FB-CLIP: Fine-Grained Zero-Shot Anomaly Detection with Foreground-Background Disentanglement

Supplementary Material

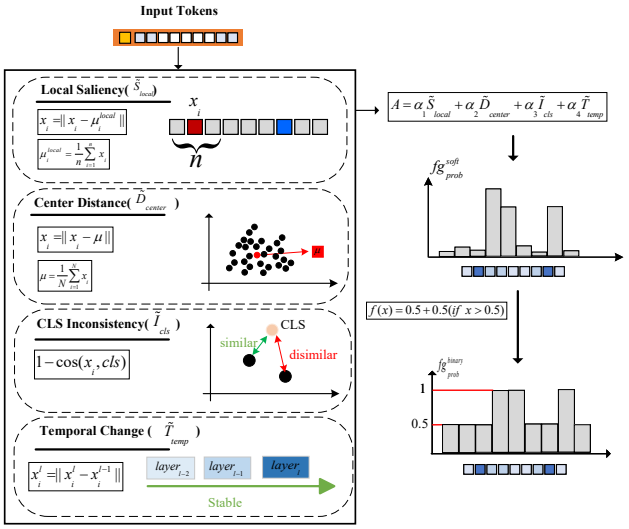

Figure 1. Generation process of the foreground–background mask. Multiple complementary anomaly indicators are first aggregated to estimate the foreground–background probability of each token.

A. Details of Multi-View Foreground-Background Enhancement

This section provides detailed formulations of the indicators used in the Multi-View Foreground-Background Enhancement (MVFBE) module described in Sec. X.

Given visual tokens $\mathbf{X} \in \mathbb{R}^{B\times(L+1)\times C}$ extracted from a Vision Transformer, including a class token $\mathbf{x}_{\text{cls}}$ and patch tokens $\mathbf{x}_{\text{img}}$, MVFBE estimates foreground likelihood using four complementary indicators.

Local saliency. Foreground regions typically exhibit stronger feature deviation from their local context compared with homogeneous background regions. Inspired by prior studies on token-level saliency estimation in Vision Transformers [2, 4, 8], we quantify the local distinctiveness of each patch token by measuring its deviation from a neighborhood-aggregated representation. Specifically, we construct a local contextual representation using 1D average pooling with kernel size 3 along the token sequence. The local saliency score is defined as the ℓ_2 distance between each token and its neighborhood average:

$$\mathcal{S}\text{local} = |\mathbf{x}\text{img} - \text{AvgPool}(\mathbf{x}_{\text{img}})|_2 \,. \tag{1}$$

This formulation highlights tokens whose representations significantly differ from their surrounding context, which often correspond to potential foreground structures.

Center distance. Foreground tokens tend to deviate from the dominant feature distribution formed by background regions. Following the common practice of distribution-based anomaly or foreground modeling [6], we characterize this deviation by computing the distance between each token and a global distribution center.

Formally, the center distance indicator is defined as

$$\mathcal{D}\text{center} = |\mathbf{x}\text{img} - \mathbf{c}|_2 \,, \tag{2}$$

where the center $\mathbf{c}$ can either be a predefined normal prototype $\mathbf{c}_{\text{normal}}$ or the empirical batch mean:

$$\mathbf{c} = \frac{1}{L}\sum_{i=1}^{L} \mathbf{x}_{\text{img}}^{(i)}. \tag{3}$$

Tokens that lie far from the global center are more likely to correspond to foreground or abnormal structures.

CLS Inconsistency. In Vision Transformers, the class token (CLS token) aggregates global semantic information across the entire image through self-attention [1, 7]. Therefore, patch tokens that are semantically inconsistent with the CLS token are likely to correspond to localized foreground or anomalous regions.

To quantify this discrepancy, we compute the cosine dissimilarity between each patch token and the CLS token:

$$\mathcal{I}_{\text{cls}}^{(i)} = 1 - \text{cosine\_similarity}\big(\mathbf{x}_{\text{img}}^{(i)}, \mathbf{x}_{\text{cls}}\big), \tag{4}$$

where $\mathbf{x}_{\text{img}}^{(i)}$ denotes the i-th patch token and $\mathbf{x}_{\text{cls}}$ denotes the global class token. A higher inconsistency score indicates a greater semantic deviation from the global image representation, highlighting potential foreground structures.

Temporal variation. Deep Transformer layers progressively refine token representations through hierarchical feature transformations [3, 5]. Tokens corresponding to foreground structures often exhibit larger representation shifts across layers due to progressive semantic enrichment.

To capture this dynamic behavior, we measure the feature variation between adjacent layers:

$$\mathcal{T}_{\text{temp}} = |\mathbf{X}^{(l)} - \mathbf{X}^{(l-1)}|_2. \tag{5}$$

This temporal variation reflects the degree of representation evolution during hierarchical feature refinement.

Indicator normalization. To ensure fair contribution from each cue, we apply per-sample min–max normalization:

$$\tilde{\mathcal{S}}_i = \frac{\mathcal{S}_i - \min_j(\mathcal{S}_i[j])}{\max_j(\mathcal{S}_i[j]) - \min_j(\mathcal{S}_i[j]) + \epsilon}. \quad (6)$$

The normalized indicators are then linearly combined to form an anomaly score:

$$\mathcal{A} = \alpha_1 \tilde{\mathcal{S}}_{\text{local}} + \alpha_2 \tilde{\mathcal{D}}_{\text{center}} + \alpha_3 \tilde{\mathcal{I}}_{\text{cls}} + \alpha_4 \tilde{\mathcal{T}}_{\text{temp}}, \quad (7)$$

B. FG/BG Token Enhancement Methods

This section presents two token enhancement strategies for foreground/background (FG/BG) refinement: **Semantic Enhancement (SEM)** and **Spatial Enhancement (SPA)**. Both methods aim to refine token features based on foreground probabilities, but they differ in focus: SEM emphasizes global semantic relations, while SPA emphasizes local spatial consistency.

B.1. Semantic Enhancement (SEM)

Core Idea: Use the CLS token to capture global semantic information and reweight tokens based on their foreground probability and semantic richness. Foreground tokens receive higher weights if they are semantically informative (less similar to CLS), while background tokens are weighted for stability.

Algorithm 1: SEM Enhancement Pseudocode

Input: tokens $X \in \mathbb{R}^{b\times(l+1)\times c}$, foreground probabilities $P_{fg} \in \mathbb{R}^{b\times l}$
Output: Refined tokens X_{out}

Extract CLS token: $CLS = X[:, 0, :]$;
Extract other tokens: $X_{tokens} = X[:, 1:, :]$;
Compute foreground weight matrix: $W_{fg} = P_{fg} \cdot P_{fg}^T$;
Compute background weight matrix: $W_{bg} = (1 - P_{fg}) \cdot (1 - P_{fg})^T$;
Compute similarity to CLS: $S = \text{cosine\_similarity}(X_{tokens}, CLS)$;
Compute information richness: $R = 1 - S$;
Compute foreground aggregation weights: $A_{fg} = \text{softmax}(R \cdot W_{fg})$;
Compute background aggregation weights: $A_{bg} = \text{softmax}(S \cdot W_{bg})$;
Aggregate tokens: $X_{fg\_agg} = A_{fg} \cdot X_{tokens}$, $X_{bg\_agg} = A_{bg} \cdot X_{tokens}$;
Combine foreground and background: $X_{agg} = \alpha(X_{fg\_agg} + X_{bg\_agg}) + (1 - \alpha)X_{tokens}$;
Concatenate CLS token: $X_{out} = [CLS, X_{agg}]$;

Analysis:

- Captures global semantic relations between tokens.
- Highlights tokens that are semantically informative (foreground).
- Background tokens are stabilized using similarity to CLS.
- Computational complexity scales with $O(l^2)$ due to pairwise token weighting.

B.2. Spatial Enhancement (SPA)

Core Idea: Incorporate spatial structure by reshaping tokens into 2D patches and aggregating local neighborhoods. Foreground and background are weighted based on local stability and information richness.

Algorithm 2: SPA Enhancement Pseudocode

Input: tokens $X \in \mathbb{R}^{b\times(l+1)\times c}$, foreground probabilities $P_{fg} \in \mathbb{R}^{b\times l}$, patch size r
Output: Refined tokens X_{out}

Extract CLS token: $CLS = X[:, 0, :]$;
Extract other tokens: $X_{tokens} = X[:, 1:, :]$;
Reshape tokens to 2D: $X_{2D} = \text{reshape}(X_{tokens}, [b, c, h, w])$, $h = w = \sqrt{l}$;
Unfold patches: $X_{patch} = \text{unfold}(X_{2D}, \text{kernel} = r)$;
Unfold foreground mask: $P_{fg\_patch} = \text{unfold}(P_{fg})$;
Compute background mask: $P_{bg\_patch} = 1 - P_{fg\_patch}$;
Compute stability score for background: $S_{bg} = \text{stability}(X_{patch} \cdot P_{bg\_patch}, CLS)$;
Compute information richness for foreground: $R_{fg} = \text{richness}(X_{patch} \cdot P_{fg\_patch}, CLS)$;
Compute weighted aggregation: $X_{bg\_agg} = \sum(X_{patch} \cdot \text{softmax}(S_{bg}))$, $X_{fg\_agg} = \sum(X_{patch} \cdot \text{softmax}(R_{fg}))$;
Combine foreground and background: $X_{agg} = X_{fg\_agg} + X_{bg\_agg}$;
Reshape to original token layout and concatenate CLS token: $X_{out} = [CLS, X_{agg}]$;

Analysis:

- Preserves spatial consistency by aggregating local patches.
- Foreground and background are weighted based on local patch statistics.
- Sensitive to token layout (requires l to be square for 2D reshaping).
- Computational complexity scales with $O(b \cdot h \cdot w \cdot r^2)$, depending on patch size.

Summary:

- SEM emphasizes *global semantic relations*, effective for highlighting informative tokens across the image.
- SPA emphasizes *local spatial consistency*, effective for preserving structure and local context.
- Both methods can be complementary in a FG/BG token refinement framework.

C. Visual comparison with other state-of-the-art methods

Figure 2 presents a qualitative comparison of anomaly localization results on various product categories, including *fryum*, *cashew*, *chewinggum*, *bottle*, *capsule*, *hazelnut*, *metal_nut*, *pill*, *screw*, and *tile*. The first row displays the original images, and the second row provides the ground-truth (GT) anomaly regions. Subsequent rows correspond to the results obtained by Anomaly CLIP, AF-CLIP, FAPrompt, and our proposed **FB-CLIP**.

As shown in the figure, **FB-CLIP** achieves more precise and concentrated anomaly localization results that align closely with the ground-truth annotations. Compared with other methods, it produces fewer false activations in background regions and captures subtle defects more effectively. For instance, in the *hazelnut* and *metal_nut* samples, FB-CLIP successfully highlights the entire defective region, while the other methods show incomplete or scattered activations. On fine-grained textures such as *tile*, FB-CLIP also demonstrates superior robustness and sensitivity to small-scale anomalies, indicating its stronger generalization capability across diverse categories.

D. Ablation Study of Multi-Strategy Text Feature Fusion (MSTFF)

Table 1 presents an ablation study of our proposed **Multi-Strategy Text Feature Fusion (MSTFF)** on two datasets (**MVTec** and **VisA**), evaluated at both image-level (AUROC and AP) and pixel-level (AUROC and PRO). The study compares single-strategy configurations (EOT, Global, Attention) against multi-strategy combinations (EOT+Attn, EOT+GP+Attn), allowing for a detailed assessment of the contribution of each textual feature and their combinations to anomaly detection and localization performance.

Considering single-strategy results, using only **EOT** or **Global** features yields limited performance. On MVTec, EOT achieves an image-level AUROC of 67.1 and AP of 84.0, but its pixel-level AUROC and PRO are only 65.2 and 32.9, respectively. Global features show similar trends, with an image-level AUROC of 62.6 and modest improvements in pixel-level metrics. These results indicate that relying solely on text-end features or global pooling is insufficient to capture fine-grained anomalies. In contrast, the **Attention** feature significantly improves performance across both datasets. On MVTec, it achieves an image-level AUROC of 80.9 and pixel-level AUROC of 79.0; on VisA, the image-level AUROC reaches 78.5 and pixel-level AUROC 92.7. This demonstrates that attention-weighted features effectively focus on the most relevant regions, enhancing lo-

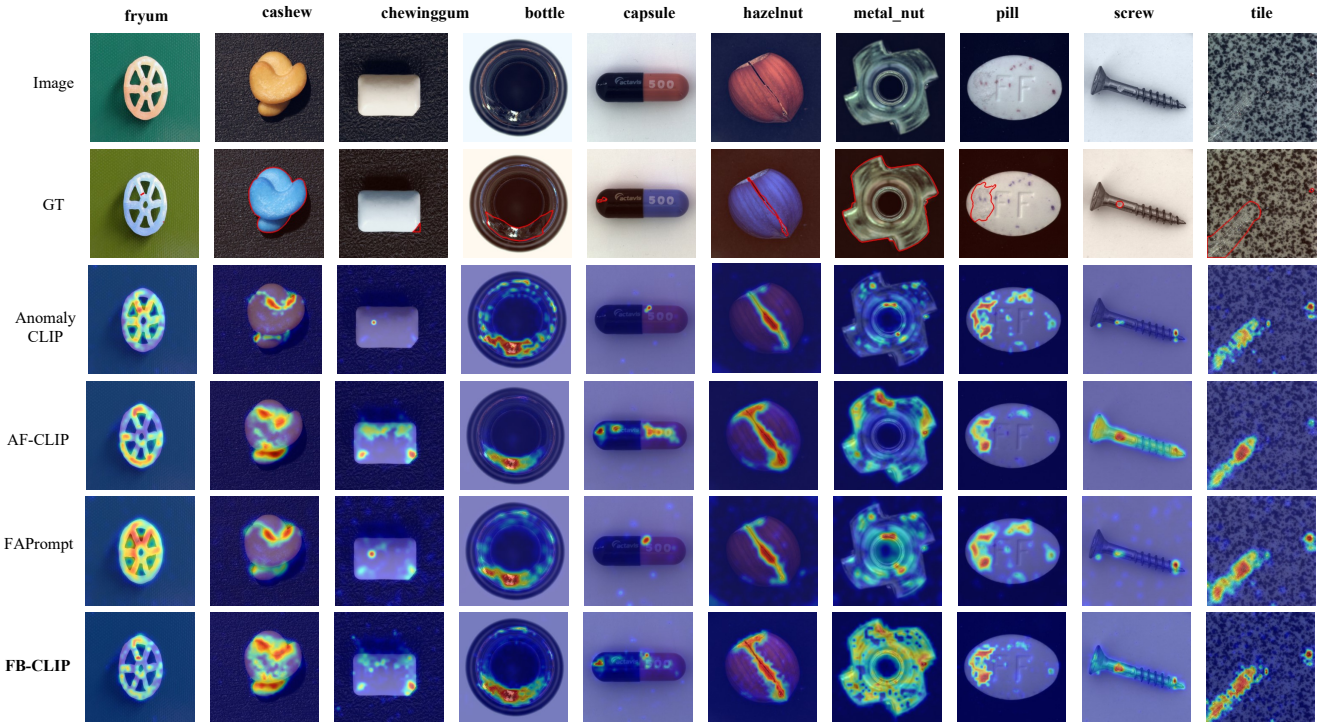

Figure 2. Visualization results of anomaly localization across multiple categories using different CLIP-based methods.

Table 1. Ablation study on **Multi-Strategy Text Feature Fusion (MSTFF)** across two datasets (**MVTec** and **VisA**). EOT: End-of-Text; GP: Global Pooling; Attn: Attention-weighted.

Strategies			MVTec		VisA	
EOT	GP	Attn	Image-level (AUROC, AP)	Pixel-level (AUROC, PRO)	Image-level (AUROC, AP)	Pixel-level (AUROC, PRO)
✓			(67.1, 84.0)	(65.2, 32.9)	(45.7, 56.2)	(88.4, 71.8)
	✓		(62.6, 81.9)	(63.3, 31.9)	(51.3, 56.6)	(83.7, 67.5)
		✓	(80.9, 90.8)	(79.0, 55.7)	(78.5, **80.7**)	(92.7, 80.3)
✓		✓	(81.1, 90.9)	(83.1, 66.4)	**(78.6, 80.7)**	(92.7, 80.4)
✓	✓	✓	**(82.0, 91.4)**	**(84.0, 68.0)**	(78.3, 80.5)	**(93.2, 81.4)**

calization precision.

Multi-strategy combinations further boost performance. Integrating **EOT with Attention (EOT+Attn)** increases the MVTec pixel-level PRO from 55.7 to 66.4 compared to Attention alone, indicating that EOT provides complementary information that enhances sensitivity to fine-grained anomalies. Incorporating global pooling as well (**EOT+GP+Attn**) yields the best overall performance: image-level AUROC reaches 82.0 and pixel-level AUROC 84.0 on MVTec, while VisA pixel-level PRO increases to 81.4. These findings suggest that fusing multiple textual feature strategies allows MSTFF to leverage complementary information effectively, resulting in more precise and robust anomaly detection and localization.

In summary, the ablation results clearly validate the effectiveness of MSTFF. While single-strategy configurations provide limited anomaly cues, attention-weighted features and multi-strategy fusion substantially improve both image-level and pixel-level metrics. The notable gains in pixel-level PRO highlight MSTFF's ability to capture fine-grained defect regions and its superior generalization across diverse datasets.

E. Ablation Study on MVFBE Components

To validate the effectiveness of each component within the proposed **Multi-View Foreground-Background Enhancement (MVFBE)** module (Figure **??**), we conduct a modular ablation study, the results of which are reported in Table 2. The MVFBE module consists of four complementary components: **Identity (ID)**, **Semantic (SEM)**, **Spatial (SPA)**, and **Foreground-Background Attention (FB-Attn)**. Each component can be independently enabled to examine its contribution to both image-level (AUROC, AP) and pixel-level (AUROC, PRO) anomaly detection performance across two datasets, **MVTec** and **VisA**.

From the single-component configurations, it is evident that **FB-Attn** alone significantly boosts image-level performance, achieving AUROC/AP of (92.5, 96.7) on MVTec and (86.7, 88.6) on VisA. However, its pixel-level PRO on MVTec is relatively low (47.9), suggesting that while FB-Attn effectively enhances global feature discrimination, precise pixel-level localization benefits from complementary components. In contrast, the traditional enhancement views (ID, SEM, SPA) individually provide moderate improvements. For instance, ID alone yields (82.0, 91.4) image-level AUROC/AP and (84.0, 68.0) pixel-level AUROC/PRO on MVTec, indicating that preserving original features maintains baseline information but lacks strong localization capability. SEM emphasizes semantic information, improving image-level AUROC (83.0) but achieving lower pixel-level PRO (48.6), highlighting its role in capturing foreground richness while background stability remains less effectively separated. SPA focuses on fine-grained spatial structures, achieving balanced improvements in both image- and pixel-level metrics, confirming its utility in capturing local contextual anomalies.

When combining traditional enhancement views, synergistic effects emerge. The combination of ID+SEM+SPA without FB-Attn increases pixel-level PRO to 64.8 on MVTec and 81.1 on VisA, while maintaining strong image-level AUROC/AP. This demonstrates that complementary perspectives contribute to more accurate anomaly localization by aggregating multiple views of foreground and background features.

Integrating **FB-Attn** with one or more enhancement views leads to substantial performance gains across all metrics. Notably, the full combination of ID+SEM+SPA+FB-Attn achieves the highest overall performance, with MVTec image-level AUROC/AP of (92.2, 96.5) and pixel-level AUROC/PRO of (91.7, 84.5), and VisA image-level (88.4,

Table 2. Ablation study with modular configuration using four components: **ID**, **SEM**,**SPA** and **FB-Attention**,. A checkmark (✓) indicates that the component is enabled.

Method				MVTec		VisA	
ID	**SEM**	**SPA**	**FB-Attn**	**Img** (AUROC, AP)	**Pix** (AUROC, PRO)	**Img** (AUROC, AP)	**Pix** (AUROC, PRO)
✓				(82.0, 91.4)	(84.0, 68.0)	(78.3, 80.5)	(93.2, 81.4)
	✓			(83.0, 92.2)	(77.3, 48.6)	(80.8, 82.9)	(92.1, 77.1)
		✓		(83.0, 91.9)	(80.8, 55.6)	(79.3, 81.7)	(91.0, 73.9)
✓	✓			(81.4, 91.1)	(84.8, 67.1)	(79.1, 81.5)	(93.4, 81.7)
✓		✓		(82.0, 91.3)	(83.1, 64.3)	(77.8, 80.5)	(93.3, 80.9)
	✓	✓		(81.6, 91.2)	(81.6, 61.2)	(78.8, 81.3)	(92.9, 79.3)
✓	✓	✓		(81.7, 91.4)	(83.7, 64.8)	(78.8, 81.0)	(93.3, 81.1)
✓			✓	(92.5, 96.7)	(90.8, 47.9)	(86.7, 88.6)	(95.8, 87 .8)
	✓		✓	**(92.7, 96.7)**	(86.2, 55.1)	(85.6, 87.5)	(95.7, 87.4)
		✓	✓	(92.2, 96.3)	(90.2, 81.8)	(88.2, 89.4)	(94.4, 83.8)
✓	✓		✓	(91.7, 96.1)	(91.4, 85.1)	(87.2, 88.8)	(95.9, 90.1)
✓		✓	✓	(92.3, 96.5)	(90.9, 83.8)	(88.2, 89.6)	(**96.0**, 90.8)
	✓	✓	✓	(92.0, 96.3)	(**91.7**, 82.8)	(88.1, 89.4)	(95.6, 87.8)
✓	✓	✓	✓	(92.2, 96.5)	**(91.7, 84.5)**	**(88.4, 89.7)**	**(96.0, 90.8)**

Table 3. Ablation study on the number of tokens for background representation. Performance is reported on MVTec and VisA datasets, including image-level (I) AUROC and AP, as well as pixel-level (P) AUROC and PRO metrics.

	MVTec		VisA	
	I-(AUROC, AP)	P-(AUROC, PRO)	I-(AUROC, AP)	P-(AUROC, PRO)
L	(92.4, 96.6)	(91.8, 85.4)	(89.3, 90.5)	(96.3, 91.4)
L/2	(92.4, 96.6)	(91.9, 85.7)	(89.5, 90.7)	(96.3, 91.4)

89.7) and pixel-level (96.0, 90.8). This confirms that FB-Attn effectively refines the aggregated features from multiple views, enhancing both global detection and fine-grained localization. The consistent improvements in pixel-level PRO when FB-Attn is included emphasize its critical role in discriminating foreground anomalies from background regions, complementing the ID, SEM, and SPA components.

In summary, the ablation results validate that each component of MVFBE contributes distinct and complementary strengths: ID preserves raw feature integrity, SEM models foreground richness and background stability, SPA captures local spatial structures, and FB-Attn refines these features through attention-based foreground-background separation. The full combination of all four components yields the most balanced and robust performance, demonstrating the effectiveness of the proposed modular design in both image-level recognition and pixel-level anomaly localization.

F. Impact of Token Number on Background Representation

We investigate how the number of tokens affects background representation by comparing the performance using L tokens versus $L/2$ tokens. Table 1 reports the results on MVTec and VisA datasets, including image-level (I) AUROC and AP, as well as pixel-level (P) AUROC and PRO metrics.

As shown, using $L/2$ tokens achieves virtually identical performance to using L tokens. On MVTec, the image-level AUROC and AP remain the same at 92.4% and 96.6%, while the pixel-level AUROC slightly improves from 91.8% to 91.9% and PRO increases from 85.4 to 85.7. On VisA, $L/2$ tokens maintain comparable or slightly better performance for image-level (AUROC/AP: 89.5/90.7 vs. 89.3/90.5) and identical pixel-level metrics.

These results demonstrate that half the number of tokens is sufficient to effectively capture background information. Reducing token count not only maintains or slightly improves anomaly detection and localization performance but also provides potential computational savings, making it an efficient strategy for multi-strategy text feature fusion and multi-view foreground-background enhancement.

G. Subject-level

To provide a more detailed evaluation, we report the subset–level performance in the following tables (Table 4, 5, 6, 7, 8, 9, 10).

Table 4. Performance on the DAGM_KAGGLEUPLOAD dataset.

Objects	Pixel		Image	
	AUROC	AUPRO	AUROC	AP
Class1	91.8	82.1	94.1	82.1
Class2	99.7	99.0	99.9	99.8
Class3	97.0	96.3	100.0	100.0
Class4	94.0	80.5	99.5	97.1
Class5	99.3	97.6	100.0	99.9
Class6	99.6	98.5	99.7	99.2
Class7	94.6	93.4	100.0	100.0
Class8	98.4	96.9	97.6	92.9
Class9	99.6	98.3	98.9	95.9
Class10	99.4	98.2	99.7	98.2
Mean	**97.3**	**94.1**	**99.0**	**96.5**

Table 5. Performance on the MPDD dataset (class-wise results).

Objects	Pixel		Image	
	AUROC	AUPRO	AUROC	AP
bracket_black	97.9	93.0	79.0	88.5
bracket_brown	94.1	86.4	57.8	73.9
bracket_white	99.6	97.5	91.2	90.6
connector	96.9	88.6	83.6	79.1
metal_plate	94.4	87.0	66.9	86.8
tubes	98.5	94.0	96.3	98.4
Mean	**96.9**	**91.1**	**79.1**	**86.2**

Table 6. Performance on the BTAD dataset (class-wise results).

Objects	Pixel		Image	
	AUROC	AUPRO	AUROC	AP
01	94.5	78.9	95.4	98.4
02	95.4	66.3	85.5	97.7
03	97.5	94.9	98.6	93.1
Mean	**95.8**	**80.0**	**93.2**	**96.4**

Table 7. Performance on the DTD dataset (class-wise results).

Objects	Pixel		Image	
	AUROC	AUPRO	AUROC	AP
Blotchy_099	99.2	95.3	99.9	100.0
Fibrous_183	99.4	97.7	100.0	100.0
Marbled_078	99.0	96.0	99.6	99.9
Matted_069	98.2	83.1	97.2	99.3
Mesh_114	97.3	86.7	92.0	96.8
Perforated_037	95.9	89.6	95.7	99.0
Stratified_154	99.5	92.6	99.8	100.0
Woven_001	99.8	98.9	100.0	100.0
Woven_068	98.8	91.9	94.6	96.9
Woven_104	98.2	93.3	99.9	100.0
Woven_125	99.5	93.4	100.0	100.0
Woven_127	95.2	93.1	96.1	97.4
Mean	**98.3**	**92.6**	**97.9**	**99.1**

Table 8. Pixel- and image-level anomaly detection performance on MVTec dataset using FB-CLIP.

Objects	Pixel		Image	
	AUROC	AUPRO	AUROC	AP
bottle	93.4	86.7	96.3	98.8
cable	79.4	69.2	73.9	84.1
capsule	96.2	91.8	93.3	98.6
carpet	99.4	94.6	100.0	100.0
grid	97.6	80.7	99.0	99.7
hazelnut	98.1	88.7	93.2	96.6
leather	99.4	97.9	100.0	100.0
metal_nut	63.3	72.4	75.1	94.7
pill	91.3	94.2	87.2	97.4
screw	98.6	92.8	86.9	95.1
tile	97.9	91.6	98.8	99.5
toothbrush	92.0	89.3	95.3	98.2
transistor	75.6	56.2	89.0	87.0
wood	98.4	95.8	98.7	99.6
zipper	97.3	83.5	98.6	99.6
Mean	**91.9**	**85.7**	**92.4**	**96.6**

Table 9. Cross-domain zero-shot performance from MVTec to VISA dataset.

Objects	Pixel		Image	
	AUROC	AUPRO	AUROC	AP
candle	98.8	96.3	92.1	93.5
capsules	97.3	91.0	94.2	97.0
cashew	96.4	97.1	94.3	97.5
chewinggum	99.4	94.0	98.6	99.4
fryum	95.9	93.6	97.0	98.7
macaroni1	99.5	96.7	89.5	90.5
macaroni2	98.8	90.9	76.0	76.5
pcb1	93.5	86.5	83.4	83.4
pcb2	92.5	80.9	81.3	80.5
pcb3	89.4	82.1	71.0	74.7
pcb4	95.4	91.0	96.8	96.7
pipe_fryum	98.1	96.2	99.4	99.7
Mean	**96.3**	**91.4**	**89.5**	**90.7**

Table 10. Cross-domain zero-shot performance on Real-IAD dataset.

Objects	Pixel		Image	
	AUROC	AUPRO	AUROC	AP
pcb	95.8	82.0	69.2	76.9
phone_battery	77.7	94.9	85.1	84.1
sim_card_set	99.8	98.5	95.6	96.3
switch	88.4	79.5	69.4	75.0
terminalblock	97.9	93.1	86.2	88.8
toothbrush	93.9	82.9	71.4	76.9
bottle_cap	98.3	92.8	78.5	78.8
end_cap	93.3	76.4	71.7	77.0
fire_hood	99.2	95.5	83.3	75.2
mounts	97.4	93.0	83.6	68.8
plastic_nut	96.7	84.0	82.7	73.4
plastic_plug	98.3	94.0	82.5	77.8
regulator	94.1	73.0	63.5	42.8
rolled_strip_base	99.2	97.4	93.8	96.9
tape	98.7	94.4	94.6	93.8
porcelain_doll	99.4	97.1	93.5	90.2
mint	94.3	84.5	77.8	78.4
eraser	99.6	96.5	89.4	89.2
button_battery	97.9	89.3	79.0	84.4
toy	83.5	76.2	76.8	84.0
transistor1	94.6	78.1	74.6	80.6
usb	95.4	83.8	70.6	69.9
usb_adaptor	98.8	93.5	81.2	74.9
zipper	96.5	90.8	91.0	94.8
toy_brick	98.2	91.1	79.7	74.8
u_block	98.9	93.5	80.4	64.6
vcpill	97.6	84.9	85.3	83.8
wooden_beads	98.3	87.5	80.3	75.1
woodstick	98.2	90.3	85.0	74.6
audiojack	95.8	77.8	63.2	48.8
Mean	**95.9**	**88.2**	**80.6**	**78.4**

H. Limitations

In this section, we analyze FB-CLIP from two perspectives: the inference stage and failure detection cases.

Table 11. Comparison with recent SOTA methods in the inference process.

Infer Metric	AF-CLIP	FAPrompt	FB-CLIP
Time	112 ms	232 ms	215 ms
Memory	2.5G	2.6G	3.3G

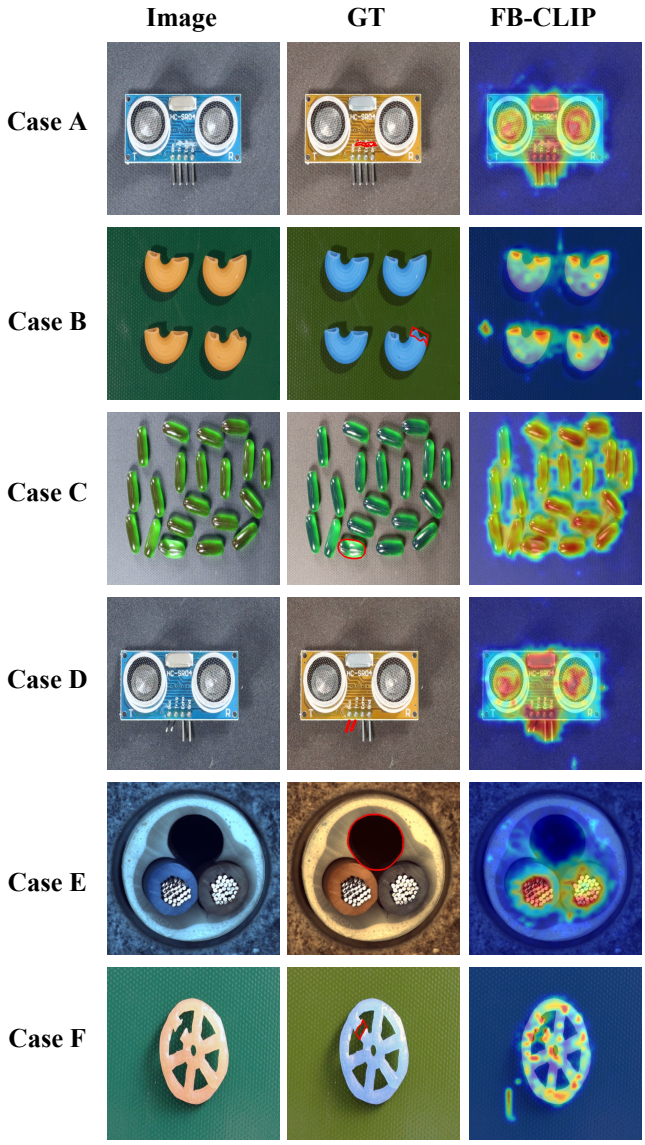

Figure 3. Visualization results of inaccurate anomaly localization using the FB-CLIP method.

In Figure 3, we present several representative bad cases of our FB-CLIP model to illustrate its limitations. Note that these examples are selected failure samples, and similar issues may also exist in other anomaly detection methods.

For the sensor module in Cases A and D, FB-CLIP over-responds to non-functional variations such as material and color changes on the PCB, while failing to capture subtle structural defects on the pins. In Case B (the crescent-shaped object), the model produces diffused activation maps and cannot localize fine-grained edge damages. In Case C (dense capsules), the method fails to distinguish the abnormal individual from the cluster, as anomalies are overwhelmed by surrounding similar instances. For Case E (the pipeline hole), FB-CLIP incorrectly treats texture and color variations inside the hole as anomalies, with poorly localized boundaries. In Case F (the wheel component), the activation spreads over the entire object rather than concentrating on the defective spoke region, leading to inaccurate defect localization.

These selected cases demonstrate that FB-CLIP still faces challenges in distinguishing functional defects from appearance variations, detecting tiny local anomalies, and maintaining precise localization under complex backgrounds. However, such limitations are commonly observed in many feature-based anomaly detection approaches rather than being unique to our method.

Beyond the above qualitative limitations, FB-CLIP also has shortcomings in computational efficiency, as shown in Table 11. Although achieving competitive detection performance, FB-CLIP consumes more memory (3.3G) and longer inference time (215 ms) than AF-CLIP, which restricts its deployment in resource-constrained real-time scenarios. These observations indicate that the performance improvement of FB-CLIP is accompanied by higher computational costs, and there is still room for optimization in efficiency.

I. Anomaly Detection in Challenging Scenarios with Physical Occlusion

While existing methods demonstrate high performance on standard benchmarks, anomaly detection in more challenging real-world scenarios—such as those involving physical occlusion or partial blockage—remains significantly difficult. As shown in Figure 4, physical occlusion can obscure critical regions, making it hard for models to capture complete feature representations, thereby reducing detection and localization accuracy.

In such complex scenarios, models are required not only to distinguish subtle differences between normal and abnormal objects but also to infer information about occluded regions. This demands approaches that can effectively integrate local and global features to enhance perception of partially visible targets. Techniques such as multi-view information fusion or context-based feature completion may offer effective solutions for these challenges.

Extending anomaly detection methods to explicitly handle physically occluded scenarios not only tests the robustness of existing models but also motivates the design of more resilient multi-strategy fusion and multi-view perception frameworks.

Inspired by the Real-IAD Dataset and the suggestions of the reviewers of this paper, we argue that another important application scenario for anomaly detection is leveraging multi-view observations to identify and localize anomalies. In practical environments, objects may be partially occluded, or critical regions may not be visible from a single viewpoint. As a result, relying solely on single-view information can limit the model's ability to accurately detect and localize abnormal patterns. By integrating complementary information from multiple viewpoints, models can obtain richer and more complete feature representations, which helps alleviate the impact of occlusion and incomplete observations. Therefore, designing anomaly detection approaches that effectively utilize multi-view information is a promising direction for improving robustness and reliability in complex real-world scenarios.

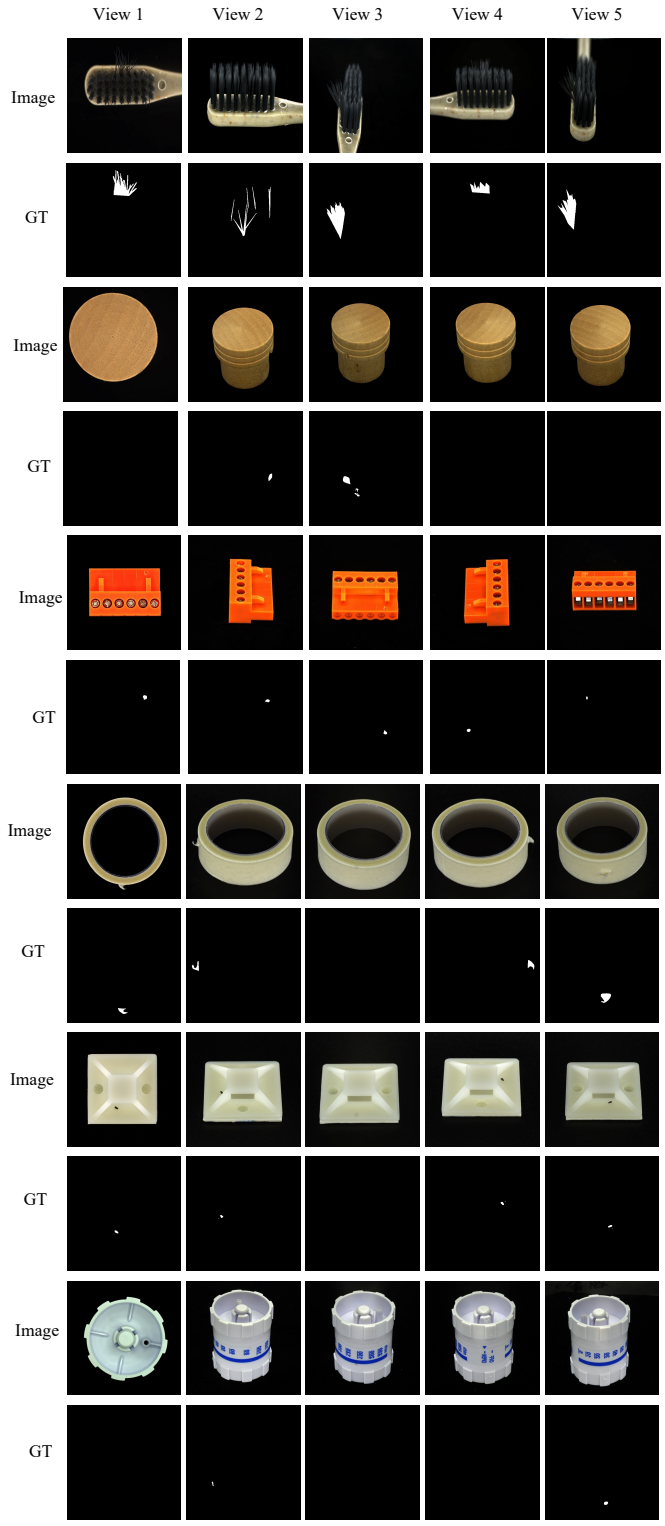

Figure 4. Visualization of the same object from different viewpoints in the Real-IAD dataset.

References

[1] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021. 1

[2] Peiyan Dong, Mengshu Sun, Alec Lu, Yanyue Xie, Kenneth Liu, Zhenglun Kong, Xin Meng, Zhengang Li, Xue Lin, Zhenman Fang, et al. Heatvit: Hardware-efficient adaptive token pruning for vision transformers. In *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 442–455. IEEE, 2023. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[4] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4722–4732, 2021. 1

[5] Yansong Peng, Hebei Li, Peixi Wu, Yueyi Zhang, Xiaoyan Sun, and Feng Wu. D-fine: Redefine regression task in detrs as fine-grained distribution refinement. *arXiv preprint arXiv:2410.13842*, 2024. 1

[6] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018. 1

[7] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020. 1

[8] Weiyan Xie, Xiao-Hui Li, Caleb Chen Cao, and Nevin L Zhang. Vit-cx: Causal explanation of vision transformers. *arXiv preprint arXiv:2211.03064*, 2022. 1