CoL-SAM : Fine-Grained Zero-Shot Anomaly Segmentation via Collaborative CLIP–SAM

Yuhang Xie[1], Xiaofei Ma[1], Lin Wang[1,2†], Bo Yang[2,1]
[1]Shandong Key Laboratory of Ubiquitous Intelligent Computing, University of Jinan, Jinan 250022, China
[2]Quan Cheng Laboratory, Jinan 250100, China
[†]Corresponding author: wangplanet@gmail.com

***Abstract*—Zero-shot anomaly segmentation (ZSAS) aims to localize defect regions on unseen industrial products without target-category training data, reference samples, or dense pixel-level annotations. Recent foundation models such as CLIP and SAM enable training-free transfer, yet their direct use remains challenging: CLIP often produces spatially coarse and noisy anomaly cues due to its global contrastive alignment, while SAM may generate redundant or over-segmented masks when driven by unreliable prompts, leading to non-trivial post-processing. In this work, we propose CoL-SAM, a collaborative CLIP-SAM framework for fine-grained ZSAS. First, we follow a WinCLIP-style multi-scale window scoring scheme to extract anomaly evidence and introduce an Adaptive Score Fusion (ASF) module to replace the fixed, manually designed scale fusion used in WinCLIP. ASF performs fusion with spatially varying weights over multi-scale anomaly maps, producing sharper anomaly guidance with reduced noise propagation. Second, we transform the fused anomaly guidance into compact prompt constraints that steer SAM toward boundary-accurate defect masks while suppressing redundant proposals. Experiments on MVTec-AD and VisA demonstrate that CoL-SAM consistently outperforms CLIP-only, SAM-only, and WinCLIP-based baselines, validating the effectiveness of adaptive multi-scale fusion and prompt-constrained refinement for zero-shot industrial anomaly segmentation.**

***Index Terms*—zero-shot, anomaly segmentation, CLIP, SAM**

I. Introduction

In industrial inspection, accurate defect masks are key to reliable quality control, but defective samples are scarce and pixel-level labeling is costly. Zero-shot anomaly segmentation (ZSAS) aims to highlight defective regions on unseen products without collecting target-category training images or manual pixel annotations, leveraging transferable priors from large pretrained models [1]–[4]. Compared with category-specific pipelines that require per-class modeling or calibration [5], [6], ZSAS demands stronger generalization to large shifts in textures, defect appearance, and background clutter across diverse industrial domains [7], [8].

Recent advances in ZSAS are largely driven by vision foundation models [9], especially Contrastive Language-Image Pretraining (CLIP) [10] and Segment Anything Model (SAM) [11], which enable zero-shot perception through prompt-based knowledge transfer. Figure 1 summarizes three paradigms. CLIP-based methods [1], [3], [12], [13], they compute patch-wise anomaly scores from image–text similarity, but the resulting heatmaps are often spatially coarse. CLIP is trained for image-level image–text matching, so its features favor global semantics over pixel-accurate boundaries. As a result, CLIP-only localization may miss thin defects or blur object edges in fine-grained segmentation [14]. SAM-based pipelines (e.g., SAA and SAA+ [15]) typically rely on text grounding to generate box or point prompts for SAM, followed by heuristic post-processing. However, when defect descriptions are vague or inconsistent, text prompts may introduce numerous irrelevant proposals, making such post-processing difficult to avoid. Within ZSAS, CLIP is effective at identifying potential anomalous regions, whereas SAM excels at refining anomaly boundaries, provided that the prompts are reliable. Recent CLIP–SAM collaboration frameworks (e.g., ClipSAM [16]) use CLIP to generate coarse anomaly cues and convert them into prompt constraints for SAM, yielding cleaner boundaries with fewer redundant masks. We first use CLIP to produce anomaly localization guidance and then transform these cues into prompt constraints to guide SAM for boundary refinement, illustrating a promising direction for accurate ZSAS. Beyond industrial inspection, CLIP–SAM coupling has also been explored in medical text-driven segmentation (MedCLIP-SAMv2 [17]) and in unified model merging (SAM-CLIP [18]), suggesting that combining CLIP's semantics with SAM's spatial priors can generalize across domains.

In this paper, We propose CoLSAM (Collaborative CLIP–SAM), a framework that stabilizes CLIP–SAM collaboration by converting multi-scale CLIP anomaly evidence into prompt-compatible priors for SAM. Many CLIP–SAM pipelines binarize CLIP heatmaps into points/boxes and then rely on SAM for mask refinement. When the CLIP signal is weak or becomes sparse after hard thresholding, SAM may be driven by noisy prompts, producing missed detections or

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61872419, 62072213, and 62403209; in part by the Shandong Provincial Natural Science Foundation under Grant Nos. ZR2022JQ30, ZR2022ZD01, ZR2023LZH015, and ZR2024QF021; in part by the Shandong Provincial Key R&D Program Project under Grant Nos. 2024CXPT084 and 2025CXPT105; in part by the Key Research Project of Quan Cheng Laboratory, China under Grant Nos. QCL20250105 and QCL20250304; in part by the Open Fund of the State Key Laboratory of Silicate Materials for Architectures under Grant No. SYSJJ2025-02; in part by the New 20 Measures for Colleges and Universities of Jinan under Grant No. 202534127; in part by the Youth Innovation Team Program of Shandong Higher Education Institutions, China under Grant No. 2024KJH104; and in part by the University of Jinan Young Faculty Interdisciplinary Convergence Development Project 2025 under Grant No. XKJC-202507.

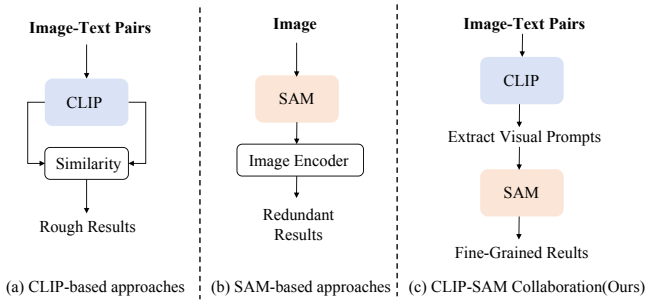

Fig. 1. Structural comparisons among different approaches for ZSAS. Left: CLIP-based approaches. Middle: SAM-based approaches. Right: Our CLIP-SAM collaboration approach.

fragmented masks [16]. To make the collaboration more stable, CoLSAM introduces Adaptive Score Fusion (ASF) module that consolidates multi-scale anomaly cues from a WinCLIP-style backbone into a dense, promptable prior that suitable for SAM. Instead of discarding low-confidence cues through hard binarization, this prior provides continuous guidance to SAM's prompt encoder, allowing SAM to refine boundaries based on more reliable spatial evidence.

Our main contributions can be summarized as follows:

- Unified CLIP–SAM framework for ZSAS. We build a WinCLIP-inspired pipeline that combines CLIP's cross-category generalization with SAM's boundary accuracy, enabling ZSAS without category-specific retraining.
- ASF as a CLIP-to-SAM bridging adapter. ASF converts CLIP-derived multi-scale anomaly evidence into stable priors, avoiding brittle hard-threshold prompts and improving segmentation stability.
- Comprehensive evaluation. We conduct extensive experiments and ablations on standard industrial benchmarks to verify the effectiveness and stability of each component.

II. Related Work

A. Zero-shot Semantic Segmentation

Foundation models such as CLIP [10] and SAM [11] enable training-free transfer. Zero-shot semantic segmentation (ZSS) targets unseen categories without class-specific training, which is useful when dense annotations are scarce [19], [20]. Recent advances are driven by vision–language alignment, where CLIP-style pretraining injects text semantics into visual features and improves generalization to novel classes [10], [21], [22]. Beyond global alignment, CLIP-RC [23] and CascadeCLIP [24] leverage region cues and multi-level alignment to enhance pixel-level semantics. Transformer-based methods further strengthen contextual reasoning via attention, as shown by ZePT [25] and OTSeg [26]. Earlier work also improves robustness via synthetic data, transductive inference, or self-supervision. Despite this progress, ZSS still faces scalability, domain shift, and robustness challenges, especially in real-world industrial settings.

B. CLIP-based Models

CLIP-based ZSAS methods form anomaly heatmaps by matching local patches with normal/abnormal prompts in the CLIP embedding space. AnomalyCLIP [4] and VCP-CLIP [27] enhance discrimination via anomaly-oriented prompts or virtual prototypes, while FiLo [28] and APRIL-GAN [1] leverage fine-grained alignment and generative priors for localization. To obtain denser cues, CLIP Surgery [29], VCP-CLIP [27], and CLIP-AD [30] refine attention or multi-level aggregation to sharpen pixel-level responses. WinCLIP [3] provides a strong baseline with window-based dense feature extraction and score aggregation. However, most CLIP-only pipelines still predict on patch grids and upsample to pixel masks, which can blur boundaries and miss thin defects [2]. This motivates an explicit boundary-refinement stage beyond coarse heatmaps.

C. SAM-based Models

SAM [11], [31] introduced promptable segmentation, producing high-quality masks from points, boxes, or coarse masks with strong zero-shot generalization. However, it is sensitive to image quality and prompt reliability, and can degrade on distorted inputs. RobustSAM [30] improves robustness via anti-degradation components that recover cleaner representations, but requires costly paired training data. SAA/SAA+ [15] generate prompts via text grounding and filter SAM masks with heuristic rules; their performance still relies on prompt design and post-processing. Overall, CLIP-only methods yield coarse, blurry localization, while SAM-only pipelines are prompt-dependent and may fail with noisy or incomplete prompts. Motivated by this complementarity, we combine CLIP and SAM to bridge zero-shot localization and high-quality segmentation through reliable prompt construction and explicit mask refinement [16].

III. Methodology

In this section, we present CoLSAM, a method that improves ZSAS by better aggregating multi-scale, multi-window anomaly evidence in WinCLIP-style pipelines. CoLSAM consists of three stages: (1) CLIP-based multi-scale window scoring, which uses a frozen CLIP model to score overlapping windows at multiple scales and produce scale-specific anomaly maps; (2) adaptive score fusion, where a lightweight predictor outputs input-dependent fusion weights to replace WinCLIP's fixed harmonic mean and yield sharper, more stable anomaly maps; and (3) prompt-constrained SAM segmentation, which converts the fused anomaly guidance into compact prompts to steer a frozen SAM model toward accurate masks with fewer redundant proposals and less post-processing. Figure 2 provides an overview of the CoLSAM pipeline.

A. CLIP-based Window Scoring

Following the WinCLIP paradigm, we use a frozen CLIP model to score local image windows for anomaly. CLIP includes an image encoder E_{img} and a text encoder E_{txt} that map images and texts into a shared embedding space.

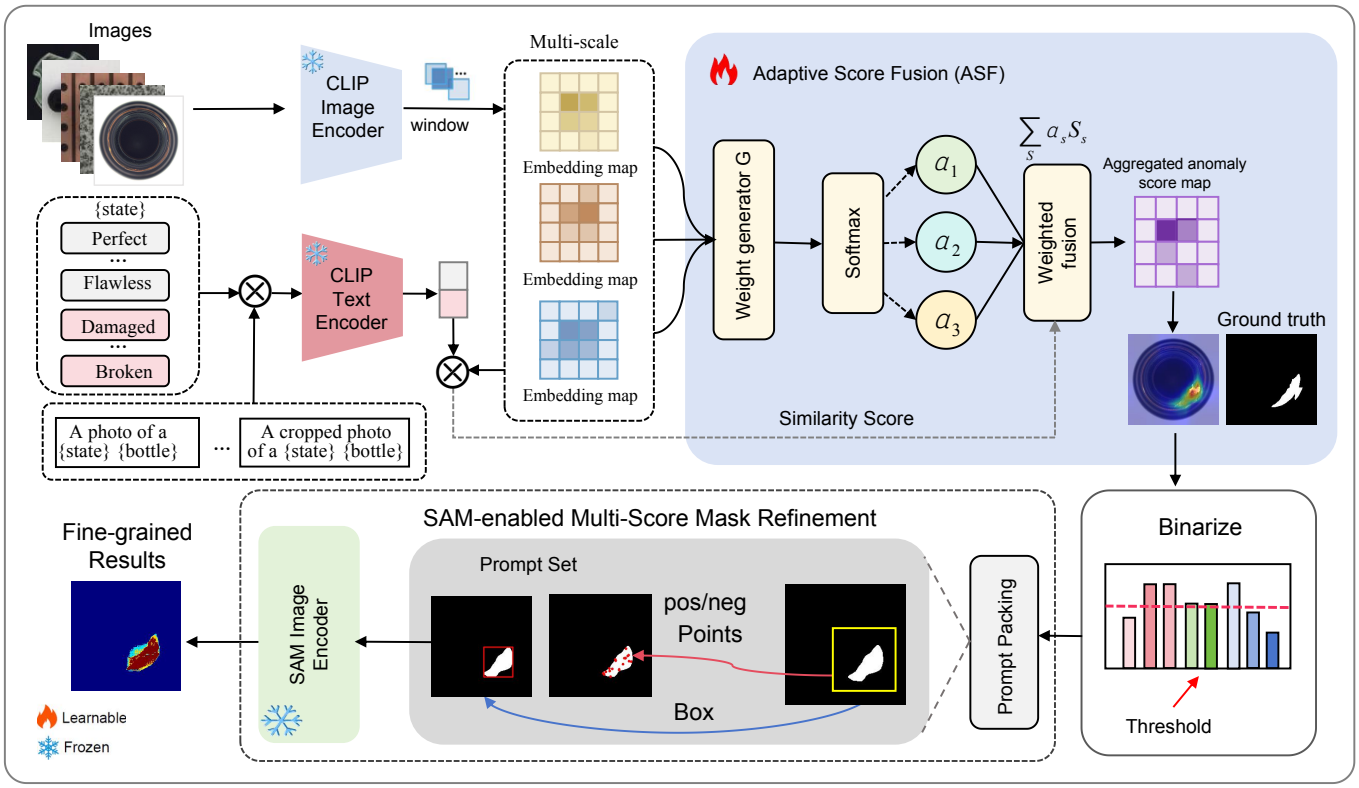

Fig. 2. Overview of CoLSAM for ZSAS. Our model consists of two stages: CLIP-based multi-scale localization and SAM-based mask refinement. First, CLIP extracts multi-scale window anomaly cues, which are fused by the learnable ASF module to produce a sharp anomaly map. Then, the map is binarized to generate box and positive/negative point prompts that guide SAM for boundary-accurate segmentation, yielding fine-grained anomaly masks.

a) Text prompts: We use two prompts (normal vs. anomalous), denoted as p_n and p_a. Their text embeddings are computed as

$$t_n = E_{\text{txt}}(p_n), \qquad t_a = E_{\text{txt}}(p_a), \tag{1}$$

We keep these embeddings fixed during both training and inference.

b) Window extraction: Given $I \in \mathbb{R}^{H\times W\times 3}$, we extract overlapping windows at multiple scales. At scale s, we resize the image and split it into windows $\{I_{s,k}\}_{k=1}^{K_s}$. We encode each window independently with the CLIP image encoder:

$$v_{s,k} = E_{\text{img}}(I_{s,k}). \tag{2}$$

c) Window-level anomaly scoring: For each $v_{s,k}$, we compute cosine similarities to t_a and t_n. We define the window anomaly score as

$$r_{s,k} = \text{sim}(v_{s,k}, t_a) - \text{sim}(v_{s,k}, t_n). \tag{3}$$

A larger $r_{s,k}$ means the window matches the anomalous prompt more than the normal prompt.

B. ASF for CLIP-to-SAM Calibration

a) Motivation: Mismatch Between CLIP and SAM Spaces.: WinCLIP-style methods produce anomaly evidence in the CLIP embedding space by comparing local image windows with normal and anomalous text prompts. However, the resulting multi-scale anomaly maps are not explicitly optimized for downstream segmentation models like SAM. In practice, we find that SAM is sensitive to the spatial distribution, contrast, and calibration of anomaly maps used to construct prompts. This mismatch between CLIP-induced anomaly scores and SAM's preferred input space motivates a learnable calibration mechanism.

b) Limitation of fixed fusion: Given multi-scale anomaly maps $\{A_s\}_{s=1}^{S}$ produced by WinCLIP methods typically adopt a fixed fusion strategy:

$$A_{\text{fixed}}(x,y) = \sum_{s=1}^{S} \alpha_s \, A_s(x,y), \qquad \sum_{s=1}^{S} \alpha_s = 1, \tag{4}$$

where α_s is typically set to a uniform value. This strategy assumes that all scales contribute equally at every spatial location, an assumption that often breaks down in industrial scenarios. More importantly, fixed fusion is agnostic to the downstream segmentation objective and cannot adapt the score distribution to suit SAM-based mask generation.

c) ASF as calibration: To overcome the limitations of fixed multi-scale fusion, we introduce an ASF module that performs content-dependent calibration of multi-scale anomaly scores. Given CLIP-generated anomaly maps $\{A_s\}_{s=1}^{S}$ at different image scales, ASF predicts spatially varying fusion weights conditioned on the local anomaly evidence.

At each spatial location (x, y), we stack the multi-scale scores into a vector

$$U(x,y) = \big[A_1(x,y), \ldots, A_S(x,y)\big] \in \mathbb{R}^S. \tag{5}$$

A lightweight calibration network $G : \mathbb{R}^S \to \mathbb{R}^S$ is applied point-wisely to produce scale-wise fusion logits

$$L(x,y) = G\big(U(x,y)\big). \tag{6}$$

Normalized fusion weights are obtained via a softmax over scales,

$$W_s(x,y) = \frac{\exp(L_s(x,y))}{\sum_{j=1}^{S} \exp(L_j(x,y))}, \tag{7}$$

and the calibrated anomaly score is computed as

$$\hat{A}(x,y) = \sum_{s=1}^{S} W_s(x,y)\, A_s(x,y). \tag{8}$$

The calibration network $G(\cdot)$ is implemented as a two-layer 1×1 convolutional module, enabling efficient operation purely in the score space. Functionally, ASF acts as a learnable calibration layer that maps CLIP-induced multi-scale anomaly evidence into a score space more compatible with SAM's prompt-driven segmentation, allowing SAM to better exploit CLIP anomaly cues without modifying either backbone.

d) SAM-driven Optimization Objective: Importantly, we optimize ASF solely through SAM's final segmentation supervision. The fused anomaly map $\hat{A}(x,y)$ acts as an intermediate representation between CLIP and SAM, and its quality is measured implicitly by the accuracy of SAM's predicted anomaly masks. Specifically, we convert $\hat{A}(x,y)$ into box and point prompts for SAM, which produces a predicted anomaly mask $\hat{Y}(x,y)$. We train ASF by minimizing a segmentation loss between $\hat{Y}(x,y)$ and the ground-truth mask $Y(x,y)$:

$$\begin{aligned}\mathcal{L}_{\text{seg}} = -\frac{1}{|\Omega|} \sum_{(x,y)\in\Omega} \Big[& Y(x,y) \log \hat{Y}(x,y) \\ & + \big(1 - Y(x,y)\big) \log\big(1 - \hat{Y}(x,y)\big)\Big].\end{aligned} \tag{9}$$

Here, Ω denotes the set of all pixel locations in the image domain.

During training, we freeze both CLIP and SAM, and update only the parameters of the ASF module $G(\cdot)$. This design encourages ASF to adaptively reweight and calibrate CLIP anomaly maps into representations that are most effective for SAM-based segmentation.

C. Prompt-Constrained Segmentation with SAM

Given the adapted anomaly evidence map $A \in \mathbb{R}^{H\times W}$, we convert it into a compact set of reliable prompts for SAM to obtain accurate boundaries while avoiding redundant mask proposals. We keep SAM frozen and use A only to (i) generate candidate prompts and (ii) apply simple constraints to filter low-confidence or redundant prompts.

a) Anomaly prior and candidate region extraction: We first normalize A to $[0, 1]$ and obtain a binary anomaly prior by thresholding:

$$\tilde{A} = \mathrm{Norm}(A), \qquad B(x,y) = \mathbb{I}\big[\tilde{A}(x,y) \geq \tau\big], \tag{10}$$

where $\mathrm{Norm}(\cdot)$ denotes min–max normalization, $\mathbb{I}[\cdot]$ is the indicator function, and τ is a fixed threshold. We extract connected components from B to obtain candidate anomaly regions $\{\mathcal{R}_i\}_{i=1}^{N}$. For each region $\mathcal{R}_i$, we compute its bounding box Box_i and the peak location

$$(x_i^\star, y_i^\star) = \arg \max_{(x,y)\in\mathcal{R}_i} \tilde{A}(x,y), \tag{11}$$

which serves as a high-confidence positive point prompt for SAM. Optionally, we sample a few negative points from low-score areas to discourage masks from spilling into normal regions.

b) Prompt constraints and redundancy suppression: Not all candidate regions are reliable. We assign each region a confidence score, e.g., the peak anomaly value within the region:

$$c_i = \max_{(x,y)\in\mathcal{R}_i} \tilde{A}(x,y). \tag{12}$$

To suppress redundant prompts, we perform non-maximum suppression on region boxes. For two candidates i and j, if

$$\mathrm{IoU}(\mathrm{Box}_i, \mathrm{Box}_j) > \gamma, \tag{13}$$

where γ is an overlap threshold, we discard the candidate with the lower confidence score c. This constraint prevents multiple prompts from triggering highly similar masks around the same area.

c) SAM inference and mask selection: For each retained candidate, we feed its prompts to SAM and obtain mask proposals. We select the final mask by combining SAM's mask quality prediction with consistency to the anomaly evidence. Specifically, for a predicted mask M, we compute an evidence-consistency score:

$$s(M) = \frac{1}{|M|} \sum_{(x,y)\in M} \tilde{A}(x,y), \tag{14}$$

where $|M|$ denotes the number of pixels inside M. We choose the mask that maximizes a weighted combination of SAM confidence and evidence consistency. The final anomaly segmentation is the selected mask, or the union of selected masks when multiple high-confidence regions remain.

This prompt-constrained design guides SAM to focus on regions supported by strong, adapted anomaly evidence, improving boundary delineation while reducing redundant mask proposals and post-processing.

IV. Experiments

A. Experimental Setup

Dataset. We evaluate on *MVTec-AD* and *VisA*, two standard benchmarks for industrial anomaly detection. *MVTec-AD* contains 15 object/texture categories with over 5,000 high-resolution images. *VisA* includes 12 object-centric subsets

TABLE I
COMPARISON OF DIFFERENT ZSAS APPROACHES ON THE MVTEC AD AND VISA DATASETS. EVALUATION METRICS INCLUDE AUROC, F1-MAX, AND PRO. BOLD INDICATES THE BEST RESULT.

Base model	Method	MVTec-AD			VisA		
		AUROC	F_1-max	PRO	AUROC	F_1-max	PRO
CLIP-based Approaches	WinCLIP [3]	85.1	31.7	64.6	79.6	14.8	56.8
	APRIL-GAN [1]	87.6	43.3	44.0	**94.2**	32.3	86.8
	SDP [30]	88.7	35.3	79.1	84.1	16.0	63.4
SAM-based Approaches	SAA+ [15]	73.2	37.8	42.8	74.0	27.1	36.8
CLIP & SAM	**CoLSAM (Ours)**	**89.2**	**45.1**	**82.3**	92.4	**34.6**	**87.1**

with 10,821 images (9,621 normal and 1,200 anomalous). Anomalies cover both surface defects (e.g., scratches, dents, stains, and cracks) and structural issues (e.g., missing or misaligned components).

Evaluation Metrics. Pixel-level performance is measured by AUROC and F_1-max (best F_1 over all thresholds). We also report Per-Region Overlap (PRO) to quantify region-level overlap between predicted defect areas and ground-truth masks.

Implementation Details. We use OpenAI's pre-trained ViT-B/16+ (24 Transformer layers) as the WinCLIP encoder [3]. Training is performed on an NVIDIA A100 GPU with AdamW [32] and a learning rate of 1×10^{-4}. We train for 6 epochs with batch size 4, and use batch size 1 for testing. For SAM [11], we adopt the official ViT-H pre-trained model.

B. Main Experimental Results

Comparison Methods. We evaluate CoLSAM on ZSAS against representative zero-shot baselines on MVTec-AD and VisA. Following the previous works, we consider only zero-shot methods and group them into three types: (1) CLIP-based ZSAS methods (e.g., WinCLIP [3], APRIL-GAN [5], SDP [11]), (2) a SAM-only prompt-driven baseline (SAA+ [20]), and (3) hybrid CLIP–SAM pipelines that use CLIP cues to guide segmentation. All methods are evaluated with pixel-level AUROC, F_1-max, and PRO, as summarized in Table I.

Zero-Shot Anomaly Segmentation. Table I compares CoLSAM with CLIP- and SAM-based baselines on MVTec-AD and VisA. CoLSAM consistently achieves the best region-aware localization and improves performance across datasets.

On MVTec-AD, CoLSAM achieves the best results on all metrics (89.2 AUROC, 45.1 F_1-max, 82.3 PRO). Compared with SDP, it improves AUROC by 0.5%, F_1-max by 9.8%, and PRO by 3.2%. Compared with SAA+, it yields larger gains: +16.0% AUROC, +7.3% F_1-max, and +39.5% PRO.

On VisA, CoLSAM achieves the best F_1-max and PRO (34.6% and 87.1%) while maintaining a competitive AUROC of 92.4%. Against APRIL-GAN, it improves F_1-max by 2.3% and PRO by 0.3%. Against SAA+, it improves AUROC by 18.4%, F_1-max by 7.5%, and PRO by 50.3%. These results show that CoLSAM generalizes well and yields more accurate anomaly-region delineation in the zero-shot setting.

Qualitative Comparisons. Figure 3 visualizes anomaly maps and masks produced by WinCLIP, APRIL-GAN, SAA+, SDP, and CoLSAM. CLIP-based methods can highlight suspicious areas but often yield coarse or noisy maps: WinCLIP and SDP tend to produce diffuse activations and respond to surrounding textures, while APRIL-GAN often generates fragmented, discontinuous regions. SAA+ produces cleaner masks, but may miss defect parts or drift from the true anomaly when prompts are unreliable. In contrast, CoLSAM generates compact and consistent predictions that align better with ground-truth masks, delivering more accurate localization and sharper boundaries, especially in complex backgrounds.

C. Ablation Study

We conduct ablation studies to examine the role of each component in CoLSAM for ZSAS, with a particular focus on the proposed ASF module. Experiments are performed on MVTec-AD and VisA. Unless otherwise specified, the CLIP image/text encoders and the SAM backbone are kept frozen. We report pixel-level AUROC and pixel-level F_1-max, where higher values indicate better performance.

Effect of ASF. Table II presents a controlled comparison of different CLIP–SAM integration strategies. Directly combining WinCLIP with SAM leads to a clear performance degradation compared to WinCLIP alone. Specifically, pixel-level AUROC drops by 2.8% on MVTec-AD and 4.0% on VisA, while F_1-max decreases by 7.1% and 4.6%, respectively. This result indicates that CLIP-generated anomaly maps are not directly compatible with SAM for prompt-driven segmentation.

In contrast, inserting the proposed ASF module between WinCLIP and SAM dramatically improves performance. Compared with the naive WinCLIP+SAM baseline, ASF yields gains of 6.9% and 16.8% in AUROC, and 20.5% and 24.4% in F_1-max on MVTec-AD and VisA, respectively. Notably, WinCLIP+ASF+SAM also surpasses WinCLIP alone, demonstrating that ASF effectively calibrates CLIP anomaly evidence into a score space that is more suitable for SAM. These results confirm that ASF is a critical component for enabling effective CLIP–SAM collaboration.

D. The Process of Fine-grained Segmentation

Figure 4 shows that introducing SAM refines WinCLIP's coarse anomaly localization into fine-grained segmentation.

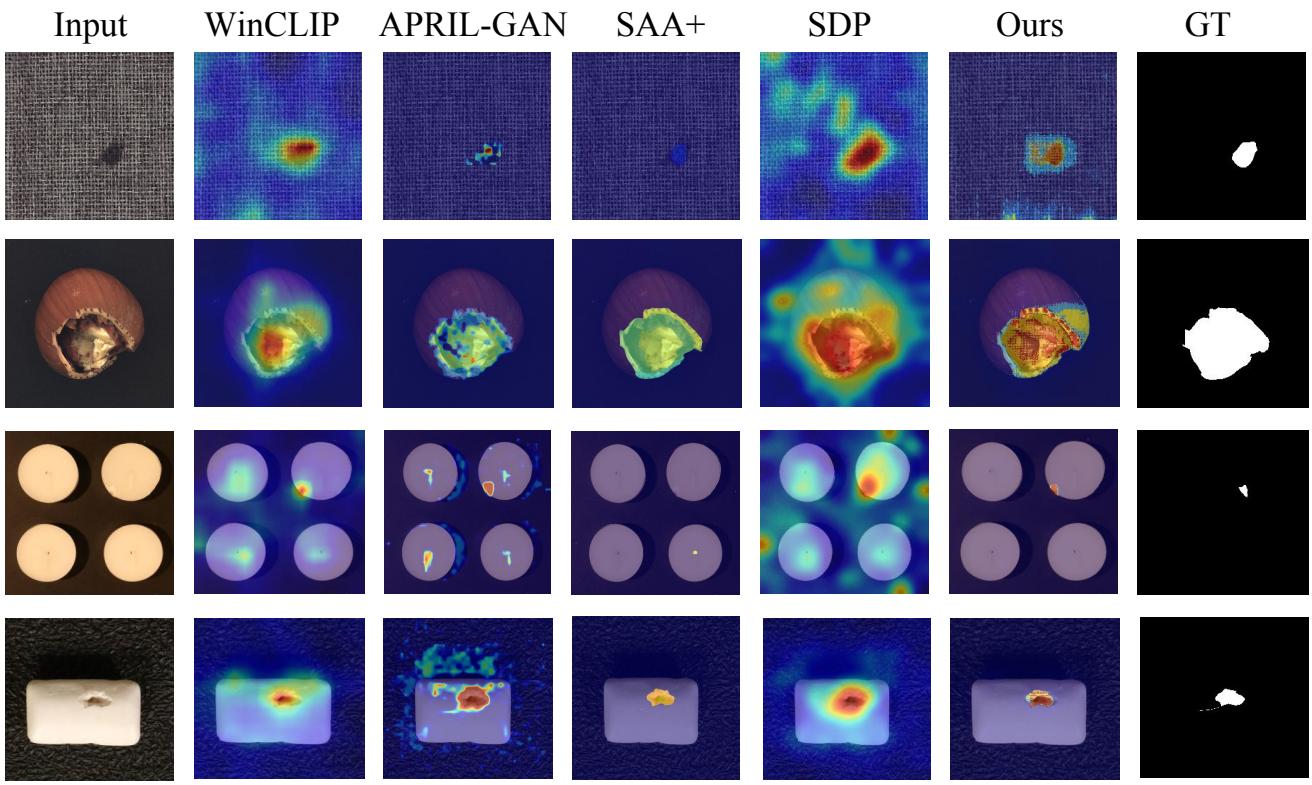

Fig. 3. Comparison of zero-shot visualization results among our model, CLIP-based, and SAM-based methods on the MVTec-AD (first two rows) dataset and VisA (last two rows) dataset. Our model performs much better on the location and boundary of the anomaly segmentation.

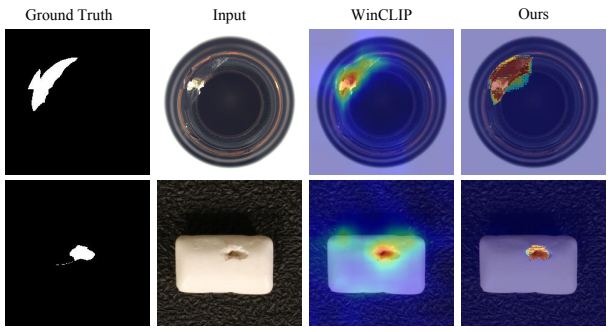

Fig. 4. Visualization of rough segmentation of WinCLIP and fine segmentation of ours. The top row shows samples from MVTec-AD, and the bottom row shows samples from VisA.

WinCLIP provides a broad and blurry response that often spills into nearby normal regions, which is sufficient for rough defect indication but inaccurate on boundaries. Guided by this coarse evidence, SAM concentrates on the true defect area and produces compact masks with sharper, boundary-aligned contours, significantly reducing background leakage. As a result, the refined masks match the ground truth more closely in both defect extent and edge precision.

TABLE II
ABLATION ON ASF IN COLSAM FOR ZSAS.

Method Variant	MVTec-AD		VisA	
	AUROC	F_1-max	AUROC	F_1-max
WinCLIP	85.1	31.7	79.6	14.8
WinCLIP + SAM	82.3	24.6	75.6	10.2
WinCLIP + ASF + SAM	**89.2**	**45.1**	**92.4**	**34.6**

V. CONCLUSION

We presented CoL-SAM, a collaborative CLIP-SAM framework for ZSAS in industrial inspection. CoL-SAM leverages CLIP for transferable anomaly localization and SAM for boundary-accurate mask generation, enabling defect segmentation on unseen categories without target-domain training or dense annotations. To stabilize CLIP–SAM collaboration, we introduced a lightweight ASF module that adaptively reweights multi-scale anomaly evidence from a WinCLIP-style window scoring pipeline, producing sharper and better-calibrated anomaly maps with reduced noise propagation. Based on the fused guidance, we further design prompt constraints to generate compact and reliable prompts for SAM, mitigating redundant mask proposals and reducing the need for cumbersome post-processing. Extensive experiments on

MVTec-AD and VisA demonstrate consistent improvements over CLIP-only, SAM-only, and WinCLIP-based baselines, highlighting the effectiveness of adaptive fusion and prompt-constrained refinement for fine-grained ZSAS.

Future work will study calibration-aware, adaptive prompt construction to reduce manual tuning, extend the framework to harder industrial settings with subtle defects and complex backgrounds, and evaluate more backbones to improve robustness and generalization.

REFERENCES

[1] X. Chen, Y. Han, and J. Zhang, "April-gan: A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad," *arXiv preprint arXiv:2305.17382*, 2023.

[2] H. Deng, Z. Zhang, J. Bao, and X. Li, "Anovl: Adapting vision-language models for unified zero-shot anomaly localization," *arXiv preprint arXiv:2308.15939*, vol. 2, no. 5, 2023.

[3] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "Winclip: Zero-/few-shot anomaly classification and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 606–19 616.

[4] Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen, "Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection," *arXiv preprint arXiv:2310.18961*, 2023.

[5] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 14 318–14 328.

[6] J. Santos, T. Tran, and O. Rippel, "Optimizing patchcore for few/many-shot anomaly detection," *arXiv preprint arXiv:2307.10792*, 2023.

[7] P. Bergmann, M. Fauser, D. Sattlegger, and C. 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*, 2019, pp. 9592–9600.

[8] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer, "Spot-the-difference self-supervised pre-training for anomaly detection and segmentation," in *European conference on computer vision*. Springer, 2022, pp. 392–408.

[9] X. Li, Z. Huang, F. Xue, and Y. Zhou, "Musc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images," in *The Twelfth International Conference on Learning Representations*, 2024.

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[11] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[12] Y. Cao, J. Zhang, L. Frittoli, Y. Cheng, W. Shen, and G. Boracchi, "Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 55–72.

[13] T. Gong, Q. Chu, B. Liu, W. Zhou, and N. Yu, "Fe-clip: Frequency enhanced clip model for zero-shot anomaly detection and segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 21 220–21 230.

[14] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, and J. Lu, "Denseclip: Language-guided dense prediction with context-aware prompting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 082–18 091.

[15] Y. Cao, X. Xu, C. Sun, Y. Cheng, Z. Du, L. Gao, and W. Shen, "Segment any anomaly without training via hybrid prompt regularization," *arXiv preprint arXiv:2305.10724*, 2023.

[16] Y. Hou, K. Xu, J. Li, Y. Ruan, and J. Qiu, "Enhancing zero-shot anomaly detection: Clip-sam collaboration with cascaded prompts," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2024, pp. 46–60.

[17] T. Koleilat, H. Asgariandehkordi, H. Rivaz, and Y. Xiao, "Medclip-samv2: Towards universal text-driven medical image segmentation," *Medical Image Analysis*, p. 103749, 2025.

[18] H. Wang, P. K. A. Vasu, F. Faghri, R. Vemulapalli, M. Farajtabar, S. Mehta, M. Rastegari, O. Tuzel, and H. Pouransari, "Sam-clip: Merging vision foundation models towards semantic and spatial understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3635–3647.

[19] M. Bucher, T.-H. Vu, M. Cord, and P. Pérez, "Zero-shot semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[20] H. Zhang and H. Ding, "Prototypical matching and open set rejection for zero-shot semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6974–6983.

[21] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," *arXiv preprint arXiv : 2201.03546*, 2022.

[22] Z. Zhou, Y. Lei, B. Zhang, L. Liu, and Y. Liu, "Zegclip: Towards adapting clip for zero-shot semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 11 175–11 185.

[23] Y. Zhang, M.-H. Guo, M. Wang, and S.-M. Hu, "Exploring regional clues in clip for zero-shot semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3270–3280.

[24] Y. Li, Z. Li, Q. Zeng, Q. Hou, and M.-M. Cheng, "Cascade-clip: Cascaded vision-language embeddings alignment for zero-shot semantic segmentation," *arXiv preprint arXiv:2406.00670*, 2024.

[25] Y. Jiang, Z. Huang, R. Zhang, X. Zhang, and S. Zhang, "Zept: Zero-shot pan-tumor segmentation via query-disentangling and self-prompting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 386–11 397.

[26] K. Kim, Y. Oh, and J. C. Ye, "Otseg: Multi-prompt sinkhorn attention for zero-shot semantic segmentation," in *European Conference on Computer Vision*. Springer, 2024, pp. 200–217.

[27] Z. Qu, X. Tao, M. Prasad, F. Shen, Z. Zhang, X. Gong, and G. Ding, "Vcp-clip: A visual context prompting model for zero-shot anomaly segmentation," in *European Conference on Computer Vision*. Springer, 2024, pp. 301–317.

[28] Z. Gu, B. Zhu, G. Zhu, Y. Chen, H. Li, M. Tang, and J. Wang, "Filo: Zero-shot anomaly detection by fine-grained description and high-quality localization," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 2041–2049.

[29] Y. Li, H. Wang, Y. Duan, J. Zhang, and X. Li, "A closer look at the explainability of contrastive language-image pre-training," *Pattern Recognition*, vol. 162, p. 111409, 2025.

[30] X. Chen, J. Zhang, G. Tian, H. He, W. Zhang, Y. Wang, C. Wang, and Y. Liu, "Clip-ad: A language-guided staged dual-path model for zero-shot anomaly detection," in *International Joint Conference on Artificial Intelligence*. Springer, 2024, pp. 17–33.

[31] L. Ke, M. Ye, M. Danelljan, Y.-W. Tai, C.-K. Tang, F. Yu *et al.*, "Segment anything in high quality," *Advances in Neural Information Processing Systems*, vol. 36, pp. 29 914–29 934, 2023.

[32] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.