NO TRUST ISSUES HERE: TECHNICAL REPORT FOR THE RAYAN AI CONTEST

Ali Nafisi
Department of Computer Engineering
Bu-Ali Sina University
a.nafisi@eng.basu.ac.ir

Sina Asghari*
Department of Computer Science
Iran University of Science and Technology
sina_asghari@mathdep.iust.ac.ir

Mohammad Saeed Arvenaghi*
Department of Computer Science
Iran University of Science and Technology
m_arvenaghi@mathdep.iust.ac.ir

Hossein Shakibania
Department of Computer Science
Technical University of Darmstadt
hossein.shakibania@stud.tu-darmstadt.de

ABSTRACT

This report presents solutions to three machine learning challenges: compositional image retrieval, zero-shot anomaly detection, and backdoored model detection. In compositional image retrieval, we developed a system that processes visual and textual inputs to retrieve relevant images, achieving 95.38% accuracy and ranking first with a clear margin over the second team. For zero-shot anomaly detection, we designed a model that identifies and localizes anomalies in images without prior exposure to abnormal examples, securing 2nd place with 73.14% accuracy. In the backdoored model detection task, we proposed a method to detect hidden backdoor triggers in neural networks, reaching an accuracy of 78%, which placed our approach in second place. These results demonstrate the effectiveness of our methods in addressing key challenges related to retrieval, anomaly detection, and model security, with implications for real-world applications in industries such as healthcare, manufacturing, and cybersecurity. Code for all solutions is available online.[2]

Keywords Compositional Retrieval · Zero-Shot Anomaly Detection · Backdoored Model Detection

1 Introduction

Traditional machine learning systems often struggle with scenarios where data distributions shift or when input data is incomplete or noisy. This report addresses three such challenges, each focusing on improving the reliability, flexibility, and security of machine learning systems. As the demand for intelligent systems grows across industries, the need for models that can process both visual and textual data, detect anomalies in unseen distributions, and safeguard against adversarial manipulations becomes increasingly critical.

This report covers three distinct tasks: compositional image retrieval, zero-shot anomaly detection, and backdoored model detection. The first task involves developing a system capable of retrieving images based on multi-modal queries, combining visual content with textual modifications. The second task focuses on detecting anomalies, particularly in industrial and medical settings. The third task aims to detect backdoor attacks in deep learning models, where a model's performance is manipulated by malicious inputs or model weights during training. Each challenge presents unique objectives, including efficient retrieval, generalization to unseen distributions, and identifying security vulnerabilities in models.

The evaluation of these solutions is based on a set of standard metrics, such as accuracy, and F1 score, to assess both the quality and robustness of the approaches. In this report, we describe the methodologies we employed, the experiments

*Equal contribution.

[2] https://github.com/safinal/rayan-ai-contest-solutions

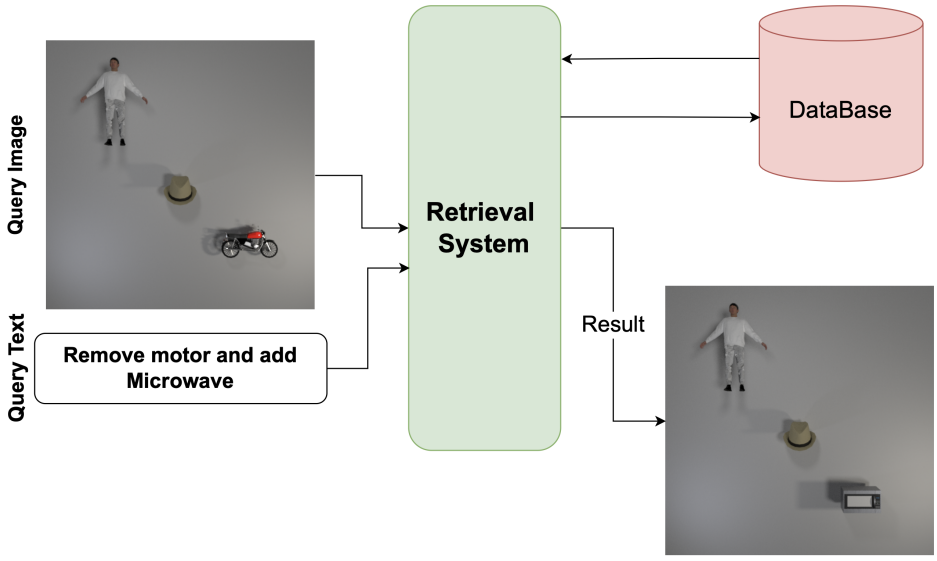


Figure 1: An overview of the Compositional Image Retrieval problem.

conducted, and the results obtained, demonstrating the effectiveness of our solutions. Our team, No Trust Issues Here, achieved first place in the competition, with a final score of 95.38% for the compositional image retrieval challenge (Q5), 73.14% for zero-shot anomaly detection (Q6), and 78% for backdoored model detection (Q7). The final leaderboard is shown in Table 1.

Table 1: Final leaderboard showing the top 10 teams and their scores for the three challenges.

Rank	Team	Q5 (%)	Q6 (%)	Q7 (%)	Final Score (Out of 300)
1	**No Trust Issues Here (Our Team)**	**95.38**	**73.14**	**78**	**298**
2	Pileh	84.61	74.92	67	292
3	AI Guardians of Trust	88.59	66.29	72	291
4	AIUoK	78.32	72.98	70	286
5	red_serotonin	85.45	63.51	65	284
6	Persistence	80.63	61.62	74	282
7	GGWP	80.97	62.25	63	277
8	Tempest	69.86	70.88	63	274
9	AlphaQ	82.89	62.15	60	272
10	Cogniverse	49.71	61.53	63	258

2 Q5: Compositional Image Retrieval

The objective of the Compositional Image Retrieval challenge is to develop an intelligent search system capable of processing multi-modal queries. The system must accept a reference image containing a visual scene and a textual modification description (e.g., "Remove motor and add Microwave") as input. The goal is to identify and retrieve the most relevant image from a database that reflects the visual changes described in the text. Figure 1 visualizes the compositional retrieval problem.

This task presents a unique challenge in that it requires the model to understand both the visual content of the reference image and the semantic intent of the textual instructions to perform accurate retrieval, all while adhering to strict resource constraints such as a 4GB model size limit and a prohibition on large language models or object detection models during inference.

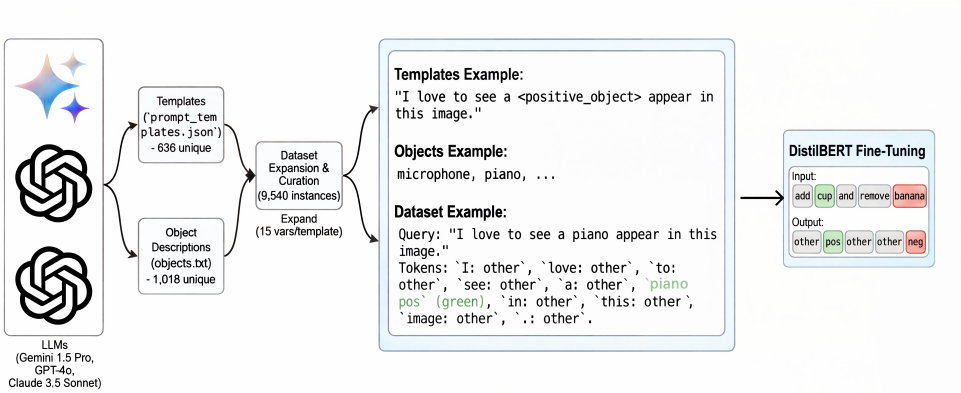


Figure 2: Overview of our token classification module.

2.1 Methodology

To address the challenge of compositional retrieval without relying on prohibited heavy architectures, we propose a two-stage pipeline: *Token Classification* and *Compositional Embedding Arithmetic*. This approach provides a lightweight method for modifying visual representations using textual instructions while remaining within the competition constraints. All code associated with our implementation is available at the provided repository. [3]

2.1.1 Token Classification for Semantic Parsing

The first stage of our pipeline is the Token Classification model, which is responsible for interpreting the query text and identifying specific modifications (i.e., which objects to add or remove), as illustrated in Figure 2. This task is framed as a token classification problem where the model classifies tokens into three categories:

- Positive (pos): Objects to be added to the scene.
- Negative (neg): Objects to be removed from the scene.
- Other: Irrelevant tokens (stopwords, punctuation, verbs).

To create training data, we generated a synthetic dataset using publicly available LLM interfaces (Gemini 1.5 Pro, GPT-4o, and Claude 3.5 Sonnet). No API access was used. We built 636 template sentences containing placeholders for positive and negative objects. We additionally generated a vocabulary of 1,018 unique object names. Each template was instantiated 15 times by replacing the placeholders with random objects from the list, producing 9,540 labeled queries.

Labels were assigned automatically. For a query such as “add apple and remove banana”, the sequence is annotated as <other> <positive_object> <other> <other> <negative_object>. This procedure ensures consistent detection of objects to add and remove.

We fine-tuned `distilbert-base-uncased` [Sanh et al., 2019]. DistilBERT was selected due to its small size and adequate performance for this restricted classification task. The model was trained for 20 epochs with a weight decay of 0.01, 500 warm-up steps, and standard token-level cross-entropy. This stage outputs two sets: detected positive objects and detected negative objects.

2.1.2 Compositional Embedding Arithmetic for Retrieval

The second stage retrieves the target image by modifying the embedding of the reference image according to the objects identified in the query. We evaluated many models from OpenCLIP [Ilharco et al., 2021] and selected ViTamin-L-384 [Chen et al., 2024] due to its superior performance in preliminary experiments. We fine-tune only the visual head (`model.feature_extractor.visual.head`), keeping the rest of the backbone fixed to remain within the resource limits.

We optimize the retrieval model using InfoNCE loss [Oord et al., 2018]. The objective is to make the transformed embedding of the query image

$$\mathbf{v}_{\text{img}} + \mathbf{t}_{\text{pos}} - \mathbf{t}_{\text{neg}}$$

[3] `https://github.com/safinal/compositional-image-retrieval`

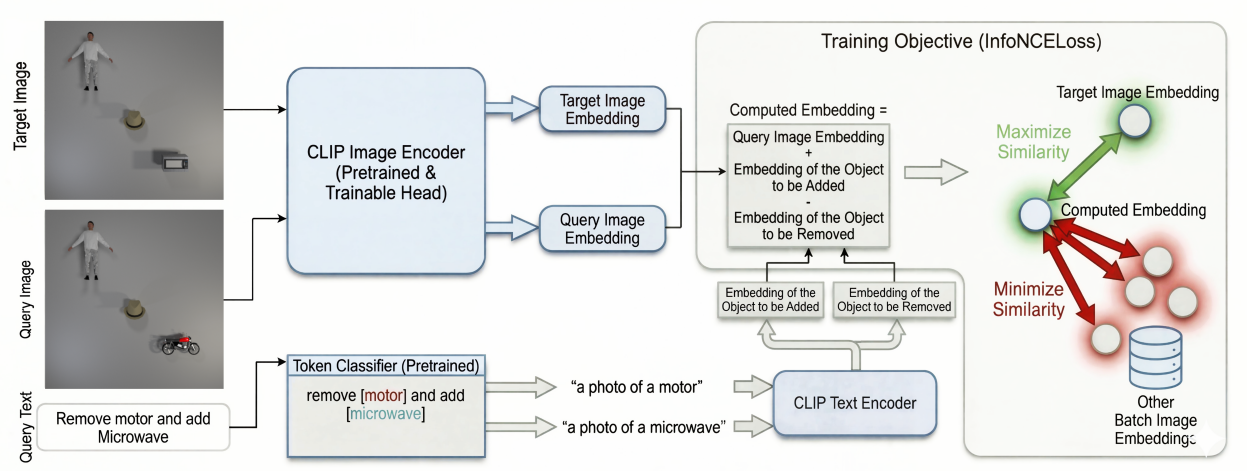


Figure 3: Overview of the proposed compositional image retrieval method. The system parses textual modifications using a lightweight token classifier and applies embedding arithmetic to adjust the reference image representation before retrieval.

similar to the embedding of the ground-truth target image. In practice, the modification embeddings are computed by encoding the text prompts "a photo of a [object]" using the model's text encoder. During training, this encourages the model to learn a consistent direction in the embedding space for each object addition or removal.

Training uses AdamW (learning rate 0.0001, weight decay 0.01), five training epochs, and a loss temperature of 0.07. We further apply a CosineAnnealingWarmRestarts scheduler with $T_0 = 5$ and $T_{\text{mult}} = 2$. Because the dataset contains repeated occurrences of the same target image, we implement a custom sampler to ensure that each batch contains unique target images, which is required for the InfoNCE formulation.

During inference, the Token Classifier identifies all positive and negative objects in the text. We then encode the corresponding text prompts to obtain the positive embeddings $\{\mathbf{t}_{\text{pos}}^{(i)}\}$ and negative embeddings $\{\mathbf{t}_{\text{neg}}^{(j)}\}$. Given the reference image embedding $\mathbf{v}_{\text{img}}$, we compute:

$$\mathbf{v}_{\text{target}} = \mathbf{v}_{\text{img}} + \sum_i \mathbf{t}_{\text{pos}}^{(i)} - \sum_j \mathbf{t}_{\text{neg}}^{(j)}.$$

We then compute cosine similarity between $\mathbf{v}_{\text{target}}$ and all database image embeddings, and return the image with the highest similarity score.

This method does not rely on object detection, captioning models, or large language models during inference. By operating directly in the latent space, the approach remains efficient while enabling precise compositional retrieval, as illustrated in Figure 2.

2.2 Experimental Results

The evaluation was conducted on a test set designed to assess generalization capability through diverse text descriptions and novel visual combinations. The primary metric for performance was *Top-1 Accuracy*.

Our solution demonstrated superior performance, achieving the highest accuracy among all participating teams. The model successfully handled complex queries involving simultaneous addition and removal of objects. Table 2 shows the top 10 leaderboard rankings for this problem.

With a final accuracy of 95.38%, our approach outperformed the second-place entry by a significant margin of 6.79%, validating the effectiveness of combining explicit semantic parsing with vector arithmetic in latent space.

3 Q6: Zero-Shot Anomaly Detection

Reliability in machine learning models is often hindered by the closed-world assumption, where test data are presumed to match training distributions. In real-world scenarios, this assumption frequently fails as models encounter outlier

Table 2: Top 10 Leaderboard Rankings for the Compositional Image Retrieval Challenge.

Rank	Team	Accuracy (%)
1	**No Trust Issues Here (Ours)**	**95.38**
2	AI Guardians of Trust	88.59
3	red_serotonin	85.45
4	Pileh	84.61
5	AlphaQ	82.89
6	GGWP	80.97
7	Persistence	80.63
8	Synapse	78.66
9	AIUoK	78.32
10	fatem17	75.76

samples. To address this, we focused on the task of "Anomaly Detection," specifically within the challenging setting of zero-shot learning.

Unlike conventional settings that rely on normal samples for training, the zero-shot approach requires the model to operate without seeing any data—normal or anomalous—from the test-time distribution during its training phase. The only source of knowledge available regarding the test distribution is the implicit information contained within the unlabeled test set itself.

The problem is formulated as follows: given a set of test images $D_{test} = \{I^1_{test}, ..., I^{n_{test}}_{test}\}$, and potentially a set of auxiliary training images D_{aux} distinct from the test distribution, the goal is to determine whether a test image is anomalous and localize the specific anomalous regions. The output for each sample consists of two components:

- Classification Score ($s_{img} \in \mathbb{R}$): An image-level scalar indicating the degree of anomaly.
- Segmentation Score ($s_{pix} \in \mathbb{R}^{H \times W}$): A pixel-level map localizing the defect.

This task targets both industrial and medical domains, requiring the model to generalize across eight industrial classes and two medical classes. In these domains, the distribution shift between normal and abnormal data is primarily a covariate shift, as they share the same semantics (e.g., a normal pill vs. a broken pill). Figure 4 shows examples of normal and anomalous images from three datasets (capsules, photovoltaic modules, and pills). For each dataset, we display a normal sample, an anomalous sample, and the corresponding anomaly mask.

3.1 Methodology

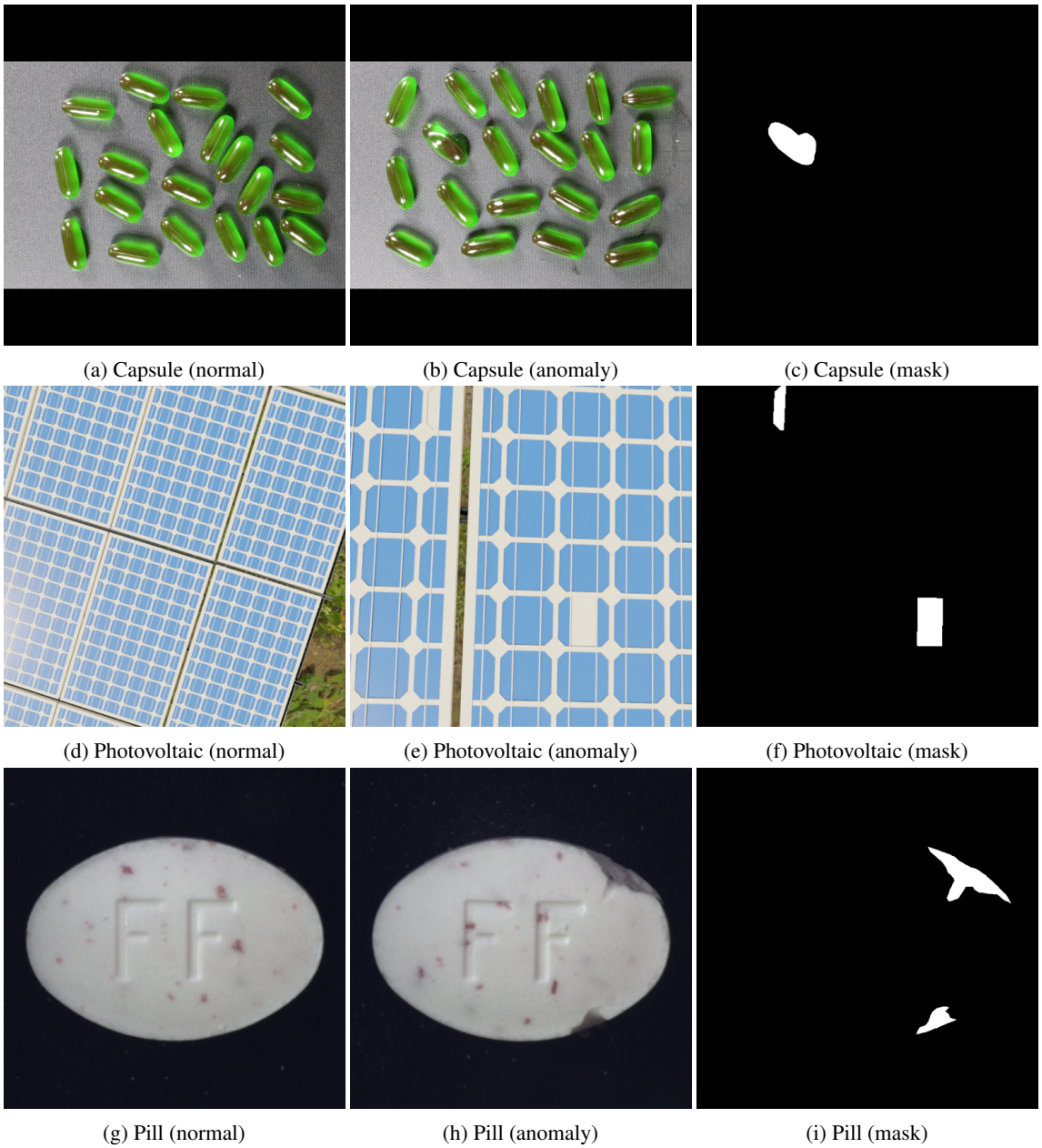
After reviewing several methods, including AnomalyCLIP [Zhou et al., 2025], we selected the MuSc method [Li et al., 2024] because it achieved the strongest results among available zero-shot anomaly detection approaches. MuSc is also the first zero-shot method that does not require class descriptions, unlike CLIP-based models.

Our approach follows the MuSc framework [Li et al., 2024], which performs zero-shot anomaly detection using only unlabeled test images. MuSc is based on the observation that normal image patches tend to recur across many test samples, whereas anomalous patches are uncommon. Patch features are extracted with a Vision Transformer, and each patch is scored by comparing it to patches from the remaining images.

MuSc contains three main components. The first is the Local Neighborhood Aggregation with Multiple Degrees (LNAMD), which aggregates patch features at different neighborhood sizes to capture anomalies at multiple scales. The second is the Mutual Scoring Mechanism (MSM), which assigns patch-level anomaly scores by measuring how often a patch finds similar counterparts in the test set, using an interval-average over the lowest similarity values to reduce noise. The third component, Re-scoring with Constrained Image-level Neighborhood (RsCIN), refines image-level anomaly scores by constructing a constrained neighborhood graph and enforcing consistency on images with similar global features.

Building on this framework, we performed a systematic search over hyperparameters and selected settings that consistently improved performance on both image-level and pixel-level metrics. Our final configuration used feature layers $\{5, 11, 17, 23\}$, $r \in \{1, 3\}$, $k_{score} \in \{1, 8, 9\}$. We also adopted a combined model design, using [Oquab et al., 2023] (dinov2-vitl14) for segmentation and [Radford et al., 2021] (ViT-L-14-336) for classification.

During evaluation, we identified one difficult class (photovoltaic modules) with weak performance. We conducted targeted error analysis on these cases and applied class-specific adjustments that improved their results without harming



(a) Capsule (normal) (b) Capsule (anomaly) (c) Capsule (mask)

(d) Photovoltaic (normal) (e) Photovoltaic (anomaly) (f) Photovoltaic (mask)

(g) Pill (normal) (h) Pill (anomaly) (i) Pill (mask)

Figure 4: Examples of normal images, anomalous images, and anomaly masks across three datasets: capsules, photovoltaic modules, and pills.

performance on other classes. We performed error analysis using only the three industrial classes provided for local development and then tuned hyperparameters on that validation set. No labels or statistics from the official blind evaluation classes were used for tuning.

We introduced a classical computer vision filter to produce a binary mask used for post-processing the anomaly maps. This step was motivated by a limitation we observed in MuSc. Because the method relies only on patch dissimilarities and has no prior information or class description, it implicitly assumes that normal samples have consistent textures. In some classes, such as photovoltaic modules in Figure 4, normal images contain background regions (grass, soil) that differ strongly from the object surface. These surrounding areas can be incorrectly treated as anomalies. To reduce this effect, the binary mask defines a narrow margin around the main object. Pixels with mask value 1 keep their original anomaly values, and all other pixels are set to the minimum of the anomaly map. This post-processing step improved stability and enhanced pixel-level localization across multiple classes.

For image-level classification, instead of taking only the highest patch anomaly score per image, we obtained better results by combining the two highest patch scores. The second-highest value is included with a weight of 0.25, which improved robustness compared to a single maximum score. All code associated with our implementation is available at the provided repository.[4]

3.2 Experimental Results

The performance of our submission was evaluated using a rigorous weighted average of seven distinct metrics, assessing precision, recall, and specificity at both the image and pixel levels. The inference phase was executed on an NVIDIA GeForce RTX 4090 GPU with a maximum time limit of 3 hours.

The final evaluation score was computed using a weighted average of all metrics:

$$Final\ score = \frac{\sum_{metric \in Metrics} \mathcal{W}_{metric} \times metric}{\sum_{metric \in Metrics} \mathcal{W}_{metric}}$$

The specific weights assigned to each metric were:

- Image-level: $w_{img\_AUROC} = 1.2$, $w_{img\_AP} = 1.1$, $w_{img\_F1} = 1.1$
- Pixel-level: $w_{pix\_AUROC} = 1.0$, $w_{pix\_AUPRO} = 1.4$, $w_{pix\_AP} = 1.3$, $w_{pix\_F1} = 1.3$

Table 3 shows the leaderboard for this problem. Our team, No Trust Issues Here, achieved the highest performance in the challenge. By leveraging the logic detailed in the methodology section, our solution attained a final accuracy of 73.14%, securing the 2nd rank on the leaderboard. This performance demonstrated superior generalization across the diverse industrial and medical classes provided in the blind test set.

Table 3: Top 10 Leaderboard Rankings for the Zero-Shot Anomaly Detection Challenge.

Rank	Team	Score (%)
1	Pileh	74.92
2	**No Trust Issues Here (Ours)**	**73.14**
3	AIUoK	72.98
4	Tempest	70.88
5	AI Guardians of Trust	66.29
6	red_serotonin	63.51
7	CortexAI	62.35
8	GGWP	62.25
9	AlphaQ	62.15
10	Persistence	61.62

4 Q7: Backdoored Model Detection

The objective of this challenge is to develop a discrimination function capable of distinguishing between clean and backdoored neural networks. As deep neural networks are increasingly deployed in mission-critical applications, the risk of attackers poisoning training data to embed imperceptible backdoors has become a significant security concern.

The threat model involved attackers capable of data poisoning and training influence to create stealthy, label-consistent, or sample-specific triggers. The defense mechanism was required to operate without prior knowledge of the attack type or trigger structure, relying only on the model parameters and a small set of clean samples (1% of the test dataset).

4.1 Methodology

Our approach builds on the Mm-Bd method [Wang et al., 2024], which reaches approximately 60% accuracy, and extends it by introducing feature-space optimization, improved initialization, image resizing, and a refined statistical detection rule. These modifications address the constraint of model-only detection and lead to substantially higher accuracy. All code associated with our implementation is available at the provided repository. [5]

[4] `https://github.com/safinal/zeroshot-anomaly-detection`

[5] `https://github.com/safinal/backdoored-model-detection`

4.1.1 Activation Perturbation Optimization

Since the original triggers are unavailable, we attempt to reverse-engineer the model's sensitivity to potential trigger patterns. The pipeline operates on the PreActResNet18 architecture [He et al., 2016] as follows:

1. **Initialization:** For each class c, we sample a batch of clean images from the verification set. Before optimization, we resize all input images to 128×128, as this resolution consistently improved accuracy compared to the default preprocessing.
2. **Forward Pass & Activation Extraction:** The resized images are passed through the model up to `model.layer1[0]` to produce the initial feature tensor. We found that initializing from these real features gives better results than the random tensor initialization used in Mm-Bd.
3. **Optimization:** We treat the extracted activations as trainable variables and freeze all deeper layers. We then optimize the activations to maximize the model's confidence for a target class. This searches for a latent trigger in the feature space; after evaluating all layers of PreActResNet18, we identified `model.layer1[0]` as the optimal point for exposing target-class over-sensitivity in backdoored models.
4. **Maximal Confidence Scoring:** After the optimization, we compute a maximal confidence score for each class, penalized by activation magnitudes of other classes to ensure specificity.

We use hyperparameters that were stable across datasets: 100 optimization steps, batch size 30, an initial learning rate of 0.1 decayed to 0.001, and 10 warmup steps.

4.1.2 Statistical Outlier Detection

Our detection hypothesis is that clean models show balanced confidence across classes, while backdoored models exhibit a single target class with unusually high maximal confidence.

Rather than using a fixed threshold, we apply a statistical anomaly test:

- **Distribution Fitting:** We collect the maximal confidence scores across classes, exclude the maximum, and fit the remaining scores to an exponential distribution. We tested all SciPy candidate distributions and found the exponential distribution produced the best separation between clean and backdoored models.
- **Hypothesis Testing:** We treat the maximum score as a candidate outlier and compute its p-value under the fitted distribution.
- **Decision Rule:** If the p-value is below 0.08, the model is classified as backdoored; otherwise, it is classified as clean.

This statistical approach provides robustness across datasets and attack types while maintaining the constraint of using only limited, fixed parameters.

Table 4: Top 10 Leaderboard Rankings for the Backdoored Model Detection Challenge.

Rank	Team	Accuracy (%)
1	AUTs	79
2	**No Trust Issues Here (Our Team)**	**78**
3	Persistence	74
4	AI Guardians of Trust	72
5	AIUoK	70
6	Pileh	67
7	My Team	66
8	Unknown	66
9	red_serotonin	65
10	DevNull	65

4.2 Experimental Results

The evaluation was conducted on a private test dataset where models were trained on various image classification datasets (e.g., CIFAR10, MNIST) using different backdoor attack types. The evaluation environment utilized a single Nvidia RTX 4090 GPU with a strict 1-minute computation limit per data sample. The primary metric for success was

Accuracy, defined as the ratio of correct predictions (identifying backdoored vs. clean models) to the total number of data samples.

Table 4 summarizes the results. Our proposed method demonstrated high efficacy, achieving the 2nd rank overall. We achieved an accuracy of 78%, finishing only a single percentage point behind the winning team. These results show that the method detects several backdoor types without prior attack information.

5 Conclusion

This report addressed three critical challenges in machine learning: compositional image retrieval, zero-shot anomaly detection, and backdoored model detection. In the compositional retrieval task, we developed a system that successfully combined visual and textual inputs to retrieve relevant images, achieving first place with a top accuracy of 95.38%. For zero-shot anomaly detection, we implemented a method that detected and localized anomalies in unseen data, securing second place with an accuracy of 73.14%. In the backdoored model detection task, our approach identified backdoor attacks in deep neural networks, achieving a 78% accuracy, placing second in the competition. These results demonstrate the effectiveness of our methodologies in addressing each problem's specific challenges.

The solutions developed in this report have significant implications for real-world applications, including medical imaging, industrial quality control, and security. By enabling models to process multi-modal queries, detect unseen anomalies, and identify backdoored networks, our work contributes to building more reliable, robust, and secure AI systems. Future work could focus on further improving model generalization, handling more complex real-world scenarios, and refining methods to detect novel attack strategies. Ultimately, these challenges highlight the need for continuous innovation to ensure that machine learning systems can operate effectively and securely across diverse environments.

References

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL `https://doi.org/10.5281/zenodo.5143773`. If you use this software, please cite it as below.

Jieneng Chen, Qihang Yu, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Vitamin: Designing scalable vision models in the vision-language era. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12954–12966, 2024.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection, 2025. URL `https://arxiv.org/abs/2310.18961`.

Xurui Li, Ziming Huang, Feng Xue, and Yu Zhou. Musc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. In *The Twelfth International Conference on Learning Representations*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Hang Wang, Zhen Xiang, David J Miller, and George Kesidis. Mm-bd: Post-training detection of backdoor attacks with arbitrary backdoor pattern types using a maximum margin statistic. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 1994–2012. IEEE, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.