Towards Smarter Segmentation: Improving SAM's Anomaly Understanding with RLHF

Wan Wang University of Minnesota wang9814@umn.edu
Yu-Tong Chuang University of Minnesota chuan120@umn.edu
Lulin Liu University of Minnesota liu02721@umn.edu
Yiu Chang University of Minnesota chan2586@umn.edu

Abstract

Industrial anomaly segmentation poses unique challenges due to the subtle, low-contrast nature of defects and the scarcity of high-quality annotations. While foundation models like Segment Anything Model (SAM) offer strong generalization in object-centric segmentation, they often fail to localize fine-grained anomalies. In this work, we make an initial attempt to adapt SAM to this setting via a two-stage pipeline: supervised fine-tuning (SFT) on a curated demonstration set, followed by preference-based finetuning using reinforcement learning from human feedback (RLHF). To facilitate this, we construct a small-scale comparison dataset with human-annotated segmentation preferences. Our method, VisionSAM, demonstrates modest yet consistent gains over the SFT baseline, achieving a +0.43% IoU and +0.69% Dice improvement on held-out categories. In contrast, directly applying RLHF to the original SAM results in performance degradation, underscoring the importance of task-specific grounding. Compared to existing fine-tuning strategies such as adapter-based or CLIP-integrated approaches, our method is the first to systematically introduce RLHF into dense prediction tasks. While results vary across prompt types and anomaly categories, our findings suggest that human-aligned feedback offers a promising supervision signal in low-data regimes, and opens a novel direction for human-in-the-loop vision foundation model adaptation.[1] [2]

1 Introduction

Problem Definition. In industrial settings, there is a growing demand for robust and interpretable anomaly detection algorithms to reduce reliance on manual inspection and enable end-to-end automation. However, a major challenge lies in the limited availability of high-quality segmentation data required for supervised learning. Industrial anomaly datasets typically suffer from severe class imbalance, where defective samples are rare, and collecting pixel-level annotations is both costly and time-consuming, often requiring domain-specific expertise. These limitations may hinder the advancement of supervised anomaly segmentation methods, especially in cases where model performance relies on access to large and diverse annotated datasets.

Objective. There are two perspectives for optimizing industrial defect detection. The model-centric perspective assumes a baseline level of data quality and focuses on improving model architectures, training strategies, or post-processing techniques to enhance performance. The data-centric perspective, on the other hand, emphasizes improving the quality and utility of the data itself. Our project focuses on the latter. We believe that obtaining higher-quality data at a lower cost can significantly improve model performance while reducing the resources required for training and iteration. To this end, we explore whether reinforcement learning from human feedback (RLHF) can inject anomaly-relevant knowledge into a vision foundation model with strong segmentation capabilities. Such a model could assist in generating more reliable pseudo-labels and facilitate scalable, data-efficient workflows for industrial inspection.

Current Gap. Existing approaches to adapting SAM for anomaly detection primarily focus on task-specific fine-tuning strategies. Notably, adapter-based methods such as HQ-SAM inject domain knowledge via lightweight modules, enabling improved performance on subtle industrial defects without retraining the entire model. Other efforts, like SAM-CLIP, incorporate knowledge distillation from multi-modal models (e.g., CLIP) to enhance segmentation by leveraging semantic cues through text-image alignment. While these

[1] Our code is available at https://github.com/slmowan/sam-finetune

[2] Please contact wang9814@umn.edu for the access permission if code check is needed.

methods have shown promising results, they still face key limitations: (1) they rely heavily on static supervision and lack mechanisms to incorporate dynamic, human-informed feedback; (2) they often struggle with precisely delineating low-contrast or ambiguous anomalies; and (3) they do not fundamentally address the model's lack of an explicit "understanding" of what constitutes an anomaly. To date, no prior work has explored the use of human preference signals—such as those enabled by reinforcement learning from human feedback (RLHF)—to guide and refine SAM's behavior in anomaly segmentation tasks, leaving a promising direction underexplored.

Significance. This work is relevant to both industry practitioners and AI researchers. In industrial settings, current anomaly detection systems often rely on fully supervised training with dense pixel-level annotations, which are costly, time-consuming, and difficult to scale due to the rarity and subtlety of many defects. Vision Foundation Models such as SAM offer general segmentation capabilities but lack sensitivity to domain-specific anomalies, often under-performing on low-contrast or fine-grained defects. For researchers, this study offers a data-efficient alternative to traditional annotation-heavy methods. By introducing a human-guided adaptation framework, this work enables more practical fine-tuning of foundation models and has the potential to improve quality control, reduce waste, and enhance automation efficiency in industrial environments.

Impact. The main contributions of this report are summarized below:

1. We make an initial attempt to explore the use of RLHF for industrial anomaly segmentation, illustrating its potential to enhance the adaptation of vision foundation models in domains where traditional supervision is limited.

2. As a first attempt, we construct a small-scale, expert-annotated anomaly segmentation dataset to begin addressing the lack of high-quality industrial data and to facilitate exploration of human-in-the-loop learning under limited supervision. The dataset can be accessed here.

2 Related Work

Vision Foundation Models. In recent years, VFMs have revolutionized computer vision by leveraging large-scale datasets and transformer-based architectures, mirroring the paradigm shift seen in Natural Language Processing (NLP) with models like GPT and BERT. Just as large language models (LLMs) have moved beyond task-specific training to enable zero-shot and few-shot learning across diverse NLP tasks, VFMs are designed to be general-purpose models capable of adapting to a wide range of downstream vision tasks, including object detection, segmentation, and image understanding. Two prominent VFMs that have significantly influenced vision-language research are SAM (Kirillov et al., 2023) and Contrastive Language-Image Pretraining (CLIP) (Radford et al., 2021)

SAM, developed by Meta AI, is a foundation model for image segmentation with unprecedented zero-shot capability. Trained on over 11 million images and 1 billion masks, it enables segmentation via prompt-based inputs such as points, bounding boxes, or textual descriptions, similar to how LLMs generate text from prompts. SAM is chosen as the pre-trained model for this study due to its pixel-level segmentation capability, which is crucial for fine-grained anomaly detection.

Current finetuning methods for Vision Language Models in anomaly detection. Existing fine-tuning approaches for vision-language models can be broadly categorized into three methods: 1) Adapter-based fine-tuning, which injects domain-specific knowledge into large models but may not fully capture anomaly characteristics such as HQ-SAM(Ke et al., 2023a). 2) Knowledge distillation, which integrates multiple models (e.g., combining SAM and CLIP into a unified vision transformer), yet balancing knowledge transfer remains a challenge such as SAM-CLIP(Wang et al., 2024a). 3) Zero-shot methods, which leverage pre-trained features but often fail to precisely segment anomalies due to limited domain adaptation such as WinCLIP(Jeong et al., 2023). Despite their effectiveness in some domains, these methods still struggle to align segmentation results with human intent, especially in anomaly detection scenarios where understanding subtle defects is critical. Empirical evidence suggests that even fine-tuned SAM models lack a deep understanding of what constitutes an anomaly and often fail to generate accurate segmentation masks.

RLHF in T2I Generation. RLHF enhances text-to-image (T2I) generation by aligning outputs with human preferences. While models like Stable Diffusion and DALL·E leverage large-scale datasets,

they often produce misaligned results, such as low fidelity, unrealistic details, or incorrect compositions. RLHF refines generation through iterative learning from human feedback, improving coherence, aesthetics, and fine-grained control. A work proposed by Liang et al.(Liang et al., 2024) introduces Rich Human Feedback (RHF) for T2I generation, proposing a fine-grained preference learning framework that refines generative outputs beyond simple binary reward signals. Instead of traditional reinforcement learning based on sparse rewards, RHF incorporates structured human feedback, allowing the model to learn nuanced quality distinctions in generated images. The results demonstrate that RLHF significantly improves image realism, adherence to textual descriptions, and user satisfaction.

3 Method

3.1 Datasets

We mainly utilize the VISION dataset for our project, which is a comprehensive and realistic benchmark specifically designed for industrial anomaly detection and segmentation. It includes 14 representative inspection categories (e.g., Cable, Capacitor), with approximately 18,000 high-resolution images captured across diverse manufacturing settings. The dataset's anomalies are often subtle, low-contrast, and embedded in complex visual contexts, making it a challenging and representative setting for evaluating the effectiveness of our proposed methods.

We also used MVTec AD dataset in our phase 1 analysis.

3.2 Motivating Fine-Tuning: Where SAM Falls Short

Prior to the generation stage of the RLHF-Based Segmentation Fine-Tuning Workflow, we first evaluate SAM's segmentation performance on the MVTec AD dataset to analyze its failure modes anMehd performance boundaries under diverse industrial anomaly scenarios. Specifically, we uniformly sample 210 test images from 14 distinct anomaly categories (e.g., Cable, Pill, Bottle, Metal Nut), with 10 to 15 representative instances per category. This sampling strategy ensures broad coverage across variations in defect shape, texture, and scale, providing a solid foundation for assessing SAM's zero-shot generalization capability.

Preliminary Results (will disclose in next section) show that despite its strong zero-shot segmentation ability, SAM struggles with anomaly detection, lacking an explicit understanding of domain-specific defects in some situations. Existing fine-tuning methods attempt to adapt SAM but still face challenges in capturing the nuanced nature of anomalies. This highlights the need for a new fine-tuning approach that integrates human expertise without relying on extensive manual annotations.

We hypothesize that human preference signals—particularly when leveraged through reinforcement learning from human feedback (RLHF)—can serve as an effective supervision alternative by guiding the model toward segmentation outcomes that align more closely with human judgment. Given SAM's strong base segmentation capability, we believe that even limited preference-based feedback can steer the model to better capture subtle anomaly patterns that traditional supervision may overlook.

3.3 RLHF-Based Segmentation Fine-Tuning Workflow

We adopt a three-stage RLHF-based fine-tuning workflow inspired by InstructGPT framework (Ouyang et al., 2022), with necessary adaptations for the segmentation task. While InstructGPT focuses on aligning language models with human preferences through reward modeling, our workflow extends this concept to the vision domain, aiming to align segmentation outputs with human perveptual judgments. As illustrated in Figure 1, the overall workflow consists of three sequential stages: (1) **Generation**, where a fine-tuned SAM model produces multiple segmentation masks in response to human-designed prompts; (2) **Annotation**, where human annotators rank these masks based on a rubric, creating preference-labeled data; and (3) **Optimization**, where we apply Direct Preference Optimization (DPO) to further align the model's predictions with human preferences.

During the Generation stage, we first construct a supervised fine-tuning (SFT) dataset to initialize a domain-adapted SAM model. To ensure high signal quality for learning, we manually filtered approximately 500 images from the VISION dataset, selecting samples that (1) exhibit clearly visible anomalies and (2) have minimal background clutter. These images span multiple defect categories and were selected to balance shape, texture, and defect scale diversity. For each selected image, we

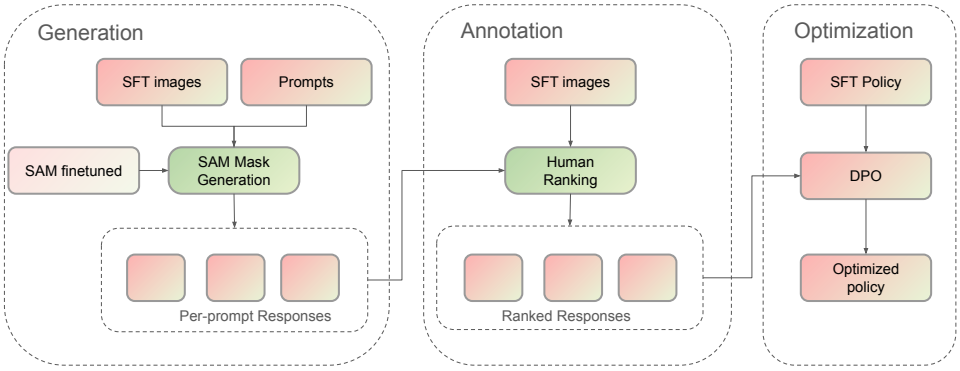


Figure 1: Overview of our RLHF-based fine-tuning workflow for segmentation. The pipeline consists of three stages: (1) Generation — SAM generates multiple masks per prompt; (2) Annotation — humans rank the masks based on quality; (3) Optimization — preference data is used to fine-tune the model via Direct Preference Optimization (DPO).

use a sparse box prompt centered around the annotated defect region, simulating minimal human input. These prompt-image pairs (see Figure 2) are then fed into a lightly modified version of SAM (ViT-B variant), where we freeze the image encoder and fine-tune only the mask decoder layers. The model is trained using classical binary cross-entropy loss. During inference, we retain SAM's native multi-mask decoding mechanism, generating three candidate masks per prompt to facilitate downstream human preference annotation.

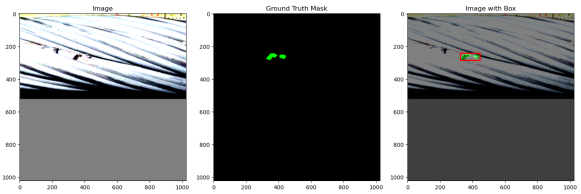


Figure 2: An example of a SAM-compatible triplet training sample.

In the Annotation stage, we incorporate human-in-the-loop feedback to create high-quality preference data. All four authors participated in the annotation process. To ensure consistency, we first reviewed several anomaly segmentation benchmarks (e.g., MVTec AD, DAGM) and conducted a brief internal calibration session to align on quality criteria. Based on this, we collaboratively designed a rubric focusing on four key aspects: (1) coverage of the anomalous region, (2) boundary accuracy, (3) over-segmentation or under-segmentation, and (4) noise or spurious artifacts. For each prompt and image pair, the SAM-based model generates three candidate masks with varying internal confidence scores. Annotators independently examined these outputs and used the rubric to rank them. From each set, we selected the most preferred and least preferred masks, forming comparison triplets in the format: (prompt, preferred mask, rejected mask). These ranked responses serve as supervision signals in the preference-based fine-tuning stage.

In the final optimization stage, the goal is to fine-tune the SAM model using the comparision dataset from the annotation stage. Each training instance consists of an input image, a bounding box prompt, and a pair of segmentation masks: a preferred mask (chosen) and a rejected mask. We frame this as a preference learning task and adopt Direct Preference Optimization (DPO) to align the model's outputs with human feedback. During training, the model receives an RGB image resize to 1024×1024, a bounding box prompt in the format $[x1, y1, x2, y2]$, and the two corresponding masks, both converted to binary tensors and resized to 256×256 for stable loss computation. The SAM mask decoder processes the image and prompt to produce a predicted segmentation mask. This prediction is then compared against both the preferred and rejected masks using binary cross-entropy (BCE) to compute scalar reward scores r_{chosen} and r_{rejected}, respectively.

These reward values are passed into the DPO loss function:

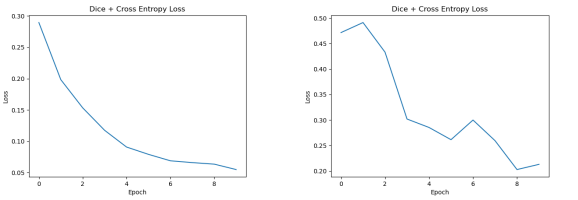
$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\sigma\left(\beta\left(r^{(i)}_{\text{chosen}} - r^{(i)}_{\text{rejected}}\right)\right)\right) \tag{1}$$

where $\sigma(\cdot)$ denotes the sigmoid function and β is a temperature hyperparameter (set to 1.0 in our experiments) that controls the sharpness of preference distinctions.

This formulation penalizes the model when its predictions align more closely with the rejected mask than with the preferred one, thereby encouraging the decoder to internalize human perceptual preferences. The output of this stage is an updated segmentation policy—represented by the fine-tuned decoder—that produces masks more consistent with our pseudo-expert judgments.

4 Experiments and Results

4.1 Experimental setup



(a) SFT-SAM training loss. (b) VisionSAM training loss.

Figure 3: Training loss curves for models in two stages.

Figure 3 presents the training loss curves for the SFT-only model (SFT-SAM) and the RLHF-enhanced model (VisionSAM). SFT-SAM is trained on our manually curated demonstration dataset, while VisionSAM is optimized using the comparison dataset derived from human preference annotations.

Evaluation Metrics. During inference, we evaluate model performance using two standard segmentation metrics: Intersection over Union (IoU) and Dice coefficient. These metrics are widely used in segmentation tasks for assessing the overlap between predicted and ground truth masks.

Baselines. We consider both the original SAM model and our fine-tuned SAM model as baselines for comparison. All models are evaluated under identical settings to ensure a fair assessment. Original SAM refers to the ViT-B variant for all experiments unless otherwise specified. While our current study focuses on ViT-B due to its favorable trade-off between performance and efficiency, future work may explore larger variants such as ViT-L and ViT-H to assess the effect of model scale on anomaly segmentation performance.

4.2 Results and Analysis

4.2.1 Baseline Behavior: SAM's Response to Anomaly Prompts

We adopt the original SAM as our baseline in this stage. To examine the impact of different prompting strategies on SAM's segmentation quality, we tested three prompt configurations: 1) Single positive point: One point indicating the center of the target defect; 2) Positive + Negative point: A positive point on the target and a negative point in a distractor region; 3) Bounding box: A tight box enclosing the target anomaly.

Figure 4 presents four representative results, covering a range of typical outcomes observed in our Phase 1 analysis:

1) Correct Segmentation, Aligned Confidence: In some cases (e.g., Row 1), SAM successfully produces a mask that closely matches the ground truth in both shape and location. Additionally, the mask with the highest confidence score corresponds to the best segmentation, indicating a strong alignment between model prediction and human judgment. These are ideal cases where SAM functions as expected;

2) Correct Segmentation, Misaligned Confidence: In cases like Row 2, SAM generates a reasonably accurate mask (e.g., Mask 2 with score = 0.512), but assigns a higher confidence to a suboptimal one (e.g., Mask 1 with score = 0.945). This discrepancy suggests a misalignment between the model's internal scoring and human perception of quality. Such behavior might hinder downstream automation systems that rely on score-based mask selection.

3) Partial Failure due to Semantic Misunderstanding: In Row 3, SAM sometimes segments entire object structures rather than focusing on the anomalous regions. For example, when tasked with segmenting a defect on a wire, SAM includes the inner cavity of the wire as part of the predicted mask. While the mask is spatially consistent, it reflects a lack of semantic understanding of what constitutes a defect, highlighting SAM's bias toward object-level rather than anomaly-level segmentation.

4) Failure Cases with Subtle Defects: Row 4

demonstrate cases where SAM fails to produce any meaningful segmentation. In Row 4, the anomaly (a slight deformation in a carpet) is barely perceptible even to human observers.

Across all experiments, we observed that SAM often segments entire objects rather than focusing on local defects, suggesting it lacks an explicit notion of "anomaly." Point-based prompts, especially those combining positive and negative clicks, generally led to better results than bounding boxes, likely due to the finer spatial guidance they provide.

4.2.2 Finetuning Results: SFT-SAM vs. VisionSAM (RLHF)

Table 1 summarizes the overall segmentation performance of the original SAM, the supervised fine-tuned variant (SFT-SAM), and the RLHF-enhanced model (VisionSAM). Compared to the original SAM, SFT-SAM achieves a substantial performance gain, with an IoU improvement of approximately +13.17% and a Dice improvement of +9.76%, demonstrating the effectiveness of domain-specific supervised finetuning. After further finetuning via RLHF, VisionSAM achieves a modest improvement over SFT-SAM, with an additional +0.43% in IoU and +0.69% in Dice. While this gain indicates that preference-based learning can further refine the model, the overall impact remains limited, suggesting that the benefits of RLHF are more subtle in the absence of large-scale or high-confidence feedback. This observation is further supported by the qualitative result shown in Figure 5 (right). Although VisionSAM attempts to capture a subtle scratch on the upper part of a wooden surface, it fails to localize the anomaly effectively, producing scattered noise instead of a coherent segmentation mask.

Table 1: Overall segmentation performance of SAM, SFT-SAM, and VisionSAM.

	IoU	Dice	Params Size
SAM	0.5818	0.7049	357.1 M
SFT-SAM	0.6584	0.7737	357.1 M
VisionSAM	0.6612 $\uparrow$	0.7790 $\uparrow$	375.1 M

Table 2 presents the results of directly applying DPO-based preference finetuning on the original SAM model, resulting in RLHF-SAM. Interestingly, this approach leads to a performance drop: IoU decreases by 2.27% and Dice by 1.67% compared to the original SAM. This indicates that direct preference-based finetuning without prior supervised adaptation may be ineffective—or even harmful—in the anomaly segmentation setting. One plausible explanation is that the base SAM model lacks task-specific alignment prior to RLHF. Without supervised grounding on domain-relevant demonstrations, DPO optimization may attempt to align preference gradients on top of a representation that is not yet sensitive to the fine-grained anomaly features. Additionally, the human-labeled preferences may conflict with SAM's original object-centric inductive biases, resulting in unstable or noisy updates when no intermediate supervised signal is provided. This result highlights the importance of using supervised finetuning as a warm start before applying RLHF techniques in dense prediction tasks like segmentation, where alignment targets are less well-defined than in language generation.

Table 2: Performance comparison between the original SAM and RLHF-SAM, where RLHF-SAM is obtained by directly applying DPO finetuning on the original SAM without prior supervised training.

	IoU	Dice	Params Size
SAM	0.5818	0.7049	357.1 M
RLHF-SAM	0.5686 $\downarrow$	0.6931 $\downarrow$	375.1 M

5 Discussion

Replicability. To ensure replicability, we can provide all necessary code, training scripts, inference utilities, and model checkpoints, along with detailed documentation of hyperparameters and experimental settings if you request. Our pipeline is built on publicly available frameworks, and we fix random seeds to reduce variability across runs.

Datasets. We reorganize a subset of the VISION anomaly segmentation dataset and introduce a small-scale preference-annotated comparison set constructed via human-in-the-loop labeling. Additionally, we use the publicly available MVTec-AD dataset for zero-shot evaluation of segmentation performance. The MVTec-AD dataset is released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which prohibits commercial use. Users who are unsure whether their application violates the non-commercial use clause are advised to contact the original dataset authors.

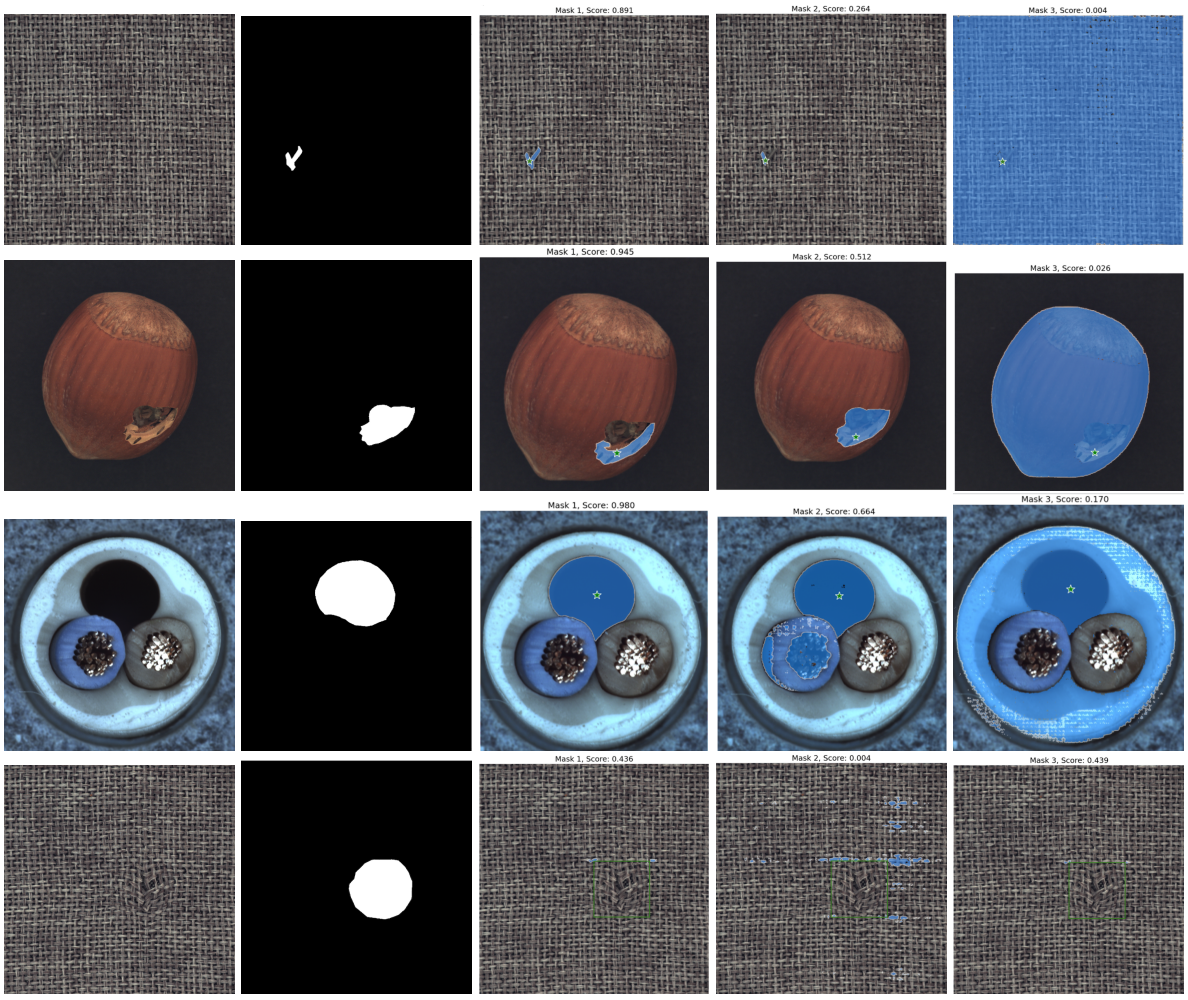
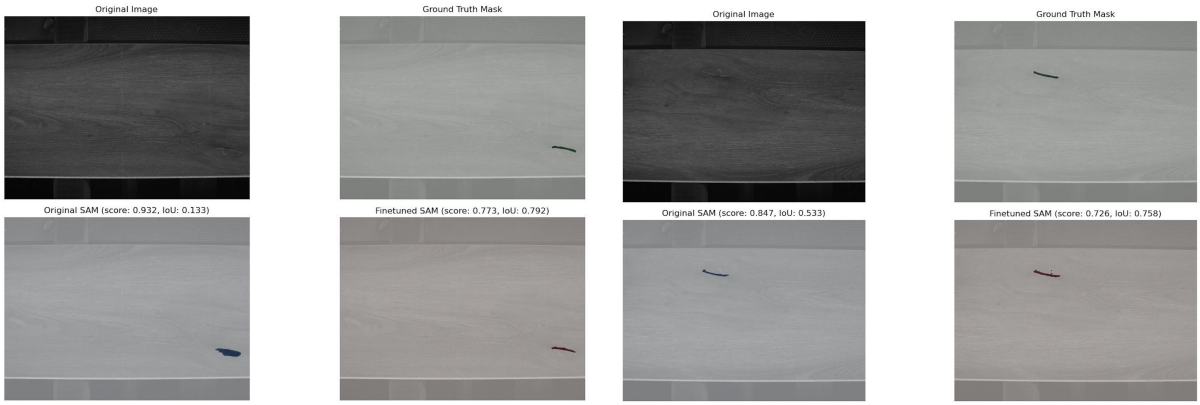


Figure 4: Representative segmentation examples from SAM across four industrial anomaly cases. Each row shows (from left to right): the original input image, the ground truth segmentation mask, and three predicted masks generated by SAM. The scores shown above each predicted mask correspond to the model's internal confidence estimate, computed as the predicted Intersection over Union (IoU) between the generated mask and its estimated ground truth. These scores reflect SAM's own assessment of mask quality.

We hope our annotations and benchmarking efforts encourage future research on incorporating human feedback into vision models in industrial settings.

Ethics. From an ethical standpoint, our work poses low direct societal risk, as it focuses exclusively on industrial anomaly segmentation without involving human subjects or sensitive personal data. However, we caution against the blind deployment of automated inspection systems, particularly in high-stakes industrial settings, and recommend incorporating human oversight in safety-critical applications. Notably, our preference annotations were collected by annotators from the same cultural and geographic background, which may introduce subtle bias in how anomalies are perceived and ranked. We encourage future research to explore more diverse human feedback and evaluate generalization across broader defect types and annotation perspectives.

Limitations. While our proposed VisionSAM demonstrates the feasibility of applying reinforcement learning from human feedback (RLHF) to industrial anomaly segmentation, several limitations remain. First, the overall performance gain from RLHF finetuning is modest, with only incremental improvements over the supervised finetuned baseline (SFT-SAM). This suggests that preference signals alone may not be sufficient to significantly shift the behavior of a model already aligned with the task via supervision. Second, although VisionSAM shows better alignment with human preferences in many cases, it occasionally struggles with capturing fine-grained structures, tends to hallucinate small disconnected components, and produces less precise boundaries compared to heavier segmentation models—particularly on challenging



(a) Comparison between the original SAM and SFT-SAM on an anomaly segmentation sample.

(b) Comparison between the SFT-SAM and VisionSAMon an anomaly segmentation sample.

Figure 5: Visualization of anomaly segmentation results across finetuning stages: (a) original SAM vs. SFT-SAM, and (b) SFT-SAM vs. VisionSAM.

categories with low contrast or irregular defect patterns. Lastly, inference speed is another practical concern. While our method remains efficient in terms of model size, generating and comparing multiple masks for each prompt during inference incurs additional computational overhead.

Future Works. Despite these limitations, our work highlights the potential of integrating RLHF into segmentation pipelines and contributes a reproducible framework for preference-based vision model optimization. Several possibilities remain for further improvement. One direction is to better inject the notion of "anomaly" into the model. Existing models like SAM are primarily object-centric and may overlook subtle defects; future work may consider pretraining on anomaly-centric datasets or designing prompts that more explicitly highlight abnormal regions. Additionally, our current preference annotations are based on binary comparisons (preferred vs. rejected), which provide limited feedback. Richer forms of supervision—such as ranked masks, region-level critiques, or confidence calibration from experts—could offer stronger learning signals. Finally, our comparison dataset is relatively small and annotated by individuals from similar cultural and professional backgrounds. Expanding both the dataset scale and annotator diversity across domains may lead to more generalizable and unbiased preference-aligned models.

Acknowledgments. We would like to thank the reviewers for their valuable feedback throughout the project. We also appreciate the insightful discussions and suggestions from James Mooney, Risako Owan, Bin Hu, and Junhan Wu, as well as the thoughtful questions raised by peers during the poster session, which greatly inspired our thinking.

References

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. 2023. Anomalygpt: Detecting industrial anomalies using large vision-language models.

Duojun Huang, Xinyu Xiong, Jie Ma, Jichang Li, Zequn Jie, Lin Ma, and Guanbin Li. 2024. Alignsam: Aligning segment anything model to open context via reinforcement learning.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. 2023. Winclip: Zero-/few-shot anomaly classification and segmentation.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. 2023a. Segment anything in high quality.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. 2023b. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36:29914–29934.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katie Collins, Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. 2024. Rich human feedback for text-to-image generation.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. 2024. Segment anything in medical images. *Nature Communications*, 15(1).

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. 2024a. Sam-clip: Merging vision foundation models towards semantic and spatial understanding.

Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. 2024b. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3635–3647.

Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. 2023. Medical sam adapter: Adapting segment anything model for medical image segmentation.

6 Appendix

6.1 Q and A for Rubric: for TA's convenience

Rubric QA is recorded in this google sheet for TA's grading convenience.