Dual Anchors, Do It Better: Hierarchical Group Merging for Zero-Shot Anomaly Detection

Supplementary Material

Overview

The supplementary material presents the following sections to strengthen the main manuscript:

– **Sec. A** shows more detailed Information, including:

- Sec. A.1: Implementation Details,
- Sec. A.2: Datasets,
- Sec. A.3: Baseline Methods.

– **Sec. B** shows additional qualitative results.

– **Sec. C** shows more analysis, including:

- Sec. C.1: Comparison of Computational Efficiency,
- Sec. C.2: Backbone Comparison,
- Sec. C.3: Effects of Grouping,

– **Sec. D** shows Failure Cases.

A. Detailed Information

In this section, we first provide additional implementation details, including the hyperparameters used in our model design. We then present dataset statistics and preprocessing protocols for the 8 industrial datasets (MVTec AD [3], VisA [24], MPDD [12], BTAD [16], RSDD [22], KSDD2 [5], DAGM [21], DTD-Synthetic [1]) and 6 medical datasets (ISIC [9], CVC-ColonDB [20], CVC-ClinicDB [4], TN3K [8], Endo [10], Kvasir [13]). Finally, we evaluate our method against a variety of vision–language anomaly detection baselines and briefly summarize their key characteristics and implementation details.

A.1. Implementation Details

Architecture. We adopt DINOv3 with a ViT-L/16 backbone as the image encoder. To align with the four selected feature layers (6th, 12th, 18th, and 24th), we employ four Hierarchical Merge Blocks. The number of learnable group tokens is initially set to 16 and is progressively reduced by a factor of two at each block, resulting in 2 group tokens in the final layer.

Optimization. All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU with 24 GB of memory. We train the model for 10 epochs with a batch size of 8, which takes approximately 1 hour in total. We use the Adam optimizer with a learning rate of 1×10^{-4} for all experiments.

A.2. Datasets

MVTec AD [3]. The MVTec Anomaly Detection (MVTec AD) dataset is a real-world benchmark for unsupervised visual anomaly detection in industrial inspection. It contains 5,354 high-resolution images from 15 object and texture categories, where 3,629 defect-free images are used for training and 1,725 images with both normal and anomalous samples are used for testing. In total, 73 defect types with pixel-accurate ground-truth masks are provided, enabling evaluation at both image and pixel level.

VisA [24]. The VisA dataset is a large-scale industrial visual anomaly detection benchmark consisting of 10,821 high-resolution images from 12 object categories across three domains, including printed circuit boards and multi-instance scenes. It contains 9,621 normal and 1,200 anomalous images with both image- and pixel-level annotations, enabling comprehensive evaluation of detection and localization methods.

MPDD [12]. The Metal Parts Defect Detection (MPDD) dataset targets visual inspection of painted metal parts in real industrial conditions. It comprises about 1.3k images of six types of metal components with pixel-precise defect masks, split into defect-free training images and a validation set containing both normal and defective samples.

BTAD [16]. The BeanTech Anomaly Detection (BTAD) dataset is a real-world industrial anomaly detection benchmark with 2,830 RGB images of three industrial products. It includes body and surface defects with a data structure similar to MVTec AD, and is widely used to evaluate industrial anomaly detection and localization methods.

RSDD [22]. The Rail Surface Defect Detection (RSDD) dataset consists of grayscale images captured from express and heavy-haul rails, containing various real rail surface defects such as cracks and squats under challenging outdoor conditions. It is designed to benchmark defect detection and localization methods for rail condition monitoring.

KSDD2 [5]. The Kolektor Surface Defect Dataset (KSDD) focuses on detecting defects on electrical commutator surfaces. It contains 399 high-resolution images, including 52 defective and 347 non-defective samples, and is commonly used as a small but challenging benchmark for industrial surface defect segmentation.

DAGM [21]. The DAGM 2007 dataset is an industrial optical inspection benchmark featuring ten classes of textured surfaces with and without defects. It provides thousands of high-resolution grayscale images with pixel-level ground truth, and is widely used for surface defect detection under complex background textures.

DTD-Synthetic [1]. The DTD-Synthetic dataset is a synthetic texture anomaly detection benchmark created by sampling diverse textures from the Describable Texture Dataset (DTD) and injecting artificial defects. It is designed to study zero-shot and few-shot texture anomaly localization under controlled anomaly patterns.

ISIC [9]. The ISIC dataset (International Skin Imaging Collaboration) is a large collection of dermoscopic images of skin lesions, covering multiple diagnostic classes such as melanoma, nevus, and seborrheic keratosis. It provides expert annotations including lesion labels and segmentation masks, and serves as a standard benchmark for skin lesion classification and segmentation.

CVC-ColonDB [20]. CVC-ColonDB is a colonoscopy dataset composed of 380 still images extracted from 15 colonoscopy videos, each with a corresponding polyp segmentation mask. It is mainly used to evaluate polyp detection and segmentation performance in challenging clinical scenarios.

CVC-ClinicDB [4]. CVC-ClinicDB is a widely used polyp segmentation benchmark consisting of 612 frames extracted from colonoscopy videos collected at Hospital Clínic de Barcelona. Each image is accompanied by a binary ground-truth mask delineating the polyp region at pixel level.

TN3K [8]. The TN3K dataset contains 3,000 thyroid ultrasound images focusing on the thyroid gland region, with expert annotations of thyroid nodules. It is designed for weakly and fully supervised thyroid nodule detection and segmentation in ultrasound imaging.

Endo [10]. The Endo dataset is an in-house gastrointestinal endoscopy dataset consisting of endoscopic images with corresponding pixel-level lesion masks. It is used to evaluate anomaly detection and segmentation performance on real clinical endoscopy data.

Kvasir [13]. The Kvasir-SEG dataset is an open-access gastrointestinal polyp dataset containing 1,000 colonoscopy images with corresponding expert-annotated polyp segmentation masks, derived from the Kvasir v2 dataset. It has become a standard benchmark for training and evaluating polyp detection and segmentation models.

A.3. Baseline Methods

WinCLIP [11]. WinCLIP is a window-based CLIP framework for zero- and few-shot anomaly classification and segmentation, which aggregates multi-scale patch features aligned with textual prompts to produce anomaly scores and maps.

APRIL-GAN [7]. APRIL-GAN extends CLIP-based anomaly detection by introducing additional adapter layers and a GAN-based refinement module to better map image features into the joint vision–language space for zero-/few-shot anomaly classification and segmentation.

AnomalyCLIP [23]. AnomalyCLIP learns object-agnostic text prompts that represent generic normality and abnormality, enabling CLIP to focus on abnormal regions rather than object semantics for accurate zero-shot anomaly detection across diverse domains.

AdaCLIP [6]. AdaCLIP adapts CLIP to zero-shot anomaly detection by introducing hybrid learnable prompts, combining static prompts shared across images and dynamic prompts generated per image, which are optimized on auxiliary anomaly detection data.

AA-CLIP [15]. AA-CLIP enhances zero-shot anomaly detection by making CLIP anomaly-aware via a two-stage adaptation that first constructs anomaly-aware text anchors and then aligns patch-level visual features with these anchors using lightweight residual adapters.

Bayes-PFL [17]. Bayes-PFL formulates prompt learning for zero-shot anomaly detection in a Bayesian manner, modeling prompts as distributions and using Monte Carlo sampling over prompt flows to improve robustness and generalization across categories and domains.

B. Additional Qualitative Results

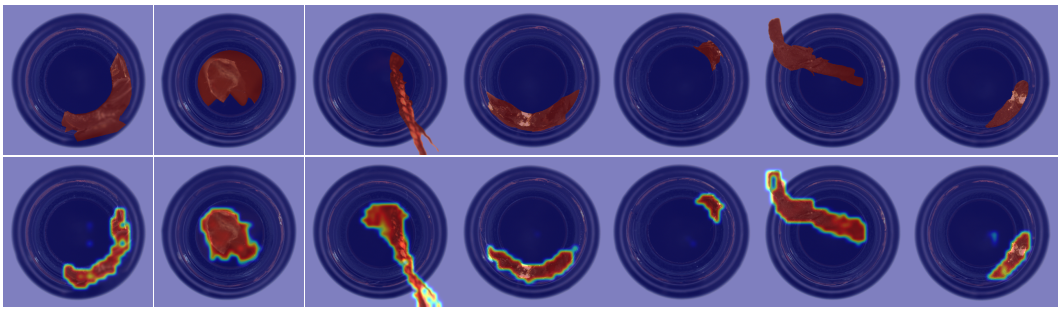

Figure S1. Segmentation results on the bottle class from MVTec-AD. Ground truth regions are marked in red (top row), and predicted anomaly maps are shown below.

Figure S2. Segmentation results on the capsule class from MVTec-AD. Ground truth regions are marked in red (top row), and predicted anomaly maps are shown below.

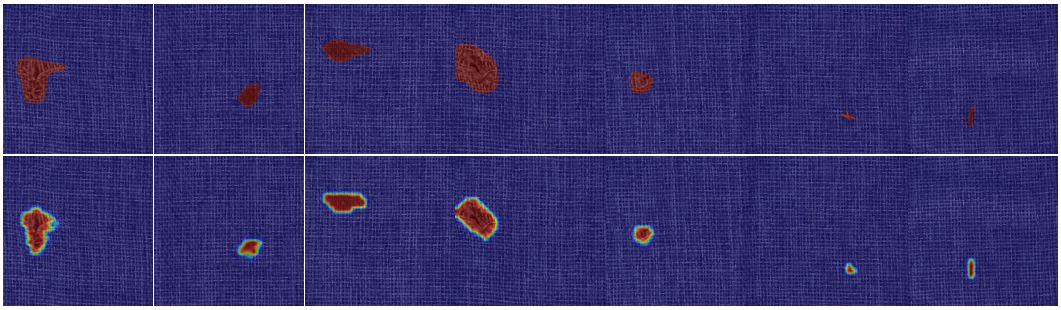

Figure S3. Segmentation results on the carpet class from MVTec-AD. Ground truth regions are marked in red (top row), and predicted anomaly maps are shown below.

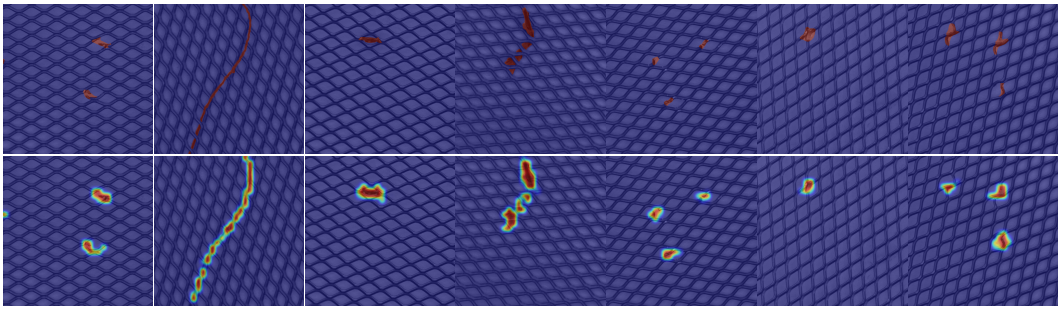

Figure S4. Segmentation results on the grid class from MVTec-AD. Ground truth regions are marked in red (top row), and predicted anomaly maps are shown below.

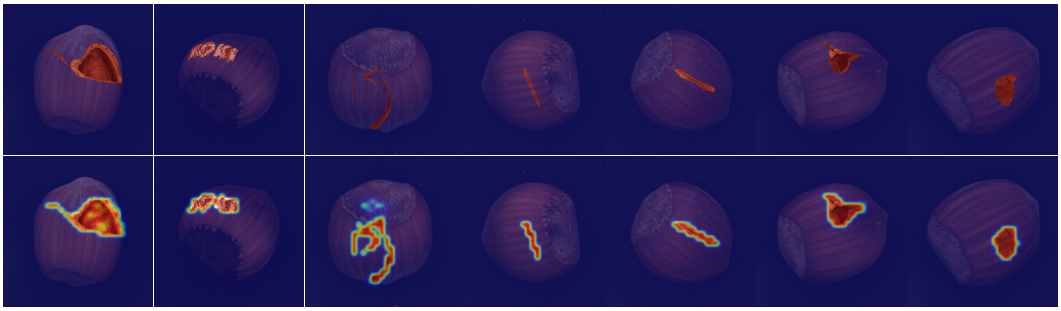

Figure S5. Segmentation results on the hazelnut class from MVTec-AD. Ground truth regions are marked in red (top row), and predicted anomaly maps are shown below.

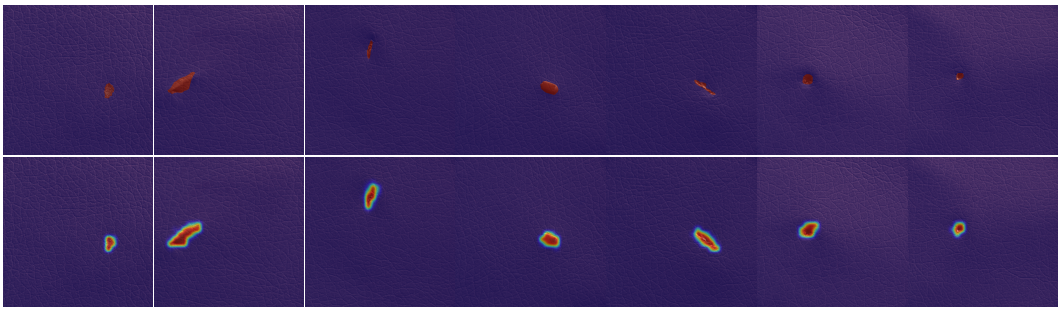

Figure S6. Segmentation results on the leather class from MVTec-AD. Ground truth regions are marked in red (top row), and predicted anomaly maps are shown below.

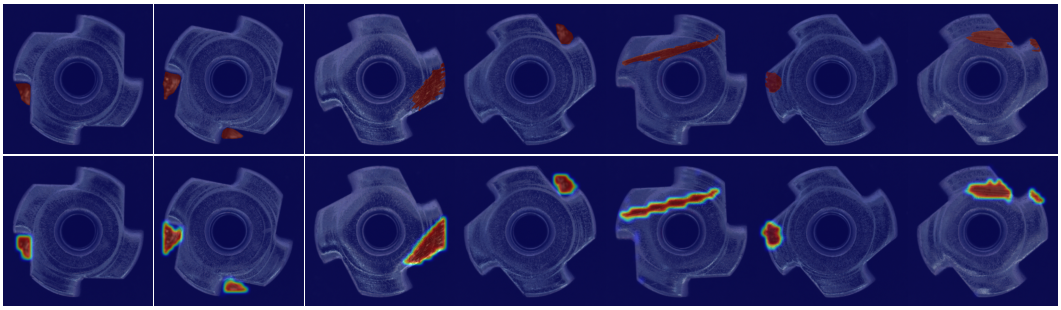

Figure S7. Segmentation results on the metal nut class from MVTec-AD. Ground truth regions are marked in red (top row), and predicted anomaly maps are shown below.

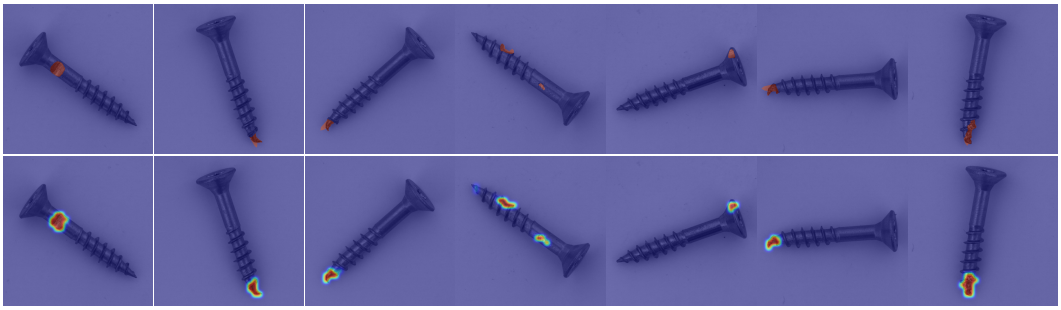

Figure S8. Segmentation results on the screw class from MVTec-AD. Ground truth regions are marked in red (top row), and predicted anomaly maps are shown below.

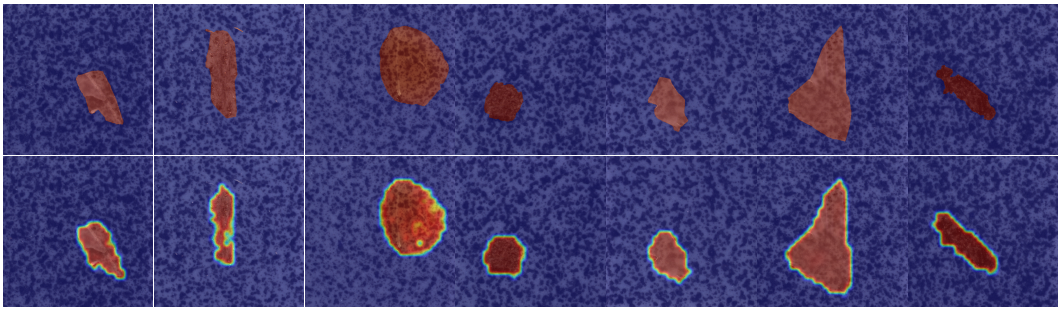

Figure S9. Segmentation results on the tile class from MVTec-AD. Ground truth regions are marked in red (top row), and predicted anomaly maps are shown below.

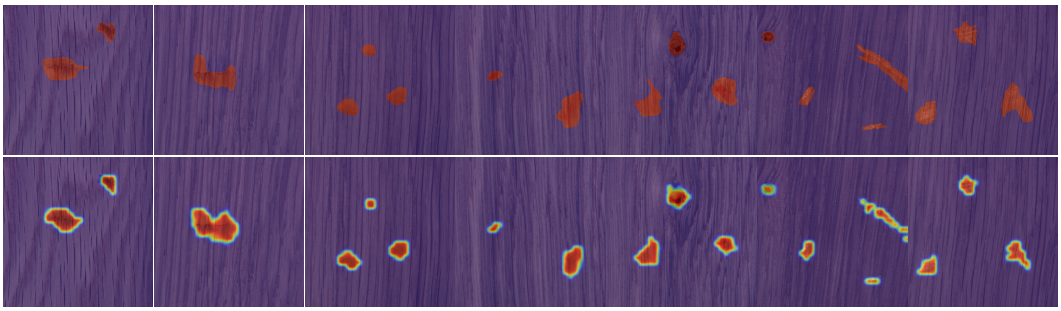

Figure S10. Segmentation results on the wood class from MVTec-AD. Ground truth regions are marked in red (top row), and predicted anomaly maps are shown below.

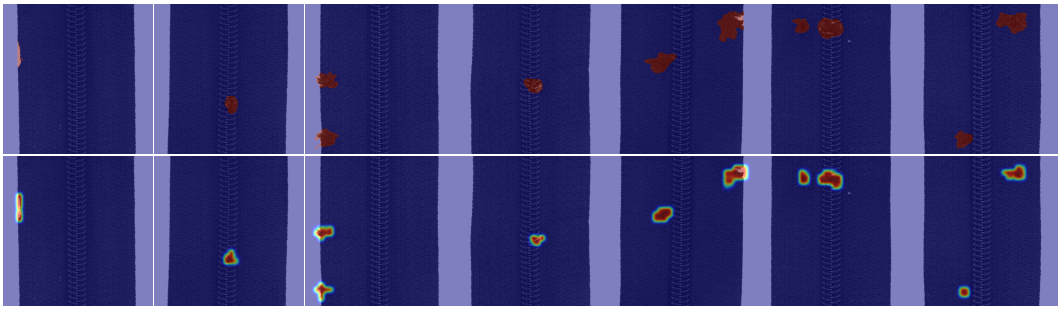

Figure S11. Segmentation results on the zipper class from MVTec-AD. Ground truth regions are marked in red (top row), and predicted anomaly maps are shown below.

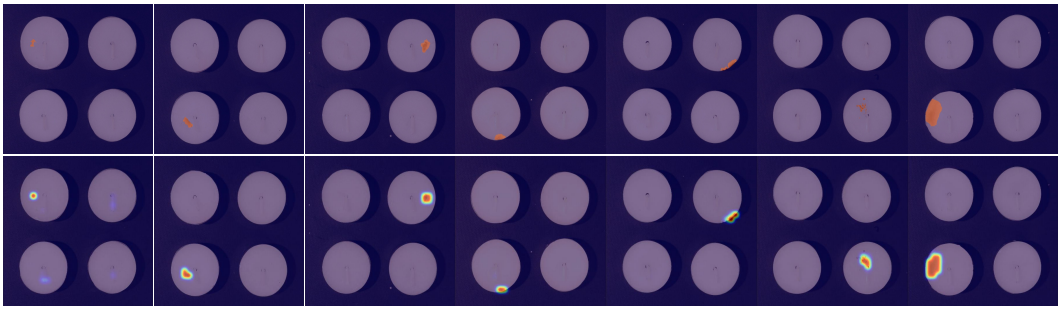

Figure S12. Segmentation results on the candle class from VisA. Ground truth regions are marked in red (top row), and predicted anomaly maps are shown below.

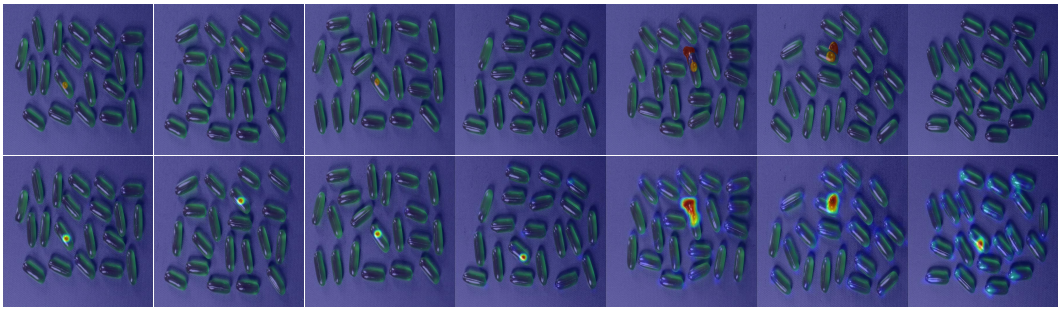

Figure S13. Segmentation results on the capsules class from VisA. Ground truth regions are marked in red (top row), and predicted anomaly maps are shown below.

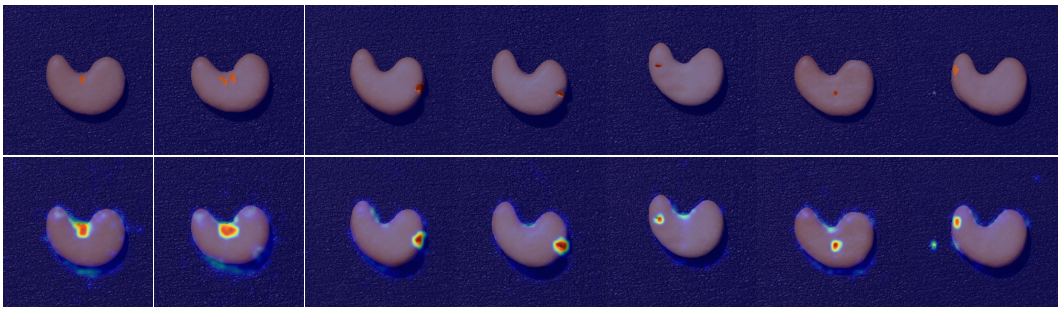

Figure S14. Segmentation results on the cashew class from VisA. Ground truth regions are marked in red (top row), and predicted anomaly maps are shown below.

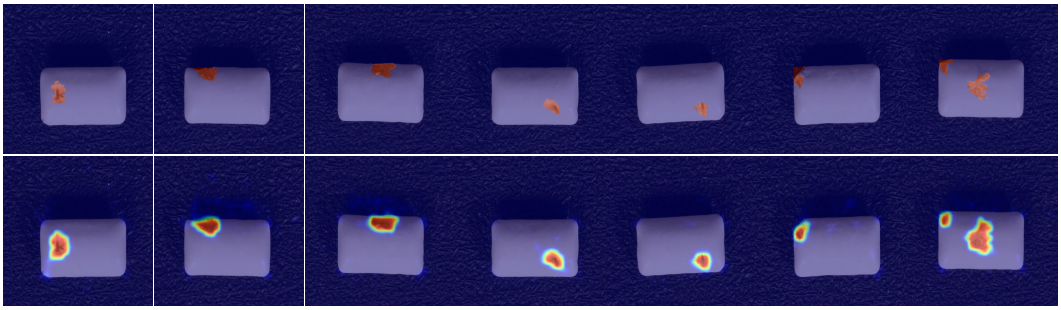

Figure S15. Segmentation results on the chewinggum class from VisA. Ground truth regions are marked in red (top row), and predicted anomaly maps are shown below.

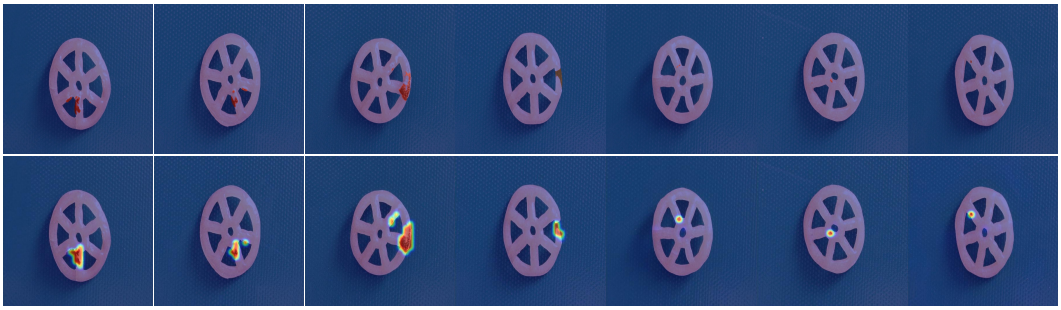

Figure S16. Segmentation results on the fryum class from VisA. Ground truth regions are marked in red (top row), and predicted anomaly maps are shown below.

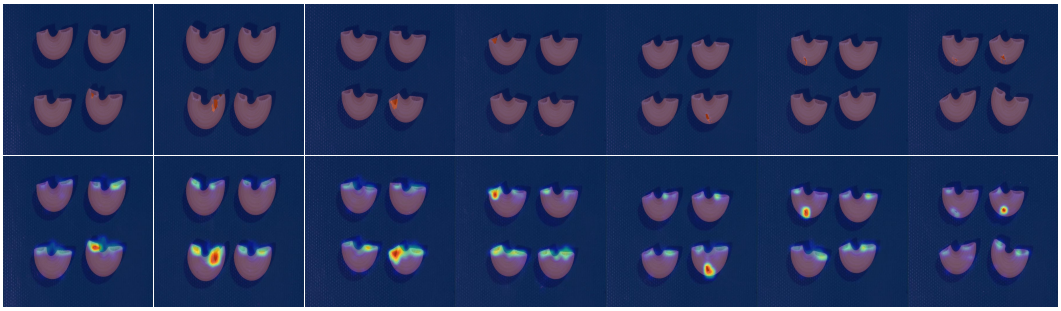

Figure S17. Segmentation results on the macaroni1 class from VisA. Ground truth regions are marked in red (top row), and predicted anomaly maps are shown below.

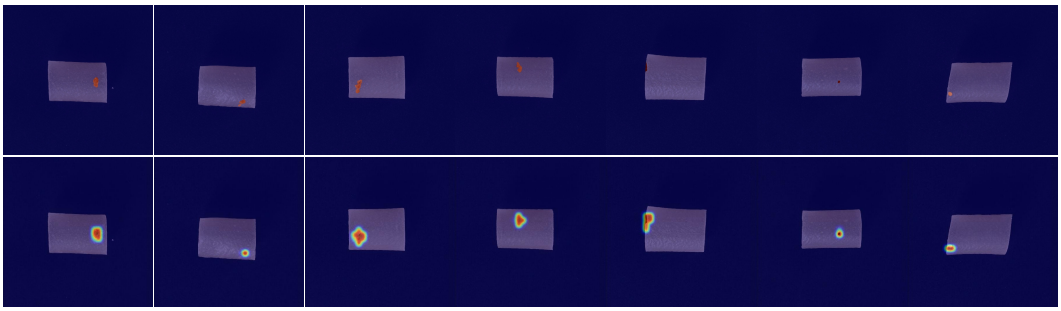

Figure S18. Segmentation results on the pipe fryum class from VisA. Ground truth regions are marked in red (top row), and predicted anomaly maps are shown below.

C. Additional Analysis of Efficiency, Backbones, and Group Merging

C.1. Comparison of Computational Efficiency

Table S1. **Comparison of GPU memory usage, training time per epoch, and inference time across different methods.**

Method	Training Memory Usage(GB)	Training Time(Min / Epoch)	Inference Time(ms)	Image-level	Pixel-Level
AnoamlyCLIP	**2.90**	4:59	**12.5**	(91.5, 92.8, 96.2)	(91.1, 34.5)
AdaCLIP	10.55	10:27	138.2	(92.0, 92.7, 96.4)	(86.8, 38.1)
Bayes-PFL	14.90	22:25	321.9	(92.3, 93.1, 96.7)	(91.8, 48.3)
Ours	4.58	**3:33**	148.0	(**92.7**, **94.0**, **97.1**)	(**92.4**, **50.4**)

Table S1 compares the computational efficiency of our method with existing vision–language anomaly detection baselines in terms of GPU memory usage, training time per epoch, and inference time. All metrics are measured with a batch size of 1, and all methods are trained on the VisA dataset and evaluated on the MVTec AD dataset. AnomalyCLIP shows the lowest training memory footprint (2.90 GB) and fastest inference (12.5 ms), but its detection performance is clearly inferior to later methods. AdaCLIP and Bayes-PFL require substantially more memory (10.55 GB and 14.90 GB, respectively) and much longer training time, while still underperforming our approach. In contrast, our method achieves competitive efficiency, using only 4.58 GB of memory and 3:33 minutes per epoch with an average inference time of 148.0 ms, while attaining the best image-level and pixel-level results (92.7/94.0/97.1 and 92.4/50.4). These results indicate that our framework strikes a favorable balance between computational cost and anomaly detection performance.

C.2. Backbone Comparison: CLIP, DINOv2, and DINOv3

Table S2. **Comparison of different image encoders used as backbones.** We report backbone parameter counts and anomaly detection performance at the image level (AUROC, F1, AP) and pixel level (AUROC, AP).

Method	Image Encoder	Back. Params (M)	**Image-level** (AUROC, F1, AP)	**Pixel-level** (AUROC, AP)
AnoamlyCLIP	CLIP-L	$\approx$ 304M	(82.1, 80.4, 85.4)	(95.5, 21.3)
AdaCLIP	CLIP-L	$\approx$ 304M	(83.0, 81.6, 84.9)	(95.1, 29.2)
Bayes-PFL	CLIP-L	$\approx$ 304M	(87.0, 84.1, 89.2)	(95.6, 29.8)
Ours-CLIP	CLIP-L	$\approx$ 304M	(86.1, 83.8, 87.7)	(95.0, 26.4)
Ours-DINOv2	DINOv2-L	$\approx$ 304M	(87.0, 83.8, 90.1)	(95.3, 31.0)
Ours-DINOv3	DINOv3-L	$\approx$ 304M	(88.3, 85.0, 91.8)	(96.6, 34.4)

As mentioned earlier, we adopt DINOv3 [19] as our image encoder. We make this choice for the following reasons. First, DINOv3 provides richer visual representations than the CLIP image encoder [18], which is beneficial for capturing fine-grained semantic cues in images. However, since DINOv3 is not explicitly aligned with text features, it can be less favorable for conventional zero-shot vision-language tasks. In our setting, we deliberately favor the former aspect of this trade-off: our primary objective is to construct an image anchor that reduces dependence on text features and to learn image representations that are semantically meaningful on their own. In this context, DINOv3 is a natural choice, as it allows us to exploit strong visual semantics while shifting the burden away from text-guided supervision. Moreover, aligning DINO-style visual features with text features is itself a non-trivial problem that has been treated as a separate research task in prior work [2, 14]. Building such alignment into our framework is a key component of our method and constitutes an essential part of our novelty.

Despite these considerations, we also compare against the original CLIP-based setting to ensure a fair evaluation. As shown in Table S2, CLIP, DINOv2, and DINOv3 have comparable numbers of parameters, yet our method consistently outperforms most existing approaches even when we replace the image encoder with CLIP or DINOv2, when training on the MVTec AD dataset and evaluating on the VisA dataset. This result not only demonstrates the robustness of our framework to the choice of backbone, but also suggests that future advances in image encoders can be readily leveraged to further improve our model's performance.

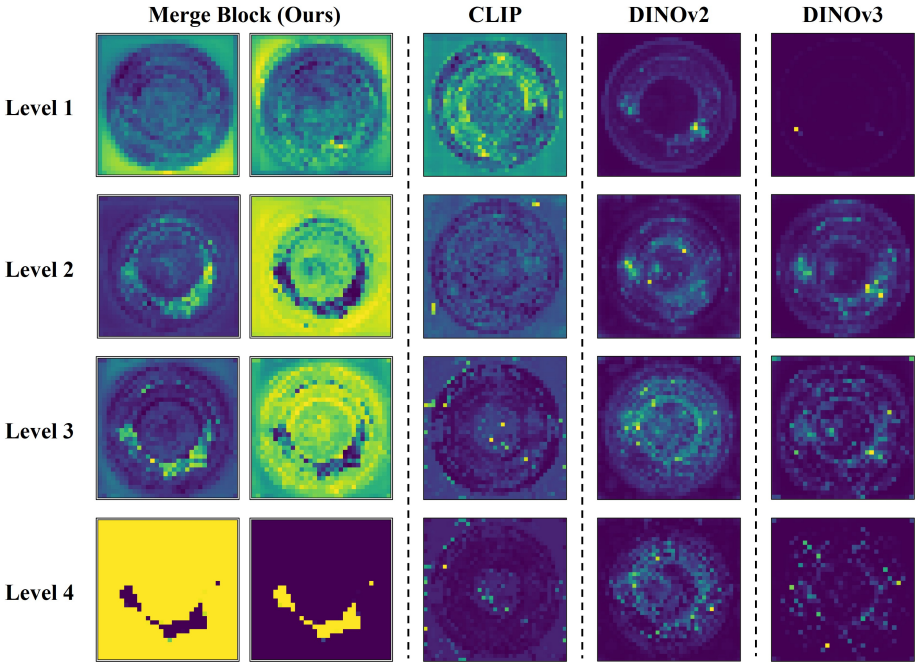

Figure S19. **Multi-level attention maps of our merge block versus CLIP, DINOv2, and DINOv3.**

C.3. Effects of Hierarchical Group Merging

The novelty of the proposed *Hierarchical Group Merging* can be summarized as follows. (1) Instead of treating patch features from the backbone network as independent tokens without explicit normal/anomaly semantics, our top-down grouping strategy progressively aggregates them into semantically meaningful normal and anomaly groups. This allows the model to form image-level semantic features that are independent of text prompts. (2) By ultimately constructing an image anchor, our method reduces the reliance on text anchors and facilitates the formation of a more robust decision boundary in unseen domains. To validate these advantages, in this section we compare the attention maps extracted from vanilla CLIP and DINOv3 with those obtained from our *Hierarchical Group Merging*.

Figure S19 qualitatively compares the attention maps of our *Hierarchical Group Merging* with those of CLIP, DINOv2, and DINOv3. For the baselines, we extract attention from the same layer used for patch tokens, using the CLS token as the query. While these backbone features produce noisy and scattered responses across levels, our merge block, which is built on hierarchically grouped tokens, progressively concentrates attention on the true anomalous region and yields compact maps that better align with the ground-truth defect.

D. Failure Cases

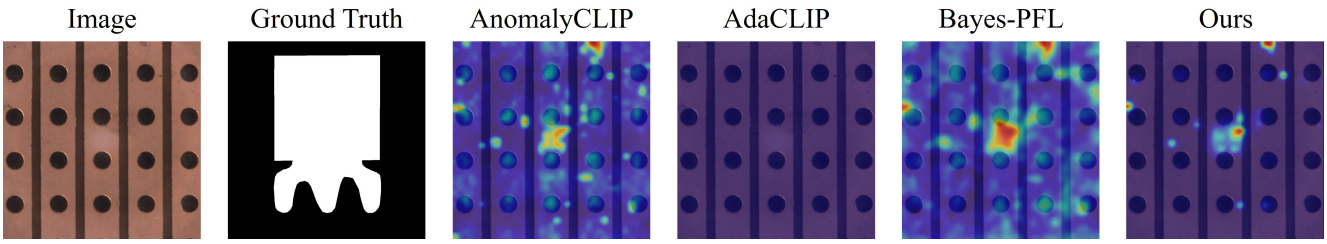

Figure S20. **Qualitative comparison of failure cases on an unseen domain.** The anomaly corresponds to a logical "misplacement" rather than a clear appearance change. All zero-shot methods, including ours, have difficulty localizing this type of anomaly, highlighting a common limitation of current zero-shot anomaly detection approaches.

Figure S20 illustrates a representative failure case. Due to the zero-shot setting, our model sometimes struggles with logical anomalies that do not manifest as clear low-level appearance changes, such as "misplaced" or subtly shifted components in unseen domains. In such cases, the visual evidence is weak and the anomaly can only be identified by understanding the global layout or functional semantics of the scene, which goes beyond what our image anchor is explicitly trained to capture. We note, however, that other zero-shot anomaly detection methods (e.g., AnomalyCLIP, AdaCLIP, and Bayes-PFL) exhibit similar failure patterns on this example, suggesting that handling such high-level logical anomalies remains an open challenge for zero-shot AD in general.

References

[1] Toshimichi Aota, Lloyd Teh Tzer Tong, and Takayuki Okatani. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5564–5572, 2023. 1, 2

[2] Luca Barsellotti, Lorenzo Bianchi, Nicola Messina, Fabio Carrara, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, and Rita Cucchiara. Talking to dino: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22025–22035, 2025. 8

[3] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4):1038–1059, 2021. 1

[4] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015. 1, 2

[5] Jakob Božič, Domen Tabernik, and Danijel Skočaj. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. *Computers in Industry*, 129:103459, 2021. 1, 2

[6] 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. 3

[7] Xuhai Chen, Yue Han, and Jiangning 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. 3

[8] Haifan Gong, Guanqi Chen, Ranran Wang, Xiang Xie, Mingzhi Mao, Yizhou Yu, Fei Chen, and Guanbin Li. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 257–261. IEEE, 2021. 1, 2

[9] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016. 1, 2

[10] Steven A Hicks, Debesh Jha, Vajira Thambawita, Pål Halvorsen, Hugo L Hammer, and Michael A Riegler. The endotect 2020 challenge: evaluation and comparison of classification, segmentation and inference time for endoscopy. In *International Conference on Pattern Recognition*, pages 263–274. Springer, 2021. 1, 2

[11] 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. 3

[12] Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT)*, pages 66–71. IEEE, 2021. 1, 2

[13] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pages 451–462. Springer, 2019. 1, 2

[14] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24905–24916, 2025. 8

[15] 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. 3

[16] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06. IEEE, 2021. 1, 2

[17] Zhen Qu, Xian Tao, Xinyi Gong, Shichen Qu, Qiyu Chen, Zhengtao Zhang, Xingang Wang, and Guiguang Ding. Bayesian prompt flow learning for zero-shot anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30398–30408, 2025. 3

[18] 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. 8

[19] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 8

[20] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015. 1, 2

[21] Matthias Wieler and Tobias Hahn. Weakly supervised learning for industrial optical inspection. In *DAGM symposium in*, page 11, 2007. 1, 2

[22] Haomin Yu, Qingyong Li, Yunqiang Tan, Jinrui Gan, Jianzhu Wang, Yangli-ao Geng, and Lei Jia. A coarse-to-fine model for rail surface defect detection. *IEEE Transactions on Instrumentation and Measurement*, 68(3):656–666, 2018. 1, 2

[23] 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. 3

[24] 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. 1