TED: Text-Axis Evidence Decomposition for Prompted Anomaly Localization

JinYoung Kim[1], **Geonho Kim**[1], **GiJeong Park**[1], **Geonu Lee**[2], **YoungJoon Yoo**[1*]
[1]Department of Artificial Intelligence, Chung-Ang University, Seoul, Korea
[2]SNUAILAB Co., Ltd., Seoul, Korea
{barraki7226,skdmlqnsrlt,rjsgh2250,yjyoo3312}@cau.ac.kr,
{geonu.lee,yjyoo3312}@snuailab.ai

Abstract

CLIP is a powerful vision-language model, but it was not designed for fine-grained defect localization; CLIP-based anomaly detectors therefore adapt it with prompts or lightweight modules to increase defect sensitivity. We show that stronger sensitivity does not necessarily make local evidence reliable: under domain shift, adapted CLIP-AD models often assign high anomaly scores to both true defects and visually complex normal regions. The issue is not simply missing defect information, but a local scoring rule that decodes defect and hard-normal evidence, having the same anomaly evidence. We propose TED (Text-Axis Evidence Decomposition), a post-hoc scoring method that asks whether each ambiguous response is better supported by source defect patches or by source normal patches mistaken as anomalous. TED compares these supports under the host's normal-versus-anomaly text response, leaves the backbone and prompts unchanged, and requires no target-domain training. It works as a train-free score for raw VLM backbones or as a source-calibrated residual correction for adapted CLIP-AD hosts. Across frozen VLM backbones, TED substantially improves pixel-level localization over raw prompt similarity; across adapted hosts, it improves most pixel-level settings over P-AUROC, P-PRO, and P-AP. Gains are largest under stronger hard-FP competition, with mean localization gain increasing from $+5.0$ in low-competition regimes to about $+10.9$ in mid/high-competition regimes. These results suggest that recoverable defect evidence can already exist in pretrained multimodal representations, but reliable localization requires decoding it against hard-normal competitors. Code will be released at `TED GitHub repository`.

1 Introduction

CLIP and related pretrained vision-language models (VLMs) provide a strong semantic interface for anomaly detection (AD), where normal and anomalous states can be described by language and compared with local visual features. However, these models were not designed to localize fine-grained defect regions. A patch can be semantically salient, structurally complex, or visually distinctive without being defective, so raw prompt similarity can assign high anomaly responses to hard normal regions. Recent CLIP-based AD methods address this mismatch with learnable prompts, prompt ensembles, or lightweight visual modules, making the representation more defect-sensitive. The key question is whether this stronger sensitivity yields cleaner defect evidence, or instead sharpens a local scoring rule in which true defects and visually complex normal regions remain difficult to rank apart.

Our analysis supports the latter interpretation. Across prompt-tuned and adapter-based CLIP-AD hosts, true defects and hard false positives often compete within the same local score regime under

*Corresponding author.

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

transfer. Normal textures, edges, reflections, and salient object parts can receive elevated anomaly responses despite being non-defective. Thus, the failure is better viewed not as semantic loss or missing defect information, but as hard-FP entanglement: defect-supported and hard-normal-supported evidence are decoded together by raw prompt similarity or host-specific local scoring.

This entanglement makes simple false-positive suppression insufficient. Hard false positives do not form a separable nuisance component that can be removed without affecting true defects. Rather, the same local response that supports defect localization can also be supported by visually complex normal regions. The relevant problem is therefore not only how to suppress false positives after they appear, but how to decide which source of evidence better explains an ambiguous local response.

To characterize this behavior, we analyze prompted VLM-based anomaly detectors through their normal-versus-anomaly text response. Empirically, the normal-versus-anomaly text response is still useful: true defect patches usually receive stronger anomaly responses than generic normal background. However, this response is not selective enough. Visually complex normal regions can receive similarly high responses, so true defects and hard false positives remain difficult to rank apart. This clarifies the role of adaptation. Prompt tuning and lightweight modules can make the model more sensitive to defects, but they do not necessarily separate true defects from visually complex hard-normal regions in the local anomaly map. Therefore, hard-normal structures can still be ranked close to true defects when anomaly maps are decoded by raw prompt similarity.

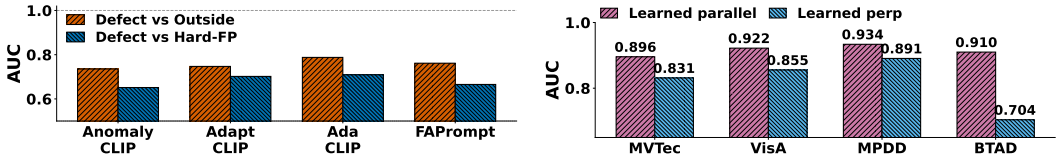
Motivated by this observation, we propose TED (Text-Axis Evidence Decomposition), a post-hoc scoring method for prompted anomaly localization. The key idea is to not trust a high anomaly score by itself. For each ambiguous patch, TED compares its response with two source banks: true defect patches and normal patches that the host previously mistook as anomalous. The comparison is made under the same normal-versus-anomaly text response used by the host, so the query and source examples are judged in the same coordinate. If the query is better supported by source defects, TED raises its score; if it is better supported by hard-normal examples, TED suppresses it. Thus, TED separates defects from hard normal regions by asking which source evidence explains the high response. It leaves the backbone and prompts unchanged, requires no target-domain training, and is used either as a train-free score for raw VLM backbones or as a bounded source-calibrated residual for adapted CLIP-AD hosts.

Because TED only re-scores local responses, it is not tightly tied to the host architecture. Many CLIP-AD methods rely on specific CLIP components, layers, or scoring rules, so changing layers or backbones can require retraining or redesign. By contrast, TED only needs patch-level visual features and a normal-versus-anomaly text response. The same source-evidence comparison can therefore be evaluated on frozen VLM backbones while keeping the host and target protocol fixed. Experiments across adapted CLIP-AD hosts, frozen VLM backbones, and cross-domain transfer benchmarks show that TED improves pixel-level localization most clearly when hard-normal evidence is ranked close to true defects. TED does not simply raise all anomaly scores; it helps most when the baseline confuses hard normal regions with true defects, and less when the host already separates them well. Overall, TED shows that prompted anomaly localization is not only about making models more sensitive to defects, but also about checking whether a high-scoring patch is a true defect or a hard normal region. Our contributions are:

- We identify a common transfer failure in CLIP-based anomaly localization: adapted models can give high anomaly scores to both true defects and visually complex normal regions, making them hard to rank apart.
- We show that the normal-versus-anomaly text response is still useful, but not precise enough: it often separates defects from ordinary background, while still confusing defects with hard normal regions.
- We propose TED, a post-hoc scoring method that compares ambiguous patches with source defect and hard-FP examples, with train-free and source-calibrated variants.

2 Related Work

Backbones, dense representations, and anomaly readouts. Industrial anomaly detection has long relied on pretrained visual backbones, with surveys covering feature-, reconstruction-, density-, and distillation-based pipelines [22, 6, 20]. Feature matching methods such as SPADE [5], PaDiM [7],



(a) Hard-FP competition.

(b) Text-axis response.

Figure 1: **Hard-FP competition and text-axis response.** (a) Across adapted CLIP-AD hosts, true defects remain much harder to separate from hard false positives than from generic outside patches. (b) Across target datasets, the learned text-axis response is consistently more informative than the orthogonal component, suggesting that the axis should be decomposed rather than discarded.

PatchCore [29], and pretrained-feature distribution modeling [27] compare test patches with normal feature statistics or nearest-neighbor structures. Other methods score anomalies through patch-distance criteria [24], normalizing flows such as FastFlow [33], CFlow-AD [12], CS-Flow [30], and SANFlow [18], reconstruction discrepancies [34, 15, 28], or teacher-student feature differences [2, 8, 31]. For transformer backbones, layer selection, feature aggregation, and attention-sink mitigation are often needed for stable dense predictions [13, 36, 38]. These works indicate that localization quality depends not only on the representation, but also on how local evidence is scored and ranked. TED follows this readout-centered view in prompted VLM pipelines, where visually complex normal regions can receive anomaly-like local responses.

Prompted vision-language anomaly detection. Vision-language models, especially CLIP [26], provide a semantic image-text space for prompt-based anomaly detection and have motivated broader studies of VLM transfer [35, 39]. WinCLIP [16] uses normal and anomalous prompt ensembles with window-level harmonic scoring for zero-shot pixel-level localization, while later work explores generalized prompts and anomaly-aware CLIP variants for data-efficient inspection [19, 23]. Recent methods improve defect sensitivity through prompt learning or lightweight adaptation: AnomalyCLIP [40] learns object-agnostic prompts, AdaCLIP [3] combines static and image-conditioned dynamic prompts, FAPrompt [41] introduces fine-grained abnormality prompts, and AdaptCLIP [9] combines textual adaptation with visual and prompt-query adapters. These methods improve benchmark performance, but stronger defect sensitivity does not necessarily yield cleaner local evidence: hard normal regions can still be ranked close to true defects. Rather than adding another prompt learner or adapter, TED keeps the host response and changes how ambiguous local evidence is decoded using source defect and hard-FP support.

Backbone flexibility and multimodal anomaly understanding. Many CLIP-AD adaptation pipelines depend on CLIP-family tokenization, text encoders, prompt learners, feature hooks, layer choices, or host-specific scoring recipes. Although they are not tied to a single image encoder, feature-layer changes or transfer to heterogeneous VLM backbones can still require re-engineering, retraining, or architecture search. TED only needs patch-level visual features and a normal-versus-anomaly text response. This lets the same source-evidence comparison work as a train-free score for raw VLM backbones such as ImageBind [10], or as a source-calibrated residual for adapted CLIP-AD hosts. This differs from multimodal large-language-model approaches such as AnomalyGPT [11], zero-shot anomaly reasoning with MLLMs [32], MMAD [17], OmniAD [37], AnomalyR1 [4], and AD-FM [21], which target reasoning, explanation, instruction following, or end-to-end multimodal decisions with heavier architectures. TED instead serves as a lightweight local evidence-decoding layer for the prompted VLM anomaly-localization pipelines evaluated here.

3 Proposed Method

3.1 From Hard-FP Competition to Evidence Decomposition

We start from a simple failure case. After adaptation, a CLIP-based anomaly detector may give high scores to true defects, but it can also give high scores to normal regions that only look suspicious. We call these confusing normal regions *hard false positives*. They often come from strong edges, repeated textures, reflections, or salient object parts. To measure this failure, we compare three kinds

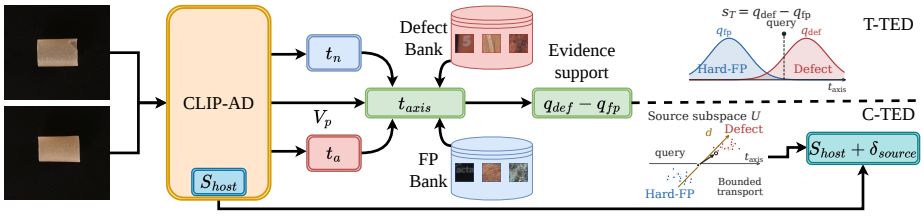


Figure 2: **Overview of TED.** The host provides patch features(V_p) and normal(t_n)/anomaly(t_a) text embeddings that define the host text axis. TED projects query patches and source defect/hard-FP features onto this axis to compute a defect-versus-hard-normal support margin. T-TED uses this margin as a train-free local score, while C-TED converts it into a bounded residual anchored to the host score.

of patches: true defect patches, ordinary normal patches, and hard false-positive patches. Fig. 1(a) shows that adapted hosts separate true defects from ordinary normal patches more easily than from hard false positives. Fig. 1(b) shows that the host's normal-versus-anomaly text response is still useful, but not enough to solve this confusion. In other words, the model has useful anomaly information, but its local score can still treat true defects and hard normal regions as similarly anomalous.

This leads to the main idea of TED. A high anomaly score should not be trusted by itself. Instead, we ask whether that high response looks more like source defect examples or source hard false-positive examples. TED keeps the host's normal-versus-anomaly response, but re-ranks ambiguous patches using this source comparison.

3.2 TED Overview

TED is a post-hoc scoring method for anomaly localization. It does not change the backbone, prompts, or host detector. For each suspicious patch, TED asks whether it looks more like source defects or source hard false positives. Defect-like patches are boosted, while hard-FP-like patches are suppressed. T-TED uses this as the local score; C-TED uses it as a small correction to the adapted host score. Neither mode uses target images, masks, labels, or target-score selection. Fig. 3 shows this effect: raw scoring mixes defects with hard false positives, while TED re-ranks them using source examples.

3.3 Text-Axis Evidence Banks and Train-Free Score

For a query image x, the host provides normal/anomaly text embeddings $e_{\mathrm{n}}(x), e_{\mathrm{a}}(x)$ and patch features $v_i^{(\ell)}(x)$ at layer ℓ. We use the normal-to-anomaly text direction as a common one-dimensional scale:

$$u(x) = \frac{e_{\mathrm{a}}(x) - e_{\mathrm{n}}(x)}{\|e_{\mathrm{a}}(x) - e_{\mathrm{n}}(x)\|_2}, \qquad c_i^{(\ell)}(x) = \langle v_i^{(\ell)}(x), u(x) \rangle. \tag{1}$$

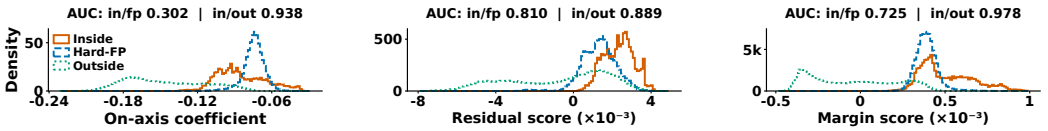
Here, $c_i^{(\ell)}(x)$ is the anomaly response of query patch i on the host text direction. TED does not modify this direction. For each layer, we build two source banks: $\mathcal{B}_{\mathrm{def}}^{(\ell)} = \{b_{\mathrm{def},j}^{(\ell)}\}_{j=1}^{N_{\mathrm{def}}}$ from source defect regions, and $\mathcal{B}_{\mathrm{fp}}^{(\ell)} = \{b_{\mathrm{fp},j}^{(\ell)}\}_{j=1}^{N_{\mathrm{fp}}}$ from source normal patches that the host scores highly. For either bank $z \in \{\mathrm{def}, \mathrm{fp}\}$, each source patch is measured on the same text direction by $c_{z,j}^{(\ell)}(x) = \langle b_{z,j}^{(\ell)}, u(x) \rangle$. Thus, query patches, source defects, and source hard-FPs are compared on the same scale. For each bank z, we compute how close the query response is to the bank responses:

$$q_z^{(\ell)}(v_i; x) = \log \frac{1}{N_z} \sum_{j=1}^{N_z} \exp\left(-\frac{(c_i^{(\ell)}(x) - c_{z,j}^{(\ell)}(x))^2}{\tau} \right), \qquad z \in \{\mathrm{def}, \mathrm{fp}\}. \tag{2}$$

Here, q_{def} is large when the query looks like source defects, q_{fp} is large when it looks like source hard-FPs, and τ controls the similarity bandwidth. The train-free TED score is

$$r_{\mathrm{TED}}^{(\ell)}(v_i; x) = q_{\mathrm{def}}^{(\ell)}(v_i; x) - \lambda_{\mathrm{fp}} q_{\mathrm{fp}}^{(\ell)}(v_i; x). \tag{3}$$

A positive score means the patch is more defect-like; a negative score means it is more hard-FP-like. We average this score over host-selected layers and use the resulting T-TED map in place of raw prompt similarity.



(a) On-axis response. (b) Orthogonal residual. (c) Source-supported margin.

Figure 3: **Text-response entanglement and source-supported re-ranking.** Raw on-axis scoring strongly overlaps true defects with hard false positives. The orthogonal residual improves defect-vs-hard-FP separation in this case, but weakens separation from generic outside regions. The source-supported margin re-ranks the same responses while preserving strong outside separation, showing that TED decomposes rather than discards the text-guided response.

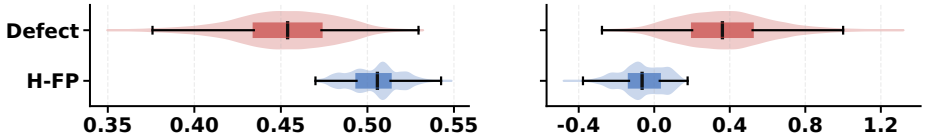


Figure 4: **Patch-level failure of raw prompt scoring.** The left panel shows the raw prompt baseline and the right panel shows TED. The baseline ranks hard-normal patches close to or above true defects, whereas TED re-ranks them using source defect and hard-FP support without backbone or prompt adaptation. H-FP denotes hard false positive.

3.4 Source-Calibrated TED Residual

Adapted CLIP-AD hosts already have useful learned scores from prompts, adapters, fusion modules, or host-specific calibration. Therefore, C-TED does not replace the host score. Instead, it uses the source defect-versus-hard-FP margin to add a small bounded correction. For each patch, we first compute the source margin

$$m_i^{(\ell)} = q_{\text{def}}^{(\ell)}(v_i) - \lambda_{\text{fp}}^{(\ell)} q_{\text{fp}}^{(\ell)}(v_i). \tag{4}$$

A large positive margin means the patch is closer to source defects; a negative margin means it is closer to source hard-FPs. C-TED converts this margin into a bounded residual:

$$\Delta s_i^{(\ell)} = \gamma_\ell \tanh(a_\ell \widehat{m}_i^{(\ell)} + b_\ell), \tag{5}$$

where $\widehat{m}_i^{(\ell)}$ is the normalized margin, and a_ℓ, b_ℓ, and γ_ℓ are learned only from source defect–hard-FP pairs. The $\tanh$ term prevents the residual from becoming arbitrarily large. The final C-TED score keeps the host score as the anchor:

$$s_{\text{TED}}^{(\ell)}(v_i) = s_{\text{host}}^{(\ell)}(v_i) + \Delta s_i^{(\ell)}. \tag{6}$$

The residual is trained so that source defect patches score above source hard-FPs, while pairs already separated by the host are preserved. After source training, all parameters are frozen. No target images, masks, anomaly labels, scores, or metrics are used.

4 Experiments

4.1 Experimental Setting

Protocol and metrics. We evaluate TED under a source-to-target anomaly localization protocol on MVTec AD, VisA, MPDD, and BTAD, with MVTec AD 2 used only for the frozen-backbone diagnostic. Source data build evidence banks or train the source-calibrated residual; no target masks, anomaly labels, target scores, or target metrics are used for calibration or inference. For class-aware baselines, class names only instantiate the fixed prompt ensemble and are not used for source-bank construction, calibration, residual-rank selection, insertion strength, model selection, or target-score selection. Target masks are used only for evaluation. We report I-AUROC and pixel-level P-AUROC, P-AP, and P-PRO, treating pixel-level localization as primary because TED modifies local anomaly maps. I-AUROC is secondary, with official host global branches preserved when available.

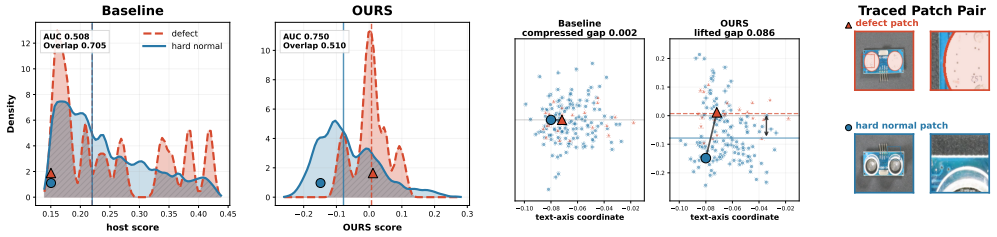


Figure 5: **Hard-FP decomposition in an adapted host.** The baseline assigns overlapping high scores to true defect and hard-normal patches. TED re-scores the same responses with source defect and hard-FP evidence, reducing overlap and lifting the defect–hard-normal ranking gap.

Table 1: Cross-dataset transfer results for frozen VLM backbones with class-aware static prompts. B denotes standard prompt similarity, T-TED train-free TED, and C-TED source-calibrated TED. All variants use the same frozen backbone and prompts; only the local anomaly score changes. Best results within each backbone and target dataset are bolded.

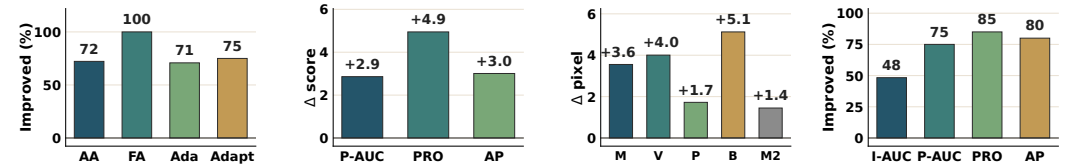
Backbone	Variant	MVTec[1]				VisA[42]				MPDD[22]				BTAD[25]				MVTec2[14]			
		I-AUC	P-AUC	PRO	AP	I-AUC	P-AUC	PRO	AP	I-AUC	P-AUC	PRO	AP	I-AUC	P-AUC	PRO	AP	I-AUC	P-AUC	PRO	AP
ViT-L/14-336	B	**88.4**	32.2	9.7	2.8	77.3	41.6	13.5	1.3	73.3	24.4	3.9	2.2	65.8	23.9	3.7	1.8	55.7	33.1	13.0	0.6
	T-TED	88.2	**69.3**	**64.0**	**21.2**	**79.4**	59.2	42.3	**6.7**	78.5	**76.4**	60.5	**5.4**	71.3	**77.9**	**50.8**	**17.2**	58.4	68.2	31.5	6.4
	C-TED	88.2	69.2	63.9	21.1	**79.4**	**59.3**	**42.7**	6.7	**78.5**	76.0	**60.6**	5.4	**71.3**	77.6	50.5	16.9	**58.5**	**68.2**	**31.5**	**6.4**
ViT-B/16+	B	89.5	17.3	1.7	2.4	73.2	21.7	3.4	0.4	**60.1**	28.0	4.1	1.4	67.1	38.3	7.5	2.2	**54.7**	35.5	5.7	0.5
	T-TED	**90.0**	82.9	67.8	17.1	**75.5**	77.4	51.9	6.1	59.4	70.8	48.4	**10.2**	**69.0**	**63.3**	34.8	**6.1**	54.5	64.5	**29.7**	**2.7**
	C-TED	**90.0**	**82.9**	**68.0**	**17.2**	**75.5**	**77.6**	**52.3**	**6.2**	59.4	**70.9**	**48.9**	10.1	**69.0**	63.1	**34.8**	6.1	54.5	**64.5**	29.3	2.6
ViT-L/14	B	**85.3**	30.7	8.2	2.7	68.1	44.6	16.3	1.3	71.6	25.1	2.8	2.0	61.0	22.8	2.7	1.8	**59.1**	36.8	12.9	0.6
	T-TED	84.7	**70.9**	**62.6**	**19.8**	**70.9**	56.2	36.0	**5.6**	**76.4**	**75.8**	**55.9**	**5.0**	**68.3**	**79.0**	**48.3**	**14.7**	57.6	64.3	26.7	4.3
	C-TED	84.7	70.9	62.4	19.7	**70.9**	**56.3**	**36.1**	5.6	76.4	75.6	55.9	5.0	**68.3**	78.5	47.7	14.2	57.6	**64.3**	**26.9**	**4.3**
ViT-H/14	B	88.8	15.2	1.0	2.0	79.1	18.0	1.5	0.4	**67.1**	14.4	0.6	1.2	77.2	23.2	3.6	1.9	57.2	29.2	6.4	0.5
	T-TED	90.4	**85.3**	**77.9**	**26.2**	**79.4**	**82.3**	**61.8**	**9.5**	65.5	**87.6**	**77.2**	19.0	**84.5**	**78.3**	**55.0**	**13.9**	**60.4**	**71.8**	38.9	**4.7**
	C-TED	**90.4**	85.2	77.7	26.2	**79.4**	82.2	61.6	9.5	65.5	87.4	77.1	**19.1**	**84.5**	78.1	54.5	13.9	**60.4**	71.6	**38.9**	4.6
ImageBind	B	**90.5**	14.4	0.7	2.0	**79.7**	16.5	1.0	0.4	65.2	19.5	1.7	1.3	**86.1**	18.6	2.9	1.8	58.0	26.9	3.9	0.4
	T-TED	90.3	**85.7**	**77.7**	25.8	79.5	**84.1**	**62.4**	**9.8**	**65.4**	**82.3**	**73.3**	17.8	84.9	**82.3**	**60.7**	**18.7**	60.7	**74.1**	**37.4**	**5.3**
	C-TED	90.3	85.6	77.7	**25.9**	79.6	84.0	62.3	9.8	65.4	82.1	73.1	**17.9**	84.9	82.1	60.3	18.6	**60.7**	73.9	37.4	5.1

Evaluation regimes. We evaluate three settings. First, raw CLIP and ImageBind with static class-aware prompts test whether frozen multimodal features contain recoverable local defect evidence without anomaly-specific adaptation. Second, adapted-host experiments keep each official CLIP-AD host unchanged and apply C-TED only as a post-hoc local evidence decoder. Third, layer/backbone recipe diagnostics test whether hard-FP competition is simply a representation-choice artifact, using both fixed-host swaps and host retraining with rebuilt source banks.

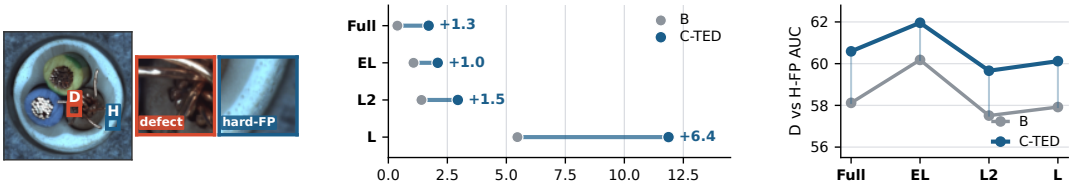
Implementation details. For each host, we preserve its official prompts, text encoder, selected layers, image resolution, post-processing, and metrics whenever possible. Defect banks are built from source anomalous patches overlapping source masks; hard-FP banks are built from source normal patches with high official host anomaly scores. T-TED uses the signed support margin directly, whereas C-TED anchors the host local score and learns only a bounded source-domain residual from source defect and hard-FP pairs. Unless otherwise stated, bank sizes, residual ranks, support bandwidths, and hard-FP mining rules are fixed before target evaluation; all variants use the same source-to-target split, and multi-seed results report mean and standard deviation.

4.2 Main Results and Diagnostics

Frozen-backbone evidence recovery. We first use raw VLM as a diagnostic setting for local evidence decoding. Table 1 evaluates frozen VLM backbones with class-aware static prompts, where B uses normal-versus-abnormal prompt similarity and T-TED/C-TED change only the local anomaly score. Across backbones and targets, direct prompt similarity gives weak pixel-level localization, especially in P-PRO and P-AP, while TED substantially improves pixel metrics. I-AUC changes are smaller and sometimes mixed, consistent with TED targeting local maps rather than image-level screening. This contrast is important: the same frozen representation can support much better localization once the local response is decoded against source defect and hard-FP evidence. In other



(a) Host-wise. (b) Metric-wise. (c) Target-wise. (d) Consistency.



(e) Traced patches. (f) Patch-rank margin. (g) Aggregate separation.

Figure 6: **Adapted-host and recipe-retuning diagnostics. Top**: source-calibrated TED improves local pixel ranking across adapted CLIP-AD hosts; M, V, P, B, and M2 denote MVTec AD , VisA, MPDD, BTAD, and MVTec AD 2, respectively. **Bottom**: changing and retraining layer recipes shifts the host response, but the same traced defect and hard-FP patches remain competitive under the baseline. C-TED increases the patch-rank margin and yields cleaner aggregate separation between defect-supported and hard-normal-supported responses.

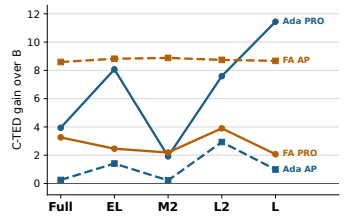
words, the frozen backbone is not necessarily missing all defect information; the problem is that raw prompt similarity ranks that information poorly. This also shows why TED does not need to update the backbone in this setting. It changes how local responses are scored, not what visual features are extracted. Fig. 4 shows the patch-level failure: raw prompt scoring can rank hard-normal patches above true defects, whereas TED reverses this ranking using source defect and hard-FP support without changing the backbone or prompts. Together, Table 1 and Fig. 4 show that frozen multimodal backbones can contain recoverable defect evidence, but direct prompt scoring may decode it poorly against hard-normal responses. Thus, this diagnostic separates representation capacity from local readout quality, supporting the evidence-decoding view behind TED.

Hard-FP decomposition in adapted hosts. Fig. 5 illustrates an adapted-host case where true defect and hard-normal patches receive overlapping high anomaly scores. TED re-scores the same ambiguous responses with source defect and hard-FP evidence, reducing their overlap and increasing the local ranking gap. Fig. 6 (top) summarizes this behavior across adapted CLIP-AD hosts, target datasets, and pixel-level metrics. C-TED improves most pixel-level settings, with gains distributed across P-AUC, P-PRO, and P-AP. The gains are host-dependent: highly calibrated hosts leave less room for correction, while settings with stronger hard-FP competition benefit more. This is expected because C-TED is designed to correct remaining hard-normal competition, not to overwrite the host's learned anomaly response. When the baseline already separates true defects from hard-normal regions, the correction tends to be smaller. When the two groups overlap, the source defect and hard-FP banks provide a clearer local re-ranking signal. This pattern indicates that C-TED complements host adaptation by correcting residual local ranking errors rather than replacing the adapted detector. All adapted-host corrections are estimated only from source-domain evidence banks; full host-wise and dataset-wise results are reported in the Appendix.

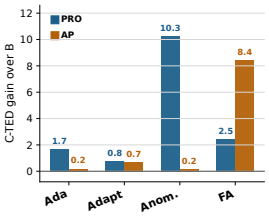
Failure-conditioned gains. We test whether hard-FP severity predicts when TED helps. If C-TED corrects hard-FP entanglement, gains should grow when the baseline assigns stronger anomaly evidence to hard-normal regions. Table 2 groups completed FAPrompt/AdaCLIP target classes into Low, Mid, and High regimes using only baseline hard-normal responses. This analysis-only grouping uses only baseline hard-normal responses, without C-TED scores, target metrics, calibration, tuning, or model selection. C-TED improves all regimes, but mean

Table 2: **Failure-conditioned gains.** Bins are grouped by baseline hard-FP severity before inspecting C-TED gains.

H-FP bin	#bins	Sev.	ΔPRO / ΔAP	ΔLoc
Low	3	0.07	+3.7 / +6.3	+5.0
Mid	3	0.10	+8.8 / +13.8	+11.3
High	3	0.19	+11.3 / +9.8	+10.6



(a) Layer retuning.

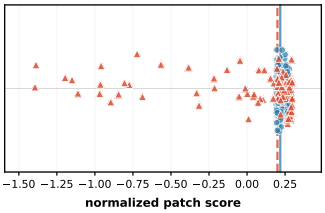


(b) Backbone retuning.

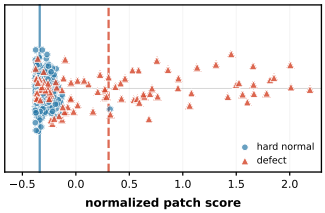
Retuning setting

Setting	Retrain	Bank	PRO / AP
Fixed swap	–	V	+3.6 / +2.8
Layer recipe	V	V	+4.7 / +5.0
Backbone recipe	V	V	+3.8 / +2.4

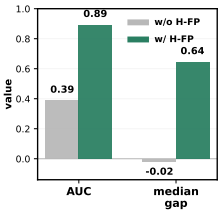
Gain over baseline; banks are source-only.



(c) Defect-only support.



(d) With hard-FP support.



(e) Separation gain.

Figure 7: **Recipe-consistent retuning and hard-FP support. Top**: C-TED remains beneficial after changing the feature-layer or backbone recipe, retraining the host when applicable, and rebuilding source banks for the resulting representation. **Bottom**: defect-only support leaves defect and hard-normal patches entangled, whereas adding hard-FP support explicitly models the competing hard-normal evidence and produces a cleaner local ranking margin.

localization gain over ΔP-PRO and ΔP-AP rises from $+5.0$ to $+10.9$ from Low to Mid/High. This makes the result useful as a diagnostic: it checks whether the measured failure mode aligns with where C-TED improves most. It also explains why we avoid presenting C-TED as a generic post-processing boost. The grouping is not used to choose any parameter, so the table should be read as post-hoc evidence for the mechanism, not as a tuning rule. Thus, TED is not a uniform score shift; it helps most when anomaly-sensitive hosts rank hard-normal evidence close to true defects. This trend is consistent with smaller gains in already well-separated settings, where less hard-FP competition remains to correct.

Recipe-consistent retuning analysis. A natural concern is that hard-FP competition reflects a poor layer or backbone recipe. We test recipe-consistent settings where the host is retrained after recipe changes and source banks are rebuilt. Fig. 6 (bottom) traces the same defect and hard-FP pair across retrained layer recipes. Although recipes shift host responses, the baseline still ranks the defect and hard-FP pair closely. In other words, changing the layer recipe can move the scores, but it does not always make the defect clearly outrank the confusing normal patch. This is why recipe search and evidence decoding address different parts of the problem. Recipe search changes which features the host uses, while C-TED changes how ambiguous high-scoring patches are compared against source defect and hard-FP examples. Fig. 7 (top) further shows that C-TED improves local PRO/AP after layer/backbone retraining and source-bank reconstruction. Together, these results suggest that recipe selection alone cannot ensure clean local ranking.

4.3 Ablation and Control Studies

Evidence components and residual rank. Table 3 ablates source evidence and residual rank. On adapted hosts, T-TED replacement can be brittle because the official score is already calibrated; C-TED instead anchors the host map with a bounded source-calibrated residual. Hard-FP support helps most under strong hard-normal competition, and the rank ablation shows that a small residual subspace is sufficient. Fig. 7 (bottom) confirms that hard-FP support separates defect and hard-normal patches more clearly than defect-only support. This means the useful signal comes from modeling the confusing normal patches, not simply from adding more correction capacity. Together, these results support hard-FP-aware residual correction over host-score replacement or larger residual capacity.

Source-supervision controls. Table 4 provides source-only controls for the residual correction. Label shuffle and label swap keep the same source split, bank size, residual capacity, and calibration

Table 3: Ablation study of TED on BTAD. (a) validates the calibrated decoding components on adapted hosts, and (b) analyzes the effect of the residual rank on AA-CLIP.

Host	Method	P-AUROC	P-PRO	P-AP
AA-CLIP	Baseline	92.48	59.78	36.55
	T-TED	86.58	53.50	27.88
	C-TED w/o hard-FP	93.58	61.76	38.40
	C-TED	**94.23**	**65.42**	**39.68**
FAPrompt	Baseline	90.52	60.32	21.54
	T-TED	90.00	60.12	20.33
	C-TED w/o hard-FP	94.99	70.93	45.01
	C-TED	**94.99**	**70.95**	**45.01**

(a) Component ablation

Rank	I-AUROC	P-AUROC	P-PRO	P-AP
Baseline	90.62	92.48	59.78	36.55
$r = 2$	91.10	**94.23**	**65.42**	**39.68**
$r = 4$	91.26	93.83	65.06	38.99
$r = 8$	**91.46**	93.51	63.67	38.26

(b) Residual rank ablation

Table 4: **Source-control ablations.** Controls corrupt source supervision or replace TED with scalar source readouts using the same evidence. Values are gains over the host baseline.

(a) Source-supervision semantics

Control	n	ΔP-AUC	ΔPRO	ΔAP
Correct C-TED	3	**+0.12**	**+2.47**	**+2.07**
Label shuffle	3	-1.42	-1.16	-2.12
Label swap	3	-6.29	-7.93	-9.32

(b) Scalar source readout

Control	n	ΔPRO / ΔAP	ΔLoc
Scalar replace	16	-19.8 / -14.1	-17.0
Scalar residual	16	+2.1 / -5.3	-1.6
C-TED	16	**+4.4 / +3.7**	**+4.1**

budget, but corrupt the defect-versus-hard-FP assignment. These variants reduce or remove the gain, suggesting that the correction depends on the semantic direction of the source evidence, not only on added residual capacity.

The second control trains simple source-supervised scalar readouts on the same source defect and hard-FP banks, either replacing the local map or adding a residual to the host map. These readouts are weaker than C-TED in the completed settings. The Appendix further tests host-score-only calibrators and additional learned readouts, showing that scalar calibration is not a clean substitute for source-supported evidence comparison. Thus, the gains are not fully explained by source-bank access, scalar supervision, or host-score recalibration alone. TED uses source banks to contrast defect and hard-FP support along the host's text axis, rather than treating full-dimensional visual similarity as the anomaly criterion. Additional same-bank controls compare full-dimensional 1-NN, prototype, and LogMeanExp readouts in both replacement and residual modes (Appendix B.6, Table 11). Across the 18 evaluated settings, none of these Full-D readouts improves all three pixel-level metrics on average, whereas C-TED does.

Efficiency and fixed-policy sensitivity. Table 5 varies the retained source-bank size while keeping the source split, target protocol, resolution, and post-processing fixed. Across tested budgets, C-TED yields nearly identical localization gains, indicating that the correction saturates with a moderate retained bank size. This reduces sensitivity to the exact retained-bank budget and avoids selecting bank sizes from target-domain performance. Because source banks are built once from the source split and then reused, the evaluation does not introduce target-specific bank construction. These results support using a fixed source-bank policy across target datasets without target-domain score selection, per-target bank retuning, or target-specific threshold search.

Table 5: **Fixed-bank sensitivity.**

Bank	Ret.	ΔPRO	ΔAP	ΔLoc
64	256	+7.17	+13.33	+10.25
128	512	+7.12	+13.32	+10.22
512	2048	+7.18	+13.33	+10.25

5 Conclusion

We presented TED, a text-axis evidence decomposition framework for prompted anomaly localization. Our analysis showed that adapted CLIP-based anomaly detectors can remain anomaly-sensitive while ranking true defects close to visually complex normal regions. Rather than replacing the host detector or discarding its text-guided response, TED re-ranks ambiguous local responses by comparing source defect support with source hard-false-positive support. This yields a train-free score for raw VLM

backbones and a source-calibrated residual correction for adapted CLIP-AD hosts, while leaving the visual backbone and prompts unchanged and requiring no target-domain training. Across frozen multimodal backbones, adapted hosts, and cross-domain transfer settings, TED improves pixel-level localization by reducing hard-FP competition and recovering defect evidence obscured by raw prompt similarity or host-specific local scoring. These results suggest that prompted anomaly localization should be treated not only as an adaptation problem, but also as a local evidence-decoding problem, with gains depending on remaining hard-FP competition and representative source evidence.

Acknowledgments and Disclosure of Funding

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University), and RS-2022-II220124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities]. This research was also supported by the AI Seoul Tech Research Support Program of the Seoul Future Foundation.

References

[1] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019.

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4183–4192, 2020.

[3] Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European conference on computer vision*, pages 55–72. Springer, 2024.

[4] Yuhao Chao, Jie Liu, Jie Tang, and Gangshan Wu. Anomalyr1: A grpo-based end-to-end mllm for industrial anomaly detection. *arXiv preprint arXiv:2504.11914*, 2025.

[5] N Cohen and Y Hoshen. Sub-image anomaly detection with deep pyramid correspondences. arxiv 2020. *arXiv preprint arXiv:2005.02357*, 2, 2005.

[6] Yajie Cui, Zhaoxiang Liu, and Shiguo Lian. A survey on unsupervised anomaly detection algorithms for industrial images. *IEEE Access*, 11:55297–55315, 2023.

[7] 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*, pages 475–489. Springer, 2021.

[8] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9737–9746, 2022.

[9] 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. *arXiv preprint arXiv:2505.09926*, 2025.

[10] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023.

[11] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 1932–1940, 2024.

[12] Denis Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 98–107, 2022.

[13] Lars Heckler, Rebecca König, and Paul Bergmann. Exploring the importance of pretrained feature extractors for unsupervised anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2917–2926, 2023.

[14] Lars Heckler-Kram, Jan-Hendrik Neudeck, Ulla Scheler, Rebecca König, and Carsten Steger. The mvtec ad 2 dataset: Advanced scenarios for unsupervised anomaly detection. *International Journal of Computer Vision*, 134(4):175, 2026.

[15] Jinlei Hou, Yingying Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, and Hong Zhou. Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8791–8800, 2021.

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

[17] Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. Mmad: A comprehensive benchmark for multimodal large language models in industrial anomaly detection. *arXiv preprint arXiv:2410.09453*, 2024.

[18] Daehyun Kim, Sungyong Baik, and Tae Hyun Kim. Sanflow: Semantic-aware normalizing flow for anomaly detection. *Advances in neural information processing systems*, 36:75434–75454, 2023.

[19] Donghyeong Kim, Chaewon Park, Suhwan Cho, Hyeonjeong Lim, Minseok Kang, Jungho Lee, and Sangyoun Lee. Genclip: Generalizing clip prompts for zero-shot anomaly detection. *arXiv preprint arXiv:2504.14919*, 2025.

[20] Geonu Lee, Yujeong Oh, Geonhui Jang, Soyoung Lee, Jeonghyo Song, Sungmin Cha, and YoungJoon Yoo. Continual-mega: A large-scale benchmark for generalizable continual anomaly detection. *Neurocomputing*, page 134460, 2026.

[21] Jingyi Liao, Yongyi Su, Rong-Cheng Tu, Zhao Jin, Wenhao Sun, Yiting Li, Dacheng Tao, Xun Xu, and Xulei Yang. Ad-fm: Multimodal llms for anomaly detection via multi-stage reasoning and fine-grained reward optimization. *arXiv preprint arXiv:2508.04175*, 2025.

[22] Jiaqi Liu, Guoyang Xie, Jinbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Jin. Deep industrial image anomaly detection: A survey. *Machine Intelligence Research*, 21(1):104–135, 2024.

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

[24] Zeqi Ma, Jiaxing Li, and Wai Keung Wong. Patch distance based auto-encoder for industrial anomaly detection. *Expert Systems with Applications*, 270:126537, 2025.

[25] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. *arXiv preprint arXiv:2104.10036*, 2021.

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[27] Oliver Rippel, Patrick Mertens, and Dorit Merhof. Modeling the distribution of normal data in pre-trained deep features for anomaly detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6726–6733. IEEE, 2021.

[28] Nicolae-Cătălin Ristea, Neelu Madan, Radu Tudor Ionescu, Kamal Nasrollahi, Fahad Shahbaz Khan, Thomas B Moeslund, and Mubarak Shah. Self-supervised predictive convolutional attentive block for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13576–13586, 2022.

[29] K Roth, L Pemula, J Zepeda, B Schölkopf, T Brox, and P Gehler. Patchcore: Towards total recall in industrial anomaly detection. In *IEEE CVPR*, 2022.

[30] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-based defect detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1088–1097, 2022.

[31] G Wang, S Han, E Ding, and D Huang. Student-teacher feature pyramid matching for unsupervised anomaly detection. arxiv 2021. *arXiv preprint arXiv:2103.04257*, 1, 2021.

[32] Jiacong Xu, Shao-Yuan Lo, Bardia Safaei, Vishal M Patel, and Isht Dwivedi. Towards zero-shot anomaly detection and reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20370–20382, 2025.

[33] Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677*, 2021.

[34] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8330–8339, 2021.

[35] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024.

[36] Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16699–16708, 2024.

[37] Shifang Zhao, Yiheng Lin, Lu Han, Yao Zhao, and Yunchao Wei. Omniad: Detect and understand industrial anomaly via multimodal reasoning. *arXiv preprint arXiv:2505.22039*, 2025.

[38] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021.

[39] 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, 2022.

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

[41] Jiawen Zhu, Yew-Soon Ong, Chunhua Shen, and Guansong Pang. Fine-grained abnormality prompt learning for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22241–22251, 2025.

[42] 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*, pages 392–408. Springer, 2022.

Table 6: **Fixed evaluation policy.** All entries are fixed before target evaluation and are not selected using target-domain scores or masks.

Item	Policy
Defect bank	Source anomalous patches overlapping source masks
Hard-FP bank	High-scoring source normal patches under the host baseline
Target usage	Target masks and scores are used only for final evaluation
Bank budget	Fixed per host; not selected per target dataset
Hard-FP mining	Fixed source-normal top fraction
Support bandwidth	Fixed before target evaluation
FP weight λ_{fp}	Fixed before target evaluation
Residual rank r	Fixed source-calibrated rank
Calibration data	Source defect and source hard-FP pairs only
Calibration budget	Fixed epochs, optimizer, and train-point budget
Host components	Backbone, prompts, layers, and post-processing unchanged

A Additional Experimental Details

A.1 Implementation Details

For each adapted CLIP-AD host, we keep the official visual backbone, text encoder, prompt format, selected feature layers, image resolution, post-processing, and metric implementation unchanged whenever possible. TED is applied only as a post-hoc local evidence decoder. For T-TED, the signed source-support margin is used directly as the local anomaly map. For C-TED, the official host local score remains the anchor, and a bounded source-calibrated residual adjusts the local ranking.

The source defect bank is constructed from source-domain anomalous patches whose spatial support overlaps source anomaly masks. The hard-FP bank is constructed from source normal patches that receive high local anomaly scores under the corresponding host baseline. Both banks are stored as visual features and projected onto the query-specific normal-versus-anomaly text response at inference time. For backbone-swap or recipe-retuning experiments, source banks are rebuilt for the resulting feature representation. No target images, target masks, target anomaly labels, target scores, or target metrics are used to select bank entries, calibration parameters, residual ranks, insertion strengths, or support bandwidths. Class names, when required by the fixed prompt protocol, are used only for prompt instantiation.

The C-TED residual calibrator is trained only from source defect and source hard-FP pairs. The calibration objective encourages source defect patches to rank above source hard-FP patches after correction, while preserving source pairs already separated by the host. All calibration parameters are frozen before target evaluation. Unless otherwise stated, the same fixed bank policy and hyperparameters are used across target datasets.

A.2 Fixed Evaluation Policy

Table 6 summarizes the fixed source-calibration policy. All entries are determined before target evaluation. No target-domain mask, anomaly label, target score, or target metric is used for policy selection. Class names are used only for fixed prompt instantiation when required by the corresponding baseline protocol, and not for source-bank construction, calibration, residual-rank selection, insertion strength, support bandwidth, hard-FP mining, or model selection.

A.3 Dataset, Metrics, and Compute

We evaluate cross-domain anomaly localization on MVTec AD, VisA, MPDD, and BTAD. MVTec AD contains industrial object and texture categories with pixel-level anomaly masks; VisA contains object-centric industrial inspection categories with diverse normal structures; MPDD contains metal-part defects with challenging texture and shape variations; BTAD contains three industrial categories with pixel-level annotations and strong class imbalance. MVTec AD 2 is additionally used as a held-out diagnostic target for frozen-backbone and adapted-host stress evaluations when completed.

We report image-level AUROC (I-AUROC) and pixel-level AUROC (P-AUROC), average precision (P-AP), and AUPRO (P-PRO). Since TED modifies local anomaly maps rather than the global image-level branch, pixel-level metrics are the primary target, and I-AUROC is reported for transparency.

Table 7: **Training-control sanity checks for C-TED.** All variants use the same source-only calibration budget and residual capacity. Panel (a) tests the effect of swapping defect and hard-normal supervision across host/transfer settings. Panel (b) compares additional corrupted-control variants on AA-CLIP MVTec→BTAD. Values report gains over the corresponding host baseline.

Host / transfer	Control	ΔP-AUC	ΔPRO	ΔAP
AA-CLIP MVTec→BTAD	C-TED	**+1.03**	**+3.88**	**+1.71**
AA-CLIP MVTec→BTAD	Label swap	-6.40	-8.28	-10.45
AA-CLIP VisA→BTAD	C-TED	-0.98	**+0.69**	**+4.09**
AA-CLIP VisA→BTAD	Label swap	-12.19	-15.95	-17.23
AdaCLIP VisA→BTAD	C-TED	**+0.31**	**+2.84**	**+0.41**
AdaCLIP VisA→BTAD	Label swap	-0.29	+0.45	-0.28

(a) Transfer sanity check

Variant	ΔP-AUC	ΔPRO	ΔAP
Correct C-TED	**+1.03**	**+3.88**	**+1.71**
Label shuffle	-2.30	-1.46	-3.95
Label swap	-6.40	-8.28	-10.45
Random subspace	+0.90	+3.34	+0.59

(b) Detailed controls

For every host and backbone, the baseline and TED variants use the same image resolution, post-processing, and metric implementation.

All experiments were run on NVIDIA RTX 6000 Ada GPUs with 48GB memory. Most jobs use a single GPU; parallel sweeps run independent host/transfer settings on separate GPUs without changing the evaluation protocol. For retained-bank sensitivity, peak GPU memory is approximately 3.8–3.9GB for AA-CLIP and 4.1GB for FAPrompt. Source banks are constructed once from the source split and reused at inference time, so the added cost scales with the retained bank budget.

B Additional Ablation and Control Studies

B.1 Training-Control Sanity Checks

Table 7 tests whether C-TED's gain can be explained by extra source-only residual capacity alone. All variants use the same source-only calibration budget and residual capacity. On AA-CLIP MVTec→BTAD, correct C-TED improves P-AUC/PRO/AP by $+1.03/+3.88/+1.71$, whereas label swap gives $-6.40/-8.28/-10.45$. On AA-CLIP VisA→BTAD, C-TED is mixed in P-AUC (-0.98) but improves PRO/AP by $+0.69/+4.09$; label swap degrades all three metrics by $-12.19/-15.95/-17.23$. On AdaCLIP VisA→BTAD, correct C-TED gives smaller positive changes of $+0.31/+2.84/+0.41$. The detailed AA-CLIP controls further show that label shuffle degrades AP by -3.95, and a random residual subspace gives a smaller AP gain ($+0.59$) than correct C-TED ($+1.71$). These controls support the modest conclusion that the source defect-versus-hard-FP direction matters, rather than residual capacity alone.

B.2 Source-Supervision Semantics

Fig. 8 visualizes the same source-supervision control. All variants keep the source split, source-bank size, residual capacity, and calibration budget fixed; only the defect-versus-hard-FP supervision semantics are corrupted. The degradation under shuffled or swapped labels indicates that C-TED is not explained by adding trainable source-only residual capacity alone.

B.3 Source Evidence in Frozen VLMs

Fig. 9 isolates evidence decoding before anomaly-specific host adaptation. Raw prompt scoring often ranks hard-normal patches close to true defects. Source-supported scoring improves defect-versus-hard-normal AUC, and the text-axis support variant is closest to TED because it keeps the normal-versus-anomaly text coordinate.

B.4 Layer-Recipe Stress Test

Table 8 tests whether layer-recipe changes alone explain the hard-FP issue. On AdaCLIP with the full recipe, C-TED improves P-PRO from 30.48 to 44.00 on MPDD, with P-AP changing slightly from 29.76 to 29.93. Under the last-layer recipe, AdaCLIP improves from $84.31/40.79/24.29$ to $89.62/54.00/25.95$ in P-AUC/PRO/AP. For AnomalyCLIP, the full recipe improves from

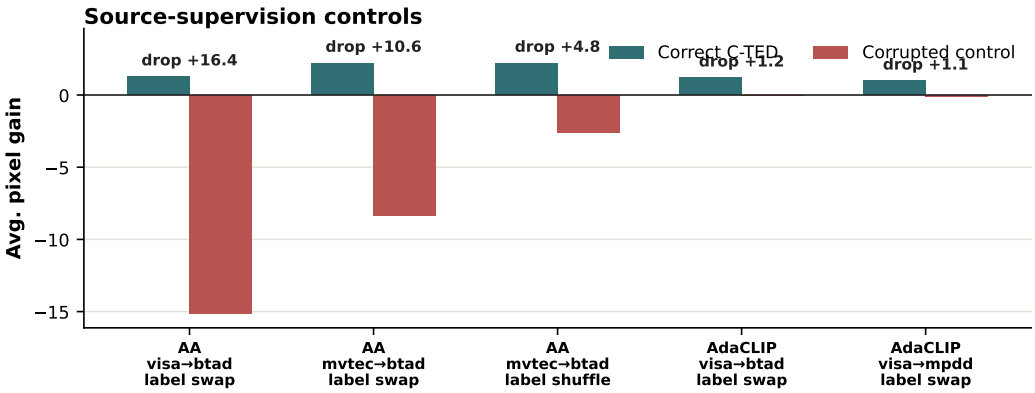


Figure 8: **Source-supervision controls for C-TED.** All variants use the same source split, source-bank size, residual capacity, and calibration budget. Correct C-TED preserves the defect-versus-hard-FP assignment, while label shuffle and label swap corrupt only the supervision semantics.

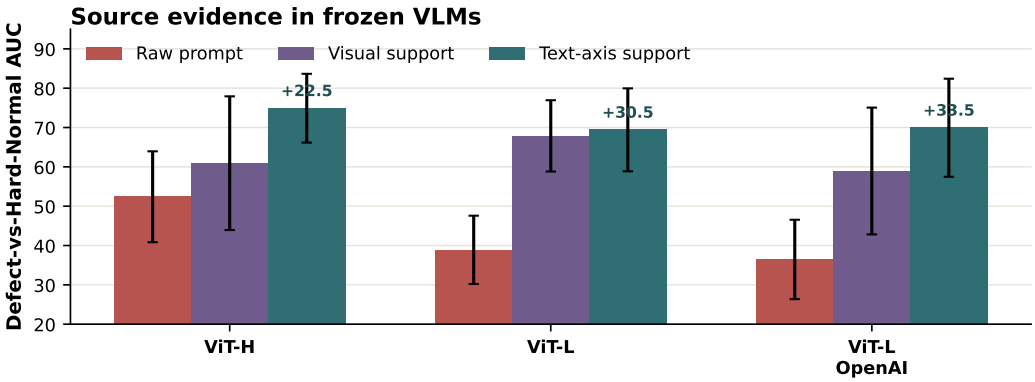


Figure 9: **Raw VLM source-support ablation.** Raw prompt scoring is compared with source-supported readouts in frozen VLM backbones. Text-axis support projects query and source-bank features onto the normal-versus-anomaly text axis, while visual support uses the source banks.

89.94/76.56/43.52 to 93.61/88.68/47.05 with C-TED, while the last-layer recipe improves P-AUC from 90.39 to 93.74 and keeps PRO high at 88.43. These numbers indicate that C-TED can help after recipe changes, although the size of the effect remains host- and metric-dependent.

B.5 Source-Bank Budget and Efficiency

Table 9 varies the retained source-bank budget while keeping the source split, target evaluation protocol, image resolution, and post-processing fixed. For AA-CLIP, increasing the bank budget from 64 to 512 changes ΔLoc only from $+2.7$ to $+2.9$, with peak memory around 3.8–3.9GB. For FAPrompt, budgets from 64 to 512 keep ΔLoc in a narrow $+8.6$–$+9.1$ range, with memory around 4.1GB. Runtime is not strictly monotonic across budgets, so we interpret the table as evidence of budget robustness rather than a precise scaling law. Overall, the correction appears to saturate with a moderate retained bank budget, supporting a fixed source-bank policy without target-domain score selection.

Table 8: Layer-recipe stress and TED recovery on MPDD dataset. B, T, and C denote the baseline host, train-free TED, and source-calibrated TED.

Host	Layers	M	I-AUC	P-AUC	P-PRO	P-AP
AdaCLIP	full	B	69.01	96.09	30.48	29.76
		T	**69.46**	96.08	30.67	29.70
		C	69.25	**96.34**	**44.00**	**29.93**
	last	B	61.59	84.31	40.79	24.29
		T	70.00	84.31	43.06	24.31
		C	**70.42**	**89.62**	**54.00**	**25.95**
AnomalyCLIP	full	B	93.89	89.94	76.56	43.52
		T	92.51	93.56	**88.73**	46.93
		C	93.89	93.61	88.68	**47.05**
	last	B	93.89	90.39	77.82	43.56
		T	90.34	93.66	88.44	43.73
		C	**93.91**	**93.74**	88.43	43.92

Table 9: Fixed source-bank budget sensitivity. Values are averaged over completed transfer settings for each host and bank budget. ΔLoc is the mean of ΔPRO and ΔAP over the host baseline.

Host	Budget	n	Retained	ms/img	Mem.	ΔLoc
AA-CLIP	64	6	6634.0	206.8	3.8GB	2.7
AA-CLIP	128	6	12180.0	182.9	3.8GB	2.9
AA-CLIP	512	6	35238.0	149.3	3.9GB	2.9
FAPrompt	64	6	256.0	1070.1	4.1GB	9.1
FAPrompt	128	6	512.0	1041.9	4.1GB	8.6
FAPrompt	256	6	1024.0	1055.2	4.1GB	9.1
FAPrompt	512	6	2048.0	1022.8	4.1GB	9.1

Table 10: Transfer-wise full-feature memory replacement control. The control uses the same source visual banks but removes the text-conditioned readout. Values are gains over the corresponding host baseline.

Host	Transfer	ΔP-AUC	ΔPRO	ΔAP
AA-CLIP	MVTec→BTAD	-5.75	-7.53	-11.35
	MVTec→MPDD	+0.58	+4.92	+3.27
	MVTec→VisA	-0.43	-1.58	-3.28
	VisA→BTAD	-13.48	-16.22	-27.88
	VisA→MPDD	-2.53	-6.18	-5.98
	VisA→MVTec	-5.45	-6.30	-12.33
FAPrompt	MVTec→BTAD	-44.60	-47.22	-15.91
	MVTec→MPDD	-37.18	-52.31	-15.68
	MVTec→VisA	-36.61	-59.25	-8.92
	VisA→BTAD	-51.69	-48.15	-15.30
	VisA→MPDD	-38.00	-44.62	-14.72
	VisA→MVTec	-40.32	-50.61	-10.58
AdaptCLIP	MVTec→BTAD	-45.82	-55.90	-39.55
	MVTec→MPDD	-68.89	-87.24	-23.90
	MVTec→VisA	-63.02	-84.15	-25.80
	VisA→BTAD	-41.15	-61.49	-43.17
	VisA→MPDD	-24.01	-45.58	-21.96
	VisA→MVTec	-25.35	-37.33	-26.88

B.6 Full-Feature Memory Replacement Control

Table 10 provides transfer-wise full-feature memory replacement controls. This control uses the same source visual banks but removes the text-conditioned evidence readout, replacing the local map with a full-feature defect-versus-hard-FP memory score. The result is mostly negative. For FAPrompt, all listed transfers degrade, for example MVTec→BTAD changes by $-44.60/-47.22/-15.91$ in P-AUC/PRO/AP, and MVTec→VisA changes by $-36.61/-59.25/-8.92$. For AdaptCLIP, degradation is also large, including MVTec→MPDD at $-68.89/-87.24/-23.90$ and MVTec→VisA at $-63.02/-84.15/-25.80$. AA-CLIP contains one positive case, MVTec→MPDD at $+0.58/+4.92/+3.27$, but most AA-CLIP transfers are still negative. This supports a cautious interpretation: generic visual memory retrieval is not a substitute for the text-conditioned defect-versus-hard-FP comparison used by TED.

Extended same-bank retrieval controls. We extend the full-feature comparison with 1-NN, prototype, and LogMeanExp readouts, each evaluated in replacement and residual modes. The comparison covers all 18 settings across AA-CLIP, AdaCLIP, and FAPrompt, with six transfers per host. Source banks, hosts, targets, and aggregation are held fixed; only the evidence readout changes.

As shown in Table 11, none of the tested Full-D readouts improves all three pixel-level metrics on average, whereas C-TED does. The prototype residual achieves a larger P-PRO gain, but reduces P-AUC and P-AP. These results indicate that access to source banks or generic full-dimensional similarity alone does not explain the joint improvements observed with C-TED.

Table 11: Extended same-bank retrieval controls across AA-CLIP, AdaCLIP, and FAPrompt, with six transfers per host (18 settings in total). Values are mean changes in percentage points relative to the corresponding host. Source banks, hosts, targets, and aggregation are fixed; only the evidence readout changes. Replacement substitutes the host local score, whereas residual mode adds a correction to it.

Readout	Mode	ΔP-AUC	ΔP-PRO	ΔP-AP
Full-D 1-NN	Replacement	-45.65	-39.39	-20.68
	Residual	-10.99	-0.22	-7.66
Full-D prototype	Replacement	-11.85	-3.13	-12.63
	Residual	-3.27	$+11.04$	-5.64
Full-D LogMeanExp	Replacement	-24.70	-18.56	-16.67
	Residual	-8.68	$+3.69$	-7.86
C-TED	Residual	$+0.87$	$+4.38$	$+5.21$

Table 12: Weak-source robustness. AA-CLIP varies the number of source images per source class used to construct banks. FAPrompt varies the retained source-bank budget per class. Values are gains over the corresponding host baseline; ΔLoc is the mean of ΔPRO and ΔAP.

Host	Source budget	n	ΔPRO	ΔAP	ΔLoc
AA-CLIP	1 img/cls	4	+3.3	+1.9	+2.6
AA-CLIP	2 img/cls	4	+3.5	+1.7	+2.6
AA-CLIP	4 img/cls	4	+3.2	+3.2	+3.2
AA-CLIP	all img/cls	4	+3.1	+2.9	+3.0
FAPrompt	8 bank/cls	3	+7.2	+13.4	+10.3
FAPrompt	16 bank/cls	3	+7.3	+13.4	+10.4
FAPrompt	32 bank/cls	3	+7.2	+13.3	+10.3
FAPrompt	64 bank/cls	3	+7.2	+13.3	+10.3

B.7 Weak-Source and Bank-Budget Robustness

Table 12 evaluates the sensitivity of TED to reduced source evidence. For AA-CLIP, we vary the number of source images per class used to construct the evidence banks. For FAPrompt, we vary the retained source-bank budget per class after bank construction. The source split, target evaluation protocol, residual rank, insertion strength, post-processing, and target-domain selection policy are kept fixed.

The gains remain positive in these reduced-source settings. For AA-CLIP, even the smallest tested source-image budget retains positive localization gains. For FAPrompt, reducing the retained bank budget from 64 to 8 entries per class gives similar localization gains in the completed settings. These results suggest that the observed improvements are not driven only by retaining a very large source bank. At the same time, this is not a zero-source setting: representative source defect and hard-FP evidence remains necessary for constructing the comparison.

B.8 Simple Source-Supervised Readout Controls

We compare C-TED with a simple source-supervised MLP rank readout trained from the same source defect and hard-FP banks. This control uses the same source split and target evaluation protocol as C-TED, but replaces the TED evidence comparison with a direct learned scalar readout. We evaluate two usage modes: replacing the local anomaly map with the learned readout, and adding the learned readout as a residual to the host map.

Table 13 shows that direct replacement is brittle because it discards the host-specific local calibration. Adding the learned readout as a residual is more stable, but remains weaker than C-TED in the completed MLP-rank settings. This suggests that the gains are not explained only by access to the same source labels and source banks. Rather, source evidence is more effective when used as a defect-versus-hard-normal comparison anchored to the host response, instead of as a generic learned scalar readout.

Table 13: Source-supervised readout control. The control uses the same source defect and hard-FP banks as C-TED, but replaces the evidence comparison with a directly learned MLP rank readout. Replace uses the learned readout as the local anomaly map, while Residual adds it to the host map. Rows average completed transfer settings for each host. Values are ΔLoc, the mean of ΔPRO and ΔAP over the host baseline.

Host	n	Replace	Residual	C-TED
AA-CLIP	4	-7.4	-4.7	**+3.0**
AdaCLIP	4	-12.3	-4.1	**+4.1**
FAPrompt	4	-42.9	+3.8	**+10.0**

Table 15: Host-score-only calibration control. Controls train a source-only rank readout using only the scalar host local score on source defect and hard-FP patches, without source defect/hard-FP support features. Each entry reports ΔPRO / ΔAP over the corresponding host baseline.

Host	n	Host replace	Host residual	C-TED
AA-CLIP	4	$+0.1/-4.5$	$-2.3/-7.7$	$\mathbf{+3.1}/\mathbf{+2.9}$
FAPrompt	4	$+3.0/+3.3$	$+4.3/+5.0$	$\mathbf{+6.4}/\mathbf{+13.5}$
AdaCLIP	4	$+5.3/-9.7$	$+0.1/-5.9$	$\mathbf{+7.4}/\mathbf{-0.3}$

B.9 Residual-Strength Boundary

We use a small strength-sweep diagnostic to check whether the source-driven residual should simply be made larger. For this diagnostic, we keep the source banks, source calibrator, residual basis, target evaluation protocol, image resolution, and post-processing fixed, and vary only the residual insertion strength α. Table 14 reports the resulting localization gain on AA-CLIP.

The sweep suggests that stronger insertion is not automatically better. Moderate residual strengths give small positive gains, while an overly strong residual can reduce localization quality. We use this result as a boundary diagnostic rather than as a universal strength-sweep claim. It supports the design choice of keeping C-TED as a bounded correction anchored to the host response, instead of allowing the source-driven residual to dominate the host map.

Table 14: **Residual-strength boundary.** AA-CLIP diagnostic with fixed calibrator and protocol.

α	0	0.25	0.5	1	2
ΔLoc	+0.0	**+0.3**	+0.3	-0.0	-5.2

B.10 Host-Score-Only Calibration Control

We additionally test whether C-TED can be reduced to a scalar recalibration of the host local score. This control uses the same source defect and hard-FP patch sets as C-TED, but removes the source support features. It trains a source-only rank readout from only the scalar host scores observed on source defect and hard-FP patches. We evaluate two variants: replacing the local anomaly map with the learned readout, and adding the learned readout as a residual to the host map.

Table 15 shows that host-score-only calibration is not a reliable substitute for C-TED. The scalar readout can improve one metric in some hosts, but it often reduces AP or weakens the overall correction. For AA-CLIP, both host-score-only variants decrease AP, whereas C-TED improves both PRO and AP. For FAPrompt and AdaCLIP, C-TED also gives the strongest AP change among the compared variants. These results suggest that C-TED is not only recalibrating the host score; the source defect and hard-FP support features provide additional evidence for resolving ambiguous local responses.

C Additional Qualitative and Diagnostic Results

C.1 Raw VLM Failure and Recovery

Fig. 10 illustrates why raw prompt scoring fails in frozen CLIP. The raw prompt map responds strongly to visually salient normal structures and does not cleanly isolate the defect. TED keeps the

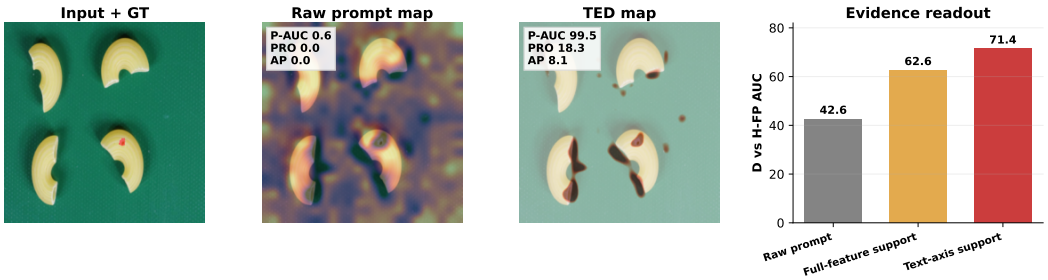


Figure 10: **Raw VLM failure and TED recovery.** Raw prompt scoring confuses defect evidence with hard-normal saliency, while TED recovers a cleaner local ranking using source-supported evidence.

Table 16: Failure-conditioned separability robustness. Intervals are setting-level percentile bootstrap 95% CIs over completed adapted-host settings. ΔAUC measures the change in AUC for ranking anomalous regions above hard normal regions. ΔLoc is the mean of ΔPRO and ΔAP.

Scope	n	ΔAUC [95% CI]	ΔLoc [95% CI]
FAPrompt	36	+8.1 [+6.3, +10.0]	+8.4 [+7.1, +9.7]
AA-CLIP	10	+2.2 [-1.9, +5.9]	+11.3 [-2.0, +25.1]
AdaCLIP	12	-2.5 [-4.3, -1.1]	+4.6 [+3.5, +5.7]
AdaptCLIP	12	+1.2 [+1.1, +1.3]	+0.3 [+0.1, +0.5]
All adapted hosts	70	+4.3 [+2.8, +5.8]	+6.8 [+4.7, +9.1]

same frozen backbone and prompts, but re-ranks the local response using source defect and hard-FP support.

C.2 Failure-Conditioned Separability Robustness

Table 16 expands the failure-conditioned diagnostic to 70 completed adapted-host settings. For each setting, hard-normal regions are selected using only the baseline response, and we compare how the baseline and TED rank anomalous regions relative to these hard-normal competitors. Across the completed settings, TED improves the AUC for separating anomalous and hard-normal regions, increases the mean score margin between the two groups, and improves localization metrics with positive setting-level bootstrap intervals. This shows that the localization gains are accompanied by better separation of anomalous regions from hard-normal competitors, rather than only by a uniform shift of anomaly scores. We therefore use this analysis as mechanism evidence that TED reduces hard-normal competition, while not interpreting baseline severity as a universal linear predictor of localization gain for every host.

C.3 Qualitative Localization on an Adapted Host

Figure 11 presents selected qualitative examples of C-TED on FAPrompt across four target datasets.

D Additional Quantitative Results

D.1 Full Adapted-Host Results

The following tables provide detailed transfer results for adapted CLIP-AD hosts. They complement the main-paper summaries by showing the behavior across host architectures, visual backbones, target datasets, and metrics. We interpret these tables as detailed evidence for a local evidence-decoding effect, not as a claim that every host, backbone, and metric improves uniformly. The clearest gains appear when the host still ranks hard-normal evidence close to true defects, while already calibrated or saturated settings can show smaller or mixed changes.

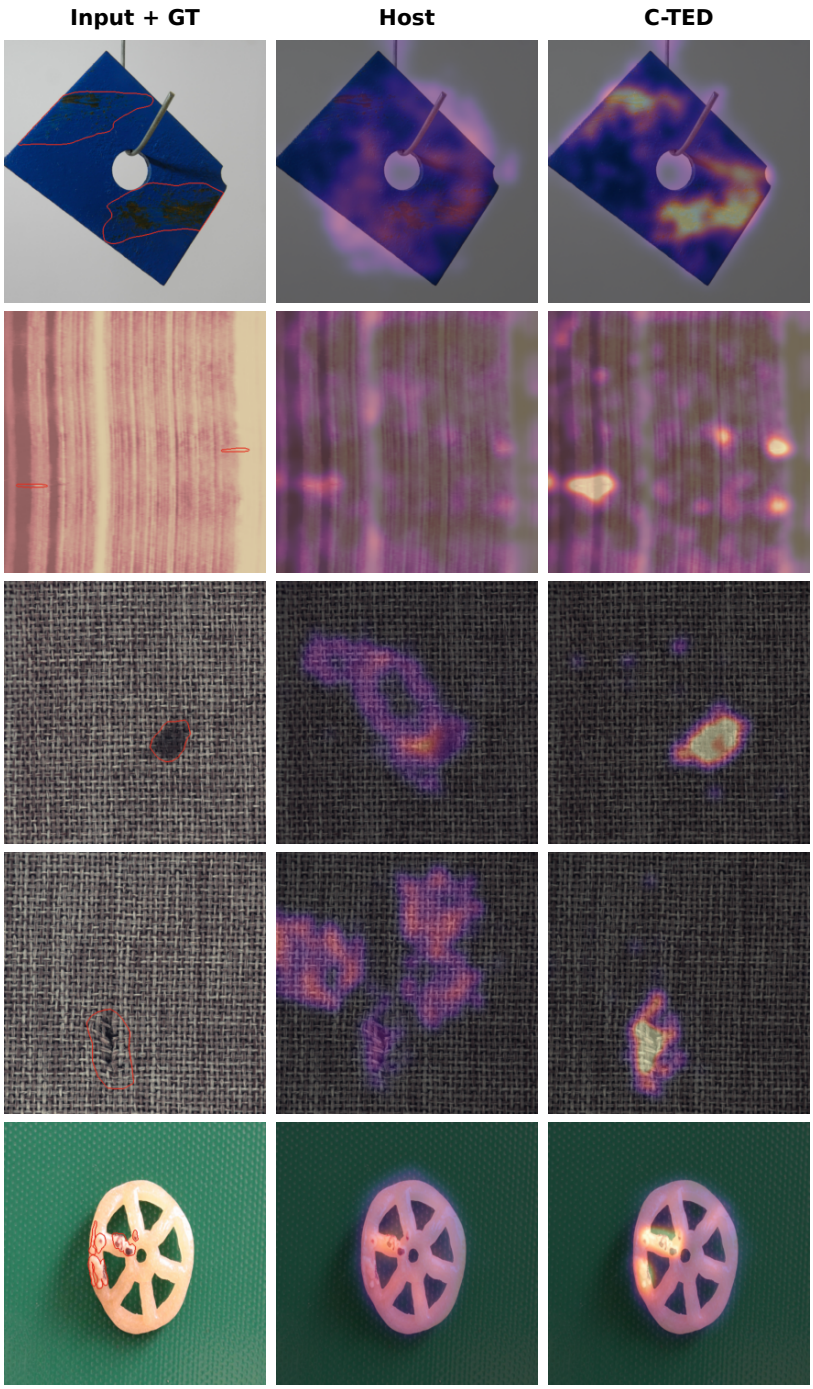


Figure 11: Qualitative comparisons of FAPrompt and C-TED across four target datasets, with ground-truth defects outlined in red and a shared color scale for each Host/C-TED pair.

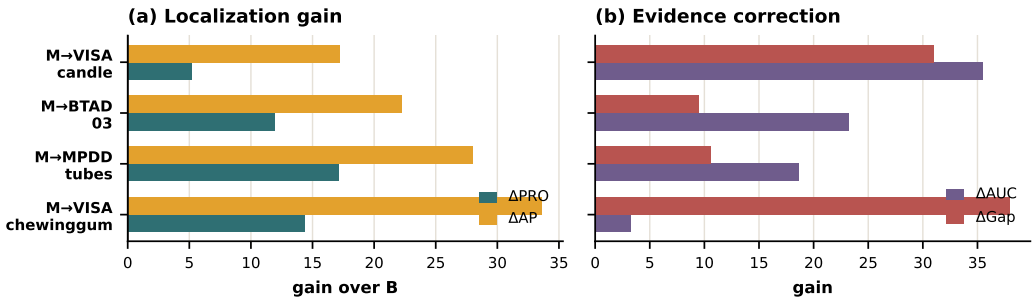


Figure 12: **Failure-conditioned class-level cases.** Representative target classes where C-TED improves localization while also increasing defect-versus-hard-FP AUC and the defect-minus-hard-FP score gap.

D.1.1 AA-CLIP

Table 17 reports AA-CLIP with host-anchored residual correction. Several settings show large pixel-level recovery from weak baseline localization. For ViT-B/16+ on MVTec, P-AUC/PRO/AP improve from $39.27/11.86/2.97$ to $69.22/38.88/10.39$. On ViT-B/16+ BTAD, the same metrics improve from $61.37/30.66/8.51$ to $76.30/41.13/13.73$. For ViT-H/14 on MVTec, C-TED improves P-AUC/PRO/AP from $51.04/16.10/4.36$ to $64.60/26.06/5.65$. These cases suggest that C-TED can recover useful local ranking when the baseline anomaly map is weak and hard-normal responses remain competitive. At the same time, the table includes boundary cases. For example, ViT-L/14-336 on BTAD has lower PRO/AP after correction despite a strong baseline. This pattern is consistent with our main claim: TED is most useful when there is hard-FP competition to correct, and more conservative or mixed when the host is already well calibrated.

D.1.2 FAPrompt

Table 18 reports FAPrompt results. C-TED gives particularly clear gains on BTAD, where hard normal structures often compete with subtle defect evidence. For ViT-L/14-336 on BTAD, P-AUC/PRO/AP improve from $90.52/60.32/21.54$ to $94.99/70.95/45.01$. The AP gain is substantial, suggesting that the correction improves sparse defect localization rather than only shifting broad pixel scores. For ViT-L/14 OpenAI on BTAD, AP improves from 16.51 to 40.96, while P-AUC/PRO improve from $87.19/54.00$ to $93.71/59.99$. These two BTAD cases are useful because the gains appear not only in ranking-based P-AUC, but also in PRO and AP, indicating improved region-level and sparse-pixel localization. This is consistent with FAPrompt already making the host defect-sensitive, while C-TED further separates hard-normal-supported responses from defect-supported responses. Some settings show smaller gains, so we treat FAPrompt as a strong but still host- and dataset-dependent case.

D.1.3 AdaptCLIP

Table 19 reports AdaptCLIP results. The gains are more modest than in some FAPrompt or AnomalyCLIP settings, which is expected for a host that already includes stronger textual, visual, and prompt-query adaptation. For ViT-L/14 OpenAI on BTAD, P-AUC/PRO/AP change from $93.80/65.41/41.15$ to $94.09/65.81/41.58$. For ViT-L/14-336 on BTAD, they change from $93.84/70.22/41.79$ to $94.31/70.66/42.40$. These are small but directionally consistent local-ranking gains, suggesting that a source-calibrated residual can still refine ambiguous responses without replacing the adapted host. Other entries are mixed, especially when the baseline is already strong or unstable across seeds. This behavior is consistent with the design of C-TED: it is intended to correct remaining hard-FP competition, not to override a well-calibrated detector or force large changes where little ambiguity remains. We include AdaptCLIP as an important boundary case that clarifies the scope of the method.

D.1.4 AdaCLIP

Table 20 reports AdaCLIP results. Several settings show useful PRO gains even when P-AUC or AP changes are mixed. For ViT-L/14 OpenAI on MVTec, PRO improves from 44.32 to 49.01. For

Table 17: AA-CLIP with host-anchored residual subspace correction. Results are averaged over three seeds; each entry is mean±std.

Host	Backbone	Target	Method	I-AUROC	P-AUROC	P-PRO	P-AP
AA-CLIP	ViT-B/16+	MVTec	Base	50.81±0.00	39.27±0.00	11.86±0.00	2.97±0.00
			OURS	**60.67±0.05**	**69.22±0.35**	**38.88±0.82**	**10.39±0.17**
		VisA	Base	46.61±0.00	54.49±0.00	22.32±0.00	1.31±0.00
			OURS	**54.41±0.14**	**63.66±0.22**	**28.74±0.29**	**2.18±0.04**
		MPDD	Base	44.86±0.00	88.93±0.00	65.37±0.00	**12.10±0.00**
			OURS	**48.04±0.21**	**88.98±0.28**	**65.44±0.67**	12.07±0.05
		BTAD	Base	63.83±0.00	61.37±0.00	30.66±0.00	8.51±0.00
			OURS	**79.78±0.07**	**76.30±0.11**	**41.13±0.08**	**13.73±0.11**
	ViT-L/14 OpenAI	MVTec	Base	84.80±0.00	**90.22±0.00**	78.94±0.00	34.51±0.00
			OURS	**85.19±0.00**	89.91±0.00	**81.45±0.08**	**35.80±0.02**
		VisA	Base	72.04±0.00	92.27±0.00	73.85±0.00	17.69±0.00
			OURS	**73.56±0.00**	**93.50±0.01**	**75.07±0.05**	**20.57±0.01**
		MPDD	Base	67.67±0.00	94.15±0.00	78.51±0.00	17.30±0.00
			OURS	**71.50±0.00**	**95.25±0.00**	**83.33±0.04**	**23.01±0.02**
		BTAD	Base	90.62±0.00	92.48±0.00	59.78±0.00	36.55±0.00
			OURS	**91.46±0.00**	**93.50±0.01**	**63.56±0.09**	**38.25±0.01**
	ViT-L/14-336	MVTec	Base	**89.78±0.00**	**91.09±0.00**	83.16±0.00	**43.57±0.00**
			OURS	89.42±0.00	90.25±0.01	**83.31±0.08**	41.66±0.02
		VisA	Base	74.26±0.00	**94.26±0.00**	**79.64±0.00**	**21.75±0.00**
			OURS	**75.22±0.01**	94.06±0.00	79.15±0.17	21.40±0.02
		MPDD	Base	58.48±0.00	96.05±0.00	85.66±0.00	21.60±0.00
			OURS	**67.88±0.04**	**96.08±0.00**	**85.94±0.01**	**22.68±0.02**
		BTAD	Base	**94.24±0.00**	**93.59±0.00**	**67.70±0.00**	**39.64±0.00**
			OURS	94.19±0.00	90.68±0.01	57.67±0.00	32.53±0.04
	ViT-H/14	MVTec	Base	46.46±1.80	51.04±1.85	16.10±0.63	4.36±0.03
			OURS	**53.25±1.78**	**72.61±4.21**	**40.36±8.33**	**9.10±2.86**
		VisA	Base	47.77±1.53	41.62±7.06	13.42±4.78	0.83±0.19
			OURS	**49.58±1.82**	**77.24±0.75**	**47.56±0.93**	**2.57±0.30**
		MPDD	Base	**51.81±1.13**	45.68±2.88	14.95±2.64	**5.83±2.78**
			OURS	50.93±0.19	**68.93±4.34**	**28.86±5.12**	3.79±0.58
		BTAD	Base	53.34±4.58	35.25±4.11	8.04±1.22	2.26±0.33
			OURS	**62.19±8.35**	**76.31±3.33**	**44.24±2.62**	**11.11±2.83**

ViT-L/14-336 on BTAD, PRO improves from 19.26 to 27.81. These gains indicate that source-conditioned correction can improve region-level localization quality in some AdaCLIP settings, even when threshold-free pixel ranking or sparse-pixel AP does not improve uniformly. This distinction is important because AdaCLIP already uses image-conditioned prompting, so the remaining errors may differ from those of simpler prompt-tuned hosts. In such cases, C-TED may help mainly by reducing region-level hard-FP competition rather than by globally reshaping the entire pixel-score distribution. However, some P-AUC and AP entries remain mixed, especially on weaker or already saturated configurations. This supports a bounded interpretation: C-TED is a local evidence-ranking correction whose benefit depends on the amount and structure of remaining hard-FP competition. We therefore use AdaCLIP as an intermediate case between strong positive hosts and boundary hosts, showing that the correction can help but should not be read as uniformly improving every metric.

Table 18: Results with **FAPrompt** as host.

Host	Backbone	Target	Method	I-AUROC	P-AUROC	P-PRO	P-AP
FAPrompt	ViT-B/16+	MVTec	Base	80.75±0.00	65.61±0.00	35.51±0.00	8.78±0.00
			OURS	80.75±0.00	**77.95±0.00**	**62.34±0.01**	**24.52±0.00**
		VisA	Base	69.16±0.00	67.98±0.00	40.10±0.00	3.61±0.00
			OURS	69.16±0.00	**71.56±0.01**	**54.98±0.08**	**8.79±0.00**
		MPDD	Base	67.35±0.00	81.33±0.00	59.09±0.00	11.07±0.00
			OURS	67.35±0.00	**82.45±0.01**	**61.23±0.17**	**12.98±0.01**
		BTAD	Base	77.86±0.00	72.46±0.00	34.31±0.00	5.92±0.00
			OURS	77.86±0.00	**79.83±0.01**	**47.19±0.06**	**19.62±0.01**
	ViT-L/14 OpenAI	MVTec	Base	85.42±0.00	88.56±0.00	68.59±0.00	18.58±0.00
			OURS	85.42±0.00	**89.80±0.00**	**74.45±0.00**	**33.34±0.00**
		VisA	Base	80.32±0.00	93.60±0.00	74.68±0.00	10.73±0.00
			OURS	80.32±0.00	**95.12±0.00**	**83.90±0.00**	**17.89±0.00**
		MPDD	Base	80.01±0.00	93.10±0.00	74.43±0.00	13.42±0.00
			OURS	80.01±0.00	**94.81±0.00**	**80.85±0.00**	**19.69±0.00**
		BTAD	Base	86.10±0.00	87.19±0.00	54.00±0.00	16.51±0.00
			OURS	86.10±0.00	**93.71±0.00**	**59.99±0.00**	**40.96±0.00**
	ViT-L/14-336	MVTec	Base	91.47±0.00	88.88±0.00	70.72±0.00	16.92±0.00
			OURS	91.47±0.00	**90.29±0.00**	**75.33±0.00**	**31.00±0.00**
		VisA	Base	82.76±0.00	94.42±0.00	77.10±0.00	10.43±0.00
			OURS	82.76±0.00	**95.61±0.00**	**82.39±0.00**	**19.21±0.00**
		MPDD	Base	79.45±0.00	94.17±0.00	78.85±0.00	17.59±0.00
			OURS	79.45±0.00	**96.07±0.00**	**83.91±0.00**	**25.31±0.00**
		BTAD	Base	89.24±0.00	90.52±0.00	60.32±0.00	21.54±0.00
			OURS	89.24±0.00	**94.99±0.00**	**70.95±0.00**	**45.01±0.00**
	ViT-H/14	MVTec	Base	71.22	75.62	44.23	10.50
			OURS	71.22	**77.23**	**53.27**	**18.87**
		VisA	Base	61.05	84.70	50.41	10.79
			OURS	61.05	**87.58**	**57.51**	**12.44**
		MPDD	Base	53.06	75.52	29.07	5.33
			OURS	53.06	**78.26**	**34.44**	**6.93**
		BTAD	Base	79.97	82.60	44.86	12.27
			OURS	79.97	**86.80**	**48.28**	**22.13**

D.1.5 BayesPFL

Table 21 reports BayesPFL results. Compared with the other hosts, BayesPFL shows more mixed pixel-level behavior. Several settings show image-level gains: for example, ViT-L/14-336 on MVTec improves I-AUC from 92.69 to 96.40, and ViT-L/14-336 on BTAD improves I-AUC from 77.89 to 89.92. Pixel-level changes are often smaller or mixed. For instance, ViT-L/14 OpenAI on BTAD has a small PRO gain from 64.28 to 65.57, while AP decreases from 35.98 to 34.30. This table is useful because it documents a boundary regime: when a host already provides relatively stable local ranking or when the remaining error is not primarily hard-FP competition, C-TED may yield limited or mixed pixel-level changes. This behavior is consistent with our failure-conditioned interpretation, since the residual correction is expected to help most when hard-normal evidence remains close to true defects. It also shows why we report image-level and pixel-level metrics separately: improvements in image-level ranking do not necessarily imply cleaner dense localization. Conversely, small or mixed pixel-level changes do not by themselves invalidate the evidence-decoding view, but indicate that the measured failure may be weaker or different in that host. BayesPFL therefore helps define

Table 19: Results with **AdaptCLIP** as host.

Host	Backbone	Target	Method	I-AUROC	P-AUROC	P-PRO	P-AP
AdaptCLIP	ViT-B/16+	MVTec	Base	66.05±4.16	**68.91±10.44**	**43.40±9.63**	15.22±3.02
			OURS	**66.06±4.14**	68.63±10.58	42.75±9.76	**15.41±3.01**
		VisA	Base	**54.46±3.63**	**50.17±14.41**	**22.72±16.49**	**2.40±2.06**
			OURS	54.21±3.27	50.01±14.65	22.33±16.42	2.36±1.99
		MPDD	Base	57.00±3.18	69.97±7.55	35.12±5.41	**6.10±2.76**
			OURS	**57.01±3.18**	**70.20±7.70**	**35.51±5.62**	6.10±2.80
		BTAD	Base	**62.52±12.53**	**61.03±12.90**	**25.20±11.36**	**8.05±7.55**
			OURS	61.71±12.09	60.84±13.13	24.53±11.17	7.88±7.36
	ViT-L/14 OpenAI	MVTec	Base	88.56±0.00	90.81±0.00	82.18±0.00	38.76±0.00
			OURS	**89.01±0.00**	**90.96±0.00**	**82.73±0.00**	**39.61±0.00**
		VisA	Base	**79.01±0.00**	95.14±0.00	85.18±0.00	23.15±0.00
			OURS	78.62±0.00	**95.22±0.00**	**85.21±0.00**	**23.45±0.00**
		MPDD	Base	**70.49±0.00**	95.33±0.00	85.95±0.00	19.94±0.00
			OURS	69.97±0.00	**95.36±0.00**	**86.40±0.00**	**20.02±0.00**
		BTAD	Base	85.72±0.00	93.80±0.00	65.41±0.00	41.15±0.00
			OURS	**85.81±0.00**	**94.09±0.00**	**65.81±0.00**	**41.58±0.00**
	ViT-L/14-336	MVTec	Base	93.48±0.00	90.93±0.00	83.54±0.00	38.30±0.00
			OURS	**93.69±0.00**	**91.15±0.00**	**84.03±0.00**	**39.25±0.00**
		VisA	Base	**84.98**±0.00	95.67±0.00	86.74±0.00	26.20 ±0.00
			OURS	84.57 ±0.00	**95.78**±0.00	**86.97** ±0.00	**26.67**±0.00
		MPDD	Base	73.57±0.00	95.94±0.00	88.47 ±0.00	25.35±0.00
			OURS	**73.61**±0.00	**95.98** ±0.00	**88.52**±0.00	**25.49**±0.00
		BTAD	Base	**91.04**±0.00	93.84±0.00	70.22 ±0.00	41.79±0.00
			OURS	90.78±0.00	**94.31**±0.00	**70.66**±0.00	**42.40**±0.00
	ViT-H/14	MVTec	Base	69.23±4.92	72.62±5.89	49.30±8.74	15.64±4.24
			OURS	**69.24±4.99**	**72.70±6.20**	**49.34±9.14**	**16.14±4.50**
		VisA	Base	**63.39±2.49**	**78.98±7.84**	**55.02±9.92**	14.92±2.67
			OURS	63.26±2.57	78.70±8.20	54.75±9.98	**15.64±2.68**
		MPDD	Base	**55.43±5.11**	**79.76±7.72**	47.91±15.70	9.55±3.77
			OURS	55.32±4.92	79.64±7.97	**48.19±16.07**	**9.70±3.79**
		BTAD	Base	**55.73±8.33**	72.71±3.65	35.63±6.81	11.26±6.63
			OURS	55.46±7.92	**73.28±3.82**	**36.06±6.86**	**11.57±6.83**

the practical scope of C-TED rather than serving as a primary positive example. We therefore use BayesPFL to clarify scope rather than to overstate uniform improvement.

D.2 Full Recipe-Retuning Results

Table 22 reports the full bidirectional recipe-retuning diagnostic. Each cell reports I-AUC / P-AUC / PRO / AP. The table complements the main-paper recipe summary by showing that C-TED can remain useful after the host is retrained under changed layer recipes and source banks are rebuilt for the resulting representation. This setting is stricter than a fixed-host swap because the host is allowed to adapt to the new recipe before C-TED is applied. Therefore, the remaining gains are less likely to be explained solely by a mismatched feature hook or an unfavorable layer choice. For FAPrompt, C-TED improves AP across all listed layer recipes in both directions. For example, in the VisA→MVTec direction, Full(B) AP improves from 22.35 to 30.94. In the MVTec→VisA direction, Full(B) AP improves from 13.61 to 19.23. For AdaCLIP, several recipes show clear PRO gains: VisA→MVTec L-r improves PRO from 34.83 to 46.27, and MVTec→VisA Full(B) improves PRO from 54.13 to 65.18. The gains appear in different metrics for different hosts: FAPrompt shows particularly clear AP improvements, while AdaCLIP often benefits more in PRO. This suggests that

Table 20: Results with **AdaCLIP** as host.

Host	Backbone	Target	Method	I-AUROC	P-AUROC	P-PRO	P-AP
AdaCLIP	ViT-B/16+	MVTec	Base	48.53±7.16	**50.32±6.37**	19.74±5.72	4.27±0.97
			OURS	**49.67±8.24**	50.26±6.45	**20.81±5.70**	**4.46±1.17**
		VisA	Base	49.19±1.25	57.15±7.57	25.60±6.46	1.16±0.35
			OURS	**50.83±1.63**	**58.70±6.52**	**28.22±5.92**	**1.32±0.39**
		MPDD	Base	**43.82±2.60**	42.46±10.95	14.05±6.18	**2.82±0.87**
			OURS	42.64±0.32	**42.57±9.39**	**15.00±5.11**	2.61±0.71
		BTAD	Base	64.80±9.50	51.09±4.34	14.84±3.01	3.58±0.78
			OURS	**66.98±10.31**	**52.24±4.57**	**15.67±2.84**	**3.68±0.64**
	ViT-L/14 OpenAI	MVTec	Base	84.66	89.54	44.32	39.36
			OURS	**87.89**	**89.55**	**49.01**	**39.51**
		VisA	Base	79.47±0.13	**95.57±0.00**	54.38±0.00	**26.45±0.00**
			OURS	**80.70±0.06**	95.49±0.00	**66.02±0.55**	25.63±0.01
		MPDD	Base	67.38±0.11	95.91±0.00	31.36±0.00	**25.44±0.00**
			OURS	**67.67±0.22**	**95.98±0.00**	**46.51±0.49**	25.26±0.00
		BTAD	Base	**87.14±0.14**	**91.88±0.00**	9.70±0.00	**42.06±0.00**
			OURS	86.19±0.07	91.51±0.04	**14.04±0.31**	40.51±0.04
	ViT-L/14-336	MVTec	Base	89.93±0.06	**89.89±0.00**	44.03±0.00	41.60±0.00
			OURS	**91.63±0.01**	89.85±0.00	**47.86±0.08**	**41.84±0.01**
		VisA	Base	85.99±0.25	**95.82±0.00**	54.13±0.00	**31.48±0.00**
			OURS	**86.38±0.02**	95.72±0.00	**65.61±0.84**	31.31±0.01
		MPDD	Base	68.96±0.23	96.09±0.00	30.48±0.00	29.76±0.00
			OURS	**69.41±0.21**	**96.33±0.00**	**44.55±0.83**	**29.92±0.01**
		BTAD	Base	**90.41±0.25**	**93.66±0.00**	19.26±0.00	**48.20±0.00**
			OURS	90.04±0.20	93.49±0.03	**27.81±0.67**	46.85±0.02
	ViT-H/14	MVTec	Base	47.24±0.48	**51.14±3.01**	**19.80±3.19**	**4.76±0.33**
			OURS	**47.46±0.64**	49.25±1.91	18.97±1.15	4.74±0.04
		VisA	Base	46.32±4.34	53.92±2.35	21.38±4.10	1.16±0.28
			OURS	**48.18±3.71**	**55.24±2.83**	**24.41±4.36**	**1.31±0.26**
		MPDD	Base	38.37±5.28	49.09±9.25	17.95±8.72	2.73±1.57
			OURS	**40.06±3.87**	**52.16±8.18**	**20.96±6.24**	**2.93±1.44**
		BTAD	Base	62.93±13.69	57.21±5.35	23.13±3.62	4.29±1.19
			OURS	**64.86±14.71**	**57.59±4.83**	**25.63±4.88**	**4.71±1.18**

C-TED does not simply optimize one metric-specific artifact, but can improve different aspects of local ranking depending on the host. At the same time, the effect is not uniform across all cells, so recipe retuning remains an important factor in the host's baseline behavior. These results support the main claim that representation selection and evidence decoding are distinct. Changing or retraining the layer recipe can shift host responses, but it does not necessarily remove hard-FP ranking errors. C-TED can still improve local ranking after recipe retuning, although the magnitude remains host-, recipe-, and metric-dependent.

E Limitations and Broader Impacts

Limitations. TED is designed for local anomaly maps, not universal image-level screening. The method preserves the host detector and modifies only the local evidence readout; therefore, image-level performance can still depend on the host's global branch, pooling rule, or score aggregation strategy. Although we report I-AUROC for transparency, the primary claim of TED is pixel-level localization under source-to-target transfer.

Table 21: Results with **BayesPFL** as host using calibrated TED.

Host	Backbone	Target	Method	I-AUROC	P-AUROC	P-PRO	P-AP
BayesPFL	ViT-B/16+	MVTec	Base	49.68±1.01	**49.73±3.18**	**17.14±3.94**	**4.29±0.80**
			OURS	**52.49±2.56**	49.67±3.19	17.00±3.86	4.26±0.78
		VisA	Base	**51.41±1.85**	**54.71±3.85**	**18.93±5.43**	**0.90±0.26**
			OURS	50.20±2.31	54.60±3.67	18.87±5.32	0.88±0.24
		MPDD	Base	50.35±4.02	**58.47±9.08**	**22.52±9.48**	**2.39±0.29**
			OURS	**51.26±4.88**	58.27±8.73	22.43±9.22	2.38±0.27
		BTAD	Base	**51.46±5.29**	**43.32±0.34**	**9.79±0.76**	2.86±0.78
			OURS	49.85±7.65	43.29±0.31	9.75±0.74	**2.87±0.79**
	ViT-L/14 OpenAI	MVTec	Base	90.42±0.16	97.28±0.03	**90.20±0.05**	**57.51±0.11**
			OURS	**94.82±0.08**	**97.28±0.03**	90.14±0.05	57.16±0.08
		VisA	Base	75.89±0.62	**94.53±0.08**	**82.14±0.07**	**18.99±0.24**
			OURS	**75.90±0.28**	94.51±0.08	82.06±0.06	18.80±0.22
		MPDD	Base	70.07±0.23	**95.73±0.06**	86.24±0.14	**24.92±0.32**
			OURS	**70.65±0.91**	95.72±0.06	**86.26±0.21**	24.81±0.31
		BTAD	Base	73.71±2.08	**87.94±0.27**	64.28±0.31	**35.98±0.28**
			OURS	**83.90±0.80**	87.89±0.28	**64.33±0.30**	35.62±0.28
	ViT-L/14-336	MVTec	Base	92.69±0.32	97.35±0.06	92.34±0.01	64.22±0.28
			OURS	**96.40±0.11**	**97.36±0.06**	**92.35±0.02**	**64.24±0.27**
		VisA	Base	79.51±0.24	95.51±0.02	89.25±0.06	**25.30±0.17**
			OURS	**80.74±0.51**	**95.52±0.02**	**89.25±0.06**	25.19±0.18
		MPDD	Base	**78.32±1.11**	**96.89±0.02**	**91.27±0.09**	**29.99±0.18**
			OURS	74.51±0.30	96.88±0.02	91.25±0.16	29.97±0.19
		BTAD	Base	77.89±0.88	86.98±0.24	**71.61±0.15**	**35.05±0.61**
			OURS	**89.92±1.14**	**86.98±0.26**	71.56±0.10	34.91±0.65
	ViT-H/14	MVTec	Base	46.82±4.08	51.74±9.31	17.43±7.07	3.86±1.50
			OURS	**53.02±0.59**	**51.83±9.36**	**17.55±7.09**	**3.87±1.52**
		VisA	Base	49.62±1.15	47.76±11.74	13.72±6.40	0.73±0.38
			OURS	**50.19±0.90**	**47.80±11.81**	**13.77±6.46**	**0.74±0.39**
		MPDD	Base	**49.60±2.81**	42.50±7.10	10.58±4.90	2.04±0.50
			OURS	49.51±7.04	**42.56±7.26**	**10.66±4.97**	**2.04±0.50**
		BTAD	Base	47.20±6.28	48.56±4.20	**12.29±5.17**	**2.84±0.19**
			OURS	**50.70±4.26**	**48.59±4.17**	12.22±4.90	2.84±0.19

TED also depends on source defect and hard-FP banks. Its correction quality can depend on whether the source split contains defect evidence and hard-normal patterns that are relevant to the target domain. If the target domain contains defect types or normal structures not represented in the source banks, the support comparison may become less reliable. Similarly, if the source anomaly masks are noisy or the mined hard-FP patches do not reflect the dominant target false positives, the residual correction may provide limited benefit. Our fixed-bank and source-supervision controls reduce some confounding explanations, but they do not eliminate the need for representative source evidence.

The method is also evaluated within the VLM backbones and CLIP-AD hosts considered in this work. TED uses a weaker interface than many prompt-learning pipelines, requiring patch-level visual features and a normal-versus-anomaly text response, but this should not be read as universal backbone-agnostic anomaly detection. Backbones without meaningful patch-level features, unstable text responses, or incompatible feature scales may require additional validation. Finally, the source-calibrated residual introduces extra source-side calibration and bank storage, even though the retained-bank sensitivity experiment suggests that moderate bank sizes are sufficient in our tested settings.

Table 22: Bidirectional layer-recipe retraining diagnostic. Each cell reports I-AUC / P-AUC / PRO / AP. B, T, and C denote the retrained host baseline, train-free TED, and source-calibrated TED, respectively. Full(B) is the official full-layer baseline recipe; EL-r, M2-r, L2-r, and L-r denote checkpoints retrained with early+last, middle-2, last-2, and last-layer recipes.

Transfer	Host	Recipe	B	T	C
VisA→MVTec	AdaCLIP	Full(B)	89.93 / **89.89** / 44.03 / 41.60	91.43 / 89.88 / 43.60 / 41.46	**91.64** / 89.86 / **47.97** / **41.85**
		EL-r	87.83 / 80.21 / 18.18 / 22.40	88.74 / 80.21 / 17.94 / 22.39	**88.84** / **80.96** / **26.25** / **23.81**
		M2-r	**89.26** / 87.62 / 23.50 / 36.18	87.74 / 87.61 / 23.63 / 36.12	87.74 / **87.79** / **25.42** / **36.40**
		L2-r	**90.58** / 89.47 / 15.19 / 36.35	89.07 / 89.47 / 15.14 / 36.31	88.69 / **90.84** / **22.78** / **39.28**
		L-r	90.40 / 88.96 / 34.83 / 35.78	**91.99** / 88.96 / 34.66 / 35.77	91.79 / **89.15** / **46.27** / **36.77**
	FAPrompt	Full(B)	**91.45** / 89.95 / 73.57 / 22.35	**91.45** / 40.41 / 25.32 / 8.16	**91.45** / **90.19** / **76.83** / **30.94**
		EL-r	**91.40** / 90.49 / 74.70 / 23.58	**91.40** / 42.46 / 26.23 / 9.25	**91.40** / **90.78** / **77.16** / **32.40**
		M2-r	**92.19** / 90.05 / 74.35 / 22.58	**92.19** / 41.66 / 25.87 / 8.79	**92.19** / **90.52** / **76.54** / **31.46**
		L2-r	**92.61** / 90.08 / 72.73 / 22.33	**92.61** / 45.33 / 30.26 / 10.78	**92.61** / **90.61** / **76.63** / **31.07**
		L-r	**90.93** / 89.63 / 74.01 / 23.01	**90.93** / 44.35 / 29.27 / 10.40	**90.93** / **89.93** / **76.08** / **31.68**
MVTec→VisA	AdaCLIP	Full(B)	85.65 / **95.82** / 54.13 / 31.48	**87.00** / **95.82** / 54.43 / **31.54**	86.36 / 95.72 / **65.18** / 31.32
		EL-r	82.05 / 90.56 / 57.48 / 20.45	**83.10** / **91.38** / **59.64** / **20.79**	82.92 / 91.33 / 57.65 / 20.63
		M2-r	78.89 / **95.17** / 58.12 / 29.03	81.56 / **95.17** / 58.10 / 29.03	**81.90** / 95.16 / **60.36** / **29.11**
		L2-r	73.87 / 92.36 / 63.13 / 22.25	**78.12** / 92.36 / 63.16 / 22.26	77.80 / **93.76** / **68.26** / **22.61**
		L-r	81.47 / 92.37 / 50.61 / 21.52	**83.08** / 92.37 / 50.44 / **21.56**	81.56 / **93.32** / **58.91** / 21.05
	FAPrompt	Full(B)	**82.23** / 95.33 / 79.34 / 13.61	**82.23** / 93.46 / 78.07 / 13.52	**82.23** / **95.60** / **82.08** / **19.23**
		EL-r	**83.25** / 95.36 / 79.45 / 13.15	**83.25** / 92.14 / 78.08 / 16.06	**83.25** / **95.59** / **80.62** / **18.45**
		M2-r	**84.34** / 95.36 / 78.13 / 13.37	**84.34** / 90.41 / 75.46 / 15.77	**84.34** / **95.64** / **81.54** / **18.57**
		L2-r	**84.40** / 95.07 / 78.82 / 13.10	**84.40** / 93.22 / 78.98 / 14.72	**84.40** / **95.39** / **82.87** / **18.66**
		L-r	**82.30** / 95.32 / 77.29 / 13.46	**82.30** / 92.68 / 79.20 / 15.81	**82.30** / **95.53** / **81.29** / **18.71**

Future work should study source-bank construction, image-level aggregation, broader host families, failure cases, and deployment-time monitoring under larger distribution shifts.

Broader impacts. TED is intended for industrial anomaly localization and may reduce false-positive burden by improving local defect ranking under domain shift. In practice, it can help human inspectors focus on suspicious regions and provide more interpretable local evidence, especially when retraining large host models is costly. At the same time, over-reliance on automated localization can be harmful: missed defects may affect product quality or safety, and false alarms may increase inspection cost or reduce operator trust. These risks are higher when source defect and hard-FP banks do not cover the target inspection conditions, including new materials, lighting, imaging artifacts, or defect types. TED should therefore be used as decision support with human oversight, periodic target-domain validation, and monitoring of false positives and missed defects. The work uses public industrial inspection datasets and does not involve personal, biometric, medical, or otherwise sensitive data; the main concern is reliability in quality- or safety-critical inspection workflows.

F Additional Analyses and Implementation Details

This appendix consolidates additional analyses conducted during the review process. It supplements, rather than repeats, the same-bank retrieval comparison provided elsewhere in the appendix. We examine the information retained by the text axis, source-bank transferability and construction, source-only calibration, applicability limits, and computational cost.

Reporting conventions. Unless explicitly identified as a diagnostic, each evaluation includes all categories and test images in its specified evaluation set, without target calibration or outcome-based filtering. Changes in P-AUC, P-PRO, P-AP, and I-AUC are absolute percentage points relative to the corresponding host. Diagnostic role AUC, per-image pixel-AUC, and dataset-level P-AUC are different quantities and are not pooled. Random-seed results are averaged within each host–transfer unit where specified. Host-native experiments, shared-recipe experiments, and shared-interface controls are reported separately.

F.1 Conditional Information Retained by the Text Axis

Protocol. We evaluate the complete three-frozen-backbone $\times$ four-transfer grid with the source banks fixed. Target Hard-FPs are the host baseline's top-1% responses on target-normal images. Target annotations are used only for this diagnostic, not to construct banks, fit a calibrator, or perform inference. In addition to category-macro Defect-vs-HFP role AUC, we condition comparisons on Full-D support in two ways: (i) 20 Full-D-score strata and (ii) non-repeating matched pairs whose Full-D scores differ by at most 1% of the category-specific interquartile range. The matched analysis contains 662,708 pairs across 102 matchable category units out of 108; unmatched units are not silently assigned scores.

Table 23: Conditional text-axis audit. AUC and accuracy are reported in percent. All-patch role AUC is a diagnostic, not dataset-level P-AUC.

Evidence readout	All-patch role AUC	Within Full-D strata	Matched-pair accuracy
Full-D evidence	96.08	70.69	49.52
Text-axis evidence	99.55	98.42	98.13

Full-D matched-pair accuracy is close to chance, whereas text-axis evidence retains discriminative information under nearly equal Full-D support (Table 23). Text-axis evidence is higher in all 12 evaluated settings. These are conditional results within this audit, not claims of an equivalent improvement in localization metrics.

Axis and role controls. Source-mean, PCA, and five-seed random axes yield role AUCs of 96.09, 81.45, and 63.61, respectively. With the other components fixed, correct, random, and swapped evidence roles yield 99.55, 47.79 ± 18.04, and 0.45, respectively. Together, these controls support the importance of a task-aligned coordinate and correct evidence roles. They do not imply that a scalar preserves arbitrary visual structure or topology.

F.2 Source-Bank Transferability and Mining Sensitivity

Reduced source-category coverage. We use AA-CLIP, AdaCLIP, and FAPrompt on MVTec AD$\rightarrow$ {BTAD, MPDD, VisA} and VisA$\rightarrow$MVTec AD, giving 12 host–transfer units. The full-category run and three random seeds at each reduced coverage level give 84 pre-specified runs. Removed source categories are not replaced. Results for random subsets are seed-averaged within each host–transfer unit.

Table 24: Reduced source-category coverage. Wins count positive changes among the 12 host–transfer units after seed averaging.

Source categories	ΔP-AUC	Wins	ΔP-PRO	Wins	ΔP-AP	Wins
All	+0.99	8/12	+6.20	12/12	+5.37	10/12
Random 50% (3 seeds)	+0.92	9/12	+5.77	12/12	+5.16	10/12
Random 25% (3 seeds)	+0.91	8/12	+5.29	12/12	+5.19	10/12

The gains remain positive on average under the tested reductions, although individual metrics do not improve in every unit. Texture-only MVTec AD evidence transferred to all VisA PCB categories gives $+0.02/+4.84/+1.82$ in P-AUC/P-PRO/P-AP. These tests reduce source coverage and introduce a material mismatch, but do not establish robustness to every unseen defect type or acquisition condition.

Distribution shift and host-interface compatibility. In a separate shared-interface audit, all 15 evaluated source–target distributions remain statistically distinguishable. Mean normalized Wasserstein-1 distances are 0.57 for defect scores and 0.91 for HFP scores. Despite this mismatch, aggregate changes are $-0.17/+0.79/+4.93$. Table 25 retains the unfavorable AdaptCLIP result. These results indicate that identical absolute score distributions are not necessary in the tested settings, but they do not guarantee transferability.

Table 25: Separate shared-interface distribution-shift audit. These controls must not be substituted for host-native results.

Host	ΔP-AUC	ΔP-PRO	ΔP-AP
AA-CLIP	+0.65	+2.45	+3.72
FAPrompt	+2.22	+3.41	+12.07
AdaptCLIP	-3.36	-3.49	-1.00
Overall	-0.17	+0.79	+4.93

The standardized AdaptCLIP control differs from its submitted host-native protocol, which gives $+0.08/+0.07/+0.34$ over 16 settings. Thus, compatibility between the correction and the host's scoring interface is an applicability condition, not an implementation detail that can be ignored.

Hard-FP mining fraction. We vary the source-normal top fraction without retuning on the three-host $\times$ four-transfer grid. All four fractions improve P-PRO in all 12 units (Table 26). The coverage audit and mining-fraction audit are separate reported experiment sets; their full/default rows are retained as reported rather than merged into a single baseline.

Table 26: Sensitivity to the source Hard-FP mining fraction without retuning.

HFP top fraction	ΔP-AUC	ΔP-PRO	ΔP-AP
0.5%	+0.96	+4.61	+5.40
1% (default)	+0.99	+6.24	+5.38
2%	+0.98	+5.95	+5.38
5%	+0.99	+5.83	+5.38

Bank diversity. Across AA-CLIP and FAPrompt on all MVTec AD-source transfers, farthest-point sampling (FPS) increases HFP/defect nearest-neighbor diversity from $0.0573/0.0753$ to $0.0766/0.1016$. Localization changes are nevertheless similar: FPS gives $+1.82/+5.15/+8.37$, while three-seed random banks give $+1.78/+5.19/+8.30$. This supports stability to the tested bank sampling choices, not independence from source-bank quality in general.

F.3 Source-Bank Construction and Budget Units

All bank-selection choices are source-only and fixed before target evaluation. Table 27 clarifies image caps, bank caps, and realized MVTec AD-source counts. Defect and HFP banks need not have identical sizes.

Table 27: Host-specific source-bank construction. Retained counts sum four layers for AA-CLIP and AdaCLIP; FAPrompt counts refer to its pooled bank.

Host	Images/category	Cap per evidence bank	HFP/defect retained
AA-CLIP	All	≤ 512/category/layer	27,096/14,480
AdaCLIP	≤ 2 normal, ≤ 4 anomalous	≤ 512/layer	1,560/2,048
FAPrompt	All	≤ 512/category; $\leq 4,096$ pooled	4,096/4,096

AA-CLIP mines top-1% source-normal HFPs and defect-overlapping top-25% patches. AdaCLIP uses top-1% HFPs and at most 64 defect patches per image. FAPrompt uses top-1% fused-score HFPs and defect-overlapping patches. FPS follows candidate mining. AA-CLIP and FAPrompt apply category caps before pooling; AdaCLIP applies image caps before constructing layer-wise banks. These caps limit category contributions but do not force equal retained counts: exhausted groups are neither duplicated nor artificially balanced.

The AdaptCLIP and BayesPFL compatibility variants use, respectively, up to 2/4 and 8/8 normal/anomalous images per category, followed by caps of 512 per branch and 1,024 per layer. These variants are distinct from the three primary host implementations.

For the submitted bank-budget sensitivity experiment, AA-CLIP budgets 64/128/512 are per category, per layer, and per bank. FAPrompt budgets are per category and per bank before a pooled $2B$ cap, where B is that experiment's bank budget; reported "Retained" counts sum both banks. This sweep-specific budget convention should not be confused with the realized counts in Table 27 or with the training-point budgets below.

F.4 C-TED Calibration and Training Controls

Source subspace and objective. The bounded correction is implemented using a source rank-r basis formed by orthonormalizing the text axis, a source defect-minus-HFP mean direction, and leading principal components of centered source pair differences. The backbone, prompts, and host parameters remain frozen. For source pair i, let $m_{0,i}$ be its original host margin and $m_i = s_{\mathrm{def},i} - s_{\mathrm{fp},i}$ its corrected margin. We optimize

$$\mathcal{L}_{\mathrm{rank}} = \frac{1}{M}\sum_{i=1}^{M} \mathrm{softplus}\left(\frac{\mu - m_i}{T}\right), \tag{7}$$

$$w_i = \sigma\left(\frac{m_{0,i} - \mu}{T}\right), \qquad \mathcal{L}_{\mathrm{pre}} = \frac{\sum_i w_i [m_{0,i} - m_i]_+}{\max(\sum_i w_i, 10^{-6})}, \tag{8}$$

$$\mathcal{L} = \mathcal{L}_{\mathrm{rank}} + \beta \mathcal{L}_{\mathrm{pre}} + \rho(\eta^2 + \gamma^2). \tag{9}$$

Here $[x]_+ = \max(x, 0)$, σ is the sigmoid, $\mu = T = 0.05$, $\beta = 1$, and $\rho = 0.001$. The denominator clamp makes explicit the numerical safeguard in the implementation. The parameters η and γ control transport magnitude and residual gain, respectively.

Pair construction and recipes. Each calibration unit forms at most half as many pairs as its sampled source point budget. The mixture is 75% hard and 25% random, with HFP assignments permuted each epoch. Hard examples prioritize high-scoring source HFPs and low-scoring source defects. Calibration is source-only. Table 28 gives the submitted host-specific recipes; the point budget is not a retained-bank size.

Table 28: Submitted target-independent C-TED recipes for the three primary hosts.

Host	Rank	Epochs	Maximum source points/unit
AA-CLIP	8	20	8,192
AdaCLIP	8	30	2,048
FAPrompt	4	30	2,048

Shared-recipe control. To assess whether different rank, epoch, and budget choices explain the gains, we rerun all 18 cells in the three-host $\times$ six-transfer intersection using rank 4, 30 epochs, at most 1,024 source points/512 pairs, and 512 entries per bank per unit. This control uses Adam with learning rate 0.02 and $\eta_{\max} = \gamma_{\max} = 0.2$; only the insertion point follows the host's score interface. These shared-control settings are not substituted for the submitted host-native recipes.

Table 29: Shared-recipe calibration over the complete 18-cell grid.

Host	ΔP-AUC	ΔP-PRO	ΔP-AP
AA-CLIP (six transfers)	+0.30	+2.47	+1.98
AdaCLIP (six transfers)	-0.14	+4.33	-0.39
FAPrompt (six transfers)	+2.20	+5.21	+12.34
Overall (18 transfers)	+0.79	+4.00	+4.64

The shared recipe improves all three metrics on average and P-PRO in 15 of 18 settings. AdaCLIP's P-AUC/P-AP trade-off is retained without retuning.

Training components and correction strength. Table 30 reports the three primary hosts on MVTec AD$\rightarrow$ {BTAD, MPDD}. Removing preservation has little effect on these aggregate results, and

random-pair training slightly reduces P-PRO. Thus these particular controls do not establish a large independent benefit from preservation. Increasing correction strength by $4\times$ reduces P-AUC and P-AP below the host, supporting conservative correction strength. The maximum reported standard deviation across the three-seed BTAD checks is 0.82 percentage points.

Table 30: Source-training and correction-strength controls.

Variant	ΔP-AUC	ΔP-PRO	ΔP-AP
Full training	+1.42	+7.69	+6.24
Without preservation	+1.42	+7.65	+6.23
Random-pair training	+1.42	+7.49	+6.23
$4\times$ correction strength	-2.62	+6.55	-3.13

F.5 Host Variation and Applicability Limits

Complete-table aggregation. Table 31 aggregates all 76 settings in the complete adapted-host result tables. It is distinct from the 70-setting diagnostic in the submitted analysis. The overall row averages settings, not the five host means. Across the 60 non-BayesPFL settings, changes are $+2.86/+4.95/+3.01$, with improvements in 45/51/48 settings for P-AUC/P-PRO/P-AP, respectively.

Table 31: Host-native changes across all 76 complete-table settings, including limited and negative results.

Host	ΔP-AUC	ΔP-PRO	ΔP-AP
AA-CLIP	+9.07	+6.24	+1.07
AdaCLIP	+0.38	+5.47	-0.14
AdaptCLIP	+0.08	+0.07	+0.34
FAPrompt	+3.47	+8.33	+10.28
BayesPFL	-0.02	-0.02	-0.08
Overall	+2.25	+3.90	+2.36

BayesPFL as a boundary case. BayesPFL combines Bayesian prompt ensembles with RCA dense alignment before TED. Its default ViT-L/14-336 configuration averages P-AUC/P-PRO of 94.18/86.12 over four targets. The all-image, all-target audit uses no target-specific gate, tuning, or outcome-selected subset. On MVTec AD, baseline Defect-vs-HFP ordering remains positive (role AUC 78.1%, score gap $+0.174$), and pixel-metric changes are $+0.011/+0.010/+0.018$. On shifted targets the role ordering is not preserved, but the largest absolute pixel-metric change is 0.14 points. These observations characterize a near-no-op in this audit, not a guarantee that a bounded residual cannot hurt performance. All 16 BayesPFL settings are retained in the complete results.

Small and rare defects. We audit the exact AA-CLIP/AdaCLIP/FAPrompt image intersection using per-image pixel-AUC, without outcome-based selection. "Small" denotes the bottom mask-area quartile per transfer. "Rare" denotes the least-frequency quartile of MPDD anomaly subtypes, retaining boundary ties. These definitions use target annotations for analysis only, not for calibration or inference.

Table 32: Small/rare-defect audit. Evaluations count host–image applications; macro changes and positive-unit counts equally weight host–transfer units.

Scope	Images	Evals.	Macro Δpixel-AUC	Positive units	Crossings/ eligible
All anomalies	3,014	9,042	+1.35	9/12	32/8,948
Smallest-area quartile	755	2,265	+0.27	8/12	22/2,233
Rare MPDD subtypes	53	159	+1.99	3/3	0/159

A crossing means that per-image pixel-AUC falls from ≥ 0.5 to < 0.5; eligible evaluations therefore have baseline pixel-AUC of at least 0.5. The rare MPDD subset improves in all three host units

with no such crossing. This does not exclude other forms of degradation or establish safety for all rare defects. The smallest-area quartile has the smallest average gain, and 22 crossings remain. Its most limited unit has mean mask area 0.092%, below one visual token in the evaluated configuration. Sub-token resolution and defect rarity should therefore be distinguished.

F.6 Runtime, Memory, and Scaling

End-to-end paired profiles. We profile the host and host+TED with GPU synchronization on one RTX 6000 Ada, using all 733 BTAD images after eight warm-up iterations per host. Table 33 reports wall-clock time per image and added GPU memory. AA-CLIP's relative overhead is larger because its baseline is faster; the absolute overhead is not negligible simply because the module is post-hoc.

Table 33: Paired host and host+TED profiles on BTAD. Times are ms/image.

Host	Host	Host+TED	Added ms	Added %	Added MB
AA-CLIP	15.93	21.90	5.97	37.47	19.46
FAPrompt	136.50	155.91	19.41	14.22	28.07
AdaCLIP	183.26	190.10	6.84	3.73	8.25

Module scaling. The separate module-level profile fixes 1,369 patches and four layers while varying the bank budget (Table 34). These module timings should not be substituted for the end-to-end overhead above. At 512 entries, the axis cache, full feature bank, and rank-8 basis require 0.016, 12, and 0.094 MB, respectively, in the reported profiling configuration.

Table 34: Module-level bank-size scaling with 1,369 patches and four layers.

Bank entries	T-TED (ms)	C-TED (ms)
128	0.98	1.72
512	0.99	1.70
2,048	1.76	2.49
8,192	6.23	7.11

Let L denote the number of selected layers, N the number of query patches, d the feature dimension, B the bank budget, and r the residual rank. With cached bank-axis projections, the reported complexity is $O(LN(d+B))$ for T-TED and $O(LN(d+B+dr))$ for C-TED. A changing text axis adds $O(LBd)$ for bank reprojection. The budget B must reflect the actual retained evidence used by the host; category-wise banks can increase this budget before pooling caps are applied. These profiles do not measure simultaneous execution of multiple hosts.

F.7 Pixel-Level Versus Image-Level Scope

Across the 76 complete-table settings, I-AUC changes by $+1.25$ points on average and improves in 41 settings. P-PRO and P-AP improve jointly in 48 settings; among these, 23 do not improve I-AUC. Correlations of I-AUC gains with P-PRO and P-AP gains are 0.029 and -0.006, respectively. Thus improved localization should not be interpreted as universal improvement in image-level screening. TED operates on local evidence; whether a local correction changes image ranking depends on the host's image-score aggregation and on whether a separate global branch is retained.

Qualitative evidence and selection. The selected adapted-host gallery already included in the appendix illustrates local defect highlighting and suppression of competing normal responses. Host and C-TED maps share each example's pooled 2nd–99.5th percentile color range. These selected examples are not an unbiased estimate of improvement frequency, and shared color limits do not themselves establish preservation of all defects. Quantitative decreases, near-neutral behavior, and small-defect limitations are retained in the analyses above.