

Received 6 December 2025, accepted 16 December 2025, date of publication 22 December 2025,
date of current version 29 December 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3646677

RESEARCH ARTICLE

Zero-Shot Anomaly Segmentation via Query-Level Uncertainty Analysis

KIMIN YUN^{1,2}, AND YUSEOK BAE¹

¹Visual Intelligence Laboratory, Electronics and Telecommunications Research Institute, Yuseong-gu, Daejeon 34129, South Korea

²University of Science and Technology (UST), Yuseong-gu, Daejeon 34113, South Korea

Corresponding author: Kimin Yun (kimin.yun@etri.re.kr)

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by [Ministry of Science and ICT (MSIT)] of Korean Government (Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities) under Grant RS-2022-II220124.

ABSTRACT We propose a training-free method for anomaly segmentation that leverages the internal signals of pretrained transformer-based segmentation models. Unlike prior approaches that rely on retraining, auxiliary outlier data, or external supervision, our method identifies out-of-distribution (OOD) regions by analyzing query-level behaviors. Specifically, we suppress confidently predicted inliers while amplifying uncertain predictions through entropy-guided selection and spatial consistency analysis. To improve spatial coherence, we introduce an object-aware refinement step that enforces consistency within class-agnostic segments, yielding anomaly maps that align more closely with object boundaries. Our approach operates in a strictly zero-shot setting and requires no modification of the base model. Experiments across standard benchmarks, including SMIYC, Fishyscapes, and RoadAnomaly, demonstrate that the proposed method achieves state-of-the-art performance among zero-shot approaches and narrows the gap to training-based methods, while providing interpretable insights into model uncertainty. This competency-driven framework demonstrates that robust anomaly segmentation can be achieved without retraining or external data, making it suitable for practical real-world deployment.

INDEX TERMS Out-of-distribution detection, anomaly segmentation, transformer-based segmentation, object-aware refinement, model competency analysis.

I. INTRODUCTION

Recent advances in deep learning have driven significant progress in visual recognition, yet evaluating model reliability beyond standard benchmarks remains a critical challenge. While modern architectures trained on large-scale datasets achieve strong performance on closed-set recognition and segmentation tasks, their predictions often fail when confronted with inputs outside the training distribution. This limitation has motivated research into methods that analyze internal model signals, including query-level uncertainty and entropy-based measures, to assess robustness under distribution shift [1], [2]. One main problem in this direction is out-of-distribution (OOD) detection, which aims to identify

The associate editor coordinating the review of this manuscript and approving it for publication was Ayman El-Baz.

inputs that do not belong to any known category. Reliable OOD detection is essential for deploying vision systems in real-world environments where unexpected objects may appear.

OOD detection is particularly critical in safety-critical applications such as autonomous driving, where failure to recognize unexpected obstacles can directly compromise collision avoidance and motion planning. However, accurate OOD detection remains challenging due to the scarcity and high variability of anomalous objects in real-world driving scenarios. This raises a fundamental question: can we detect anomalies by analyzing what a model does not know, rather than explicitly teaching it what anomalies look like?

Figure 1 illustrates this principle. In the road-driving scene (a), a large log has fallen across the road, lying outside the training distribution of the segmentation model.

The baseline semantic segmentation (b) incorrectly predicts the center of the log as truck and its boundaries as road, revealing how conventional models force every pixel into the closest in-distribution class. Our method (c) takes a different approach: rather than relying on the segmentation output alone, we analyze the internal query-level signals of the transformer-based model. Queries that cannot confidently resolve to known object categories exhibit characteristic uncertainty patterns, which we exploit to identify the entire log as anomalous without any additional training or annotations.

This uncertainty-driven perspective offers several advantages over conventional approaches. By examining how the model internally represents ambiguous regions, we gain interpretable insights into where and why predictions may be unreliable, without requiring auxiliary anomaly data or model retraining. We ground our approach in the concept of *model competency*, which we define as the ability of a pretrained segmentation model to confidently and consistently assign input regions to known semantic categories. Regions where the model exhibits low competency, characterized by high uncertainty in class predictions and inconsistent query behaviors, become strong candidates for out-of-distribution detection. This competency-based perspective differs fundamentally from methods that require external anomaly exposure. Rather than learning what anomalies look like through synthetic outliers or auxiliary data, we identify regions where the model's own knowledge is insufficient.

Building upon transformer-based architectures, our framework extends rejection-only approaches such as RbA [3] by incorporating entropy-guided query selection, spatial consistency analysis, and object-aware refinement. The main contributions of this work can be summarized as follows:

- We propose a **zero-shot anomaly segmentation framework** that detects OOD regions by exploiting the internal signals of pretrained transformer-based segmentation models, without requiring retraining or auxiliary data.
- We introduce **query-level competency analysis with rejection and entropy cues** to identify anomalous regions in a principled and interpretable manner.
- We incorporate **object-aware refinement with class-agnostic segmentation** to enforce structural consistency and improve localization of unknown objects.
- We further discuss how our framework can provide **insights into model generalization**, offering a practical perspective on how segmentation models behave under open-world conditions.

For clarity, we state that our method targets transformer-based segmentation backbones that provide query-level representations, as these signals form the core interface for our competency-based reasoning.

II. RELATED WORK

Research on anomaly segmentation and out-of-distribution (OOD) detection has developed along multiple directions, aiming to identify and localize regions that deviate from

the training distribution. Existing approaches can be broadly grouped into five categories: (1) discriminative methods, (2) generative approaches, (3) outlier-exposure methods, (4) ensemble and Bayesian approaches, and (5) foundation models. We review each category below, with emphasis on transformer-based and zero-shot approaches most relevant to our work.

A. DISCRIMINATIVE APPROACHES

Discriminative approaches leverage the output statistics of pretrained models, such as softmax scores, logits, or feature-based distances, to identify anomalous regions. Early works include maximum softmax probability [4] and ODIN [5]. For dense prediction tasks, pixel-level extensions such as Mahalanobis distance [6], Dirichlet Prior Networks [7], and entropy-based measures like Standardized Max Logits (SML) [8] have been explored.

Recent transformer-based methods have advanced this line of work by exploiting structured object queries. For example, rejection-based approaches such as RbA (Reject by All) [3] suppress regions confidently explained by known classes to reveal unknown objects, while Maskomaly [9] performs zero-shot anomaly segmentation by identifying object masks that cannot be reliably assigned to any known semantic category. Related work also includes EAM (Ensemble over Anomaly scores of Mask-wide predictions) [10], which leverages mask-level recognition signals for open-set and zero-shot segmentation. More recent efforts consider attribute-level cues inside outliers [11], offering finer granularity in anomaly detection.

B. GENERATIVE AND OUTLIER-EXPOSURE APPROACHES

Generative methods explicitly model the distribution of normal data and detect anomalies as deviations from this learned distribution. Autoencoder-based reconstruction and image re-synthesis techniques [12], [13] have been widely studied. For example, SynBoost [13] combines re-synthesis with uncertainty cues to improve anomaly localization. While generative approaches can capture previously unseen anomalies, they typically require substantial computational resources and are often impractical for real-time deployment.

Outlier Exposure (OE) takes a complementary approach by introducing external anomalous data during training to improve model robustness. Examples include MetaOOD [14], DenseHybrid [15], and PEBAL [16], which leverage copy-paste augmentation or explicit “unknown” classes. More recent work explores weakly supervised extensions [17] to broaden anomaly coverage while reducing annotation cost. Although OE methods achieve strong performance on standard benchmarks, their effectiveness depends on the diversity and quality of the auxiliary anomaly data used during training, which limits their applicability in true zero-shot scenarios.

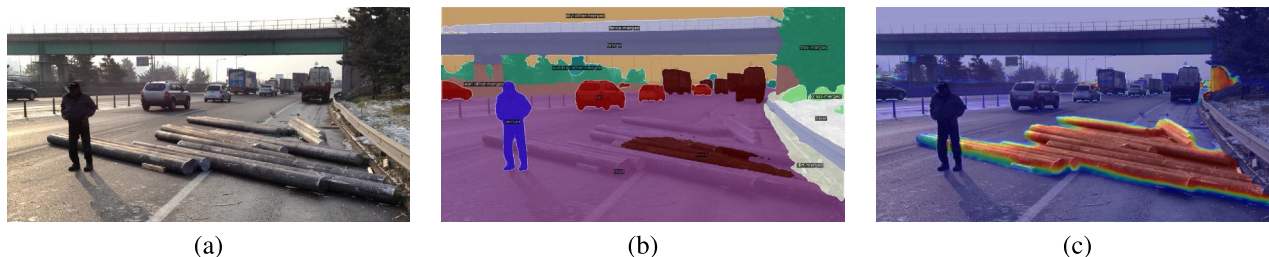


FIGURE 1. Example of zero-shot anomaly segmentation. (a) Input image with a large log fallen across the road. (b) Semantic segmentation result: the center of the log is incorrectly labeled as “truck”, and the edges as “road”. (c) Our method identifies the entire log as an anomaly using internal model signals without any additional training.

C. ENSEMBLE AND BAYESIAN APPROACHES

Ensemble-based methods [18] and Bayesian models [19] quantify predictive variance through multiple forward passes or weight distributions to identify anomalies. MC Dropout [20] offers a computationally lighter approximation to Bayesian inference by treating dropout as approximate posterior sampling. While effective at capturing model uncertainty, these techniques increase inference cost proportionally to the number of samples or ensemble members, and may produce false positives on noisy inputs.

D. HYBRID ARCHITECTURES

Hybrid approaches decouple anomaly detection from the primary segmentation task through auxiliary modules. ConfidNet [21] and ObsNet [22] employ lightweight auxiliary networks to predict per-pixel anomaly status alongside the main segmentation output. Residual Pattern Learning [23] introduces dedicated residual modules to capture outlier patterns without interfering with inlier segmentation. These hybrid architectures often integrate multiple signals, including discriminative outputs, generative reconstructions, and uncertainty estimates, to achieve more robust anomaly detection.

E. FOUNDATION MODELS FOR ANOMALY SEGMENTATION

Recent foundation models offer new capabilities for zero-shot anomaly segmentation through their broad pretraining on diverse data. Vision-language models (VLMs) such as CLIP [24] and Molmo [25] can generalize to unseen concepts through natural language prompts and have been adapted for anomaly detection [26], [27]. Segmentation foundation models like SAM [28] generate class-agnostic object masks and have been integrated with anomaly scoring methods to improve region extraction [29], [30].

However, these approaches face limitations in anomaly segmentation. VLMs require carefully designed textual prompts that describe potential anomalies, which contradicts the open-world assumption where anomaly types are unknown in advance. Consequently, while foundation models enable rapid adaptation to new domains, they require additional guidance mechanisms for precise anomaly localization and remain most effective when combined with task-specific anomaly detection modules.

III. PROPOSED METHOD

Figure 2 shows the overall framework for our zero-shot anomaly segmentation method. The process consists of three stages: (1) *Rejection by Inliers*, which removes regions confidently predicted as known classes by the pretrained segmentation model; (2) *Selection for Anomalous Regions*, which identifies ambiguous regions by analyzing query-wise uncertainty and spatial consistency; and (3) *Anomaly Map Fusion and Object-Aware Refinement*, which combines the two signals into a unified anomaly map and enforces spatial consistency within object boundaries through class-agnostic segmentation. The entire framework operates in a strictly zero-shot manner without any retraining, relying solely on internal signals of the pretrained model.

A. REJECTION BY INLIERS

Our first stage suppresses regions that are confidently classified as inliers by the backbone segmentation model. Similar rejection strategies have been explored in prior work [3], showing that discarding high-confidence inliers can improve anomaly separation. Here, we adapt this idea as a lightweight pre-filter in a broader zero-shot pipeline, designed to isolate potential anomalies without task-specific fine-tuning. The intuition is analogous to background subtraction in motion detection [31], [32], where stable regions are explained away to reveal unexpected structures.

We adopt Mask2Former [33] as the backbone segmentation model, which produces N object queries. Each query q yields a class probability vector $p_q \in \mathbb{R}^{K+1}$ and a corresponding spatial membership mask $M_q(x)$, where x denotes a pixel coordinate. Here, K denotes the number of trained (known) foreground classes, and class index $K + 1$ corresponds to background (void).

Following [3], we treat each query as an independent classifier. The image is initially assumed anomalous, and we progressively reject regions where any query confidently predicts a known class. The rejection score is computed as:

$$s_{\text{reject}}(x) = \min_{q \in Q} \left(1 - M_q(x) \cdot \max_{1 \leq k \leq K} p_q[k] \right), \quad (1)$$

where $p_q[k]$ is the predicted probability of known class k and $M_q(x)$ denotes the membership mask value at location x for query q .

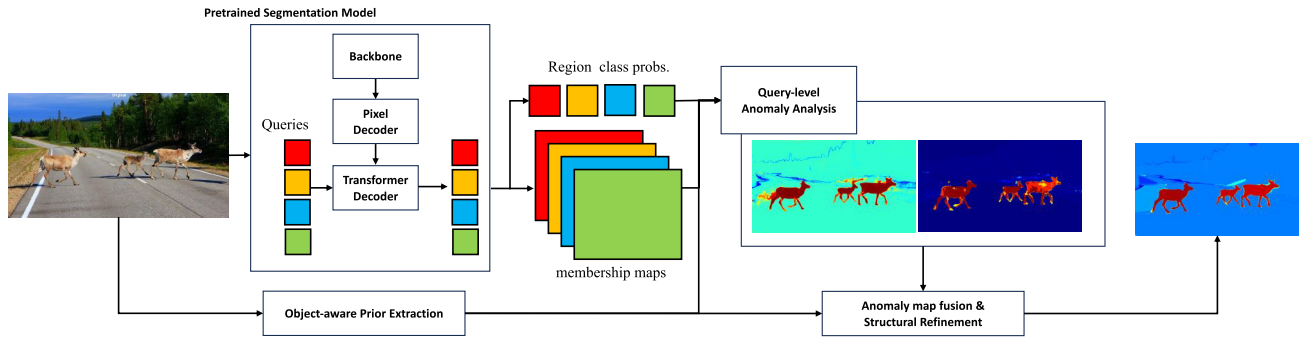


FIGURE 2. Overall framework of the proposed method.

B. SELECTION FOR ANOMALOUS REGIONS

The rejection stage filters out regions confidently explained by inliers, but additional cues are needed to highlight queries that are more likely to correspond to unknown objects. To this end, we analyze the predictive uncertainty of each query by examining its class distribution entropy. For each query q , we first normalize its class probabilities over the K foreground classes, explicitly excluding the background class ($K + 1$):

$$p_{q,i}^{fg} = \frac{p_q[i]}{\sum_{j=1}^K p_q[j] + \epsilon}, \quad i = 1, \dots, K. \quad (2)$$

In segmentation models, background predictions typically capture void regions, occluded areas, or uncertain boundaries rather than coherent objects. A query that correctly assigns high probability to void space should not be considered uncertain. By normalizing only over foreground classes, we isolate the model’s uncertainty about which semantic object category is present. High entropy in this normalized distribution indicates genuine ambiguity among meaningful object classes, which is the precise signal we seek for anomaly detection. We quantify this predictive uncertainty through the entropy of the foreground class distribution:

$$S_q^{init} = - \sum_{i=1}^K p_{q,i}^{fg} \log(p_{q,i}^{fg} + \epsilon) / \log K, \quad (3)$$

where the normalization by $\log K$ ensures $S_q^{init} \in [0, 1]$. A query with S_q^{init} close to 1 cannot confidently distinguish between known object types, suggesting the region may contain an out-of-distribution object. This entropy-based uncertainty score serves as our initial anomaly cue.

Entropy alone, however, is insufficient for reliable anomaly detection. High-entropy queries can spatially overlap with regions that other queries confidently classify as known objects. To address this issue, we incorporate spatial consistency by measuring each query’s alignment with confidently predicted inlier regions. We first define a binarized support mask for each query:

$$M_q^b(x) = \begin{cases} 1 & \text{if } M_q(x) > 0.5 \\ 0 & \text{otherwise,} \end{cases} \quad (4)$$

and construct a confident inlier region M^{conf} as the union of masks from low-entropy queries:

$$M^{conf} = \bigcup_{q: S_q^{init} < \tau} \{x \mid M_q^b(x) = 1\}, \quad (5)$$

where we set $\tau = 0.01$ as a conservative threshold to ensure that only highly confident queries contribute to the inlier region.

Each query’s spatial overlap with this confident region is then measured as:

$$r_q = \frac{|\{x \mid M_q^b(x) = 1\} \cap M^{conf}|}{|\{x \mid M_q^b(x) = 1\}| + \epsilon}. \quad (6)$$

A high value of r_q indicates that query q exhibits spatial behavior consistent with known objects. We therefore attenuate its anomaly score proportionally to its overlap with the confident inlier region:

$$S_q^{adj} = S_q^{init} \cdot (1 - r_q). \quad (7)$$

This adjustment ensures that queries overlapping with confident predictions receive reduced anomaly scores, even if their class distributions exhibit high entropy.

Finally, we compute a pixel-wise anomaly score by aggregating the adjusted query scores:

$$w_q(x) = \frac{\exp(S_q^{adj}/T) \cdot M_q^b(x)}{\sum_{q'} \exp(S_{q'}^{adj}/T) \cdot M_{q'}^b(x)}, \quad (8)$$

$$s_{accept}(x) = \sum_q w_q(x) \cdot S_q^{adj}, \quad (9)$$

where T is a temperature parameter controlling the weighting sharpness. This two-stage selection process enables uncertain queries to contribute to anomaly detection only when they provide spatial evidence distinct from confident inlier predictions, yielding anomaly maps that are both uncertainty-aware and spatially consistent.

TABLE 1. Comparison of anomaly segmentation performance on the SMIYC anomaly track test set. We report Average Precision (AP) and False Positive Rate at 95% recall (FPR95), following the official test server. Methods are grouped into *training-based* approaches (requiring auxiliary data or fine-tuning) and *zero-shot* methods (no auxiliary training). \uparrow indicates higher is better; \downarrow indicates lower is better.

Method	Segmentation Framework	Aux. Data	AP \uparrow	FPR95 \downarrow
Training-based Methods				
PEBAL [16]	DeepLabV3+ [34]	✓	49.1	40.8
SynBoost [13]	VPLR [35]	✓	56.4	61.9
DenseHybrid [15]	DeepLabV3+ [34]	✓	78.0	9.8
Maximized Entropy [14]	DeepLabV3+ [34]	✓	85.5	15.0
EAM [10]	Mask2Former [33]	✓	93.8	4.1
RbA [3]	Mask2Former	✓	94.5	4.6
RPL+CoroCL [23]	Mask2Former [33]	✓	83.5	11.7
CSL [36]	Mask2Former [33]	✓	80.1	7.2
ContMAV [37]	Mask2Former [33]	✓	90.2	3.8
UNO [38]	Mask2Former [33]	✓	96.1	2.3
PixOOD [39]	DeepLabV3+ [34]	✓	68.9	54.3
FlowCLAS [40]	Mask2Former [33]	✓	94.3	6.6
Zero-Shot Methods (No Auxiliary Training)				
JSRNet [41]	DeepLabV3+ [34]	✗	33.6	43.9
DenseHybrid [15]	DeepLabV3+ [34]	✗	51.5	33.2
Image Resynthesis [12]	PSPNet	✗	52.3	25.9
ObsNet [42]	DeepLabV3+ [34]	✗	75.4	26.7
EAM [10]	Mask2Former [33]	✗	76.3	93.9
RbA [3]	Mask2Former [33]	✗	86.1	15.9
Proposed Method	Mask2Former [33]	✗	89.2	11.3

C. ANOMALY MAP FUSION AND OBJECT-AWARE REFINEMENT

We first obtain a pixel-wise anomaly heatmap by combining the rejection and selection scores:

$$s_{\text{fuse}}(x) = \lambda \cdot s_{\text{reject}}(x) + (1 - \lambda) \cdot s_{\text{accept}}(x), \quad (10)$$

where $\lambda \in [0, 1]$ balances confident inlier suppression and uncertainty-based enhancement. The rejection and selection terms capture complementary aspects of model competency: the former removes regions confidently explained by known classes, while the latter highlights regions where the model exhibits high query-level uncertainty. Balancing these two signals avoids overly aggressive inlier suppression or noisy uncertainty amplification, leading to more stable anomaly localization.

Although this fusion produces an informative anomaly map, directly thresholding s_{fuse} often leads to noisy, fragmented predictions. Pixel-level fluctuations result in inconsistencies across parts of the same object, which contradicts the assumption that anomalous objects form coherent, bounded regions. To enforce such coherence, we incorporate class-agnostic object segmentation to provide spatial structure priors. This approach leverages the principle that anomalous objects, like known objects, occupy coherent spatial regions with well-defined boundaries. In our implementation, we use SAM [28] to extract object masks, though any class-agnostic segmentation or edge-based grouping method could serve this role.

We formulate refinement as a variance-penalized regularization that encourages anomaly scores to remain faithful to the fused map while promoting intra-segment consistency and

suppressing unreliable segments:

$$E_{\text{var}}(s) = \frac{1}{2} \sum_x (s(x) - s_{\text{fuse}}(x))^2 + \frac{\lambda}{2} \sum_i \sum_{x,y \in r_i} (s(x) - s(y))^2 + \mu \sum_i \text{Var}_{z \in r_i}[s(z)]. \quad (11)$$

In practice, this regularization leads to a variance-aware interpolation between the segment-level average and the original fused score:

$$s_{\text{refined}}(x) = \alpha_i \bar{s}(r_i) + (1 - \alpha_i) s_{\text{fuse}}(x), \quad x \in r_i, \quad (12)$$

where $\alpha_i = \exp(-\beta v(r_i))$ depends on the variance $v(r_i)$ within segment r_i . High-variance segments are assigned lower weights, preventing unreliable masks from dominating the refinement. This design ensures that the final anomaly map respects object boundaries while remaining robust to segmentation errors.

IV. EXPERIMENTS

A. EXPERIMENTAL SETUP

Our experiments evaluate the effectiveness of the proposed method for detecting out-of-distribution (OOD) objects in urban driving scenes. We adopt Mask2Former [33] with a Swin-L backbone [43], pretrained on the Cityscapes dataset [44], without any further fine-tuning. This configuration enables direct comparison with existing methods under a strictly zero-shot setting, as no retraining or adaptation is performed. We report Average Precision (AP) and False Positive Rate at 95% recall (FPR95), following standard evaluation protocols for anomaly segmentation benchmarks.

All results are measured on validation sets to ensure consistency with prior work. We set key hyperparameters empirically and keep them fixed across all experiments. The entropy threshold $\tau = 0.01$ defines the confident inlier region M^{conf} in Eq. (5), and the fusion coefficient $\lambda = 0.6$ in Eq. (10) balances rejection and selection signals. For reference, training-based OOD methods generally require multi-epoch fine-tuning and auxiliary datasets, whereas our method is fully training-free and only performs inference on a pretrained backbone. As a training-free approach, our method avoids the substantial computational cost of training-based pipelines. It introduces no auxiliary networks, iterative sampling, or ensemble computation during inference, resulting in a lightweight and practical zero-shot segmentation framework. We note that the choice of class-agnostic segmentation model is not restricted to SAM, and alternative lighter-weight grouping mechanisms could be substituted if required.

B. BENCHMARK DATASETS

We evaluate our method on three widely used benchmarks for anomaly segmentation, each covering complementary aspects of out-of-distribution (OOD) detection in driving scenarios.

Segment Me If You Can (SMIYC) [50] is a benchmark designed for anomaly segmentation under domain shift, collected from diverse web sources to introduce substantial variability compared to standard training distributions like Cityscapes. It provides two evaluation tracks: the *anomaly track* features clearly OOD objects such as animals, furniture, or debris, while the *obstacle track* focuses on small road-level obstacles including traffic cones or fallen objects. Test annotations are withheld, and performance is measured through a centralized evaluation server.

RoadAnomaly [12] is constructed from natural driving scenes with unexpected obstacles such as overturned vehicles or animals on highways. Its realistic sourcing and diverse anomaly types provide a valuable complement to synthetic or controlled benchmarks, though some images overlap with SMIYC.

Fishyscapes [51] builds upon Cityscapes and provides two complementary evaluation scenarios. The *Lost & Found* subset features real physical anomalies on the road such as boxes and tires, while the *Static* subset overlays synthetic foreign objects from Pascal VOC onto Cityscapes scenes. Despite remaining visually closer to Cityscapes, the diverse and localized anomaly types offer additional robustness evaluation.

C. QUANTITATIVE RESULTS

Table 1 presents results on the SMIYC Anomaly Track test set, evaluated by the official benchmark server. Our method operates in a strictly zero-shot setting, using only a publicly available segmentation backbone without auxiliary data or retraining. Within the zero-shot category, our method achieves state-of-the-art performance, substantially outperforming prior approaches and narrowing the gap to training-based methods.

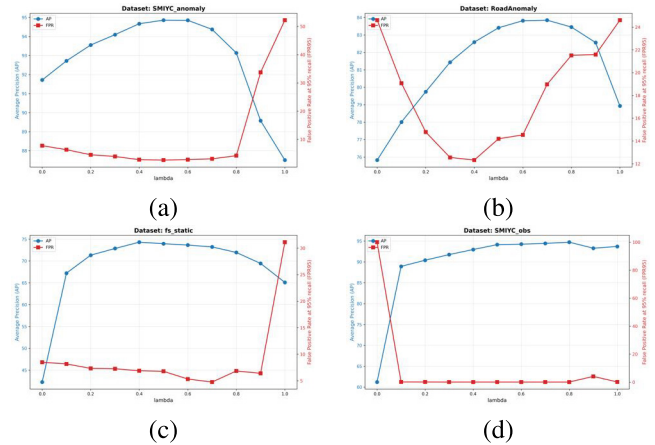


FIGURE 3. Effect of fusion weight λ in Eq. (10) on anomaly segmentation performance across four benchmarks. (a) SMIYC Anomaly, (b) RoadAnomaly, (c) Fishyscapes Static, (d) SMIYC Obstacle.

While methods such as UNO [38] achieve the strongest absolute results through additional supervision, our results demonstrate that competitive performance can be achieved through a lightweight, training-free design.

Tables 2 and 3 summarize results on RoadAnomaly, Fishyscapes, and SMIYC Obstacle benchmarks. For clarity, we report both the originally published results (Maskomaly (paper) and RbA (paper)) and our re-evaluated scores (Maskomaly* and RbA*), obtained by running the official Mask2Former implementation with Cityscapes-pretrained weights. This separation ensures a fair comparison among zero-shot methods under a shared pretrained model. Our method achieves strong performance among zero-shot approaches across all datasets. On RoadAnomaly, the method remains robust under natural distribution shifts, while on Fishyscapes Static, it handles synthetic overlays despite their challenging small-scale and partial occlusion characteristics. The SMIYC Obstacle and Fishyscapes Lost & Found results further confirm robustness to small road-level anomalies.

As summarized in Table 4, the proposed selection stage (Rejection + Selection) already outperforms the rejection-only RbA* [3] baseline by a large margin, confirming its core contribution. The subsequent object-aware refinement further enhances boundary coherence and reduces false positives—particularly on Fishyscapes Static—while yielding comparatively smaller gains. These results indicate that the primary performance improvement stems from the query-level uncertainty analysis, and that enforcing spatial consistency within object boundaries provides complementary benefits rather than dominating the overall performance.

D. ABLATION STUDY

1) EFFECT OF HYPERPARAMETER

To examine the stability of the proposed fusion step in Eq. (10), we perform a sensitivity analysis on the mixing coefficient λ , which balances the rejection score s_{reject} and

TABLE 2. Evaluation results on RoadAnomaly and Fishyscapes static. We report Average Precision (AP) and False Positive Rate at 95% recall (FPR95). Methods are grouped into *training-based* and *zero-shot*. * indicates reproduced results using official implementations with Cityscapes-pretrained Mask2Former weights. \uparrow / \downarrow denote higher / lower is better. \dagger indicates that FPR95 is undefined as the model failed to reach 95% recall coverage.

Method	Framework	Aux.	RoadAnomaly		Fishyscapes Static	
			AP \uparrow	FPR95 \downarrow	AP \uparrow	FPR95 \downarrow
Training-based Methods						
SynBoost [13]	VPLR [35]	✓	38.2	64.8	66.4	25.6
PEBAL [16]	DeepLabV3+ [34]	✓	45.1	44.6	92.1	1.5
DenseHybrid [15]	DeepLabV3+ [34]	✓	63.9	43.2	60.0	4.9
Max. Entropy [14]	DeepLabV3+ [34]	✓	79.7	19.3	76.3	7.1
Zero-Shot Methods (No Auxiliary Training)						
ML [45]	ResNet-101 [46]	✗	19.0	70.5	38.6	18.3
SML [8]	DeepLabV3+ [34]	✗	25.8	49.7	48.7	16.8
DenseHybrid [47]	DeepLabV3+ [34]	✗	35.1	43.2	54.7	15.5
ObsNet [42]	DeepLabV3+ [34]	✗	54.7	60.0	9.4	47.7
GMMSeg [48]	SegFormer [49]	✗	57.7	44.3	82.6	–
EAM [10]	Mask2Former [33]	✗	66.7	13.4	87.3	2.1
Maskomaly (paper) [9]	Mask2Former [33]	✗	70.9	11.9	69.5	14.4
Maskomaly* [9]	Mask2Former [33]	✗	71.1	17.0	53.2	40.6
RbA (paper) [3]	Mask2Former [33]	✗	78.5	11.8	–	–
RbA* [3]	Mask2Former [33]	✗	73.9	31.9	56.5	32.3
Molmo [25] + SAM [28]	VLM [25]	✗	42.4	– \dagger	31.2	– \dagger
Proposed method	Mask2Former [33]	✗	82.5	10.5	74.5	5.39

TABLE 3. Evaluation results on SMIYC obstacle and Fishyscapes Lost & Found (L&F) validation sets. All methods operate in a zero-shot setting without auxiliary anomaly data or fine-tuning. \uparrow / \downarrow denote higher/lower is better. \dagger indicates that FPR95 is undefined as the model failed to reach 95% recall coverage of ground-truth regions. * indicates reproduced results using official implementations with Cityscapes-pretrained Mask2Former weights.

Method	SMIYC Obstacle		Fishyscapes L&F	
	AP \uparrow	FPR95 \downarrow	AP \uparrow	FPR95 \downarrow
RbA (paper) [3]	87.3	3.3	60.1	10.6
RbA* [3]	94.5	0.25	49.6	49.9
Maskomaly* [9]	88.5	2.85	16.9	35.8
Molmo [25] + SAM [28]	14.3	– \dagger	21.3	– \dagger
Proposed method	94.6	0.09	59.5	36.9

the uncertainty-driven selection score s_{accept} . We sweep λ from 0 to 1 with a step size of 0.1. Figure 3 reports both Average Precision (AP) and FPR95 according to λ . Figure 3(a)-(d) correspond to SMIYC Anomaly, RoadAnomaly, Fishyscapes Static, and SMIYC Obstacle datasets, respectively. Across all four benchmarks, AP steadily increases up to the range $\lambda \approx 0.5 - 0.7$, after which performance plateaus or slightly declines, indicating that $\lambda = 0.6$ provides a consistent near-optimal balance.

This behavior aligns with the complementary nature of the two fused signals. The rejection term emphasizes confidently predicted inlier regions and suppresses overly broad anomaly responses, which benefits datasets such as Fishyscapes Static and SMIYC Obstacle, where anomalous objects are extremely small and are often not reliably represented by individual queries. In contrast, when anomalous objects are large enough to be captured by queries as in SMIYC Anomaly, a stronger contribution from the selection term becomes advantageous. Consequently, the intermediate value

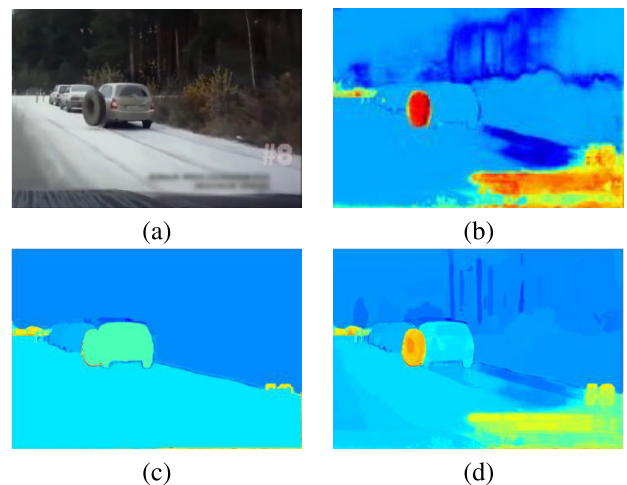


FIGURE 4. Effect of the variance-penalized regularization: (a) Input image, (b) pre-refinement anomaly map, (c) refinement without variance term, (d) refinement with variance term.

$\lambda = 0.6$ achieves a robust compromise, enabling stable performance across datasets without requiring dataset-specific tuning.

To further evaluate the stability of the entropy threshold τ in Eq. (5), we conduct an ablation study on the RoadAnomaly dataset (Table 5). Across the tested range $\tau \in [0.0, 0.05]$, both AP and FPR95 vary only marginally, indicating that the threshold has limited influence on the global scoring structure. This behavior is expected because τ governs only the selection of extremely low-entropy queries that form the confident inlier region. Based on this stability, we adopt $\tau = 0.01$ as a conservative default that reliably filters out only the most confident inlier queries while keeping the overall method robust across datasets.

TABLE 4. Ablation study on the effect of query-level selection and structural refinement across four benchmarks. The comparison explicitly includes RbA [3] to isolate the contribution of the proposed selection stage. * indicates reproduced results using official implementations with Cityscapes-pretrained Mask2Former weights.

Method	RoadAnomaly		Fishyscapes Static		SMIYC Obstacle		Fishyscapes L&F	
	AP ↑	FPR95 ↓	AP ↑	FPR95 ↓	AP ↑	FPR95 ↓	AP ↑	FPR95 ↓
RbA* [3] (Rejection only)	73.9	31.9	56.5	32.3	94.5	0.25	49.6	49.9
Ours (Rejection + Selection)	79.0	13.9	71.6	21.5	94.0	0.06	52.7	39.6
Ours (+ Object-aware refinement)	82.5	10.5	74.5	5.39	94.6	0.09	59.5	36.9

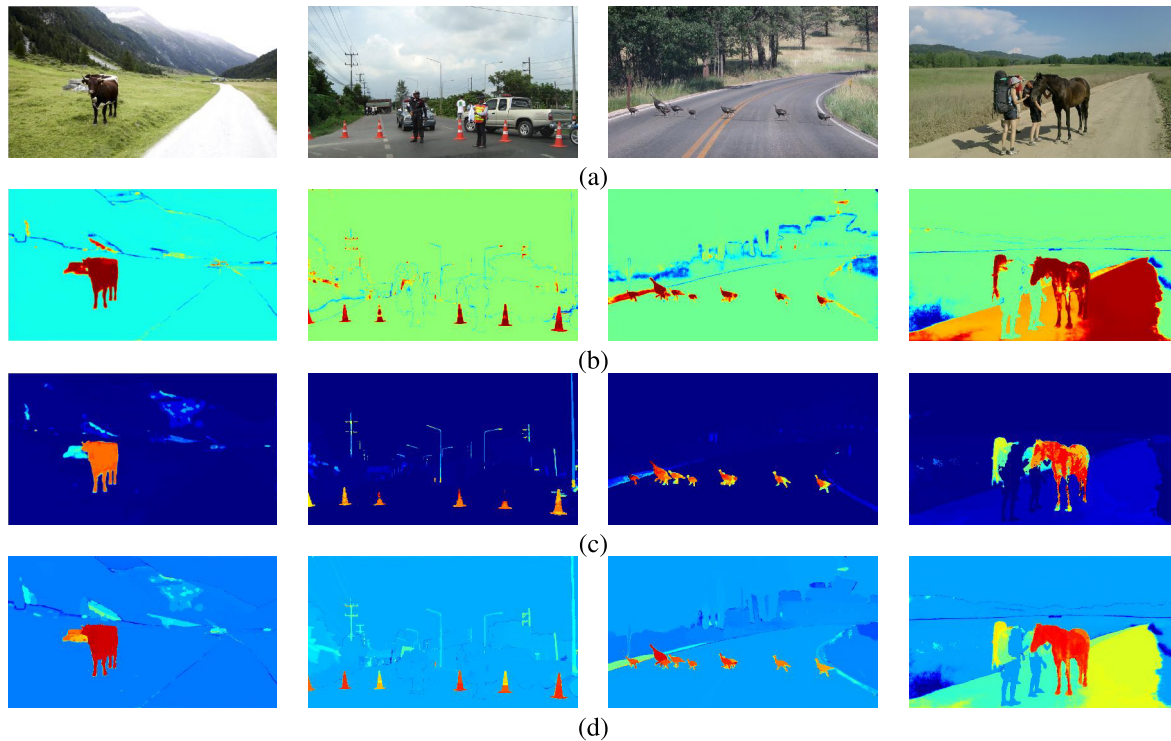


FIGURE 5. Qualitative illustration of intermediate outputs:(a) input image,(b) inlier suppression s_{reject} .(c) anomaly selection s_{accept} .(d) refined anomaly map s_{refine} .

TABLE 5. Ablation study of threshold τ in Eq. (5) on RoadAnomaly dataset.

τ	AP ↑	FPR95 ↓
0.0	81.2	21.5
0.01	82.5	10.5
0.02	81.9	9.2
0.05	82.0	9.9

2) VARIANCE-PENALIZED REGULARIZATION INTERPOLATION SAM typically provides reliable object-level boundaries that help smooth the anomaly map. However, when SAM produces imperfect segments due to ambiguous or imprecise boundaries, the refinement stage may incorrectly merge anomalous and normal regions. To address this issue, we introduce a variance-penalized regularization that reduces the influence of SAM segments whose internal anomaly values disagree with the initial fused map. Figure 4 illustrates this behavior. As shown in Figure 4(a), the input image contains an

TABLE 6. Ablation study of variance-penalized regularization for object-aware refinement on RoadAnomaly dataset.

	AP ↑	FPR95 ↓
without variance-penalized regularization	81.2	11.9
with variance-penalized regularization	82.5	10.5

isolated tire next to a vehicle. The initial anomaly map in (b) correctly highlights the isolated tire. The naive refinement in (c) mistakenly merges the tire with the nearby vehicle because they are grouped into a single SAM segment. In contrast, the variance-aware refinement in (d) identifies the inconsistency between the SAM mask and the initial prediction and preserves a clearer separation between the two objects. This leads to more reliable refinement overall, and Table 6 shows that the variance-penalized regularization improves both AP and FPR95 on the RoadAnomaly dataset.

E. QUALITATIVE RESULTS

Figure 5 illustrates the intermediate outputs of our method. Given the input image in Figure 5(a), the inlier suppression map (b) identifies confidently predicted regions for removal, while the anomaly selection map (c) highlights uncertain areas likely to contain unknown objects. The final refinement (d) combines these signals with object-level spatial consistency, producing coherent anomaly predictions aligned with object boundaries.

Figure 6 provides insight into query-level behavior through an example containing a cow on the road. Given the input image in Figure 6(a), we visualize in (b) the queries with the highest and lowest anomaly scores. This figure also highlights the contribution of the proposed Selection stage by showing that query-level uncertainty alone can localize anomalous regions even before the refinement step, complementing the quantitative results in Table 4. High-scoring queries align with the anomalous object, capturing regions where the model cannot confidently resolve to known classes. In contrast, low-scoring queries correspond to background areas with stable predictions. This visualization demonstrates how query-level signals naturally distinguish between regions of competency and uncertainty, supporting our approach of combining rejection and selection mechanisms.

F. DISCUSSION: COMPETENCY-BASED VS. CONTEXTUAL ANOMALY DETECTION

Our method is grounded in a competency-based perspective: it detects anomalies by identifying regions where the pretrained segmentation model exhibits uncertainty or conflicting predictions across object queries, rather than learning from external outlier examples. This approach effectively identifies what the model does not know, but raises an important question about the definition of anomalies in driving scenarios.

1) COMPETENCY-DEFINED ANOMALIES

Because our approach reflects the model's knowledge boundaries, it identifies anomalies relative to the training distribution rather than contextual abnormality. If an object such as a giraffe or an overturned car is within the pretrained model's label space, the method may confidently classify it as an inlier, even though it is contextually unusual in a driving scene. Figure 7 illustrates such cases.

This behavior highlights a fundamental tension in anomaly segmentation benchmarks: "anomalous" is typically defined with respect to dataset-specific classes (e.g., Cityscapes training set) rather than scene-level risk or semantic appropriateness. From a competency perspective, correctly recognizing a giraffe as "giraffe" demonstrates model capability, not failure. However, from a safety perspective, the presence of a giraffe on a highway represents a critical anomaly regardless of recognition accuracy.

We clarify that all benchmarks follow a semantic OOD protocol based on the Cityscapes label space, where anomalies are defined by class membership rather than contextual

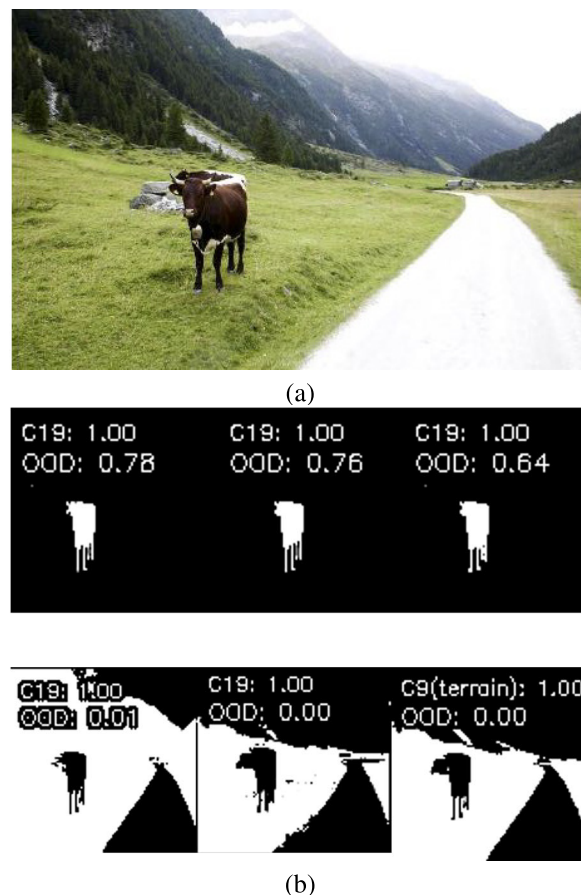


FIGURE 6. Example of query-wise anomaly analysis. (a) Input image. (b) Top-3 queries with high anomaly scores and bottom-3 with low scores, illustrating how queries highlight anomalous versus known regions.

plausibility. As a result, contextual outliers are typically not annotated, making direct quantitative evaluation of such cases infeasible. The examples in Fig 7 therefore serve as qualitative evidence of this limitation, highlighting the gap between competency-based and context-aware anomaly definitions.

2) EXPLORING CONTEXTUAL REASONING WITH FOUNDATION MODELS.

To explore whether foundation models can provide contextual anomaly detection, we experimented with Molmo-72B [25] combined with SAM [28]. Figure 8 shows that this pipeline can identify large, contextually abnormal objects such as animals or fallen trees through high-level vision-language reasoning. However, it struggles with occlusions, small anomalies, and precise localization, achieving substantially lower benchmark performance than query-based zero-shot methods. We note that the performance of foundation models can vary significantly depending on prompt design, SAM configuration, and interaction strategy. In this study, we intentionally limit our comparison to a single-pass, training-free setup to examine how users might directly apply vision-language models without additional tool chaining or reasoning modules. This suggests that while foundation

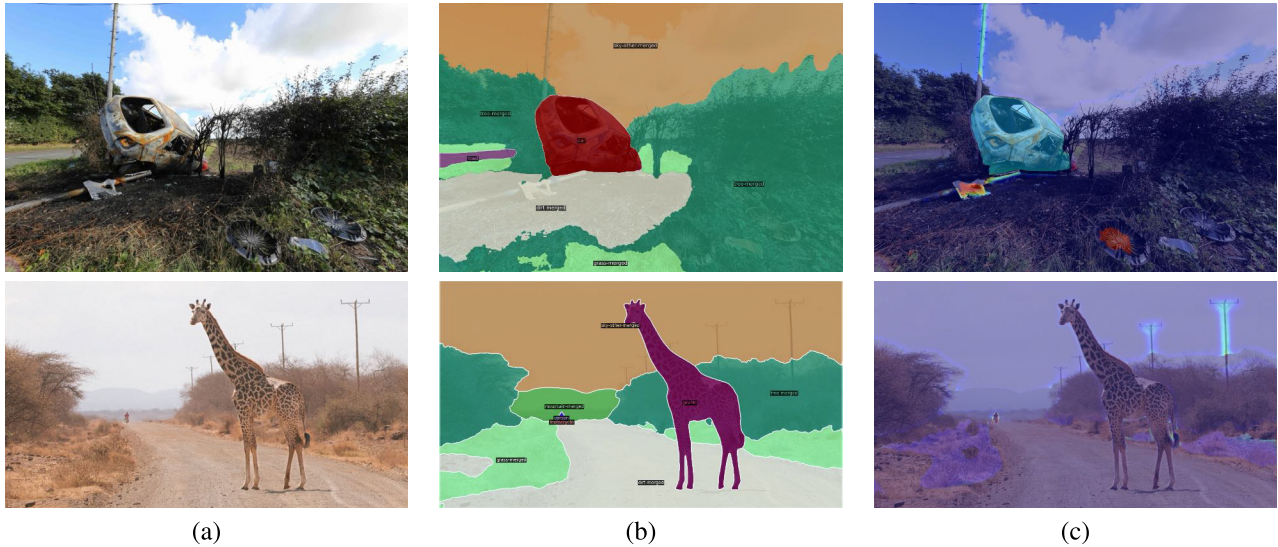


FIGURE 7. Contextual anomalies missed by competency-based detection. (a) Input image, (b) COCO-pretrained segmentation, (c) our anomaly map. Objects like overturned cars or giraffes are confidently classified as known classes, showing that class membership, rather than contextual abnormality, defines anomalies under current benchmarks.

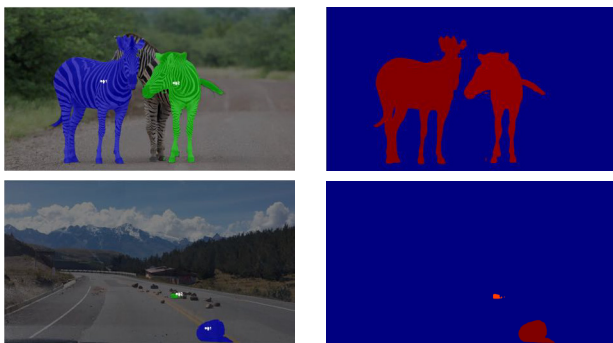


FIGURE 8. Illustration of the Molmo+SAM pipeline for contextual anomaly detection. While capable of highlighting large abnormal objects, it struggles with small or occluded cases and yields limited benchmark performance.

models offer contextual awareness, they remain insufficient for precise anomaly segmentation without additional task-specific adaptation. Recent studies have also explored multi-agent or reasoning-based utilization of vision-language models, where VLMs iteratively generate hypotheses and collaborate with detectors or segmentation tools [30], [52]. Such multi-step pipelines, while promising, are beyond the scope of this paper but represent a complementary direction toward integrating contextual reasoning and visual grounding. For clarity, we include the exact prompts used for this experiment and the corresponding Molmo-72B pointing outputs, including both successful and failed cases, in the Appendix.

Our findings indicate that competency-based and context-based anomaly detection address different aspects of the problem. Competency-based methods excel at identifying distribution shifts and model uncertainty, providing interpretable signals about recognition boundaries. Contextual approaches could complement this by reasoning about

semantic appropriateness, though current foundation models require further development for this purpose. Future work may integrate both perspectives, combining model competency analysis with contextual understanding to better align anomaly detection with real-world safety requirements. More broadly, this highlights an open question: in modern vision-language systems, “unknown” may need to account not only for unseen classes but also for contextual inconsistency. Integrating competency- and context-based signals remains an important direction for future work.

G. LIMITATIONS AND FUTURE WORKS

Although the proposed framework achieves consistent results across multiple driving-scene benchmarks, it has several inherent limitations. First, our method focuses on detecting regions of low model competency rather than contextual abnormality. As a result, objects that lie within the model’s training distribution but appear in unusual contexts may not be identified as anomalous. Second, our approach relies on transformer-based segmentation architectures with object-query decoders, since query-level representations are central to the competency-based formulation. Models without such query tokens cannot be directly integrated without additional components such as pseudo-query generators or region-level descriptors.

Beyond these limitations, the competency-based perspective suggests opportunities for more fine-grained anomaly analysis. A natural extension is a part-level competency framework for industrial inspection or medical imaging, where objects can be over-segmented into meaningful parts and anomaly cues emerge from local competency failures. Such a formulation could bridge training-free segmentation with

fine-grained zero-defect inspection, extending the applicability of competency-based reasoning to broader domains.

Another valuable direction concerns integrating competency-based signals with contextual reasoning mechanisms. As discussed in Section IV-F, vision-language models and open-set detectors offer complementary capabilities for identifying semantically inappropriate objects, suggesting that competency cues alone may be insufficient in context-driven anomaly scenarios. A future research direction is to couple query-level uncertainty analysis with a complementary contextual pathway by incorporating a global scene-reasoning branch that provides high-level semantic priors and modulates the entropy-based query selection process. Such a dual-stream design would enable the model to capture both fine-grained uncertainty patterns and broader contextual inconsistencies, offering a principled route toward unifying competency-based and context-based anomaly detection.

V. CONCLUSION

We have presented a zero-shot anomaly segmentation framework that detects out-of-distribution regions by analyzing the internal signals of pretrained transformer-based segmentation models. Our approach operates without retraining or auxiliary data, integrating query-level uncertainty analysis through rejection and selection mechanisms with object-aware spatial refinement. Experiments across multiple benchmarks demonstrate state-of-the-art performance among zero-shot methods while remaining competitive with supervised approaches. By exploiting model-internal competency signals, our framework provides a scalable solution for anomaly detection that can identify distribution shifts without task-specific adaptation.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work the authors used OpenAI ChatGPT and Claude in order to improve the readability and clarity of the manuscript text. After using this tool, they reviewed and edited the content as needed and take full responsibility for the content of the publication.

APPENDIX COMPARISON TO VISION-LANGUAGE MODELS

A. PROMPT DESIGN

Vision-language models such as Molmo-72B offer the ability to reason about scene semantics and to describe objects that do not belong in typical road environments. However, in order to evaluate such models in a fair and reproducible manner within the zero-shot anomaly segmentation setting, we designed a prompt that focuses the model strictly on identifying anomalous objects rather than describing the entire scene. Molmo-72B supports pointing functionality, where the model attempts to localize objects by predicting approximate image coordinates. We therefore crafted the prompt to explicitly request: (i) concise object names for abnormal items, and (ii) the corresponding center coordinates normalized to a

```

prompt = ""
This is a scene of a road-driving scenario. Focus ONLY on identifying the
PRECISE CENTER COORDINATES of UNUSUAL or ABNORMAL
objects on the road.

IGNORE normal road elements such as:
- Cars, trucks, buses, motorcycles, bicycles
- Pedestrians, cyclists
- Traffic signs, traffic lights
- Road markings, guardrails
- Buildings, trees, lampposts
- Normal road debris like small rocks or branches

FOCUS ONLY on clearly abnormal/unusual objects like:
- Large animals (cows, deer, bears, etc.)
- Large obstacles (boulders, fallen trees, large debris)
- Unusual items (furniture, appliances, large objects)
- Flooding, fire, or other hazardous conditions
- Damaged infrastructure that creates a hazard

For each ABNORMAL object, provide ONLY:
1. Brief, specific object name
2. EXACT center coordinates in (x, y) format with one decimal place
precision

IMPORTANT NOTES ABOUT COORDINATES:
- Coordinates must be between 0.0 and 100.0
- (0.0, 0.0) is the TOP-LEFT corner of the image
- (100.0, 100.0) is the BOTTOM-RIGHT corner of the image
- Use decimal points for precision (e.g., 45.5, not just 45)

Format your response EXACTLY as:
- [Object name]: (x.x, y.y)
- [Object name]: (x.x, y.y)
- ...

If no ABNORMAL objects are detected, simply state "No anomalies detected".
""

```

FIGURE 9. Structure of the prompt used for anomaly localization with Molmo-72B.

fixed 0–100 range. The prompt also instructs the model to ignore all common inlier categories in road-driving scenes (for example, cars, pedestrians, buildings, road furniture) and to focus only on unusual or hazardous entities such as large animals, unexpected obstacles, or damaged infrastructure. This design encourages the model to output structured anomaly candidates that can later be paired with SAM masks for segmentation. Figure 9 illustrates the overall structure of the prompt used in our evaluation. The same prompt was applied consistently across all datasets and all images, without per-image tuning or manual intervention.

B. QUALITATIVE RESULT OF MOLMO-72B

Figure 10 and Figure 11 present representative qualitative results illustrating the strengths and limitations of Molmo-72B when used with the structured pointing prompt introduced in Figure 9. The positive examples (Figure 10) demonstrate that Molmo reliably identifies large, salient anomalous objects, including cows, hikers with animals, fallen structures, and crossing wildlife. In these cases, the anomalies are visually distinct from the surrounding road environment, and Molmo's built-in pointing mechanism produces stable coordinate predictions.



FIGURE 10. Successful anomaly localizations from Molmo-72B using the proposed prompt.



FIGURE 11. Failure cases of Molmo-72B, showing incorrect or missing anomaly localization.

However, the failure cases in Figure 11 reveal several systematic challenges. For instance, when animals are densely clustered (that is, herds of sheep or cows) or partially occluded, Molmo often predicts only a single point for the entire group, produces inconsistent coordinate outputs, or completely misses relevant anomalous instances. Similarly, for small or distant hazards, such as cones, small debris, or compact animals, the model fails to produce stable coordinates, confirming that its pointing mechanism is not designed for fine-grained anomaly localization. These observations provide additional qualitative evidence that general-purpose VLMs are not yet effective as one-shot segmentation engines for anomaly detection.

Recent work suggests that multi-agent pipelines, in which a VLM performs scene reasoning while a specialized open-set object detector (for example, Grounding DINO) performs localization, substantially improve robustness.

While promising, such hybrid systems require multi-stage coordination and architectural modification, which is beyond the scope of our training-free, transformer-internal approach. Thus, these results reaffirm that VLM pointing alone is insufficient for pixel-level anomaly segmentation and further justify evaluating Molmo-72B only at a qualitative level in this study.

REFERENCES

- [1] S. Pohland and C. Tomlin, "Understanding the dependence of perception model competency on regions in an image," in *Proc. World Conf. Explainable Artif. Intell.* Cham, Switzerland: Springer, 2024, pp. 130–154.
- [2] H.-I. Kim, K. Yun, J.-S. Yun, and Y. Bae, "Task-specific adaptation of segmentation foundation model via prompt learning," 2024, *arXiv:2403.09199*.
- [3] N. Nayal, M. Yavuz, J. F. Henriques, and F. Güney, "RbA: Segmenting unknown regions rejected by all," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 711–722.

- [4] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2017.
- [5] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2018.
- [6] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018.
- [7] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018.
- [8] S. Jung, J. Lee, D. Gwak, S. Choi, and J. Choo, "Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15425–15434.
- [9] J. Ackermann, C. Sakaridis, and F. Yu, "Maskomaly: Zero-shot mask anomaly segmentation," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2023.
- [10] M. Grcić, J. Šarić, and S. Šegvić, "On advantages of mask-level recognition for outlier-aware segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 2937–2947.
- [11] Y. Lei, L. Ji, and P. Liu, "Mining in-distribution attributes in outliers for out-of-distribution detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 17, Apr. 2025, pp. 18181–18188. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/34000>
- [12] K. Lis, K. K. Nakka, P. Fua, and M. Salzmann, "Detecting the unexpected via image resynthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2152–2161.
- [13] G. Di Biase, H. Blum, R. Siegwart, and C. Cadena, "Pixel-wise anomaly detection in complex driving scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16913–16922.
- [14] R. Chan, M. Rottmann, and H. Gottschalk, "Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5108–5117.
- [15] M. Grcić, P. Bevanđić, and S. Šegvić, "DenseHybrid: Hybrid anomaly detection for dense open-set recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 500–517.
- [16] Y. Tian, Y. Liu, G. Pang, F. Liu, Y. Chen, and G. Carneiro, "Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes," in *Proc. Eur. Conf. Comput. Vis.*, 2021, pp. 246–263.
- [17] Y. Lei, L. Ji, and P. Liu, "WSOE: Weakly supervised outlier exposure for object-level out-of-distribution detection," *Expert Syst. Appl.*, vol. 270, Apr. 2025, Art. no. 126507.
- [18] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016.
- [19] A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," 2015, *arXiv:1511.02680*.
- [20] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1050–1059.
- [21] C. Corbière, N. Thome, A. Bar-Hen, K. Bailly, and P. Pérez, "Addressing failure prediction by learning model confidence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019.
- [22] T. Besnier, G. Csurka, M. Bolaños, R. Valle, and M. Salzmann, "Triggering failures: Out-of-distribution detection by learning from local adversarial attacks in semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15701–15710.
- [23] Y. Liu, C. Ding, Y. Tian, G. Pang, V. Belagiannis, I. Reid, and G. Carneiro, "Residual pattern learning for pixel-wise out-of-distribution detection in semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 1151–1161.
- [24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
- [25] M. Deitke et al., "Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models," 2024, *arXiv:2409.17146*.
- [26] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer, "WinCLIP: Zero-/few-shot anomaly classification and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19606–19616.
- [27] M. Lee and J. Choi, "Text-guided variational image generation for industrial anomaly detection and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 26509–26518.
- [28] A. M. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Lo, P. Dollár, and R. Girshick, "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Apr. 2023, pp. 4015–4026.
- [29] W. Zhao, J. Li, X. Dong, Y. Xiang, and Y. Guo, "Segment every out-of-distribution object," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 3910–3920.
- [30] J. Song, K. Yun, D. Jo, J. Kim, and Y. Yoo, "CoT-segmenter: Enhancing OOD detection in dense road scenes via chain-of-thought reasoning," in *Proc. IEEE Int. Conf. Adv. Vis. Signal-Based Syst. (AVSS)*, Aug. 2025, pp. 1–6.
- [31] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Mar. 1999, pp. 246–252.
- [32] K. Yun, H.-I. Kim, K. Bae, and J. Moon, "Background memory-assisted zero-shot video object segmentation for unmanned aerial and ground vehicles," *ETRI J.*, vol. 45, no. 5, pp. 795–810, Oct. 2023.
- [33] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1280–1289.
- [34] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 833–851.
- [35] Y. Zhu, K. Sapra, F. A. Reda, K. J. Shih, S. Newsam, A. Tao, and B. Catanzaro, "Improving semantic segmentation via video propagation and label relaxation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8856–8865.
- [36] H. Zhang, F. Li, Q. Lu, M.-H. Yang, and N. Ahuja, "CSL: Class-agnostic structure-constrained learning for segmentation including the unseen," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 7, pp. 7078–7086.
- [37] M. Sodano, F. Magistri, L. Nunes, J. Behley, and C. Stachniss, "Open-world semantic segmentation including class similarity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 3184–3194.
- [38] A. Delić, M. Grcić, and S. Šegvić, "Outlier detection by ensembling uncertainty with negative objectness," 2024, *arXiv:2402.15374*.
- [39] T. Vojří, J. Šochman, and J. Matas, "PixOOD: Pixel-level out-of-distribution detection," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 93–109.
- [40] C. W. Lee, S. Leveugle, S. Stolpner, C. Langley, P. Grouchy, J. Kelly, and S. L. Waslander, "FlowCLAS: Enhancing normalizing flow via contrastive learning for anomaly segmentation," 2024, *arXiv:2411.19888*.
- [41] T. Vojří, T. Šipka, R. Aljundi, N. Chumerin, D. O. Reino, and J. Matas, "Road anomaly detection by partial image reconstruction with segmentation coupling," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15631–15640.
- [42] V. Besnier, A. Bursuc, D. Picard, and A. Briot, "Triggering failures: Out-of-distribution detection by learning from local adversarial attacks in semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15681–15690.
- [43] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
- [44] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
- [45] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, and D. Song, "Scaling out-of-distribution detection for real-world settings," 2019, *arXiv:1911.11132*.
- [46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [47] M. Grcić and S. Šegvić, "Hybrid open-set segmentation with synthetic negative data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 10, pp. 6748–6760, Oct. 2024.

- [48] L. Chen, W. Wang, J. Miao, and Y. Yang, "GMMSeg: Gaussian mixture based generative semantic segmentation models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 31360–31375.
- [49] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Álvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.
- [50] R. Chan, K. Lis, S. Uhlemeyer, H. Blum, S. Honari, R. Siegwart, M. Salzmann, P. Fua, and M. Rottmann, "SegmentMeIfYouCan: A benchmark for anomaly segmentation," in *Proc. 35th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track*, 2021.
- [51] H. Blum, P.-E. Sarlin, J. Nieto, R. Siegwart, and C. Cadena, "The fishyscapes benchmark: Measuring blind spots in semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3119–3135, Nov. 2021.
- [52] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang, "Grounded SAM: Assembling open-world models for diverse visual tasks," 2024, *arXiv:2401.14159*.



KIMIN YUN received the B.S. degree in electrical engineering and the integrated M.S./Ph.D. degree in electrical and computer engineering from Seoul National University, Republic of Korea, in 2010 and 2017. Since 2017, he has been with the Visual Intelligence Research Section of the Artificial Intelligence Research Laboratory, Electronics and Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea. In 2025, he joined the University of Science and Technology (UST),

as an Associate Professor of artificial intelligence. His current research interests include machine learning and computer vision, with a focus on real-world challenges and the development of practical applications.



YUSEOK BAE received the B.S., M.S., and Ph.D. degrees in computer science from Kyungpook National University (KNU), Daegu, Republic of Korea, in 1995, 1997, and 2011, respectively. Since 1997, he has been with the Electronics and Telecommunications Research Institute (ETRI), where he has been involved in various research projects, including distributed computing, home-network and IPTV middleware, big-data analytics, and visual artificial intelligence. He is a Principal Researcher with the Visual Intelligence Research Section, Artificial Intelligence Research Laboratory, ETRI. His current research interests include computer vision, machine learning, deep learning, and artificial intelligence.

• • •