	99.77±0.06	78.37±0.31	74.30±0.46
	Woven 068	96.80±0.66	98.37±0.35	94.67±0.91	99.10±0.00	70.37±0.61	65.87±0.12
	Woven 104	99.57±0.06	99.80±0.00	98.53±0.38	96.40±0.61	70.77±1.00	67.50±0.62
	Woven 125	100.00±0.00	99.90±0.00	100.00±0.00	99.80±0.00	87.87±0.57	80.33±0.60
	Woven 127	96.60±0.35	97.67±0.23	95.87±0.40	95.30±0.10	61.93±0.70	63.70±0.36
	Average	**98.60±0.17**	**99.27±0.06**	**97.97±0.25**	**98.37±0.12**	**79.43±0.40**	**74.57±0.45**

context while highlighting anomalous regions. Once trained on a source domain, AdaptCLIP requires no target-domain fine-tuning, supporting both zero-shot and few-shot inference. Beyond its standalone use, AdaptCLIP also serves as the foundation for our **AnomalyCLIP+** baseline, where its

Table 18: Per-category zero-shot (0-shot) and few-shot (1-shot, 2-shot, 4-shot) performance of **CrossCLIP** on the MPDD dataset. Few-shot results are reported as mean±std over three runs. All metrics are reported in %.

Shot	Category	Image-level			Pixel-level		
		AUROC	AP	F_1-max	AUROC	AP	F_1-max
0-Shot	Bracket Black	87.0	92.7	86.4	98.9	34.7	44.2
	Bracket Brown	43.2	62.6	80.3	96.1	11.6	21.1
	Bracket White	95.3	95.8	91.2	99.9	31.8	41.0
	Connector	90.7	85.5	78.6	97.5	23.5	27.3
	Metal Plate	69.4	89.0	84.7	92.0	54.9	57.1
	Tubes	91.2	96.0	90.8	98.2	54.4	51.9
	Average	**79.5**	**86.9**	**85.3**	**97.1**	**35.2**	**40.4**
1-Shot	Bracket Black	72.47±0.83	76.77±1.22	77.57±1.37	97.77±0.23	15.10±0.17	26.13±1.21
	Bracket Brown	72.20±7.28	83.70±4.31	84.70±1.61	94.53±0.21	5.27±0.06	10.43±0.21
	Bracket White	82.53±3.14	81.47±6.06	80.20±2.02	99.80±0.00	21.97±0.67	33.83±0.91
	Connector	89.60±5.38	82.50±5.60	76.47±8.83	96.63±0.12	31.93±0.55	34.13±0.15
	Metal Plate	99.73±0.15	99.73±0.06	98.60±0.70	98.37±0.12	89.37±0.86	81.27±0.83
	Tubes	97.73±1.01	98.90±0.46	94.87±2.15	99.10±0.17	72.80±2.19	69.77±1.46
	Average	**85.73±1.17**	**87.17±1.19**	**85.37±0.57**	**97.70±0.10**	**39.40±0.36**	**42.63±0.25**
2-Shot	Bracket Black	73.80±0.53	74.10±1.42	79.50±0.95	97.83±0.21	13.57±0.40	22.60±1.59
	Bracket Brown	71.20±3.89	80.47±4.27	82.60±0.30	94.90±0.30	5.90±0.20	11.67±0.50
	Bracket White	87.63±0.97	86.23±0.15	84.60±2.11	99.87±0.06	19.67±1.07	30.40±0.95
	Connector	92.00±6.16	87.77±7.38	85.37±7.18	96.83±0.12	31.77±0.61	32.67±0.23
	Metal Plate	100.00±0.00	99.80±0.00	100.00±0.00	98.60±0.20	91.13±1.11	82.67±1.37
	Tubes	98.27±0.64	99.10±0.26	95.50±0.70	98.83±0.23	71.27±2.59	68.40±2.02
	Average	**87.13±0.47**	**87.90±0.60**	**87.93±0.76**	**97.83±0.06**	**38.90±0.17**	**41.40±0.30**
4-Shot	Bracket Black	74.20±1.61	74.27±1.04	80.30±1.61	98.00±0.20	11.80±1.30	20.73±2.22
	Bracket Brown	76.13±2.16	84.03±2.40	82.87±1.00	95.63±0.25	7.77±0.38	15.23±0.68
	Bracket White	94.33±1.78	93.73±1.35	90.23±1.46	99.90±0.00	20.00±1.01	30.73±0.83
	Connector	93.03±3.48	90.03±4.61	85.17±5.70	97.10±0.10	32.13±1.14	33.60±0.44
	Metal Plate	100.00±0.00	99.80±0.00	100.00±0.00	98.77±0.21	92.37±1.00	84.07±1.37
	Tubes	98.57±0.64	99.23±0.25	96.13±1.06	98.87±0.21	72.77±2.48	69.50±2.27
	Average	**89.37±0.38**	**90.20±0.60**	**89.10±0.72**	**98.03±0.06**	**39.47±0.15**	**42.33±0.35**

patch-level association mechanism is incorporated into AnomalyCLIP to support few-shot anomaly detection.

A.10 More Visualization Results

In our main paper, we only visualize some randomly selected examples from MVTec, VisA, BTAD, MPDD, DTD, Kvasir, Endo,ColonDB and ClinicDB to illustrate CrossCLIP's anomaly localization capability. Here, we show more visualizations covering every category of each industrial dataset (Figures 5–9).

A.11 Additional Qualitative Results

We provide additional feature visualizations on BTAD, MPDD, MVTec AD, and VisA in Figures 10–13. Each group presents input images, CLIP backbone spatial feature output and frequency-domain feature visualization from AdaCLIP, AA-CLIP, AdaptCLIP, and CrossCLIP. The frequency maps are presented by Fourier transform with logarithmic compression, as low-frequency components centered in each map. These qualitative examples suggest that CrossCLIP attenuates irrelevant frequency information, making potentially anomaly-related spectral structures more prominent.

A.12 Impacts

As a method for learning robust and discriminative representations for anomaly detection, the proposed approach does not raise specific ethical concerns or negative social impacts. All datasets used in this work are publicly available. In addition, the qualitative visualizations are conducted on in-

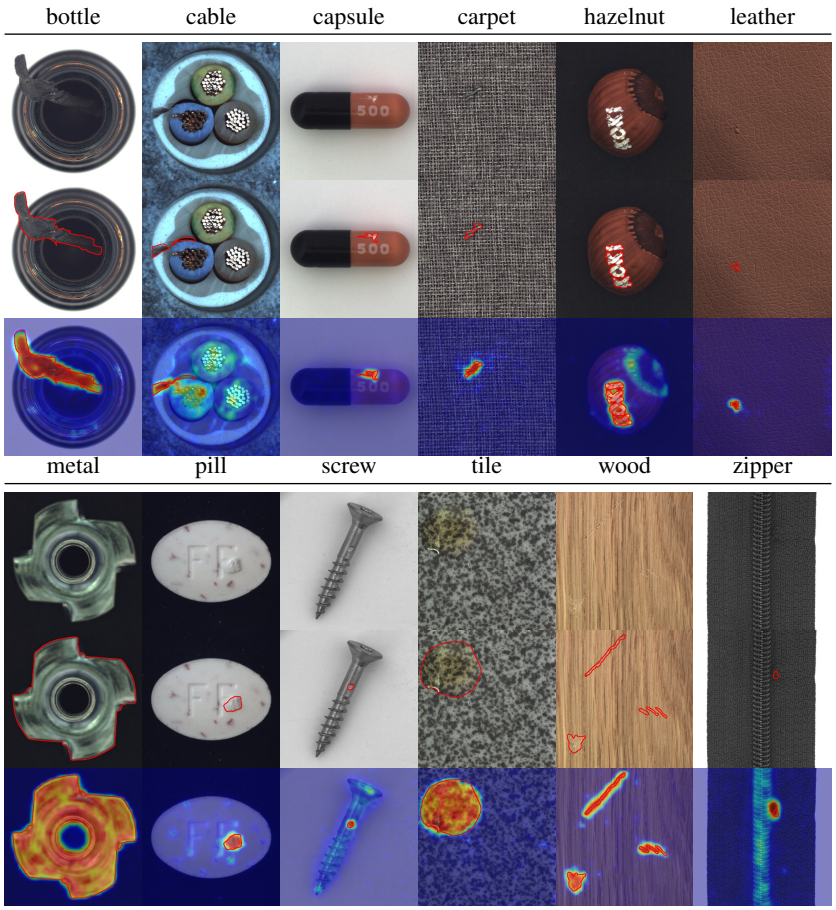

Figure 5: Visualization results of MVTec-AD dataset

dustrial product images and do not involve any personal or sensitive information, thereby avoiding issues related to privacy infringement.

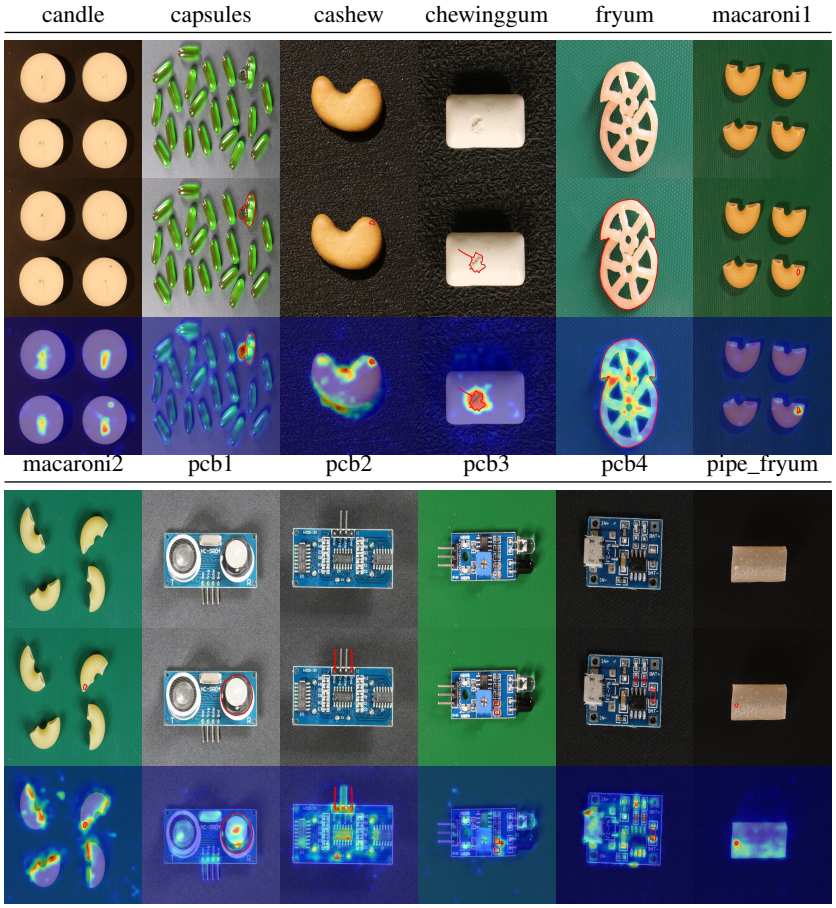

Figure 6: Visualization results of VisA dataset

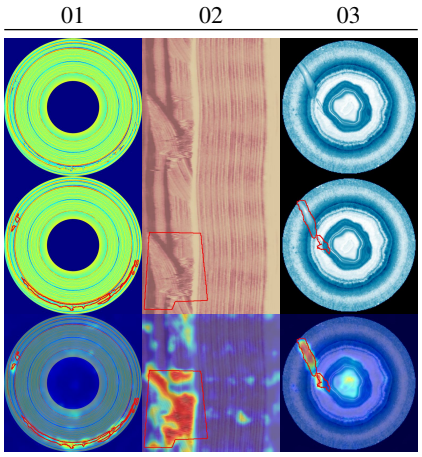

Figure 7: Visualization results of BTAD dataset

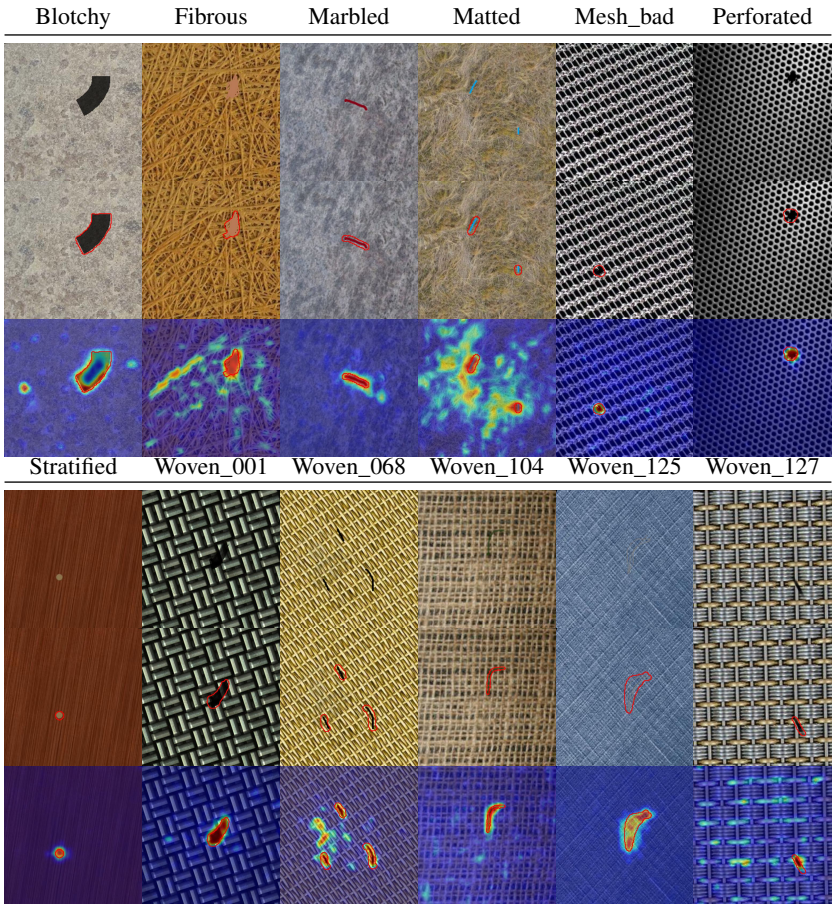

Figure 8: Visualization results of DTD dataset

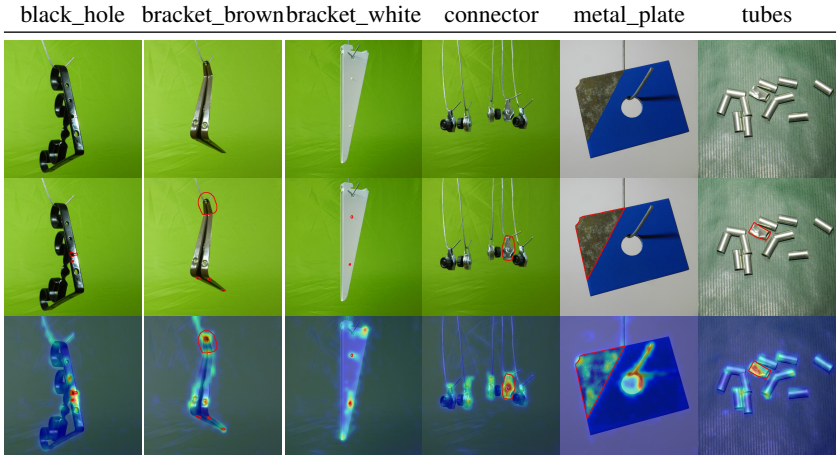

Figure 9: Visualization results of MPDD dataset

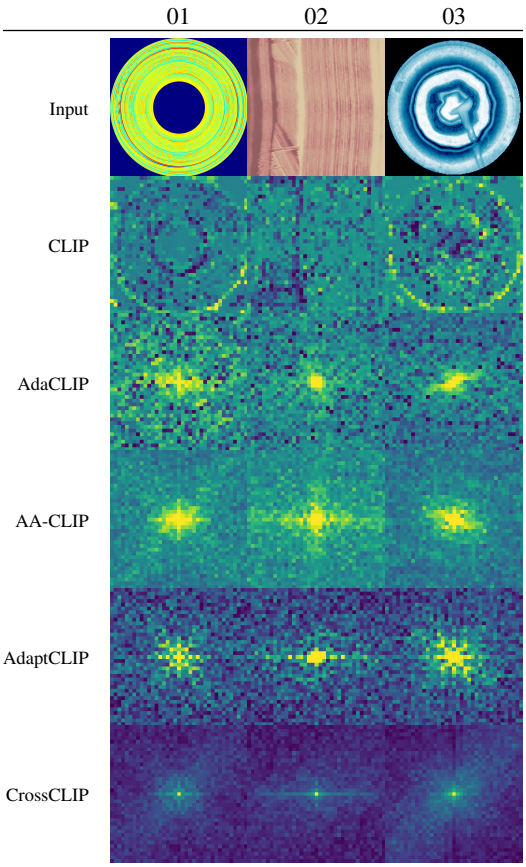

Figure 10: Feature visualizations on BTAD: CLIP spatial features and AdaCLIP/AA-CLIP/AdaptCLIP/CrossCLIP frequency features.

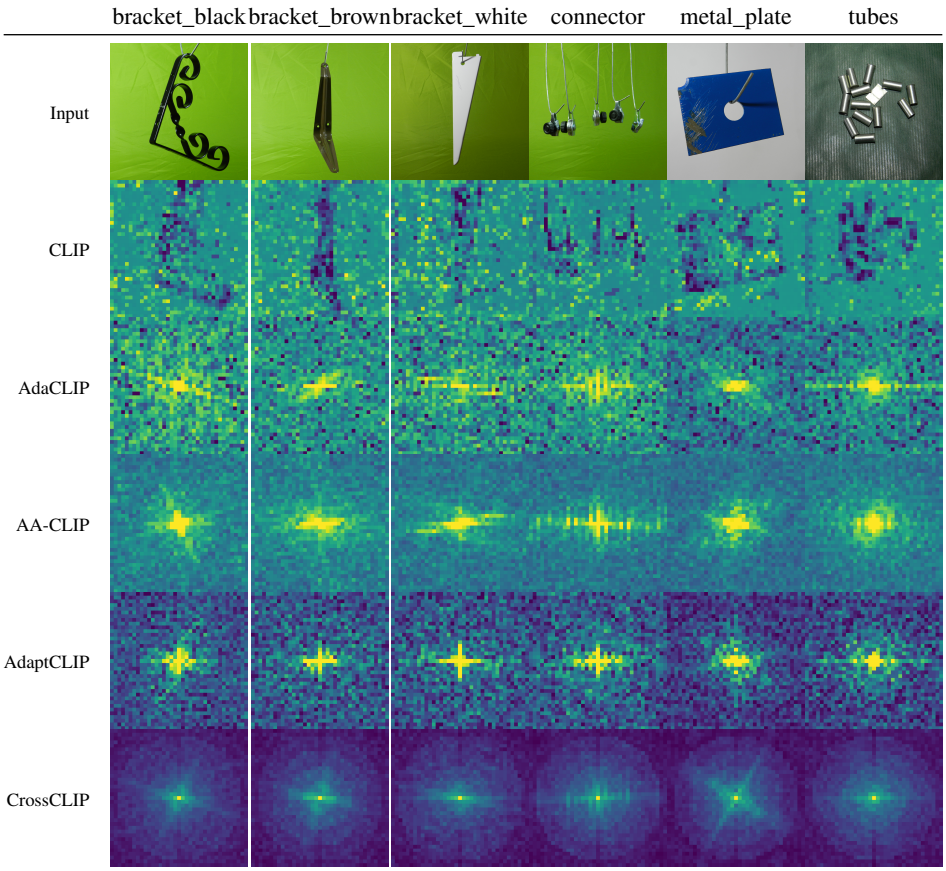

Figure 11: Feature visualizations on MPDD: CLIP spatial features and AdaCLIP/AA-CLIP/AdaptCLIP/CrossCLIP frequency features.

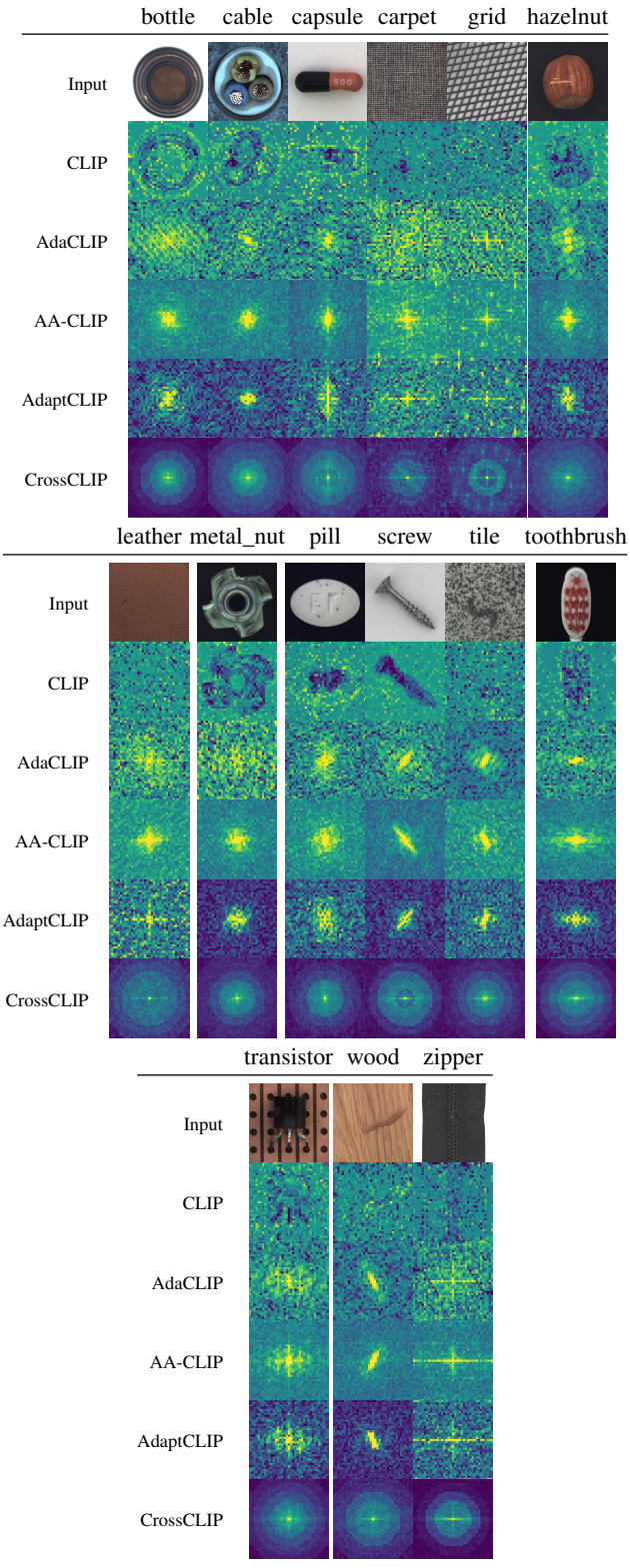

Figure 12: Feature visualizations on MVTec: CLIP spatial features and AdaCLIP/AA-CLIP/AdaptCLIP/CrossCLIP frequency features.

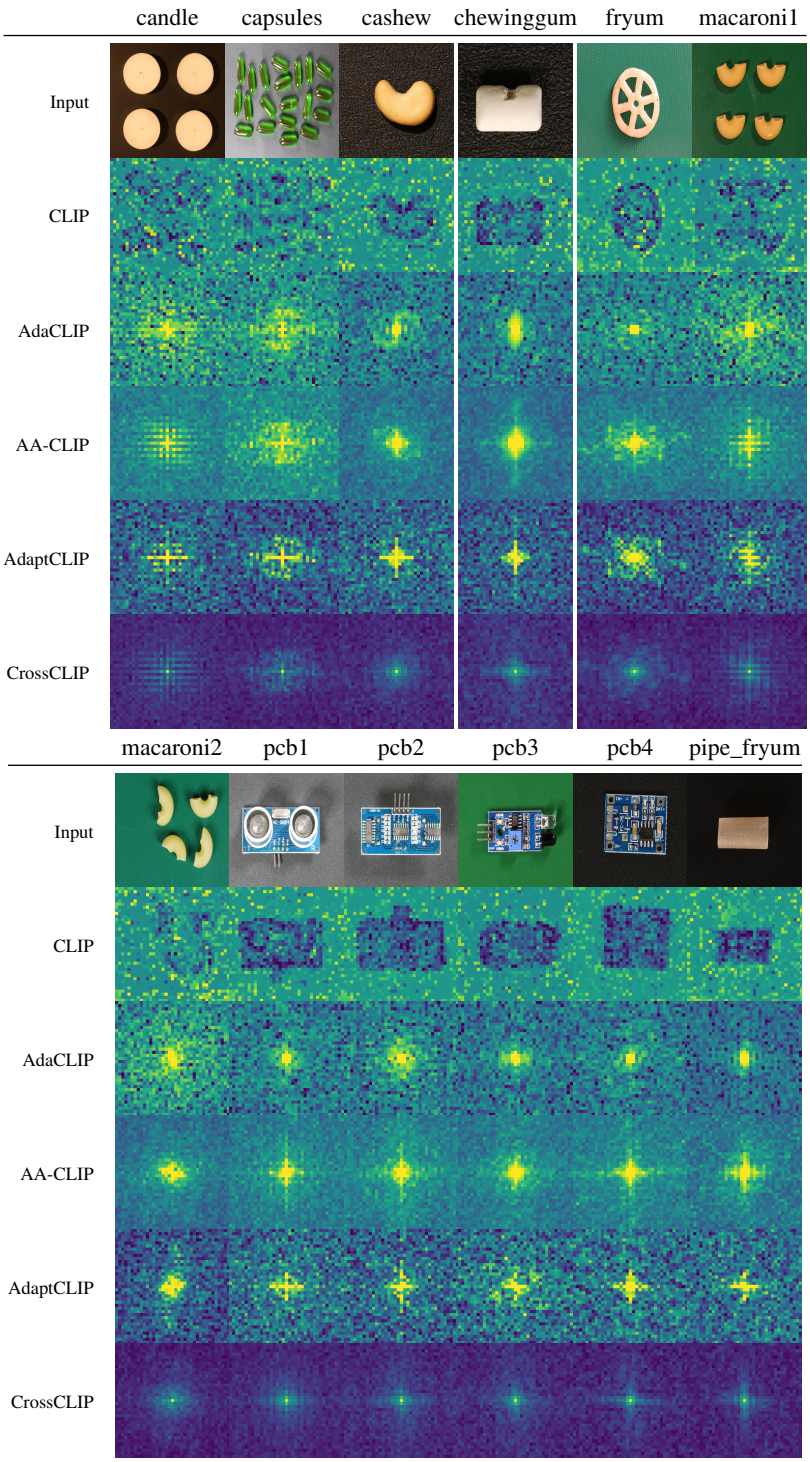

Figure 13: Feature visualizations on VisA: CLIP spatial features and AdaCLIP/AA-CLIP/AdaptCLIP/CrossCLIP frequency features.