electronics

MDPI

Article

AnomalyNLP: Noisy-Label Prompt Learning for Few-Shot Industrial Anomaly Detection

Li Hua and Jin Qian *

College of Information Engineering, Taizhou University, Taizhou 225300, China; huali@tzu.edu.cn
* Correspondence: qianjin@tzu.edu.cn

Abstract

Few-Shot Industrial Anomaly Detection (FSIAD) is an essential yet challenging problem in practical scenarios such as industrial quality inspection. Its objective is to identify previously unseen anomalous regions using only a limited number of normal support images from the same category. Recently, large pre-trained vision-language models (VLMs), such as CLIP, have exhibited remarkable few-shot image-text representation abilities across a range of visual tasks, including anomaly detection. Despite their promise, real-world industrial anomaly datasets often contain noisy labels, which can degrade prompt learning and detection performance. In this paper, we propose **AnomalyNLP**, a new Noisy-Label Prompt Learning approach designed to tackle the challenge of few-shot anomaly detection. This framework offers a simple and efficient approach that leverages the expressive representations and precise alignment capabilities of VLMs for industrial anomaly detection. First, we design a Noisy-Label Prompt Learning (NLPL) strategy. This strategy utilizes feature learning principles to suppress the influence of noisy samples via Mean Absolute Error (MAE) loss, thereby improving the signal-to-noise ratio and enhancing overall model robustness. Furthermore, we introduce a prompt-driven optimal transport feature purification method to accurately partition datasets into clean and noisy subsets. For both image-level and pixel-level anomaly detection, AnomalyNLP achieves state-of-the-art performance across various few-shot settings on the MVTecAD and VisA public datasets. Qualitative and quantitative results on two datasets demonstrate that our method achieves the largest average AUC improvement over baseline methods across 1-, 2-, and 4-shot settings, with gains of up to 10.60%, 10.11%, and 9.55% in practical anomaly detection scenarios.

Keywords: industrial anomaly detection; few-shot learning; optimal transport; vision-language model

Academic Editor: George A. Papakostas

Received: 26 August 2025
Revised: 9 October 2025
Accepted: 12 October 2025
Published: 13 October 2025

Citation: Hua, L.; Qian, J. AnomalyNLP: Noisy-Label Prompt Learning for Few-Shot Industrial Anomaly Detection. *Electronics* **2025**, *14*, 4016. https://doi.org/10.3390/electronics14204016

1. Introduction

Industrial anomaly detection (IAD) [1], a critical task in computer vision, plays a vital role in modern manufacturing [2]. It focuses on identifying and localizing defects within images of industrial products. Acquiring and labeling these defects is typically costly and labor-intensive, as real-world anomalies vary significantly across applications in texture, color, shape, and size. Consequently, standard IAD models are often trained exclusively on defect-free ("normal") samples to identify anomalies [3]. However, this ideal scenario is rarely attainable in practice. For instance, when introducing inspection into a new industry chain, the relevant training data may be scarce or entirely absent, rendering conventional model training infeasible.

To address these challenges of high annotation costs and novel anomaly detection, Few-Shot Industrial Anomaly Detection (FSIAD) has emerged. FSIAD aims to detect anomalies in a query image by leveraging a small number of normal support images from the same object class [4]. Although few-shot learning typically assumes that the limited labeled samples are accurate and reliable, this assumption often breaks down in industrial settings. Due to vague defect boundaries, diverse material appearances, and human annotation inconsistency, even a few support samples can contain mislabeled or ambiguous regions. Under such scarcity, a single noisy label may dominate the feature space, amplifying its negative effect on model generalization. Therefore, it becomes crucial to design few-shot learning frameworks that are not only data-efficient but also noise-robust. This limitation reveals a clear research gap: existing few-shot prompt learning methods largely ignore the impact of label noise and assume clean support sets, which may not hold in realistic industrial scenarios. Therefore, it becomes crucial to design few-shot learning frameworks that are not only data-efficient but also robust to noisy labels, motivating the development of our proposed AnomalyNLP framework.

Recently, large-scale pre-trained vision-language models (VLMs), particularly CLIP, have shown remarkable few- and zero-shot recognition performance across a wide range of visual tasks, including anomaly detection. Trained on vast datasets of image–text pairs, CLIP provides strong generalization for downstream applications [5–8]. Specifically, WinCLIP [9] pioneered the use of CLIP for FSIAD, enhancing anomaly detection performance in few-shot settings. Nevertheless, a fundamental limitation exists: VLMs such as CLIP are mainly trained to capture the semantic alignment of foreground objects rather than the subtle distinctions between normal and abnormal regions within images. As a result, their generalization ability for recognizing visual anomalies is limited, which can lead to reduced performance in FSIAD tasks. Other methods, such as SAA+ [10] and AnoCLIP [11], also rely on prompt engineering to boost performance, making it a common practice in prompt-guided anomaly detection. However, prompt engineering necessitates manual effort and meticulous design, which conflicts with the automation demands of industrial deployment. Furthermore, real-world applications frequently contend with noisy labels in annotated datasets, demanding robust learning strategies. Prior research [9,12,13] indicates that prompt tuning exhibits greater resilience to label noise compared to other fine-tuning approaches like adapter tuning. Despite this advantage, prompt tuning remains susceptible to overfitting on noisy labels when optimized using cross-entropy loss, as the model may memorize corrupted supervision signals instead of learning robust feature alignments. To mitigate this issue, we propose a noise-robust prompt-learning framework that leverages Mean Absolute Error (MAE) loss and optimal transport-based feature purification to suppress the influence of corrupted labels.

In this paper, we introduce **AnomalyNLP**, an optimal transport-based method as shown in Figure 1, designed to enhance the robustness of prompt learning in vision-language foundation models for few-shot industrial anomaly detection across different domains. Optimal transport (OT) [14–16] provides a mathematical framework for measuring the minimal cost required to align one probability distribution with another. Its ability to establish fine-grained correspondences between feature distributions makes it particularly suitable for separating clean and noisy samples in our prompt-learning framework. AnomalyNLP leverages text description features as prototypes for the transportation matrix, facilitating robust prompt learning by partitioning the dataset into clean and noisy subsets. We further introduce noisy-label prompt learning to harmonize the strengths of both MAE and CE loss under noisy conditions. To this end, our model improves the expressive representation and alignment capabilities of vision-language models, offering a simple and efficient solution for robust prompt learning in the presence of noisy labels.

Specifically, by integrating optimal transport into the prompt-learning process, AnomalyNLP adaptively purifies feature distributions and mitigates the propagation of label noise through distribution-level alignment. In summary, our contributions are threefold:

- We propose a Noisy-Label Prompt-Learning framework grounded in feature learning theory. By decoupling latent representations into task-relevant and task-irrelevant components, our method effectively suppresses noise propagation while enhancing robustness in vision-language prompt optimization.
- We introduce a prompt-driven optimal transport purification mechanism that leverages the expressive representation power and semantic alignment capabilities of vision-language foundation models, enabling robust prompt learning against label corruption.
- We perform comprehensive experiments on MVTecAD and VisA, two publicly available industrial anomaly detection datasets, and the results show that AnomalyNLP attains state-of-the-art performance in few-shot anomaly detection.

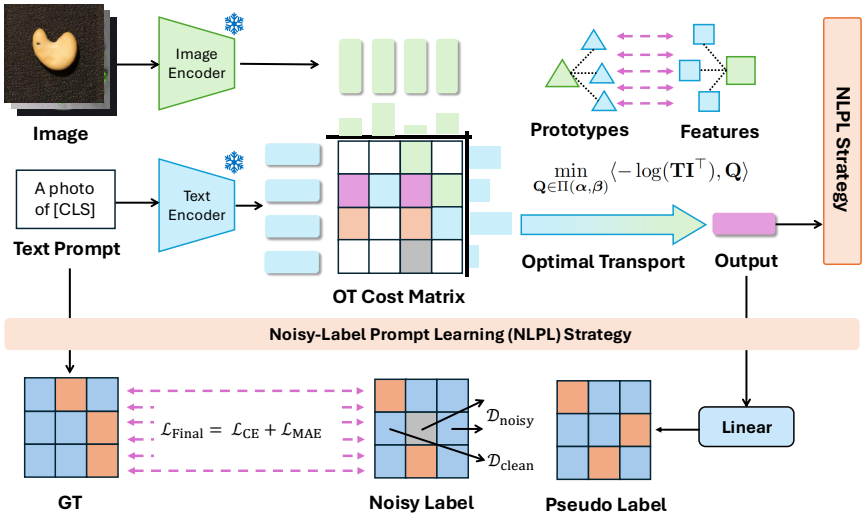

Figure 1. Overview of the proposed AnomalyNLP framework. It consists of three components: (1) Vision-Language Feature Representation-extracts aligned image–text embeddings using frozen CLIP encoders with learnable prompts; (2) Prompt-Driven OT Purification—computes a semantic transport plan to separate clean and noisy samples; and (3) Noisy-Label Prompt Learning (NLPL) which optimizes prompts via a hybrid CE–MAE loss for noise-robust anomaly detection.

2. Related Work

2.1. Vision-Language Foundation Models

In recent years, vision-language models (VLMs) [17–19] have achieved remarkable advancements, significantly enhancing the integration of visual and textual understanding across various tasks. These foundation models demonstrate powerful capabilities across diverse applications, including zero-shot classification [20,21], and medical image analysis [5,22]. A pioneering example is CLIP [23], which is trained on massive image–text pairs through contrastive learning, which achieves exceptional zero-shot classification performance across domains. Concurrently, large language models (LLMs), dominant in natural language processing (NLP), are increasingly being adapted to address vision-related tasks This trend is exemplified by models like LLaVA [24,25], which integrates a vision encoder with an LLM into an end-to-end pretrained large multimodal model (LMM) for comprehensive vision-language understanding and efficient offline inference. Similarly, BLIP-2 [26] employs a Q-Former to bridge a Vision Transformer's [27] visual features to the Flan-T5 LLM [28], while PandaGPT [13] connects ImageBind [29] to Vicuna [30] via a linear layer for

multimodal input processing. A core objective shared by these approaches is the automatic learning of improved prompts through contrastive learning to enhance CLIP-based image classification guidance. These foundation models demonstrate powerful capabilities across diverse applications, including zero-shot classification [20,21], and medical image analysis [5,22]. A pioneering example is CLIP [23], trained on large-scale image–text pairs via contrastive learning, which achieves exceptional zero-shot classification performance across domains. Concurrently, large language models (LLMs), dominant in NLP, are increasingly adapted for visual tasks. This trend is exemplified by models like LLaVA [24,25], which integrates a vision encoder with an LLM into an end-to-end pretrained large multimodal model (LMM) for comprehensive vision-language understanding and efficient offline inference. Similarly, BLIP-2 [26] employs a Q-Former to bridge a Vision Transformer's [27] visual features to the Flan-T5 LLM [28], while PandaGPT [13] connects ImageBind [29] to Vicuna [30] via a linear layer for multimodal input processing. A core objective shared by these approaches is the automatic learning of improved prompts through contrastive learning to enhance CLIP-based image classification guidance. However, despite their strong performance in general vision tasks, these models often encounter difficulties in fine-grained anomaly detection where subtle visual deviations must be identified, and they rely heavily on large-scale pretraining, which may not transfer effectively to domain-specific industrial settings. Despite these advances, recent studies indicate that even state-of-the-art LMMs struggle to capture crucial compositional aspects of visual reasoning, such as object attributes and relationships.

2.2. Industrial Anomaly Detection

Industrial Anomaly Detection (IAD) is a critical task in computer vision, focused on detecting samples that deviate significantly from the normal ones. Existing IAD approaches generally fall into three main paradigms: feature embedding paradigm, knowledge distillation paradigm, and reconstruction-based paradigm. Feature embedding-based methods aim to model the feature representations of normal samples. For example, PatchSVDD [31] seeks to define a hypersphere that closely encloses normal samples. Cflow-AD [32] and PyramidFlow [33] employ normalizing flows to map normal samples onto a Gaussian distribution. Meanwhile, methods such as PatchCore [1] and CFA [34] construct a memory bank of patch embeddings derived from normal samples, identifying anomalies by calculating the distance between a test sample embedding and its nearest normal embedding within the memory bank. The knowledge distillation paradigm [35–38] enables a student network to learn only the representations of normal samples from a teacher network, and anomalies are detected by measuring the discrepancies between the teacher and student outputs. In contrast, reconstruction-based methods focus on reconstructing anomalous samples into their corresponding normal forms and identify anomalies based on the reconstruction error. Methods such as RIAD [39], SCADN [40], InTra [41] and AnoDDPM [42] adopt various reconstruction architectures, including autoencoders, Generative Adversarial Networks (GANs), Transformers, and diffusion models. Although these methods achieve good performance on standard benchmarks, they generally require abundant normal samples per class, making them less suitable for few-shot scenarios and novel object categories. Moreover, most methods do not explicitly handle label noise, limiting their applicability in real-world industrial settings. These approaches generally adopt a "one-class-one-model" learning paradigm, which depends on a large number of normal samples for each object category to accurately capture its distribution. This reliance limits their practicality for novel object classes and reduces adaptability in dynamic production environments. In contrast, our method facilitates in-context learning for novel object categories, enabling inference with only a few normal samples.

2.3. Few-Shot Industrial Anomaly Detection

Recent research has increasingly explored approaches that rely on a limited number of normal samples to perform Industrial Anomaly Detection (IAD), a task known as Few-Shot Industrial Anomaly Detection (FSIAD). The concept of FSIAD was first introduced by RegAD [43], which trained an image registration network to align test images with normal samples, then calculated similarity scores between corresponding patches. PatchCore [1] , when adapted to few-shot settings, constructs a memory bank from only a few normal samples, leading to a noticeable drop in performance. WinCLIP [9], on the other hand, utilizes CLIP [23] to measure the similarity between images and textual descriptions that encode normal and anomalous semantics, identifying anomalies based on their relative similarity scores. APRIL-GAN [44] utilizes learnable linear layers to align patch-level image features with textual features, effectively overcoming the inefficiencies associated with WinCLIP's multiple window design and achieving improved performance. AnoamlyGPT [12] proposes a decoder that matches visual and textual features to produce pixel-level anomaly localization, using both the original image and the decoder output as inputs to an LVLM for anomaly detection, thereby eliminating the need for manual threshold tuning. PromptAD [45] generates numerous negative samples by combining normal prompts with abnormal suffixes, facilitating more effective text prompt learning, and further incorporates the notion of explicit abnormal edges to enhance detection precision. Despite these innovations, current FSIAD approaches often rely on manually designed prompts, limited pretraining, or heuristic-based alignment, which may underperform when facing noisy labels or extreme few-shot scenarios. Our work addresses these limitations by combining noise-robust prompt learning with optimal transport-based feature purification, providing a principled approach for improving generalization under label scarcity and noise.

3. Preliminaries

Notation. We adhere to the following conventions: vectors in bold lowercase (e.g., $\mathbf{v}$), matrices in bold uppercase (e.g., $\mathbf{A}$), and scalars in regular type. The ℓ_2-norm and Frobenius norm are denoted as $\|\mathbf{v}\|_2$ and $\|\mathbf{A}\|_F$, respectively. Integer sequences use $[n] = \{1, \ldots, n\}$, and element sequences are denoted as $\mathbf{v}_{[n]} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$.

Prompt Learning. We formalize learnable prompt tuning in vision-language pre-trained (VLP) models for classification. Prompt tuning has emerged as a powerful parameter-efficient fine-tuning (PEFT) alternative to full model fine-tuning. Its core premise is the introduction of a small set of learnable parameters (the prompts) while keeping the massive pre-trained model weights frozen. This approach is particularly crucial for few-shot industrial anomaly detection, where training data is scarce, and overfitting on large models is a significant risk.

In this preliminary formulation, we adopt a multi-class classification framework, where each image belongs to one of C normal categories. The objective here is *not* to detect anomalies directly, but rather to identify and mitigate the influence of noisy or mislabeled samples among the normal classes. This step aims to purify the feature space and improve the reliability of representations that will later be used for downstream few-shot anomaly detection.

Given an image $\mathbf{x}$ with label $y \in [C]$ (C classes total), we leverage the inherent text–image feature alignment that is the cornerstone of VLP models like CLIP: semantically matched prompts and images maximize their feature similarity in a joint embedding space. This alignment is the foundational principle that enables effective few-shot transfer learning in VLMs.

In our implementation, both the CLIP image encoder $g(\cdot)$ and text encoder $h(\cdot)$ are **frozen** during training. Only the prompt tokens are learnable, serving as lightweight

adapters that interact with the frozen encoders to adapt the pre-trained model to the noisy few-shot setting.

Our framework integrates a learnable prompt $\mathbf{p} \in \mathbb{R}^d$ with fixed class-specific prompts $\{\mathbf{p}_c\}_{c=1}^{C} \subset \mathbb{R}^d$. The frozen text encoder $h(\cdot)$ processes the combination $(\mathbf{p}, \mathbf{p}_c)$ to generate class-specific textual features, allowing semantic adaptation without modifying CLIP's internal weights. The text encoder h generates class c's feature

$$\mathbf{h}_c = h(\mathbf{p}, \mathbf{p}_c) \in \mathbb{R}^m,$$

while the image encoder g produces

$$\mathbf{g} = g(\mathbf{x}) \in \mathbb{R}^m.$$

The similarity vector is computed as

$$\boldsymbol{\rho} = [\cos(\mathbf{g}, \mathbf{h}_1), \ldots, \cos(\mathbf{g}, \mathbf{h}_C)]^\top \in \mathbb{R}^C,$$

where Equation (1) captures the alignment between visual and textual embeddings obtained from the frozen CLIP encoders. The learnable prompts softly adapt this alignment, enhancing robustness to label noise while maintaining CLIP's rich semantic structure. And this is optimized via classification loss $\ell(\boldsymbol{\rho}, \mathbf{e}_y)$, where $\mathbf{e}_y$ is the one-hot encoding of y. This formulation provides a simple yet effective differentiable framework for optimizing prompt representations through standard gradient-based methods, directly steering the model's predictions towards the correct class.

By performing noise-aware prompt optimization in this multi-class setting, the model learns to suppress the effect of corrupted supervision signals, which in turn benefits the subsequent anomaly detection stage built upon these purified embeddings.

4. Methodology

We summarize the entire workflow of AnomalyNLP in Figure 1. The framework follows a progressive pipeline: first, image–text feature representations are extracted by the CLIP-based dual encoder; then, an Optimal Transport (OT) purification mechanism partitions the data into clean and noisy subsets; finally, a Noisy-Label Prompt Learning (NLPL) strategy jointly optimizes the learnable prompts using a hybrid CE–MAE loss. This workflow explicitly indicates feature extraction, purification, and adaptive learning, ensuring that each module contributes transparently to the model's robustness against noisy labels.

4.1. Vision-Language Feature Representation

Text Description Encoder. We adopt a dual-prompt encoding scheme comprising a learnable context vector $\mathbf{p} \in \mathbb{R}^d$ and fixed class-specific prompts $\{\mathbf{p}_c\}_{c=1}^{C}$. This design is motivated by the need to balance adaptability with stability. The learnable context vector $\mathbf{p}$ allows the model to adapt to the specific nuances of the industrial anomaly detection task, while the fixed class tokens preserve the rich semantic knowledge already encoded in the pre-trained VLM, preventing catastrophic forgetting.

In our framework, the text encoder $h(\cdot)$ is inherited from a pre-trained CLIP model and remains frozen during training. Only the learnable prompt tokens $\mathbf{p}$ are updated, serving as lightweight adapters that guide the frozen encoder to produce class-specific representations under noisy supervision.

The text encoder h generates class representations through a symmetric transformation:

$$\mathbf{h}_c = \sigma(\mathbf{W}\mathbf{p} + \mathbf{W}\mathbf{p}_c) - \sigma(-\mathbf{W}\mathbf{p} + \mathbf{W}\mathbf{p}_c), \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{m \times d}$ denotes the projection matrix. The symmetric design is not arbitrary; it ensures gradient stability during optimization, mitigates internal covariate shift, and empirically has been shown to enhance representation capacity and model robustness, which is critical when dealing with noisy data.

To provide theoretical insight into the feature learning dynamics, we conceptually decompose $\mathbf{W}$ along its row space:

$$\mathbf{W} = \left[\boldsymbol{\mu}, \boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_L\right]^\top. \tag{2}$$

Here, $\boldsymbol{\mu}$ captures task-relevant semantic information (e.g., the concept of a "scratch" or "dent"), while $\{\boldsymbol{\xi}_\ell\}_{\ell=1}^{L}$ correspond to nuisance components such as background texture, lighting variations, or sensor noise. Although this decomposition is not used in the forward computation, it provides a conceptual lens to understand why our noise-robust prompt learning—via MAE loss and optimal transport-based feature purification—effectively suppresses nuisance features, allowing the model to focus on task-relevant representations.

Image Feature Encoder. The image encoder g produces latent representations $\mathbf{g}_i = g(\mathbf{x}_i) \in \mathbb{R}^m$ that are aligned with the text feature space. We model this alignment through a structured generative decomposition:

$$\mathbf{g}_i = [y_i, x_{i,1}, \cdots, x_{i,L}]^\top, \quad x_{i,l} \sim \mathcal{N}(0, \sigma_p^2)\ \forall l \in [L], \tag{3}$$

where task-relevant and task-irrelevant components are explicitly separated. This explicit separation is crucial for our theoretical analysis, as it allows us to precisely track how the prompt learning process amplifies or suppresses different signal and noise components throughout training, especially under the influence of incorrect labels. Similarity computation and probability estimation follow

$$\boldsymbol{\rho} = [\langle \mathbf{g}_i, \mathbf{h}_1 \rangle, \ldots, \langle \mathbf{g}_i, \mathbf{h}_C \rangle]^\top, \tag{4}$$

$$\mathbf{s}_i = \text{softmax}(\boldsymbol{\rho}), \tag{5}$$

with corresponding loss functions

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} y_{i,c} \log s_{i,c}, \tag{6}$$

$$\mathcal{L}_{\text{MAE}} = \|\mathbf{y}_i - \mathbf{s}_i\|_1. \tag{7}$$

The Cross-Entropy (CE) loss provides strong, sharp gradients for clean samples, facilitating fast and accurate learning. However, its sensitivity to outliers makes it unsuitable for noisy labels. The MAE loss, being more symmetric and bounded, offers inherent robustness against label noise but can lead to slower convergence and less sharp decision boundaries. Our framework harmonizes these strengths, using CE where labels are reliable and MAE where they are suspect.

4.2. Prompt-Driven Optimal Transport Purification

Optimal Transport Framework. Optimal transport (OT) provides a powerful geometric framework for comparing and aligning probability distributions while preserving their intrinsic mass structure. It is particularly suited for our task because it operates on the global geometry of the data distribution, unlike many other methods that make local,

point-wise decisions. This makes it inherently more robust to local perturbations like label noise. OT solves the Monge–Kantorovich problem:

$$\min_{\mathbf{Q}\in\Pi(\boldsymbol{\alpha},\boldsymbol{\beta})} \quad \langle \mathbf{C}, \mathbf{Q} \rangle, \tag{8}$$

where $\Pi(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \{\mathbf{Q} \in \mathbb{R}_+^{n\times m} : \mathbf{Q}\mathbf{1}_m = \boldsymbol{\alpha}, \mathbf{Q}^\top \mathbf{1}_n = \boldsymbol{\beta}\}$ is the transport polytope of coupling matrices between the two distributions. The entropic-regularized variant enables efficient computation via the fast and differentiable Sinkhorn algorithm:

$$\min_{\mathbf{Q}\in\Pi(\boldsymbol{\alpha},\boldsymbol{\beta})} \langle \mathbf{C}, \mathbf{Q} \rangle - \epsilon H(\mathbf{Q}), \quad H(\mathbf{Q}) = \sum_{ij} Q_{ij}(\log Q_{ij} - 1). \tag{9}$$

This regularization transforms a linear program into a strictly convex problem, ensuring a unique solution and allowing the use of iterative scaling algorithms. The resulting probabilistic coupling $\mathbf{Q}^\star$ provides a soft, global assignment that is far more informative and reliable than hard, greedy pseudo-labeling based solely on prediction confidence, which is highly susceptible to confirmation bias in noisy settings.

OT-based Feature Purification. We leverage the powerful semantic alignment inherent in vision-language models to construct a cost matrix $\mathbf{C}$ that is semantically grounded. Traditional pseudo-labeling methods rely solely on the model's often overconfident and erroneous predictions, creating a feedback loop that amplifies noise. Our OT-based approach instead incorporates the global feature distribution geometry, making it robust to local perturbations and outliers. Given image features $\mathbf{I} = [\mathbf{g}_1, \ldots, \mathbf{g}_N]^\top \in \mathbb{R}^{N\times d}$ and text prototypes $\mathbf{T} = [\mathbf{h}_1, \ldots, \mathbf{h}_C]^\top \in \mathbb{R}^{C\times d}$, we formulate the cost as the negative log-similarity, transforming high similarity into low transport cost:

$$\min_{\mathbf{Q}\in\Pi(\boldsymbol{\alpha},\boldsymbol{\beta})} \langle -\log(\mathbf{T}\mathbf{I}^\top), \mathbf{Q} \rangle, \quad \boldsymbol{\alpha} = \tfrac{1}{C}\mathbf{1}_C,\ \boldsymbol{\beta} = \tfrac{1}{N}\mathbf{1}_N. \tag{10}$$

The logarithmic transformation enhances numerical stability and sharply penalizes matches between highly dissimilar features. The optimal transport plan $\mathbf{Q}^\star$, which minimizes the total cost of moving mass from the text prototype distribution to the image feature distribution, yields purified and globally consistent pseudo-labels:

$$\tilde{y}_i = \arg\max_c \ \mathbf{Q}^\star_{ci}. \tag{11}$$

This label is not based on a single model prediction but on the consensus of how the entire set of image features is optimally aligned with the entire set of text prototypes under a global distributional constraint. This process effectively filters out outliers and corrects many local labeling errors.

4.3. Noisy-Label Prompt Learning

Noisy-Label Modeling. Label noise is an unavoidable reality in real-world industrial datasets. It arises from numerous sources: subjective annotation guidelines (e.g., different thresholds for what constitutes a "minor scratch"), human fatigue and error, and imperfections in automated labeling systems. To design a robust algorithm, we must first formalize this corruption. We adopt a Rademacher noise model, which captures the symmetric label flipping commonly encountered in practice, where an annotation is equally likely to be mistakenly marked as positive or negative. Let $\tilde{y}$ denote the observed noisy label and $y \in \{-1, +1\}$ the latent ground truth label. The corruption process follows

$$\tilde{y} = \begin{cases} y & \text{with probability } 1-p \\ -y & \text{with probability } p \end{cases}, \quad p \leq \tfrac{1}{2}.$$

The dataset naturally partitions into clean and corrupted subsets based on this generative assumption:

$$\mathcal{S}_+ = \{i \mid \tilde{y}_i = y_i\}, \quad \mathcal{S}_- = \{i \mid \tilde{y}_i = -y_i\}.$$

This simple yet powerful probabilistic model provides a rigorous foundation for developing robust learning strategies that explicitly account for the uncertainty present in the training labels, moving beyond the naive assumption of fully clean data.

Feature Representation Dynamics. This section aims to conceptually illustrate how prompt weights evolve during training and to highlight the separation between task-relevant and task-irrelevant components, providing insight into the effect of noisy labels. Building on recent advances in theoretical machine learning and feature learning theory [46,47], we decompose the evolution of the prompt weights throughout the training process into its constituent components. This decomposition is analytical and provides profound insights into how the model learns under the influence of both clean and noisy signals. By projecting the prompt's trajectory onto an orthogonal basis derived from the data, we can precisely quantify what is being learned:

$$\mathbf{p}^{(t)} = \underbrace{\alpha^{(t)}\mathbf{p}^{(0)}}_{\text{initialization residual}} + \underbrace{\beta^{(t)}\|\boldsymbol{\mu}\|_2^{-2}\boldsymbol{\mu}}_{\text{task-relevant}} + \underbrace{\sum_{\ell=1}^{L} \phi_\ell^{(t)}\|\boldsymbol{\xi}_\ell\|_2^{-2}\boldsymbol{\xi}_\ell}_{\text{task-irrelevant}}, \tag{12}$$

where

- $\alpha^{(t)}$: Projection coefficient onto the initialization subspace, capturing the optimization inertia or "memory" of the starting point. Its decay rate is an indicator of optimization efficiency.
- $\beta^{(t)}$: Alignment strength with the true semantic feature $\boldsymbol{\mu}$. The growth of this coefficient is the primary signal of meaningful learning progress and generalization.
- $\phi_\ell^{(t)}$: Susceptibility to nuisance feature $\boldsymbol{\xi}_\ell$. The premature or excessive growth of these coefficients is a telltale sign of overfitting and noise memorization, a common failure mode in late-stage training on noisy datasets.

The normalization $\|\cdot\|_2^{-2}$ renders these coefficients interpretable as approximate inner products, e.g., $\beta^{(t)} \approx \langle \mathbf{p}^{(t)}, \boldsymbol{\mu} \rangle$, providing a clear geometric interpretation. Although this decomposition is not used in forward computation, it provides theoretical intuition for why our noise-robust prompt-learning framework (e.g., MAE loss and optimal transport-based purification) is effective: emphasizing task-relevant features and suppressing nuisance components helps mitigate the influence of noisy labels and improve downstream few-shot anomaly detection. This decomposition not only facilitates a rigorous analysis of feature learning trajectories but also informs the design of our loss function, as we can now explicitly aim to maximize the growth of $\beta^{(t)}$ while suppressing the growth of $\phi_\ell^{(t)}$.

Dual Loss Function Optimization. The final step is to integrate the purified pseudo-labels from OT with the original noisy labels in a principled learning framework. Using the high-confidence pseudo-labels $\hat{y}_i$ generated by the global optimal transport process, we partition the dataset into clean and noisy subsets based on the consensus between the OT predictions and the original labels:

$$\mathcal{D}_{\text{clean}} = \{i \mid \hat{y}_i = \tilde{y}_i\}, \tag{13}$$

$$\mathcal{D}_{\text{noisy}} = \{j \mid \hat{y}_j \neq \tilde{y}_j\}. \tag{14}$$

This consensus-based partitioning is a powerful filtering mechanism. A match suggests the original label is likely correct and reinforced by global feature structure. A mismatch flags the sample as highly suspect; the original label is likely wrong, and the OT label, based on global geometry, is probably more reliable. Distinct loss functions are then applied, tailored to the inferred reliability of each subset: Cross-Entropy (CE) for $\mathcal{D}_{\text{clean}}$ to exploit the high-confidence samples and achieve fast, accurate learning, and Mean Absolute Error (MAE) for $\mathcal{D}_{\text{noisy}}$ to ensure robustness against the potential label errors that reside in this subset. The composite final loss is

$$\mathcal{L}_{\text{Final}} = \sum_{i \in \mathcal{D}_{\text{clean}}} \mathcal{L}_{\text{CE}}(\mathbf{s}_i, \mathbf{y}_i) + \sum_{j \in \mathcal{D}_{\text{noisy}}} \mathcal{L}_{\text{MAE}}(\mathbf{s}_j, \mathbf{y}_j) \tag{15}$$

$$= \sum_{i \in \mathcal{D}_{\text{clean}}} -\mathbf{y}_i^\top \log \mathbf{s}_i + \sum_{j \in \mathcal{D}_{\text{noisy}}} \|\mathbf{y}_j - \mathbf{s}_j\|_1, \tag{16}$$

where $\mathbf{y}_k$ is the one-hot encoded label and $\mathbf{s}_k$ the predicted probability distribution for the k-th sample. This hybrid, sample-adaptive approach dynamically modulates the loss landscape based on the estimated reliability of each training sample. It allows the model to learn aggressively from clean data while learning cautiously and robustly from noisy data, thereby achieving superior generalization performance in the presence of widespread label corruption.

4.4. Experiment Setup

4.4.1. Datasets

MVTecAD is extensively utilized as a benchmark in anomaly detection, addressing both logical and structural defects. The dataset consists of 3644 images, with a division of 1772 for training, 304 for validation, and 1568 for testing. It encompasses five object categories, including breakfast box and juice bottle, and features image resolutions from 700 × 700 to 1024 × 1024. The VisA dataset constitutes a collection of 9621 normal and 1200 anomalous color images, spanning 12 objects across three distinct domains: complex structures, single instances, and multiple instances. Anomalies manifested in the dataset constitute a variety of defects, which include surface imperfections such as scratches, dents, color spots, and cracks, in addition to logical anomalies like component misplacement or absence.

4.4.2. Evaluation Metrics

Consistent with prior IAD approaches, we use the Area Under the Receiver Operating Characteristic (AUC) as the evaluation metric, employing image-level AUC to measure anomaly detection performance and pixel-level AUC for anomaly localization. Notably, our method enables anomaly identification without relying on manually defined thresholds.

4.4.3. Implementation Details

We adopt the publicly available CLIP model (https://github.com/mlfoundations/open_clip) (`VIT-L/14@336px`) as our backbone, with all CLIP parameters frozen. The length of learnable word embeddings E is set to 12. The learnable token embeddings are integrated into the first 9 layers of the text encoder to refine the textual feature space, with λ set to 4. All experiments are implemented in PyTorch 2.0.0 and executed on a single NVIDIA RTX 4090 24GB GPU.

5. Results

5.1. Image-Level Comparison Results

The image-level comparison results of AnomalyNLP and existing methods are presented in Table 1. Here, SPADE [48], PaDiM [49], and PatchCore [1] represent adaptations of traditional full-shot methods to few-shot settings, and their image-level IAD performance

remains limited. Methods such as WinCLIP [9] and AnomalyGPT [12], which incorporate CLIP [23], substantially enhance performance under few-shot scenarios. In comparison, AnomalyNLP [45] demonstrates significant improvements across three evaluation settings on both benchmarks. Specifically, under the 1-, 2-, and 4-shot settings of MVTecAD, AnomalyNLP significantly outperforms WinCLIP and AnomalyGPT. Moreover, our approach achieves this performance using fewer prompts than both WinCLIP and AnomalyGPT.

Table 1. Performance comparisons of our model on the MVTecAD and VisA datasets. **Bold** indicates the best performance.

Setup	Method	Public	MVTecAD		VisA		AVG
			Image-AUC	Pixel-AUC	Image-AUC	Pixel-AUC	
1-shot	SPADE	arXiv'2020	81.0 ± 2.0	91.2 ± 0.4	79.5 ± 4.0	95.6 ± 0.4	86.8
	PaDiM	ICPR'2020	76.6 ± 3.1	89.3 ± 0.9	62.8 ± 5.4	89.9 ± 0.8	79.7
	PatchCore	CVPR'2022	83.4 ± 3.0	92.0 ± 1.0	79.9 ± 2.9	95.4 ± 0.6	87.7
	WinCLIP	CVPR'2023	93.1 ± 2.0	95.2 ± 0.5	83.8 ± 4.0	96.4 ± 0.4	92.1
	APRIL-GAN	arXiv'2023	92.0 ± 1.0	95.1 ± 0.3	91.2 ± 2.0	96.0 ± 0.7	93.6
	AnomalyGPT	AAAI'2024	94.1 ± 1.1	95.3 ± 0.1	87.4 ± 0.8	96.2 ± 0.1	93.3
	PromptAD	CVPR'2024	94.6 ± 0.3	95.9 ± 1.1	86.9 ± 0.8	96.7 ± 2.0	93.5
	KAG-prompt	AAAI'2025	95.8 ± 0.2	96.2 ± 1.1	91.6 ± 0.4	97.0 ± 3.0	95.2
	AnomalyNLP	——	**96.3 ± 0.6**	**96.9 ± 1.2**	**92.8 ± 1.3**	**97.8 ± 2.1**	**96.0**
2-shot	SPADE	arXiv'2020	82.9 ± 2.6	92.0 ± 0.3	80.7 ± 5.0	96.2 ± 0.4	88.0
	PaDiM	ICPR'2020	78.9 ± 3.1	91.3 ± 0.7	67.4 ± 5.1	92.0 ± 0.7	82.4
	PatchCore	CVPR'2022	86.3 ± 3.3	93.3 ± 0.6	81.6 ± 4.0	96.1 ± 0.5	89.3
	WinCLIP	CVPR'2023	94.4 ± 1.3	96.0 ± 0.3	84.6 ± 2.4	96.8 ± 0.3	93.0
	APRIL-GAN	arXiv'2023	92.4 ± 2.3	95.0 ± 1.5	92.2 ± 0.1	96.2 ± 1.6	94.0
	AnomalyGPT	AAAI'2024	95.5 ± 0.8	95.6 ± 0.2	88.6 ± 0.7	96.4 ± 0.1	94.0
	PromptAD	CVPR'2024	95.7 ± 1.0	96.2 ± 0.3	88.3 ± 1.3	97.1 ± 2.0	94.3
	KAG-prompt	AAAI'2025	96.6 ± 1.3	96.5 ± 0.5	92.7 ± 2.0	97.4 ± 2.0	95.8
	AnomalyNLP	——	**97.9 ± 0.4**	**97.4 ± 1.2**	**94.0 ± 1.7**	**98.2 ± 0.6**	**96.9**
4-shot	SPADE	arXiv'2020	84.8 ± 2.5	92.7 ± 0.3	81.7 ± 3.4	96.6 ± 0.3	89.0
	PaDiM	ICPR'2020	80.4 ± 2.5	92.6 ± 0.7	72.8 ± 2.9	93.2 ± 0.5	91.3
	PatchCore	CVPR'2022	88.8 ± 2.6	94.3 ± 0.5	85.3 ± 2.1	96.8 ± 0.3	91.3
	WinCLIP	CVPR'2023	95.2 ± 1.3	96.2 ± 0.3	87.3 ± 1.8	97.2 ± 0.2	94.0
	APRIL-GAN	arXiv'2023	92.8 ± 2.0	95.9 ± 0.5	92.2 ± 1.7	96.2 ± 3.0	94.3
	AnomalyGPT	AAAI'2024	96.3 ± 0.3	96.2 ± 0.1	90.6 ± 0.7	96.7 ± 0.1	95.0
	PromptAD	CVPR'2024	96.6 ± 0.4	96.5 ± 2.1	89.1 ± 0.5	97.4 ± 1.1	94.9
	KAG-prompt	AAAI'2025	97.1 ± 0.4	96.7 ± 1.3	93.3 ± 0.9	97.7 ± 2.7	96.2
	AnomalyNLP	——	**98.1 ± 0.1**	**98.6 ± 2.0**	**94.4 ± 1.5**	**98.8 ± 0.3**	**97.5**

5.2. Pixel-Level Comparison Results

The pixel-level comparison results are presented in Table 1. It is observed that CLIP-based methods like WinCLIP [9] and other approaches exhibit similar performance in pixel-level anomaly detection, indicating that the benefits of incorporating CLIP [23] are less pronounced at the pixel level compared to image-level detection. AnomalyNLP attains the highest performance on MVTecAD and VisA in the 1-shot and 2-shot settings, surpassing WinCLIP [9] by 1.79%/1.45% and 1.46%/1.45%, respectively. In the 4-shot setting, AnomalyNLP ranks first on VisA but comes second on MVTecAD, narrowly trailing APRIL-GAN [44] by 1.87%. Compared with PatchCore [1] and WinCLIP [9], AnomalyNLP demonstrates superior anomaly localization capabilities for both objects and textures in the 1-shot scenario. Additionally, our model is capable of accurately detecting very small anomalous regions.

5.3. Compared with Many-Shot Methods

Table 2 compares the performance of AnomalyNLP under few-shot settings with other methods evaluated under many-shot or full-shot settings. The results demonstrate that

AnomalyNLP achieves superior image-level performance compared to several methods operating under many-shot conditions, while its pixel-level results remain competitive. This strongly validates the capability of AnomalyNLP in few-shot anomaly detection. Furthermore, AnomalyNLP outperforms early full-shot AD methods such as MKD [36] and P-SVDD [31]. However, a noticeable gap persists between AnomalyNLP and state-of-the-art full-shot methods like PatchCore [1] and SimpleNet [50].

Table 2. Comparison with exiting many-shot methods in AUROC (image and pixel level) on MVTecAD. Results below our 1-shot are marked in red, and those below our 4-shot are marked in blue.

Model	Public	Setting	Image-AUC	Pixel-AUC
AnomalyNLP	-	1-shot	96.3	96.9
AnomalyNLP	-	4-shot	98.1	98.6
DiffNet [51]	WACV'2021	16-shot	87.3	-
TDG [52]	ICCV'2021	10-shot	78.0	-
RegAD [43]	ECCV2022	8-shot	91.2	96.7
FastRecon [53]	ICCV'2023	8-shot	95.2	97.3
MKD [36]	CVPR'2021	full-shot	87.8	90.7
P-SVDD [31]	ACCV'2021	full-shot	95.2	96.0
PatchCore [1]	CVPR'2022	full-shot	99.1	98.1
SimpleNet [50]	CVPR'2023	full-shot	99.6	98.1

5.4. Ablation Study

To evaluate the effectiveness of each component of our method, we conduct ablation studies on the MVTecAD and VisA two datasets. The experimental results are shown in Table 3. To validate the effectiveness of prompt-driven optimal transport purification, we designed two sets of experiments: one without using OT for data purification and the other using OT for data purification. The full experimental design is as follows: (A) simple standard FSIAD baseline WinCLIP; (B) Use CE loss; (C) Use CE loss and MAE loss for improved optimization; (D) Use proposed OT-based purification; (E) Use all proposed strategies combination. The average results show that (B) outperforms (A), and (C) outperforms (A) which validates the effectiveness of our Noisy-Label Prompt Learning strategy via the improved dynamic loss functions. Moreover, the average results show that (D) outperforms (A), and (E) outperforms (C), further validating the effectiveness of Prompt-Driven Optimal Transport in the data purification process. Of all methods, our AnomalyNLP achieves the best performance, with significant improvements over other baselines, further validating the effectiveness of each component for FSIAD.

Table 3. Image-level/pixel-level results (AUROC) of ablation study under 1-shot setting. The best and second-best results are in **bold** and underlined, respectively.

Model	NLPL		OT	MVTecAD		VisA		AVG
	$\mathcal{L}_{CE}$	$\mathcal{L}_{MAE}$		Image-AUC	Pixel-AUC	Image-AUC	Pixel-AUC	
A	✗	✗	✗	93.1	95.2	83.8	96.4	92.1
B	✓	✗	✗	93.4	95.7	84.3	96.9	92.6
C	✓	✓	✗	95.3	96.0	87.6	97.1	94.0
D	✗	✗	✓	95.1	96.3	91.3	97.4	95.0
E	✓	✓	✓	**96.3**	**96.9**	**92.8**	**97.8**	96.0

5.5. Visualization Results

Figure 2 presents qualitative visualization results of our proposed AnomalyNLP on the MVTec AD and VisA datasets. The model produces anomaly localization maps that closely align with the corresponding ground truth annotations. Notably, AnomalyNLP not only accurately localizes prominent anomalous regions but also captures subtle and

easily overlooked anomalies, underscoring its sensitivity to fine-grained defect patterns. These results compellingly demonstrate the superior anomaly localization capability of our approach. The consistent performance under few-shot conditions further attests to the robustness and generalization capability of AnomalyNLP, highlighting its substantial promise for practical deployment in real-world industrial scenarios.

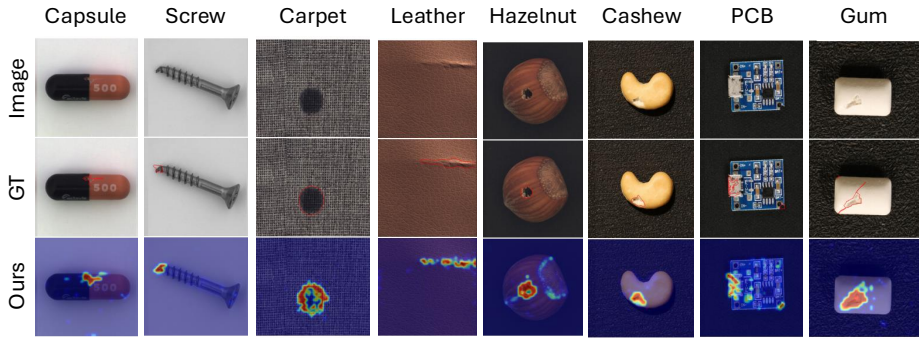

Figure 2. Visualization of representative results for pixel-level anomaly detection of our proposed method on two datasets. The first, second, and last row indicate the input image, ground truth, and the detection result from our AnomalyNLP.

6. Conclusions

This paper addresses the critical challenge of noisy labels in Few-Shot Industrial Anomaly Detection (FSIAD). We propose **AnomalyNLP** details refer Appendix A, a novel framework that synergizes feature learning theory with prompt optimization to enhance vision-language models' robustness against label corruption. Our core innovations include the following: (1) a Noisy-Label Prompt Learning strategy that decouples representations into task-relevant/irrelevant components while suppressing noise propagation via MAE loss; (2) and a prompt-driven optimal transport purification mechanism that leverages VLMs' alignment capabilities for accurate noise identification. Extensive experiments on MVTecAD and VisA datasets demonstrate that AnomalyNLP achieves state-of-the-art performance across diverse few-shot settings for both image-level and pixel-level anomaly detection. The framework provides an efficient, automated solution requiring minimal manual intervention which significantly advances industrial deployment readiness. While our experiments are conducted on standard benchmarks such as MVTecAD and VisA, which feature clean test sets, we acknowledge that this setup does not fully capture the noise encountered in real-world industrial applications. In practice, noisy inputs may appear at inference time due to automated inspection logs, ambiguous defect boundaries, or human labeling errors. Nevertheless, demonstrating robustness to noisy labels during training is a crucial first step, as industrial datasets often contain imperfect annotations. Future work should investigate end-to-end robustness under realistic noisy conditions, potentially by evaluating on real-world inspection logs or introducing controlled noise to benchmark test sets.

Author Contributions: Conceptualization, L.H. and J.Q.; methodology, L.H.; software, L.H.; validation, L.H.; formal analysis, L.H.; investigation, L.H.; resources, L.H.; data curation, L.H.; writing—original draft preparation, L.H. and J.Q.; writing—review and editing, L.H. and J.Q.; visualization, L.H.; supervision, J.Q.; project administration, J.Q.; funding acquisition, J.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Jiangsu Province's Blue and Young Project. Grant Number 601201800102

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The datasets used in this work are all publicly available, MVTecAD: https://www.mvtec.com/company/research/datasets/mvtec-ad/downloads (accessed on 25 August 2025), VisA: https://github.com/amazon-science/spot-diff (accessed on 25 August 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Algorithm A1 AnomalyNLP: Optimal Transport-based Noisy-Label Prompt Learning (Detailed Version)

Require: Pre-trained CLIP encoders: image encoder g, text encoder h; support set $\mathcal{S} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$; class prompts $\{\mathbf{p}_c\}_{c=1}^{C}$; learnable context prompt $\mathbf{p}$; hyperparameters ϵ (OT regularization), learning rate η, loss weight λ_{MAE}, confidence τ, epochs E.

Ensure: Optimized prompt $\mathbf{p}^\star$ and anomaly scoring function $A(x)$.

Initialization: Freeze g, h; randomly initialize $\mathbf{p}$.

for $t = 1$ to E **do**

Step 1: Feature Extraction. For each $(x_i, \tilde{y}_i) \in \mathcal{S}$, extract $\mathbf{v}_i = g(x_i)$. For each class c, compute textual feature $\mathbf{h}_c = h(\mathbf{p}, \mathbf{p}_c)$.

Step 2: Optimal Transport-based Alignment. Compute similarity $S_{c,i} = \cos(\mathbf{h}_c, \mathbf{v}_i)$ and cost $C_{OT} = -\log(S + \delta)$. Solve for transport matrix

$$Q^\star = \arg\min_{Q} \langle Q, C_{OT} \rangle + \epsilon H(Q),$$

where $H(Q)$ is the entropy regularization term. $Q^\star$ captures soft correspondences between image and text prototypes.

Step 3: Noise-aware Soft Partition. For each sample i, compute confidence $\omega_i = \max_c Q^\star_{c,i}$ and pseudo-label $\hat{y}_i = \arg\max_c Q^\star_{c,i}$. Define

$$\mathcal{S}_{clean} = \{i \mid \omega_i > \tau \wedge \hat{y}_i = \tilde{y}_i\}, \quad \mathcal{S}_{noisy} = \{i \mid \omega_i \leq \tau \text{ or } \hat{y}_i \neq \tilde{y}_i\}.$$

Clean samples guide precise alignment, while noisy ones contribute regularized gradients.

Step 4: Prompt-based Prediction. For each x_i, compute logits $\rho_i = [\cos(\mathbf{v}_i, \mathbf{h}_1), \ldots, \cos(\mathbf{v}_i, \mathbf{h}_C)]$ and probability $p_i = \text{Softmax}(\rho_i)$.

Step 5: Dual-branch Optimization. Apply different losses for each subset:

$$\mathcal{L}_{CE} = -\sum_{i \in \mathcal{S}_{clean}} \log p_{i,\tilde{y}_i}, \quad \mathcal{L}_{MAE} = \sum_{i \in \mathcal{S}_{noisy}} \|p_i - \mathbf{e}_{\hat{y}_i}\|_1.$$

Combine and update prompts

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_{MAE} \mathcal{L}_{MAE}, \quad \mathbf{p} \leftarrow \mathbf{p} - \eta \nabla_{\mathbf{p}} \mathcal{L}.$$

Step 6: Iterative Refinement. Every K epochs, recompute $Q^\star$ and re-estimate clean/noisy partitions, enabling dynamic purification as $\mathbf{p}$ evolves.

end for

Step 7: Inference. Given query x_q, compute $\mathbf{v}_q = g(x_q)$ and class logits $\rho_{q,c} = \cos(\mathbf{v}_q, h(\mathbf{p}^\star, \mathbf{p}_c))$.

Anomaly score:

$$A(x_q) = 1 - \max_c \text{Softmax}(\boldsymbol{\rho}_q)_c.$$

Higher scores indicate stronger anomaly likelihood.

return $\mathbf{p}^\star$ and $A(x)$.

References

1. Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; Gehler, P. Towards total recall in industrial anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14318–14328.
2. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv. (CSUR)* **2009**, *41*, 1–58. [CrossRef]
3. Ahmed, M.; Mahmood, A.N.; Hu, J. A survey of network anomaly detection techniques. *J. Netw. Comput. Appl.* **2016**, *60*, 19–31. [CrossRef]
4. Wang, Z.; Zhou, Y.; Wang, R.; Lin, T.Y.; Shah, A.; Lim, S.N. Few-shot fast-adaptive anomaly detection. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 4957–4970.
5. Zhao, Z.; Liu, Y.; Wu, H.; Wang, M.; Li, Y.; Wang, S.; Teng, L.; Liu, D.; Cui, Z.; Wang, Q.; et al. Clip in medical imaging: A comprehensive survey. *arXiv* **2023**, arXiv:2312.07353. [CrossRef]
6. Zhang, R.; Guo, Z.; Zhang, W.; Li, K.; Miao, X.; Cui, B.; Qiao, Y.; Gao, P.; Li, H. Pointclip: Point cloud understanding by clip. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022 2022; pp. 8552–8562.
7. Hafner, M.; Katsantoni, M.; Köster, T.; Marks, J.; Mukherjee, J.; Staiger, D.; Ule, J.; Zavolan, M. CLIP and complementary methods. *Nat. Rev. Methods Prim.* **2021**, *1*, 20. [CrossRef]
8. Pan, B.; Li, Q.; Tang, X.; Huang, W.; Fang, Z.; Liu, F.; Wang, J.; Yu, J.; Shi, Y. NLPrompt: Noise-Label Prompt Learning for Vision-Language Models. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 19963–19973.
9. Jeong, J.; Zou, Y.; Kim, T.; Zhang, D.; Ravichandran, A.; Dabeer, O. Winclip: Zero-/few-shot anomaly classification and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19606–19616.
10. Cao, Y.; Xu, X.; Sun, C.; Cheng, Y.; Du, Z.; Gao, L.; Shen, W. Segment any anomaly without training via hybrid prompt regularization. *arXiv* **2023**, arXiv:2305.10724. [CrossRef]
11. Deng, H.; Zhang, Z.; Bao, J.; Li, X. Bootstrap fine-grained vision-language alignment for unified zero-shot anomaly localization. *arXiv* **2023**, arXiv:2308.15939.
12. Gu, Z.; Zhu, B.; Zhu, G.; Chen, Y.; Tang, M.; Wang, J. Anomalygpt: Detecting industrial anomalies using large vision-language models. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 1932–1940.
13. Su, Y.; Lan, T.; Li, H.; Xu, J.; Wang, Y.; Cai, D. Pandagpt: One model to instruction-follow them all. *arXiv* **2023**, arXiv:2305.16355. [CrossRef]
14. Villani, C. *Optimal Transport: Old and New*; Springer: Berlin/Heidelberg, Germany, 2008; Volume 338.
15. Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013; Volume 26.
16. Montesuma, E.F.; Mboula, F.M.N.; Souloumiac, A. Recent advances in optimal transport for machine learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *47*, 1161–1180. [CrossRef] [PubMed]
17. Bao, H.; Wang, W.; Dong, L.; Liu, Q.; Mohammed, O.K.; Aggarwal, K.; Som, S.; Piao, S.; Wei, F. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 32897–32912.
18. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021; PMLR: 2021; pp. 4904–4916.
19. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; PMLR: 2022; pp. 12888–12900.
20. Abdelhamed, A.; Afifi, M.; Go, A. What do you see? Enhancing zero-shot image classification with multimodal large language models. *arXiv* **2024**, arXiv:2405.15668. [CrossRef]
21. Naeem, M.F.; Khan, M.G.Z.A.; Xian, Y.; Afzal, M.Z.; Stricker, D.; Van Gool, L.; Tombari, F. I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 15169–15179.
22. Lai, Z.; Li, Z.; Oliveira, L.C.; Chauhan, J.; Dugger, B.N.; Chuah, C.N. Clipath: Fine-tune clip with visual feature fusion for pathology image analysis towards minimizing data collection efforts. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 2374–2380.
23. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021; PMLR: 2021; pp. 8748–8763.

24. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 26296–26306.
25. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 34892–34916.
26. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; PMLR: 2023; pp. 19730–19742.
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
28. Chung, H.W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.* **2024**, *25*, 1–53.
29. Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K.V.; Joulin, A.; Misra, I. Imagebind: One embedding space to bind them all. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 15180–15190.
30. Chiang, W.L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J.E.; et al. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* Chatgpt Quality. 2023. Available online: https://vicuna.lmsys.org (accessed on 14 April 2023).
31. Yi, J.; Yoon, S. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In Proceedings of the Asian Conference on Computer Vision, Kyoto, Japan, 30 November–4 December 2020.
32. Gudovskiy, D.; Ishizaka, S.; Kozuka, K. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 98–107.
33. Lei, J.; Hu, X.; Wang, Y.; Liu, D. Pyramidflow: High-resolution defect contrastive localization using pyramid normalizing flow. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 14143–14152.
34. Lee, S.; Lee, S.; Song, B.C. Cfa: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access* **2022**, *10*, 78446–78454. [CrossRef]
35. Gu, Z.; Liu, L.; Chen, X.; Yi, R.; Zhang, J.; Wang, Y.; Wang, C.; Shu, A.; Jiang, G.; Ma, L. Remembering normality: Memory-guided knowledge distillation for unsupervised anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 16401–16409.
36. Salehi, M.; Sadjadi, N.; Baselizadeh, S.; Rohban, M.H.; Rabiee, H.R. Multiresolution knowledge distillation for anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 14902–14912.
37. Wang, G.; Han, S.; Ding, E.; Huang, D. Student-teacher feature pyramid matching for anomaly detection. *arXiv* **2021**, arXiv:2103.04257. [CrossRef]
38. Zhang, X.; Li, S.; Li, X.; Huang, P.; Shan, J.; Chen, T. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 3914–3923.
39. Zavrtanik, V.; Kristan, M.; Skočaj, D. Reconstruction by inpainting for visual anomaly detection. *Pattern Recognit.* **2021**, *112*, 107706. [CrossRef]
40. Yan, X.; Zhang, H.; Xu, X.; Hu, X.; Heng, P.A. Learning semantic context from normal samples for unsupervised anomaly detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 3110–3118.
41. Pirnay, J.; Chai, K. Inpainting transformer for anomaly detection. In Proceedings of the 21st International Conference on Image Analysis and Processing, Lecce, Italy, 23–27 May 2022; Springer: Cham, Switzerland, 2022; pp. 394–406.
42. Wyatt, J.; Leach, A.; Schmon, S.M.; Willcocks, C.G. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 21–24 June 2022; pp. 650–656.
43. Huang, C.; Guan, H.; Jiang, A.; Zhang, Y.; Spratling, M.; Wang, Y.F. Registration based few-shot anomaly detection. In Proceedings of the 17th European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 303–319.
44. Chen, X.; Han, Y.; Zhang, J. April-gan: A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv* **2023**, arXiv:2305.17382.
45. Li, X.; Zhang, Z.; Tan, X.; Chen, C.; Qu, Y.; Xie, Y.; Ma, L. Promptad: Learning prompts with only normal samples for few-shot anomaly detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 16838–16848.

46. Li, Y.; Wang, S.; Tian, Q.; Ding, X. Feature representation for statistical-learning-based object detection: A review. *Pattern Recognit.* **2015**, *48*, 3542–3559. [CrossRef]
47. Pérez, D.; Alonso, S.; Morán, A.; Prada, M.A.; Fuertes, J.J.; Domínguez, M. Comparison of network intrusion detection performance using feature representation. In Proceedings of the 20th International Conference on Engineering Applications of Neural Networks, Crete, Greece, 24–26 May 2019; Springer: Cham, Switzerland, 2019; pp. 463–475.
48. Cohen, N.; Hoshen, Y. Sub-image anomaly detection with deep pyramid correspondences. *arXiv* **2020**, arXiv:2005.02357.
49. Defard, T.; Setkov, A.; Loesch, A.; Audigier, R. Padim: A patch distribution modeling framework for anomaly detection and localization. In Proceedings of the International Conference on Pattern Recognition, Virtual, 10–15 January 2021; Springer: Cham, Switzerland, 2021; pp. 475–489.
50. Liu, Z.; Zhou, Y.; Xu, Y.; Wang, Z. Simplenet: A simple network for image anomaly detection and localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, USA, 18–22 June 2023; pp. 20402–20411.
51. Rudolph, M.; Wandt, B.; Rosenhahn, B. Same same but differnet: Semi-supervised defect detection with normalizing flows. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 1907–1916.
52. Sheynin, S.; Benaim, S.; Wolf, L. A hierarchical transformation-discriminating generative model for few shot anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 8495–8504.
53. Fang, Z.; Wang, X.; Li, H.; Liu, J.; Hu, Q.; Xiao, J. Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 17481–17490.