# Reasoning-Driven Anomaly Detection and Localization with Image-Level Supervision

## Supplementary Material

Table 4. Ablation study of the CGRO hyperparameters, including the number of top-$k$ reasoning tokens, the Jaccard threshold $\delta_1$ and $\delta_2$. Base model is *Qwen2.5-VL-7B + CGRO*. Results are image-level and pixel-level metrics on the WFDD dataset.

| hyperparameter | | | Image-level | | | Pixel-level | | |
|---|---|---|---|---|---|---|---|---|
| $\delta_1$ | $\delta_2$ | $k$ | AUROC | AUPR | ACC | AUROC | AUPR | ACC |
| 0.3 | 0.1 | 2 | 76.0 | 76.0 | 75.6 | 67.9 | 5.6 | 97.5 |
| 0.3 | 0.1 | 4 | 76.1 | 76.6 | 75.8 | 67.0 | 5.4 | 97.6 |
| 0.3 | 0.3 | 3 | 77.8 | 77.7 | 77.1 | 68.7 | 5.9 | 97.6 |
| 0.5 | 0.1 | 3 | 73.3 | 74.4 | 72.1 | 69.5 | 6.3 | 97.5 |
| 0.3 | 0.1 | 3 | **79.9** | **78.8** | **79.8** | **70.5** | **7.4** | **97.8** |

Table 5. Ablation study comparing different threshold configurations for $\tau_{\text{i}}$ and $\tau_{\text{t}}$. Base model is *Qwen2.5-VL-7B + CGRO*. Results are pixel-level metrics on the WFDD dataset.

| $\tau_{\text{i}}$ | $\tau_{\text{t}}$ | AUROC | AUPR | ACC |
|---|---|---|---|---|
| Maximum Curvature | Maximum Curvature | **70.7** | 6.5 | 97.5 |
| Median | Maximum Curvature | 70.4 | 6.3 | 97.5 |
| Median | Median | 70.4 | **7.5** | 97.6 |
| Maximum Curvature | Median | 70.5 | 7.4 | **97.8** |

## 7. Additional Qualitative Results

**Impact of consistency reward.** As shown in Fig. 5, additional qualitative results highlight the impact of the consistency reward. By introducing the consistency reward, the model aligns reasoning with visual evidence, focusing attention on true anomalies. This alignment leads to consistent improvements in both detection and localization, demonstrating that the consistency reward effectively enforces coherent, visually grounded reasoning, which enhances anomaly localization performance.

**Comparison with other methods.** Fig. 6 presents a qualitative comparison of GPT-4.1 [3], Triad [5], AnomalyGPT [1], EIAD [6], and our method. GPT-4.1 [3] and Triad [5] exhibit strong reasoning capabilities but fail to localize anomalies at the pixel level. In contrast, our method, trained solely with image-level supervision, achieves accurate anomaly detection and interpretable reasoning, while generating high-fidelity pixel-level localization maps. Notably, our method yields qualitative pixel-level localization on par with AnomalyGPT [3] and EIAD [6], despite their reliance on additional vision modules and pixel-level supervision.

## 8. Additional Ablation Studies

**Hyperparameter ablation for CGRO.** The CGRO module has three hyperparameters: (1) $k$, the number of top reasoning tokens used for spatial consensus, (2) $\delta_1$, a Jaccard threshold enforcing high consensus on anomalous images, and (3) $\delta_2$, a threshold enforcing low consensus on normal images. Larger $k$ increases token coverage but risks including noisy reasoning; $\delta_1$ and $\delta_2$ jointly shape the reward landscape to favor discriminative reasoning patterns.

Tab. 4 reports performance over varying $(k, \delta_1, \delta_2)$. We select $k = 3$, $\delta_1 = 0.3$, and $\delta_2 = 0.1$, which yield the most consistent gains across both image-level and pixel-level anomaly detection metrics.

**Hyperparameter ablation for ReAL.** The ReAL module involves two thresholds: $\tau_{\text{i}}$, which filters tokens by spatial entropy (lower entropy $\Rightarrow$ more spatially concentrated attention), and $\tau_{\text{t}}$, which filters by semantic relevance (higher $S_{\text{T}}^r \Rightarrow$ stronger alignment with the anomaly query). To select these thresholds, we sort the tokens in ascending order of $S_{\text{I}}^r$ and $S_{\text{T}}^r$. We then determine the thresholds based on the maximum curvature and median strategies.

As shown in Tab. 5, we ultimately select $\tau_{\text{i}}$ as the threshold determined by maximum curvature and $\tau_{\text{t}}$ as the median.

**Ablation on ReAL and CGRO.** To complement the averaged results reported in Table 2 of the main paper, we provide the complete dataset-wise performance metrics for the ablation study on ReAL and CGRO. As shown in Tab. 6, ReAL consistently improves localization by extracting anomaly-related tokens from the autoregressive reasoning process and aggregating their attention responses to produce pixel-level anomaly maps, which effectively enhances the model's pixel-level segmentation capability. Building on this, CGRO leverages reinforcement learning to further strengthen both detection and localization by aligning the reasoning process with visual evidence. The dataset-level breakdown shows clear gains at both the image level and the pixel level, demonstrating the effectiveness of our method and the complementary roles of ReAL and CGRO within the overall framework.

**Ablation on token selection in ReAL.** To complement the averaged results reported in Table 3 of the main paper, we provide the full dataset-wise pixel-level localization metrics for the ablation on token selection in the ReAL module. As shown in Tab. 7, combining both semantic relevance $S_{\text{T}}$ and spatial entropy $S_{\text{I}}$ with composite weighting leads to the strongest and most stable performance across all datasets, confirming that spatial concentration and semantic relevance provide complementary cues for selecting anomaly-relevant reasoning tokens.


Figure 5. Qualitative comparison of *Qwen2.5-VL-7B*, *Qwen2.5-VL-7B+R1*, and *Qwen2.5-VL-7B+CGRO*.

## 9. Discussion

**MVTec-COT Construction.** Following the established VisionR1 pipeline [2], our data generation process incorporates a rigorous human-verification stage to filter reasoning signals and minimize noise, thereby ensuring high-quality data synthesis. All anomaly-related concepts—including scene context, defect types, and spatial locations—are grounded in the publicly annotated MMAD benchmark [4] to guaranty biological and physical validity. To ensure the integrity of our results, the MVTec-COT dataset is strictly reserved for evaluation and is never utilized during training or optimization, effectively preventing data leakage. Furthermore, the complete dataset will be publicly released to


Figure 6. Comparison of GPT-4.1, Triad, AnomalyGPT, EIAD, and our method.

support reproducibility and further research in the community.

**Failure cases.** Although our framework achieves competitive performance using only image-level supervision, we observe several typical failure cases (Fig. 7). These include the model's attention being distracted by salient but non-defective regions, coarser predicted anomaly boundaries compared to dense-supervision methods, and missed detections in multi-spot scenarios. Furthermore, incorrect image-level predictions inherently prevent the generation of

Table 6. Dataset-wise results image-level and pixel-level performance for the ReAL and CGRO ablation study. Ablation on ReAL and CGRO. *Vanilla* denotes the original *Qwen2.5-VL-7B*. All metrics are macro-averaged across four AD datasets. Detection results are reported as (AUROC, AUPR).

| Model | Image-level | | | | | Pixel-level | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AVG | SDD | DTD | WFDD | MVTec | AVG | SDD | DTD | WFDD | MVTec |
| Vanilla | 63.4, 59.4 | 65.0, 13.7 | 68.3, 82.3 | 60.5, 63.7 | 59.6, 77.7 | 64.7, 2.6 | 60.6, 0.1 | 79.7, 4.7 | 58.3, 1.6 | 60.1, 4.1 |
| Vanilla + ReAL | 63.4, 59.4 | 65.0, 13.7 | 68.3, 82.3 | 60.5, 63.7 | 59.6, 77.7 | 61.7, 5.1 | 56.6, 0.1 | 76.0, 13.2 | 56.5, 2.0 | 57.5, 5.0 |
| Vanilla + CGRO | **83.9**, **76.7** | 81.4, 44.3 | 94.5, 96.4 | 79.9, 78.8 | 79.8, 87.1 | 72.7, 5.9 | 72.5, 0.2 | 87.1, 13.1 | 63.3, 2.7 | 67.9, 7.5 |
| Full | **83.9**, **76.7** | 81.4, 44.3 | 94.5, 96.4 | 79.9, 78.8 | 79.8, 87.1 | **80.7**, **13.3** | 82.3, 2.6 | 94.2, 26.4 | 70.5, 7.4 | 75.6, 16.6 |

Table 7. Dataset-wise pixel-level results for the ablation on token selection in the ReAL module. Base model is *Qwen2.5-VL-7B + CGRO*. We compare four configurations: using all tokens, filtering with spatial entropy $S_{\mathrm{I}}$, filtering with semantic relevance $S_{\mathrm{T}}$, and combining both criteria. Results are pixel-level metrics reported on MVTec-AD, WFDD, SDD, and DTD, complementing Table 3 in the main paper.

| Spatial entropy $S_{\mathrm{I}}$ | Semantic relevant $S_{\mathrm{T}}$ | AVG | | | SDD | | | DTD | | | WFDD | | | MVTec | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUROC | AUPR | ACC | AUROC | AUPR | ACC | AUROC | AUPR | ACC | AUROC | AUPR | ACC | AUROC | AUPR | ACC |
| | | 72.7 | 5.9 | 92.6 | 72.5 | 0.2 | 97.0 | 87.1 | 13.1 | 97.9 | 63.3 | 2.7 | 95.6 | 67.9 | 7.5 | 80.0 |
| ✓ | | 74.1 | 11.8 | 89.2 | 75.0 | 1.0 | 99.7 | 92.4 | 31.1 | 98.6 | 55.0 | 1.7 | 68.6 | 74.0 | 13.4 | 90.0 |
| | ✓ | 76.7 | **14.0** | 90.2 | 75.2 | 1.5 | 99.8 | 93.5 | 35.0 | 98.6 | 60.4 | 3.3 | 72.1 | 77.5 | 16.3 | 90.3 |
| ✓ | ✓ | **80.7** | 13.3 | **97.1** | 82.3 | 2.6 | 99.8 | 94.2 | 26.4 | 98.6 | 70.5 | 7.4 | 97.8 | 75.6 | 16.6 | 92.2 |

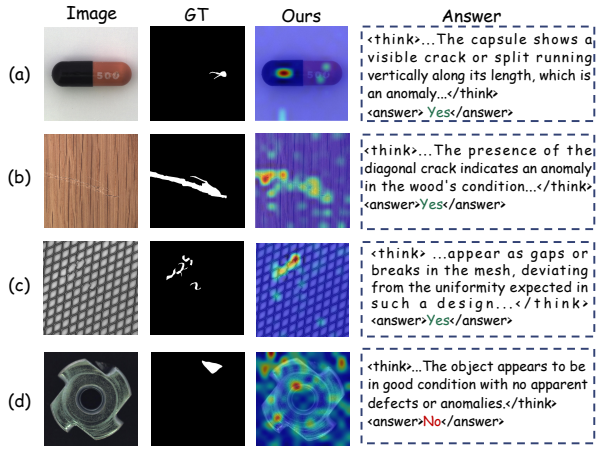


Figure 7. Typical failure cases. (a) Attention distracted by salient but non-defective regions; (b) coarser predicted anomaly boundaries; (c) missed detections in multi-spot scenarios; and (d) overall localization failure caused by incorrect image-level predictions and the subsequent lack of valid reasoning.

a valid reasoning process, ultimately leading to an overall localization failure.

**Limitations.** Our framework reflects a trade-off between annotation efficiency and localization performance. While relying solely on image-level supervision and avoiding auxiliary vision modules improves scalability, a performance gap remains compared with fully supervised methods, particularly in boundary fidelity and attention misalignment in complex scenes. In addition, real-time deployment is constrained by the autoregressive decoding latency of MLLM backbones, as reasoning-based inference remains slower than conventional industrial anomaly detection pipelines.

## References


[1] 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*, pages 1932–1940, 2024. 1

[2] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025. 2

[3] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1

[4] 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. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[5] Yuanze Li, Shihao Yuan, Haolin Wang, Qizhang Li, Ming Liu, Chen Xu, Guangming Shi, and Wangmeng Zuo. Triad: Empowering lmm-based anomaly detection with expert-guided region-of-interest tokenizer and manufacturing process. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21917–21926, 2025. 1

[6] Zongyun Zhang, Jiacheng Ruan, Xian Gao, Ting Liu, and Yuzhuo Fu. Eiad: Explainable industrial anomaly detection via multi-modal large language models. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2025. 1