Journal of Machine Learning Research 27 (2026) 1-38 Submitted 11/25; Revised 9/26; Published 10/26

Provable Hierarchical Martingale-based Testing for Visual Anomaly Detection

Jie Wu JIEWU0907@GMAIL.COM
School of Computer Science
Wuhan University
Wuhan, Hubei, China

Xinsong Ma MAXINSONG1018@GMAIL.COM
Department of Computer Science
City University of Hong Kong
Kowloon, Hong Kong SAR

Xin Zou ZOUXIN2021@GMAIL.COM
School of Computer Science
Wuhan University
Wuhan, Hubei, China

Xiuwen Gong* GONGXIUWEN@GMAIL.COM
School of Computer Science
Wuhan University
Wuhan, Hubei, China

Kaizhu Huang KAIZHU.HUANG@DUKEKUNSHAN.EDU.CN
Digital Innovation Research Center
Duke Kunshan University
Suzhou, Jiangsu, China

Weiwei Liu* LIUWEIWEI863@GMAIL.COM
Institute of Big Data
Fudan University
Shanghai, China

Editor: Samy Bengio

Abstract

Existing visual anomaly detection approaches primarily focus on designing robust score functions but often neglect investigating the decision-making process, which plays a key role in reliable and safety-critical applications. Currently, the decision-making process suffers from two key challenges: distribution shift and theoretical guarantee. To address these issues, this paper proposes a novel approach called Hierarchical Martingale-based Testing framework (HiMT). Specifically, our approach leverages martingale theory to mitigate distribution shift issue and employs a hierarchical multiple testing procedure to rigorously control decision errors at both image level and pixel level. Extensive experimental results verify the superior decision-making capabilities of our proposed HiMT framework at both image level and pixel level.

Keywords: visual anomaly detection, hypothesis testing, martingale

*. Corresponding Author.

©2026 Jie Wu, Xinsong Ma, Xin Zou, Xiuwen Gong, Kaizhu Huang and Weiwei Liu.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/25-2884.html.

1. Introduction

Visual anomaly detection (VAD) aims to identify anomalous images (image-level binary classification) and locate the regions that do not conform to the normal patterns (pixel-level binary classification). It is a crucial aspect of computer vision with numerous applications, such as product quality control (Bergmann et al., 2019), healthcare monitor system (Li et al., 2018) and video surveillance (Lv et al., 2023). Due to the scarcity of anomalous samples, unsupervised paradigms are widely adopted.

In the current literature (Guo et al., 2023; Zhang et al., 2024; Zavrtanik et al., 2021; Roth et al., 2022; Zhang et al., 2023; Tien et al., 2023), for a test image set $\{I_1, I_2, \dots\}$, the VAD task is formulated as the following two subtasks:

$$\begin{aligned} \text{image level:} \quad & \mathbf{y_i} = \Phi(s(I_1)_*, s(I_2)_*, \dots), \\ \text{pixel level:} \quad & \mathbf{y_p} = \Phi(s(I_1), s(I_2), \dots), \end{aligned} \tag{1}$$

where both $\mathbf{y_i}$ and $\mathbf{y_p}$ are sets of binary classification results, $s(\cdot)$ is a score function which outputs continuous confidence scores, $s(I_i)$ is the pixel-level score map, $s(I_i)_*$ is an image-level score, and $\Phi(\cdot)$ denotes the decision-making process as

$$\Phi(s) = \begin{cases} \text{normal} & \text{if } s < \tau \\ \text{anomalous} & \text{if } s \geq \tau \end{cases}, \tag{2}$$

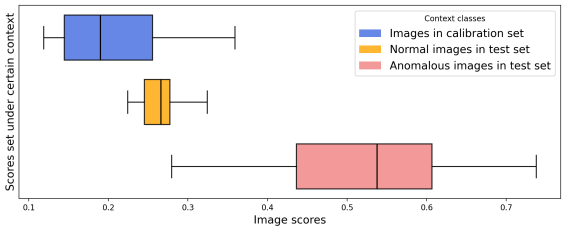
where τ is an empirical threshold. Currently, various powerful score functions are designed to tackle this challenging task (Yao et al., 2023; Li et al., 2024b; Schwartz et al., 2024; Li et al., 2024a; Gu et al., 2024; Li et al., 2023). Unfortunately, little research has been done to investigate the decision-making process, despite its vital role in VAD task. This paper identifies two essential aspects that the decision-making process should take into account.

A. Distribution Shift Issue. Extensive experiments are conducted to demonstrate the widespread presence of distribution shift issue in VAD task. We present some representative results in Figure 1.

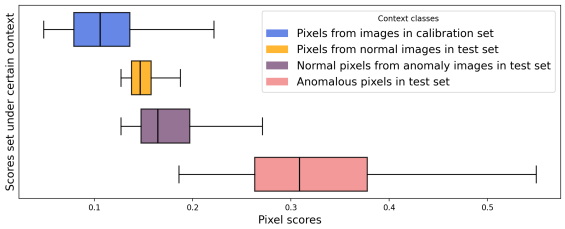
At image level, as shown in Figure 1 (a), the score distribution for images in the calibration set (blue box plot) significantly deviates from that of normal images in the test set (yellow box plot). At pixel level, illustrated in Figure 1 (b), a similar discrepancy is observed between the calibration pixels and the pixels from normal test images. Moreover, even within the test set, there is a notable discrepancy between the scores of pixels from normal images (yellow box plot) and those from anomalous images (purple box plot). This discrepancy arises due to the presence of anomalous regions in the input images. Similarly, this distribution shift phenomenon is clearly observed across different data sets and score functions, as shown in Figure 1 (c)-(l). Consequently, decisions based on the calibration set may underperform on the test set due to the distribution shift issue.

B. Theoretical Guarantee. An effective decision-making process should yield results that both have strong empirical performance and rigorous theoretical guarantees, especially critical in some sensitive applications such as medical diagnosis (Fernando et al., 2021).

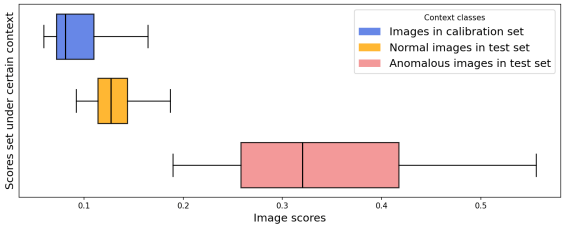
An effective decision-making process should address the two aspects simultaneously, yet existing approaches fall short. The empirical decision-making (EDM) process assumes identical score distributions for normal images in calibration and test sets, selecting a



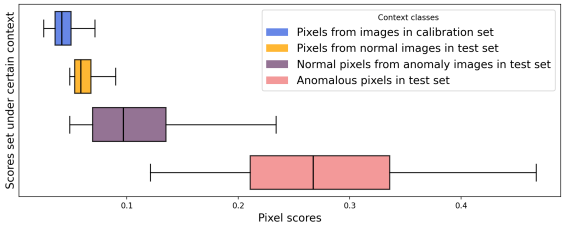
(a) RRD, MVTec, Image-level.



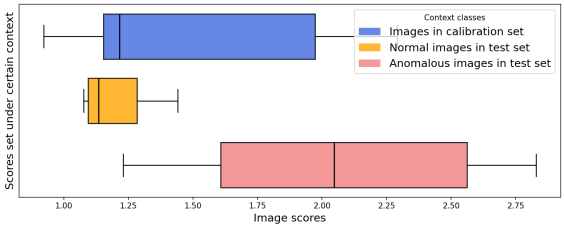
(b) RRD, MVTec, Pixel-level.



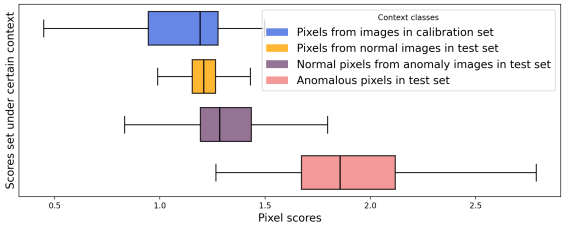
(c) RRD, BTAD, Image-level.



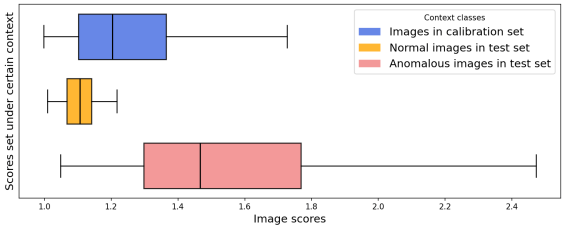
(d) RRD, BTAD, Pixel-level.



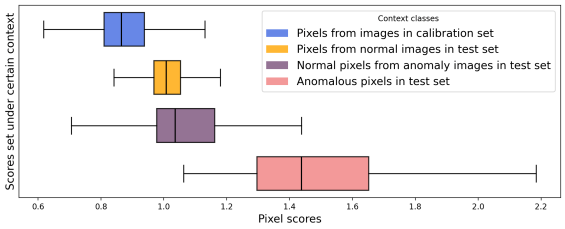
(e) PatchCore, MVTec, Image-level.



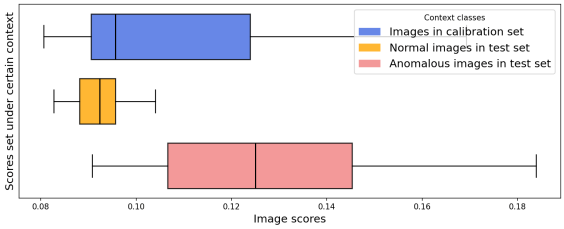
(f) PatchCore, MVTec, Pixel-level.



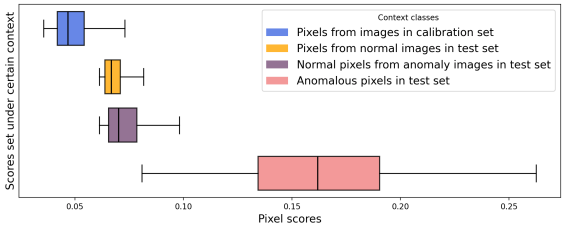
(g) PatchCore, VisA, Image-level.



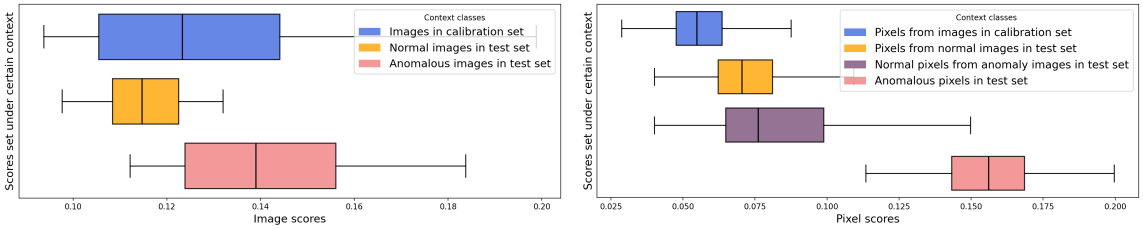
(h) PatchCore, VisA, Pixel-level.



(i) CFA, VisA, Image-level.



(j) CFA, VisA, Pixel-level.



(k) CFA, BTAD, Image-level.

(l) CFA, BTAD, Pixel-level.

Figure 1: Visualization of distribution shift issue. 'RRD, MVTec, Image-level' denotes the image-level score distribution produced by RRD on MVTec data set. The blue box plot in (a) shows the score distribution of images in calibration set. The training and calibration sets are split at a ratio of $4:1$ (Arlot and Celisse, 2010; Li et al., 2022). Experiments are conducted on the MVTec (Bergmann et al., 2019), VisA (Zou et al., 2022), and BTAD (Mishra et al., 2021) data sets, using RRD (Tien et al., 2023), PatchCore (Roth et al., 2022), and CFA (Lee et al., 2022) as score functions.

threshold to correctly classify 95% of the calibration set (Hendrycks et al., 2022; Wang et al., 2022; Yang et al., 2021). However, this approach is prone to distribution shift issue. The adaptive threshold tuning (ATT) method uses Markov chains to model correlations and employs tracking technology to adjust detection thresholds (Ali et al., 2013). Some researchers frame thresholding in anomaly detection as a Markov decision process, proposing an agent-based dynamic thresholding (ADT) framework using a deep Q-network (Yang et al., 2025). The adaptive thresholding heuristic (ATH) method incorporates local data properties and periodicity constraints intuitively (Isaac and Sharma, 2024). Despite their careful design, these methods lack rigorous theoretical guarantees. Recently, researchers introduce a generalized Benjamini-Hochberg (g-BH) procedure to theoretically control errors (Ma et al., 2024; Yu et al., 2023), but both methods depend on a calibration set, making them vulnerable to distribution shifts. The ATCS method developed the adaptive anomaly thresholding with confidence sequences, providing statistical guarantees on false positive and false negative rates (Sun et al., 2024), but it assumes score independence—an assumption violated by spatial correlations in pixels.

This paper proposes a novel decision-making framework, the Hierarchical Martingale-based Testing framework (HiMT), to effectively address the aforementioned issues in VAD task. Specifically, we first model the VAD task as a multiple testing problem (Casella and Berger, 2002), where rejecting a null hypothesis indicates that the corresponding image or pixel is identified as anomalous. We then serialize an image to construct a test supermartingale process (Shafer et al., 2011), using Ville's inequality (Ville, 1939) to provide statistical evidence for rejecting or accepting the null hypothesis. This process relies solely on the test image, thereby avoiding the impact of distribution shifts on decision-making. Furthermore, we integrate the hierarchical structure of visual data with the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), introducing the Hierarchical Benjamini-Hochberg procedure. This procedure is rigorously proven to control the false discovery rate at both image and pixel levels. The proposed HiMT framework is compared with

existing decision-making approaches, and experimental results demonstrate that it achieves state-of-the-art decision-making performance. For instance, compared to the ATCS method with DMAD score function, our proposed HiMT framework achieves 56.1% and 20.2% improvements in terms of F1-score on MVTec data set at image level and pixel level, respectively.

Overall, the proposed HiMT framework achieves promising performance on the VAD task and can be readily applied using existing score functions without requiring complex modifications. The rest of this paper is organized as follows: Section 2 discusses related work on VAD and multiple testing. Section 3 introduces the formulation of the VAD task and related statistical concepts. Section 4 presents the HiMT framework and proves that it simultaneously controls the FDR at both image level and the pixel level under a specified threshold. Section 5 reports extensive comparative experiments and ablation studies on various components. The omitted proofs are provided in Appendix A.

2. Related Work

In this section, we review related work on visual anomaly detection and multiple testing.

Visual Anomaly Detection. VAD is a key problem in computer vision and industrial applications. Current approaches can be broadly divided into four categories. Reconstruction-based methods focus on training a generative model to capture the patterns and features of normal images (Ma et al., 2025b; Wu et al., 2026; Ma et al., 2025c). The reconstruction error is usually considered as the indicator for anomalies (Zavrtanik et al., 2021). Knowledge distillation-based methods are a promising class of solutions for unsupervised anomaly detection (Bergmann et al., 2020; Liu et al., 2025), in which the student model is trained on only anomaly-free examples, and the teacher student (T-S) model is expected to produce divergent features for anomalous samples during inference. Tien et al. (2023) and Park et al. (2025) improve reverse distillation by reducing student generalization to anomalies and introducing skip connections for more robust anomaly detection. Embedding-based methods (Rudolph et al., 2022; Luo et al., 2025) encode the initial anomaly-free images into a compressed feature space, in which the features of the anomalous image deviate from the normal clusters. Many studies use the pre-trained networks on large-scale data sets to construct this feature space (Defard et al., 2021; Roth et al., 2022).

These works have greatly advanced the development of VAD. However, they mainly focus on designing or training the powerful score function but overlook investigating the decision-making based on the proposed score function.

Recently, foundation models have also been explored for VAD. Jeong et al. (2023) propose a CLIP-based framework that uses compositional prompts and window-level multi-scale features for anomaly classification and segmentation. AnomalyVLM uses hybrid prompts to guide GroundingDINO (Liu et al., 2024) and SAM (Kirillov et al., 2023) in generating refined anomaly maps (Cao et al., 2025). VLMDiff (Hicsonmez et al., 2026) integrates an off-the-shelf vision-language model (VLM) as a conditioning signal for a latent diffusion model to perform anomaly detection. Although these VLM-based methods benefit from rich semantic priors and potential interpretability, they mainly aim to improve the quality of anomaly scores and do not explicitly design a decision-making process. AnomalyGPT pioneers threshold-free industrial anomaly detection by adapting VLMs to judge anomalous

images (Gu et al., 2024). However, its engineering-based decision mechanism lacks rigorous statistical guarantees.

Overall, existing VAD methods lack a systematic study of decision-making at both image and pixel levels. Most of them still rely on empirical or adaptive thresholding decisions, which are vulnerable to distribution shifts and lack theoretical guarantees.

Multiple Testing. Multiple testing is a classical statistical framework for simultaneously testing a family of hypotheses and deciding which null hypotheses should be rejected (Shaffer, 1995; Romano et al., 2018). Central goal of multiple testing is to control the decision error over the entire set of hypotheses (Dudoit and Van Der Laan, 2008). Early studies mainly focus on controlling the family-wise error rate (FWER) (Shaffer, 1995). However, FWER control is often conservative and can lead to low detection power when the number of hypotheses is large. To better balance error control and statistical power, Benjamini and Hochberg (1995) introduce the false discovery rate (FDR), which controls the expected proportion of false discoveries among all rejected hypotheses.

Despite its theoretical appeal, applying multiple testing to practical VAD tasks remains challenging. First, valid multiple testing requires p-values that satisfy super-uniformity (Casella and Berger, 2002). Bates et al. (2023) and Yu et al. (2023) construct conformal p-values based on a calibration set. Fischer et al. (2026) construct anytime-valid Monte Carlo p-values for large-scale multiple testing using a valid null simulator. Permutation-based methods construct p-values by comparing the observed test statistic with a null distribution generated from valid permutations under the null hypothesis (Ramdas et al., 2023a; Winkler et al., 2014). The validity of permutation-based methods relies on a correct exchangeability structure (Phipson and Smyth, 2010). In VAD tasks, directly applying these p-value construction methods relies on calibration data, where distribution shift can violate validity and lead to unreliable p-values.

Second, FDR control depends on both the testing procedure and its dependence assumptions (Sarkar, 2002). The classical BH procedure controls FDR under independence or certain positive dependence conditions (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). It is widely used since it provides a powerful and less conservative way to control FDR in multiple testing. Ma et al. (2026, 2025a) further remove the constraints on the dependence between p-values and propose the g-BH procedure to control the FDR. But this relaxation is achieved through a more conservative correction. In VAD tasks, pixel-level data are inherently dependent, making it challenging to achieve theoretical FDR control while retaining statistical power comparable to the BH procedure.

3. Preliminaries

3.1 Visual Anomaly Detection

Let $\mathcal{X}_P$ denote the pixel instance space and $\mathcal{X}_I = \mathcal{X}_P^m$ denote the image instance space, where m is the number of pixels in each image. We consider a test set of n images $\mathcal{S}^{\text{test}} = \{I_i\}_{i=1}^n$, where each image $I_i = (x_{i1}, x_{i2}, \ldots, x_{im}) \in \mathcal{X}_I$ is sampled from an unknown image-level distribution $\mathcal{D}_I$. Here, $x_{ij} \in \mathcal{X}_P$ denotes the j-th pixel of image I_i, and $[n] = \{1, \ldots, n\}$, $[m] = \{1, \ldots, m\}$. $\mathcal{D}_I$ is a distribution over image instances, and pixel-level distribution $\mathcal{D}_{P|I}$ is a conditional distribution over pixels within given image I, induced by the corresponding image-level sample.

Let $Y_i^I \in \{0,1\}$ and $Y_{ij}^P \in \{0,1\}$ denote the image-level and pixel-level ground-truth labels, respectively, where 1 represents anomaly and 0 represents normality. During testing, these labels are unknown, and the goal of VAD is to produce image-level decisions $\hat{y}_i^I$ and pixel-level decisions $\hat{y}_{ij}^P$.

Let $s(\cdot)$ denote a score function. For image I_i, the score function produces a pixel-level score map

$$A_i = s(I_i) = (a_{i1}, a_{i2}, \ldots, a_{im}) \in \mathbb{R}^m,$$

where a_{ij} is the anomaly score of pixel x_{ij}. The corresponding image-level score is denoted by a_i^*, which can be directly produced by the score function or obtained by aggregating the score map A_i. Larger scores indicate stronger evidence of anomaly.

In contrast to existing solutions in Equation (1), we formulate VAD as a hierarchical multiple testing problem. First, the proposed Martingale-based P-value Construction (MPC) algorithm $\Gamma(\cdot,\cdot)$, described in Section 4.1, transforms each queried score into a valid p-value by using the score map of its corresponding image:

$$p_i^I = \Gamma(A_i, a_i^*), \qquad p_{ij}^P = \Gamma(A_i, a_{ij}).$$

Here, p_i^I is the image-level p-value for image I_i, and p_{ij}^P is the pixel-level p-value for pixel x_{ij}.

Then, the proposed Hierarchical Benjamini-Hochberg (HiBH) procedure $\Phi(\cdot)$, described in Section 4.2, takes all image-level and pixel-level p-values as input and outputs the rejected hypotheses under the prescribed levels α_{il} and α_{pl}:

$$\left(\hat{y}_i^I, (\hat{y}_{i1}^P, \ldots, \hat{y}_{im}^P)\right) = \Phi\left(\{p_i^I\}_{i=1}^n, \{p_{ij}^P\}_{i\in[n],j\in[m]}; \alpha_{il}, \alpha_{pl}\right).$$

The final image-level and pixel-level binary decision sets on $\mathcal{S}^{\text{test}}$ are

$$\hat{\mathbf{y}}_I = \{\hat{y}_i^I\}_{i=1}^n, \qquad \hat{\mathbf{y}}_P = \{(\hat{y}_{i1}^P, \ldots, \hat{y}_{im}^P)\}_{i=1}^n. \tag{3}$$

3.2 Test Supermartingale

Let triplet $(\Omega, \mathcal{F}, P)$ be a probability space, where Ω is a set, $\mathcal{F}$ is a σ-algebra on Ω, and P is a probability measure defined on $\mathcal{F}$. Let T denote an ordered set. Let $\{\mathcal{F}_n, n \in T\}$ be an increasing sequence of σ-algebra of $\mathcal{F}$ sets. We introduce the concept of supermartingale which is a widely used tool of modern abstract mathematics.

Definition 1 (Supermartingale; Liptser and Shiryayev, 1989) *Let $\{X_n, n \in T\}$ be a sequence of random variables on Ω. X_n is measurable with respect to $\{\mathcal{F}_n\}$. If X_n satisfies*

i. $\mathbb{E}[|X_n|] < \infty$,

ii. $\mathbb{E}[X_n|\mathcal{F}_m] \leq X_m$ for all $m < n, m, n \in T$,

then the sequence $\{X_n\}$ is a supermartingale with respect to $\{\mathcal{F}_n, n \in T\}$.

According to Definition 1, it follows inductively that $\mathbb{E}[X_t] \leq \mathbb{E}[X_0]$ for any fixed value of t (Coquio, 2006). Consider a sequence of random variables $\{X_n\}$ that initially has the value one and is a nonnegative supermartingale under a probabilistic hypothesis P. We call such a random process a test supermartingale (Shafer et al., 2011; Malek et al., 2017). In statistics, the values of a test supermartingale measure the changing evidence against P. A large value can be interpreted as discrediting P (Ramdas et al., 2023b).

3.3 Multiple Testing

Let $\mathcal{D}_i$ denote the distribution of normal pixels on image I_i, $\mathcal{D}_*$ denote the distribution of normal images. For simplicity, we let triplet $(X, N, \mathcal{D}) \in \{(I, n, \mathcal{D}_*), (p^i, m, \mathcal{D}_i)\}$. We then formulate the VAD task as the following multiple testing problem:

$$\begin{aligned} H_{1;0} : X_1 \sim \mathcal{D}, \quad & H_{1;1} : X_1 \nsim \mathcal{D}; \\ H_{2;0} : X_2 \sim \mathcal{D}, \quad & H_{2;1} : X_2 \nsim \mathcal{D}; \\ & \vdots \\ H_{N;0} : X_N \sim \mathcal{D} \quad & H_{N;1} : X_N \nsim \mathcal{D}, \end{aligned} \tag{4}$$

where $H_{i;0}$ and $H_{i;1}$ are called null hypothesis and alternative hypothesis, respectively. In VAD setting, if $H_{i;0}$ is rejected, we declare that X_i is anomalous.

In statistics, the decision to accept or reject the null hypothesis is determined by the concept of p-value. Its general definition is presented as follows:

Definition 2 (*P*-value; Casella and Berger, 2002) *Given a sample set X, a statistic $p(X)$ is called a p-value corresponding to the null hypothesis H_0, if $p(X)$ satisfies $0 \leq p(X) \leq 1$ and for every $0 \leq \alpha \leq 1$,*

$$\mathbb{P}[p(X) \leq \alpha | H_0] \leq \alpha. \tag{5}$$

Small values of $p(X)$ usually give evidence that H_1 is true. To control the error rate, we introduce the false discovery rate (FDR). In multiple testing, we usually refer to the rejected hypotheses as discoveries.

Definition 3 (FDP & FDR; Benjamini and Hochberg, 1995) *Let V denote the number of rejected true null hypotheses and R denote the number of rejected hypotheses. The FDP is the proportion of errors committed by falsely rejecting null hypotheses and the FDR, false discovery rate, is the expected FDP*

$$\mathrm{FDR} = \mathbb{E}\left[\mathrm{FDP}\right] = \mathbb{E}\left[\frac{V}{\max(R, 1)}\right]. \tag{6}$$

In the VAD task, anomalous images or pixels are considered as positive samples. The true positive (TP) and false positive (FP) are key metrics derived from the confusion matrix (Sammut and Webb, 2017). Precision (Pr) is a standard metric for evaluating decision-making performance. FDR is a prospective error-control criterion at test time. The relationship between the FDP and Pr can be written as

$$\mathrm{FDP} = \frac{V}{R \vee 1} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}} = \frac{1}{\frac{\mathrm{TP}}{\mathrm{FP}} + 1} = 1 - \mathrm{Pr}. \tag{7}$$

Equation (7) illustrates how the FDR control contributes to decision-making in VAD. Multiple testing formulation provides a statistical guarantee for controlling the FDR at a prescribed level, thereby offering a principled decision-making criterion for VAD.

4. Hierarchical Martingale-based Testing Framework

In this section, we propose the HiMT framework, comprising two key components: the MPC algorithm in Section 4.1 addresses the distribution shift issue, while the HiBH procedure in Section 4.2 provides theoretical error control at both image and pixel levels. The overall structure of the proposed HiMT framework is illustrated in Figure 2.

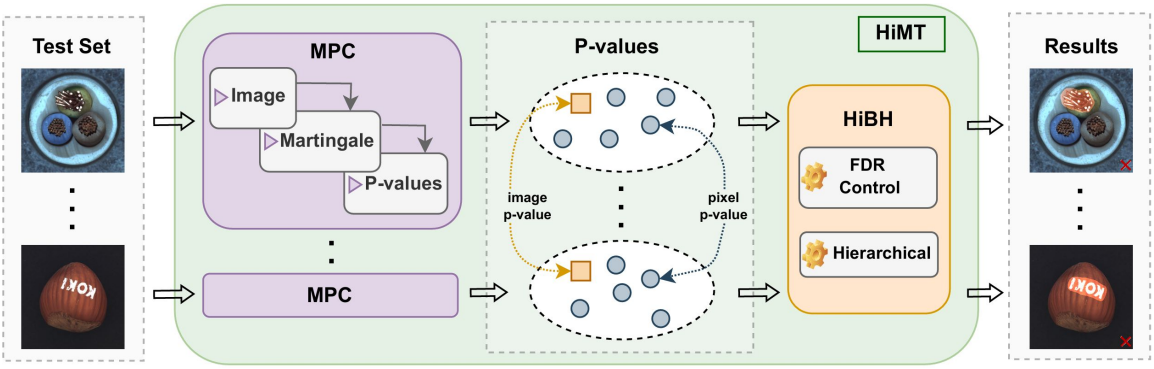


Figure 2: Illustration of the HiMT framework.

4.1 MPC: Martingale-based *P-value* Construction

Given an input image I, let $s(\cdot)$ denote a score function, S denote the image-level or pixel-level score to be queried, and H_0 denote the null hypothesis that S is a score for a normal image or a normal pixel. We formulate the MPC algorithm as the following three steps.

STEP-1: Construct a sequence of variables with a desirable property. We first split score map $s(I)$ into a sequence of disjoint patches $s(Z_1), s(Z_2), \ldots, s(Z_k)$, where k is the number of patches.

Let $Z \in \{Z_1, Z_2, \ldots, Z_k\}$ denote a patch sampled from image I, and $\mathcal{P}_Z$ denote the set of pixels in Z. Let $U \sim \text{Unif}(0,1)$ be an auxiliary random variable independent of all score variables, where $\text{Unif}(0,1)$ is the uniform distribution on $[0,1]$. For any score a, we introduce a randomized comparison function

$$\mathbb{1}_U\{a \preceq S\} = \mathbb{1}\{a < S\} + U\mathbb{1}\{a = S\}, \tag{8}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. When the score distribution is continuous, ties occur with probability zero, and $\mathbb{1}_U\{a \preceq S\} = \mathbb{1}\{a \leq S\}$. Based on Equation (8), we define

$$Y = \frac{\sum_{x \in \mathcal{P}_Z} \mathbb{1}_U\{s(x) \preceq S\}}{|\mathcal{P}_Z|}, \tag{9}$$

where $|\cdot|$ denotes the set size. The statistic Y is the empirical proportion of scores in patch Z that do not exceed S. Its distribution is induced by the score map of the input image, the patch-splitting procedure, and the randomized comparison rule in Equation (8).

Consider a latent partition of $\mathcal{P}_Z$ into the normal pixels $\mathcal{N}_Z = \{x_1, \ldots, x_{r_0}\}$ and the anomalous pixels $\mathcal{A}_Z = \{\tilde{x}_1, \ldots, \tilde{x}_{r_1}\}$, where $\mathcal{P}_Z = \mathcal{N}_Z \cup \mathcal{A}_Z$. Y can be formulated as

$$Y = \frac{\sum_{j=1}^{r_0} \mathbb{1}_U\{s(x_j) \preceq S\} + \sum_{j=1}^{r_1} \mathbb{1}_U\{s(\tilde{x}_j) \preceq S\}}{r_0 + r_1}. \tag{10}$$

In particular, when the input image is normal, all pixels in Z are normal, and Y reduces to the empirical randomized cumulative distribution function (CDF) of normal scores within the sampled patch, evaluated at S.

Let X denote a normal pixel under the null hypothesis H_0, and let $S = s(X)$ be its score. The score function $s(\cdot)$ induces a null score distribution over S. With the randomized comparison rule $\preceq$, we define the randomized CDF:

$$F_s^{\text{rand}}(t) = \mathbb{P}[S \preceq t] = \mathbb{P}(S < t) + U\mathbb{P}(S = t). \tag{11}$$

Based on Equation (11), we derive that Y in Equation (10) satisfies the following property:

Lemma 4 *Given an image I with patches $\{Z_i, 1 \le i \le k\}$, let Z be an arbitrary patch in I. If S is generated under the image-level or pixel-level null hypothesis, then*

$$\mathbb{E}_Z\left[Y \mid S\right] \le F_s^{\text{rand}}(S). \tag{12}$$

In particular, if the corresponding image is normal, equality holds.

The proof of Lemma 4 is presented in Appendix A.1. For all patches in image I, we obtain a sequence of random variables $Y_1, Y_2, \ldots, Y_k$.

STEP-2: Build a test supermartingale. Based on Lemma 4, we construct a test supermartingale as a function of $\{Y_1, Y_2, \ldots, Y_k\}$.

Theorem 5 (Construction of Test Supermartingale) *Given an image I with patches $\{Z_i, 1 \le i \le k\}$, let S be a score for a normal image or a normal pixel, and let $\{Y_i, 1 \le i \le k\}$ be a sequence of variables as defined in Equation (10). Suppose that $\{Y_i, 1 \le i \le k\}$ are jointly independent under the null hypothesis.*[1] *Let f be a positive, concave, and nondecreasing function with integral 1 over $[0, 1]$. For all integers $n \in [0, k]$, if we define Λ_n as*

$$\Lambda_n = \begin{cases} 1 & \text{if } n = 0 \\ \prod_{i=1}^{n} f(Y_i) & \text{if } 1 \le n \le k \end{cases}, \tag{13}$$

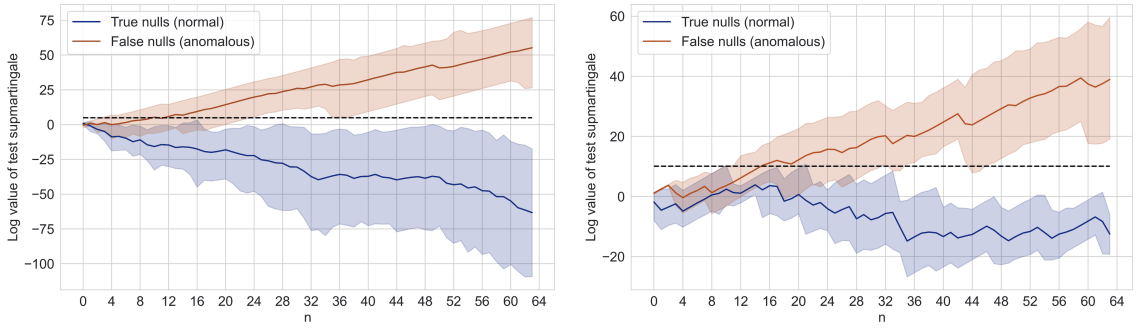
then $\{\Lambda_n, 0 \le n \le k\}$ is a test supermartingale.

The proof of Theorem 5 is presented in Appendix A.2. According to Theorem 5, if the null hypothesis H_0 holds (i.e., S is a score for a normal image or a normal pixel), we can construct a test supermartingale $\{\Lambda_n, 0 \le n \le k\}$ using the information from the input image. According to Ville's inequality (Ville, 1939), for every constant $c \ge 1$, the test supermartingale Λ_n satisfies

$$\mathbb{P}\left[\sup_{0 \le n \le k} \Lambda_n \ge c\right] \le \frac{1}{c}. \tag{14}$$

1. This assumption is empirically supported in the VAD setting. See Appendix A.2 for a detailed justification.

Ville's inequality in Equation (14) indicates that the values of test supermartingale Λ_n provide evidence against the null hypothesis H_0. The visualizations of test supermartingale Λ_n are shown in Figure 3. As depicted in Figure 3, the test supermartingale shows significantly larger values when testing anomalous samples (the red plots), and conversely, remains smaller for normal samples (the blue plots).



(a) MVTec, Bottle category

(b) MVTec, Hazelnut category

Figure 3: Visualization of test supermartingale Λ_n. Each plot depicts the evolution of test supermartingale over time $n \in [0, 64]$, with shaded areas indicating the 95% confidence interval. The blue and red plots indicate the supermartingale process under true and false null hypotheses, respectively. The black dashed line represents our proposed method's decision boundary. All scores are produced by ReContrast.

STEP-3: Calculate p-value based on test supermartingale. Next, we show that the inverse of supremum of a test supermartingale is a valid p-value.

Theorem 6 (Test Supermartingale to *P*-value; Dellacherie and Meyer, 1988) *Let $\{\Lambda_n, n \geq 0\}$ be a test supermartingale, and $\Lambda_t^* = \sup_{i \leq t} \Lambda_i$ be the greatest value up to time t. Then $1/\Lambda_t^*$ is a p-value.*

According to Theorem 6, for a test supermartingale $\{\Lambda_n, 0 \leq n \leq k\}$, we obtain a sequence of p-values $\{1/\Lambda_1^* \geq 1/\Lambda_2^* \geq \cdots \geq 1/\Lambda_k^*\}$. According to the literature (Demortier, 2008), smaller p-values tend to give a more precise decision. Thus, we choose $1/\Lambda_k^*$ as the p-value for the following HiBH procedure. We summarize the proposed MPC algorithm in Algorithm 1.

4.2 HiBH: Hierarchical Benjamini-Hochberg Procedure

We first introduce the classical Benjamini-Hochberg (BH) procedure, one of the most widely used procedures for multiple testing (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001; Blanchard and Roquain, 2008; Basu et al., 2018).

Definition 7 (BH Procedure; Benjamini and Hochberg, 1995) *Given the p-values $p_1, p_2, \cdots, p_n$ corresponding to the null hypotheses $H_{1;0}, H_{2;0}, \cdots, H_{n;0}$, let $p_{(i)}$ be the i-th order statistic from the smallest to the largest. For a pre-specified level $\alpha \in (0,\ 1)$, define*

$$i_{BH}^* = \max \left\{ i \in [1:n] : p_{(i)} \leq \frac{i}{n} \alpha \right\}, \tag{15}$$

then, the null hypothesis $H_{(i),0}$ is rejected if $i \leq i^*_{BH}$.

In statistics, α is usually set to 0.05. While the BH procedure solves the multiple testing problem for single-level data, it fails to account for the hierarchical structure of visual data.

Algorithm 1 MPC

Input: Image I, score function $s(\cdot)$, parameters k_{il}, k_{pl}.
Calculate score map $s(I)$ and image score $s(I)_*$.
Initialize an empty set $\mathcal{Q} = \{\}$.
for $(\mathcal{S}, k)$ in $\{(\{s(I)_*\}, k_{il}), (s(I), k_{pl})\}$ **do**
  Split score map $s(I)$ into k disjoint patches $s(Z_1), s(Z_2), \ldots, s(Z_k)$.
  for S in $\mathcal{S}$ **do**
    Calculate sequence $\{Y_1, Y_2, \ldots, Y_k\}$ according to Equation (9).
    Calculate test supermartingale $\{\Lambda_0, \Lambda_1, \ldots, \Lambda_k\}$ according to Equation (13).
    Add p-value $p = 1/\sup_{0 \leq i \leq k}(\Lambda_i)$ into set $\mathcal{Q}$.
  end for
end for
Output: Image p-value $\mathcal{Q}[0]$, pixel p-values $\mathcal{Q}[1:]$.

In VAD, the image-level p-value summarizes the global anomaly evidence of an image. Images not rejected by the image-level BH step usually have weak anomaly evidence, and pixel-level discoveries in these images are more likely to be caused by local score fluctuations. Therefore, we use image-level BH as a screening step to determine whether pixel-level BH should be performed. By this mechanism, we can suppress low-confidence pixel-level discoveries.

Compared with the standard BH procedure, incorporating the hierarchical structure into multiple testing yields a tighter upper bound on pixel-level FDR, as established in the following section. Moreover, the potential pixel-level sensitivity loss caused by this mechanism can be mitigated by relaxing the image-level BH step.

To handle the hierarchical multiple testing problem in VAD, we extend the BH procedure and introduce the HiBH procedure in Algorithm 2. We prove that the HiBH procedure controls the FDR at both image and pixel levels in Section 4.3.

By integrating the MPC algorithm and the HiBH procedure, the complete HiMT framework is detailed in Algorithm 3, aligning with the structure illustrated in Figure 2.

4.3 Theoretical Guarantee for HiMT

We investigate the theoretical results of our proposed HiMT framework, beginning with a review of known theoretical results for the classical BH procedure.

Theorem 8 (FDR Control for BH Procedure; Benjamini and Hochberg, 1995) *If the p-values $p_1, p_2, \ldots, p_n$ are independent or satisfy positive regression dependency on the subset corresponding to true null hypotheses (PRDS), the BH procedure applied at level $\alpha \in (0, 1)$ controls the FDR at level α:*

$$\mathrm{FDR}_{\mathrm{BH}} \leq \pi\alpha \leq \alpha, \tag{16}$$

where π is the proportion of true null hypotheses.

Algorithm 2 HiBH Procedure

Input: Ordered image p-values $\{p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}\}$, ordered pixel p-values $\{p^i_{(1)} \leq p^i_{(2)} \leq \cdots \leq p^i_{(m)}\}_{i\in[n]}$, pre-specified levels $\alpha_{il}, \alpha_{pl} \in (0,1)$.
Initialize empty sets $\mathcal{R}_1$ and $\mathcal{R}_2$.
Calculate $i^*_{BH} = \max\{i \in [n] : p_{(i)} \leq \frac{i}{n}\alpha_{il}\}$.
for $i = 1$ **to** n **do**
  if $i \leq i^*_{BH}$ **then**
    Add $p_{(i)}$ into set $\mathcal{R}_1$.
    Calculate $j^*_{BH} = \max\{j \in [m] : p^i_{(j)} \leq \frac{j}{m}\alpha_{pl}\}$.
    Merge set $\{p^i_{(j)} : j \leq j^*_{BH}\}$ into set $\mathcal{R}_2$.
  end if
end for
Output: Reject null hypotheses corresponding to the p-values in $\mathcal{R}_1 \cup \mathcal{R}_2$.

Algorithm 3 HiMT Framework

Input: Test set $S^{Test} = \{I_i\}_{i\in[1,n]}$ where image $I_i = \{x^i_j\}_{j\in[1,m]}$, score function $s(\cdot)$, pre-specified levels $\alpha_{il}, \alpha_{pl} \in (0,1)$.
Call Algorithm 1 to get image p-values $\{p_i\}_{i\in[1,n]}$ and pixel p-values $\{p^i_j\}_{i\in[1,n],j\in[1,m]}$.
Sort the image p-values and pixel p-values, respectively.
Call Algorithm 2 to get p-value sets $\mathcal{R}_1$ and $\mathcal{R}_2$.
Output: Declare image I_i anomalous if $p_i \in \mathcal{R}_1$; otherwise, declare it normal. Declare pixel x^i_j anomalous if $p^i_j \in \mathcal{R}_2$; otherwise, declare it normal.

Theorem 8 requires that p-values are either independent or PRDS. We prove that the p-values in our proposed HiMT framework satisfy these conditions.

Lemma 9 *Let $p_1, p_2, \ldots, p_n$ denote the p-values calculated by Algorithm 1 based on scores $s_1, s_2, \ldots, s_n$, and let $\mathcal{N}_0$ and $\mathcal{N}_1$ be the indices for image scores and pixel scores. Then the p-values $\{p_i\}_{i\in\mathcal{N}_0}$ are independent and p-values $\{p_j\}_{j\in\mathcal{N}_1}$ are PRDS on the subset corresponding to normal scores.*

The proof of Lemma 9 is provided in Appendix A.3. Next, we show that our proposed HiMT framework also controls the FDR at both image level and pixel level.

Theorem 10 (FDR Control for HiMT Framework) *Let $p_1, p_2, \ldots, p_n$ be the p-values for images, $p^i_1, p^i_2, \ldots, p^i_m$ be the p-values for pixels within any image I_i. The HiMT framework applied at levels α_{il} (image level) and α_{pl} (pixel level) controls the FDR as follows:*

$$\begin{aligned} \textit{image-level}: &\quad \mathrm{FDR}_{\mathrm{HiMT}} \leq \alpha_{il}, \\ \textit{pixel-level}: &\quad \mathrm{FDR}_{\mathrm{HiMT}} \leq \alpha_{pl}. \end{aligned} \tag{17}$$

Theorem 10 shows that the HiMT framework simultaneously controls the FDR at two levels: i) at image level, it controls the FDR across all images in the data set, and ii) at pixel level, it controls the FDR across all pixels within each image. The proof of Theorem 10 is presented in Appendix A.4. We further derive test-set-level control for HiMT.

Proposition 11 (Averaged Pixel-level FDP Control) *Let G_i indicate whether image I_i is selected by image-level BH step, and $K = \sum_{i=1}^{n} G_i$. Let FDP_i^P denote pixel-level FDP within image I_i. Given a test set $\{I_i\}_{i=1}^n$, HiMT satisfies:*

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{FDP}_i^P\right] \leq \alpha_{pl}\frac{\mathbb{E}[K]}{n} \leq \alpha_{pl}. \tag{18}$$

The proof of Proposition 11 is provided in Appendix A.5. According to Proposition 11, the upper bound on the averaged pixel-level FDP adapts to the proportion of images rejected at image level. It is capped by the small constant α_{pl} and therefore cannot increase indefinitely.

5. Experiment

The HiMT framework is extensively evaluated and compared with existing decision-making methods in VAD task. Experiments are performed on three challenging data sets with ten popular score functions.

5.1 Experimental Settings

Score Functions. We choose ten popular VAD models as score functions, including CFA (Lee et al., 2022), CFLOW (Gudovskiy et al., 2022), DMAD (Liu et al., 2023), DRAEM (Zavrtanik et al., 2021), PaDiM (Defard et al., 2021), PatchCore (Roth et al., 2022), ReContrast (Guo et al., 2023), RRD (Tien et al., 2023), AnomalyGPT (Gu et al., 2024), and Dinomaly (Guo et al., 2025).

Baselines. We use four existing decision-making methods as baselines, including EDM (Hendrycks et al., 2022; Wang et al., 2022; Yang et al., 2021), ATT (Ali et al., 2013), ATH (Isaac and Sharma, 2024) and ATCS (Sun et al., 2024).

Data sets. We conduct experiments on three popular data sets, MVTec (Bergmann et al., 2019), VisA (Zou et al., 2022) and BTAD (Mishra et al., 2021). MVTec contains a total of 5354 images from 15 different objects and texture categories. VisA comprises 10821 images covering 12 categories in 3 domains. BTAD contains 2830 images of 3 industrial products.

Evaluation Metrics. We use the recall (Rec), false positive rate (FPR), precision (Pr) and F1-score (F1) to evaluate the performance. We use AUROC and AUPRO to evaluate the separability.

Implementation Details. The prescribed levels α_{il} and α_{pl} are set to 0.05. In all experiments, we use the betting function $f(y) = \phi(y; 1, 1)/c$ where $\phi(\cdot; \mu, \sigma)$ denotes the Gaussian density with mean μ and standard deviation σ, and $c = \int_0^1 \phi(y; 1, 1)\, dx$ is a normalization constant. All input images are resized to 256×256. No padding is used during patch splitting, and the stride is set equal to the patch size. All baseline methods and score functions in experiments use default settings. All experiments are conducted on a workstation with eight NVIDIA GeForce RTX 4090 GPUs and two 2.2GHz Intel CPUs.

5.2 Decision-making Performance of the HiMT Framework

Quantitative Results. We evaluate our proposed HiMT framework against four baseline methods using ten score functions on three data sets, demonstrating its superior performance and

generalization. Image-level and pixel-level results are shown in Tables 1 and 2, respectively. From Tables 1 and 2, we can see that:

1) Our method achieves superior decision-making performance in terms of F1-score. For instance, compared to ATCS using DMAD as score function, our method achieves 56.1% F1-score improvement on MVTec at image level.

2) The inferior F1-score of EDM indicates that distribution shift can severely degrade decision-making performance. For instance, in Table 2, when using RRD as score function on MVTec, EDM exhibits overly aggressive decision-making performance, tending to classify samples as anomalous (i.e., high recall, low precision). This occurs since EDM relies heavily on the calibration set, making it vulnerable to distribution shift—as shown in Figure 1 (b), where scores of calibration set are significantly lower than those of test set. Our method resolves this issue, delivering a 34.4% improvement in F1-score.

3) Compared to baseline methods unaffected by distribution shift, our method maintains better decision performance. For instance, in Table 1, ATCS prefers to guarantee the performance for anomalous data (i.e., high precision) but sacrifices the performance in normal data (i.e., low recall). In contrast, our method achieves a balance between precision and recall by controlling the FDR.

In-depth Analysis for Quantitative Results. The ATH method relies on periodicity condition and anomaly proportion, set to 0.01 in the original paper. Only samples that violate the periodicity condition while maintaining this anomaly proportion are flagged as anomalous. This conservative design favors classifying samples as normal, typically resulting in low FPR and high precision as shown in the third row of Table 1. However, the reduced number of TPs lowers recall, leading to poor decision performance (i.e., low F1). The ATCS method computes confidence intervals $[p-u, p+u]$ for each sample sequence, where $p \geq 0.9$ is a quantile and u is interval width. Samples below the lower bound are classified as normal, those above the upper bound as anomalous, and samples within the interval are abstained (for fairness, abstained samples are excluded from metric computation). The high quantile increases interval coverage over high-score regions, leading to high precision and low FPR. However, this also leaves some anomalous samples below the lower bound, reducing recall as shown in the fourth row of Table 1. Both the ATH and ATCS rely on information about the distribution of anomalous data, which is not available in an unsupervised setting. Consequently, their parameters must be heuristically determined based on prior knowledge. The ATT method models a Markov chain by initializing the state space and transition probability matrix using normal data. This makes its decisions vulnerable to distribution shifts, leading to high recall and low precision as shown in the second row of Table 1.

In contrast, our method makes decisions directly from the input image, naturally robust to distribution shift and does not rely on inaccessible prior information. As a result, it achieves superior performance, as shown in the fifth row of Table 1.

Method	Score Func	MVTec				VisA				BTAD			
		F1↑	Rec↑	FPR↓	Pr↑	**F1↑**	Rec↑	FPR↓	Pr↑	**F1↑**	Rec↑	FPR↓	Pr↑
EDM	CFLOW	72.8	65.4	14.0	94.8	56.4	43.3	9.1	87.4	48.4	52.4	5.4	81.0
ATT	CFLOW	82.8	**98.6**	99.8	72.1	71.5	**99.0**	100	56.3	63.6	**98.3**	100	55.2

Continued on next page

Method	Score Func	MVTec				VisA				BTAD			
		F1↑	Rec↑	FPR↓	Pr↑	**F1↑**	Rec↑	FPR↓	Pr↑	**F1↑**	Rec↑	FPR↓	Pr↑
ATH	CFLOW	73.2	65.8	14.0	94.8	56.4	43.3	9.1	87.4	48.4	52.4	5.4	81.0
ATCS	CFLOW	43.9	30.1	**0.0**	**98.1**	28.1	16.8	**0.0**	**100**	42.2	29.9	**0.0**	**100**
HiMT	CFLOW	**90.1**	95.0	45.7	86.2	**80.7**	90.4	45.2	74.6	**94.1**	96.7	33.9	91.7
EDM	Dinomaly	88.8	**99.8**	66.8	80.1	89.6	93.7	19.1	85.9	66.7	55.4	9.1	83.6
ATT	Dinomaly	92.3	98.3	39.2	87.1	84.3	74.3	2.5	**97.4**	51.5	74.4	95.1	39.4
ATH	Dinomaly	90.2	99.4	56.5	82.6	88.6	82.8	5.1	95.3	52.9	37.2	2.8	**91.7**
ATCS	Dinomaly	1.8	0.9	**0.0**	**100**	2.9	1.5	**0.6**	64.3	1.1	0.5	**0.1**	70.0
HiMT	Dinomaly	**97.6**	98.0	9.6	97.1	**90.6**	**95.0**	22.8	86.5	**76.1**	**85.5**	34.1	68.5
EDM	DMAD	22.7	16.2	**0.0**	**100**	45.5	31.6	0.4	**99.0**	64.2	48.3	0.1	98.7
ATT	DMAD	51.8	38.7	16.0	86.4	49.6	37.2	13.6	80.0	43.1	39.5	46.0	58.9
ATH	DMAD	22.7	16.2	**0.0**	**100**	45.9	31.8	0.4	**99.0**	64.2	48.3	0.1	98.7
ATCS	DMAD	39.0	27.5	3.3	98.1	27.5	16.6	**0.1**	97.9	35.8	23.9	**0.0**	**100**
HiMT	DMAD	**95.1**	**97.3**	16.4	93.7	**88.7**	**87.5**	10.5	92.6	**90.5**	**95.5**	36.0	86.4
EDM	DRAEM	90.2	**98.8**	46.4	84.6	82.1	**90.8**	45.7	78.4	63.8	56.8	11.9	82.9
ATT	DRAEM	90.1	91.2	24.6	90.2	66.2	64.8	30.3	83.8	45.4	37.9	**0.0**	**100**
ATH	DRAEM	90.2	**98.8**	46.4	84.6	82.1	**90.8**	45.7	78.4	63.8	56.8	11.9	82.9
ATCS	DRAEM	46.6	34.4	**4.2**	**97.8**	33.3	21.1	**0.0**	**100**	41.9	34.7	**0.0**	**100**
HiMT	DRAEM	**95.8**	97.4	10.9	94.6	**86.0**	90.4	27.9	83.4	**85.0**	**91.5**	37.1	80.3
EDM	PaDiM	37.2	27.1	**0.2**	**99.9**	22.4	15.1	0.8	85.6	26.4	30.9	1.4	89.7
ATT	PaDiM	82.7	**98.5**	99.8	72.0	71.5	**99.0**	100	56.3	63.6	**98.3**	100	55.2
ATH	PaDiM	38.1	28.1	**0.2**	**99.9**	22.4	15.1	0.8	85.6	26.4	30.9	1.4	89.7
ATCS	PaDiM	35.3	23.5	1.7	97.3	32.7	20.1	**0.2**	**97.4**	25.2	15.7	**0.0**	**100**
HiMT	PaDiM	**92.7**	98.0	31.3	88.6	**83.1**	94.2	41.4	75.8	**89.7**	87.1	10.6	92.7
EDM	ReContrast	95.1	97.3	16.4	93.4	88.4	84.3	6.8	94.0	79.2	82.2	4.1	84.2
ATT	ReContrast	89.9	94.5	33.9	87.0	68.6	76.6	56.4	63.3	66.0	88.6	53.0	66.2
ATH	ReContrast	95.1	97.3	16.4	93.4	88.4	84.3	6.8	94.0	79.2	82.2	4.1	84.2
ATCS	ReContrast	41.4	28.9	**5.3**	97.3	31.8	20.2	**0.0**	**100**	36.0	24.5	**0.0**	**100**
HiMT	ReContrast	**97.6**	**97.7**	**5.3**	**97.6**	**91.9**	**91.2**	8.2	93.0	**92.1**	**96.3**	26.6	88.8
EDM	CFA	59.2	48.9	**0.0**	**100**	26.7	18.8	0.1	95.8	78.9	68.7	0.3	97.2
ATT	CFA	68.3	64.2	55.0	74.4	49.1	54.4	83.6	45.9	58.0	69.6	60.8	60.8
ATH	CFA	60.4	50.0	**0.0**	**100**	26.7	18.8	0.1	95.8	78.9	68.7	0.3	97.2
ATCS	CFA	40.4	28.6	3.3	96.2	29.8	18.0	**0.0**	**100**	39.8	28.1	**0.0**	**100**
HiMT	CFA	**94.0**	**97.1**	31.5	91.8	**84.5**	**95.0**	42.1	77.7	**91.4**	**89.3**	33.4	94.6
EDM	PatchCore	68.1	59.6	**0.7**	**99.7**	64.5	52.2	1.3	92.3	81.0	74.7	0.6	95.1
ATT	PatchCore	83.8	97.9	87.0	74.4	71.7	**96.6**	93.7	57.3	65.3	96.3	82.5	59.3
ATH	PatchCore	68.1	59.6	**0.7**	**99.7**	64.5	52.2	1.3	92.3	81.0	74.7	0.6	95.1
ATCS	PatchCore	44.1	31.4	3.8	96.7	35.8	22.8	**0.3**	**95.8**	39.4	27.7	**0**	**100**
HiMT	PatchCore	**95.3**	**98.0**	18.8	93.1	**86.9**	92.3	27.8	83.4	**95.8**	**96.5**	33.3	95.7
EDM	RRD	91.0	**98.2**	42.5	85.4	88.5	**92.6**	23.8	85.3	71.2	**87.7**	23.0	71.1
ATT	RRD	85.8	85.9	29.7	87.0	64.6	69.0	54.6	61.2	63.6	78.2	53.6	64.2
ATH	RRD	91.0	**98.2**	42.5	85.4	88.5	**92.6**	23.8	85.3	71.2	**87.7**	23.0	71.1
ATCS	RRD	39.6	26.8	**5.6**	**97.8**	25.8	15.5	**0.0**	**100**	35.6	25.0	**0.0**	**100**
HiMT	RRD	**95.7**	97.2	12.4	94.5	**91.0**	91.9	12.6	90.3	**87.9**	82.5	3.6	96.7
EDM	AnomalyGPT	91.6	**98.3**	37.8	86.4	45.1	32.2	5.6	84.7	61.1	48.0	1.1	99.7
ATT	AnomalyGPT	86.3	79.3	**3.2**	**97.7**	21.2	14.8	**4.8**	**95.8**	84.7	78.2	53.6	92.7
ATH	AnomalyGPT	91.5	**98.3**	38.0	86.4	44.5	31.7	5.4	84.7	61.1	48.0	1.1	99.7

Continued on next page

Method	Score Func	MVTec				VisA				BTAD			
		F1↑	Rec↑	FPR↓	Pr↑	**F1↑**	Rec↑	FPR↓	Pr↑	**F1↑**	Rec↑	FPR↓	Pr↑
ATCS	AnomalyGPT	31.2	25.8	15.4	86.5	16.9	10.6	27.6	47.4	36.4	26.5	**0.0**	**100**
HiMT	AnomalyGPT	**95.6**	94.0	5.6	97.5	**71.9**	**100**	100	56.6	**94.8**	**100**	98.9	90.2

Table 1: Experimental results (%) of decision-making performance at **image level**. The ↑ indicates that larger values are better and vice versa. F1 is the primary metric, and Rec, Pr, and FPR are reported to further explain the source of F1.

Method	Score Func	MVTec				VisA				BTAD			
		F1↑	Rec↑	FPR↓	Pr↑	**F1↑**	Rec↑	FPR↓	Pr↑	**F1↑**	Rec↑	FPR↓	Pr↑
EDM	CFLOW	23.3	88.0	15.9	15.3	14.0	83.0	6.1	8.0	19.5	94.4	19.1	11.4
ATT	CFLOW	6.3	**99.9**	99.6	3.3	1.2	**99.6**	99.4	0.6	5.9	**100**	99.2	3.1
ATH	CFLOW	12.3	13.7	**0.3**	**65.1**	11.9	9.2	**0.1**	**34.7**	18.7	31.1	**0.4**	**44.8**
ATCS	CFLOW	33.4	54.6	2.1	31.3	12.9	68.0	2.7	7.9	15.3	31.3	2.5	16.3
HiMT	CFLOW	**41.0**	75.0	6.1	32.5	**28.3**	46.3	1.5	22.7	**40.4**	65.2	6.3	31.3
EDM	Dinomaly	27.0	**97.8**	23.3	15.7	10.7	**96.4**	13.0	5.7	5.8	95.9	7.5	3.0
ATT	Dinomaly	59.6	91.7	5.2	44.2	47.6	72.3	1.1	35.4	2.4	**98.7**	19.2	1.2
ATH	Dinomaly	69.1	82.6	2.5	59.4	50.9	48.7	**0.3**	53.2	35.9	24.6	**0.0**	**66.5**
ATCS	Dinomaly	37.6	34.4	**1.8**	41.4	19.3	50.0	2.6	12.0	11.7	84.6	2.8	6.3
HiMT	Dinomaly	**72.1**	85.0	2.5	**62.4**	**60.9**	61.6	1.9	**60.3**	**48.3**	56.6	0.2	42.1
EDM	DMAD	36.9	44.0	1.1	51.3	32.4	71.0	1.4	30.4	31.2	97.1	11.9	19.1
ATT	DMAD	23.4	**98.3**	27.9	14.1	9.6	**93.1**	33.3	5.9	14.1	**99.0**	27.6	7.8
ATH	DMAD	2.1	1.5	**0.0**	**82.5**	1.9	1.0	**0.0**	**71.9**	0.4	0.2	**0.0**	**84.3**
ATCS	DMAD	34.8	58.4	2.0	32.4	13.5	74.9	2.7	8.1	20.6	40.7	2.4	20.8
HiMT	DMAD	**55.0**	71.1	2.7	47.3	**40.6**	70.6	1.2	32.8	**57.0**	67.1	2.5	50.5
EDM	DRAEM	19.2	**97.3**	28.8	11.2	5.2	**92.9**	33.9	2.8	19.1	**55.0**	13.1	15.4
ATT	DRAEM	48.4	86.7	5.9	36.3	21.9	62.7	9.3	17.7	13.8	26.7	3.6	**19.9**
ATH	DRAEM	54.6	61.8	**1.4**	**66.6**	27.9	45.5	2.8	33.4	9.7	18.6	**2.8**	16.1
ATCS	DRAEM	38.4	59.6	1.9	36.8	12.1	62.0	2.7	7.5	6.1	6.8	**2.8**	7.5
HiMT	DRAEM	**65.4**	67.5	**1.4**	65.2	**38.5**	48.2	**1.2**	**35.5**	**25.1**	43.7	8.3	18.3
EDM	PaDiM	41.9	**77.1**	5.4	35.4	23.8	76.0	2.6	15.1	27.6	**97.4**	12.6	17.8
ATT	PaDiM	5.0	76.1	94.7	2.7	1.2	**89.7**	88.8	0.6	6.3	85.1	76.0	3.3
ATH	PaDiM	3.8	2.2	**0.0**	**67.8**	1.9	1.0	**0.0**	**78.1**	11.4	7.3	**0.0**	**93.2**
ATCS	PaDiM	34.4	56.1	2.0	32.3	12.7	67.0	2.7	7.8	19.2	38.3	2.4	19.5
HiMT	PaDiM	**44.5**	51.1	2.1	50.5	**31.4**	55.6	1.6	28.3	**47.5**	46.8	1.9	51.7
EDM	ReContrast	24.3	96.9	17.9	14.8	9.8	96.9	9.7	5.3	18.0	99.4	22.0	10.4
ATT	ReContrast	6.3	**100**	99.9	3.3	1.2	**99.2**	99.9	0.6	5.8	**100**	99.9	3.1
ATH	ReContrast	54.3	66.7	**1.8**	**53.3**	38.4	42.9	**0.4**	**42.7**	55.1	70.8	2.2	50.0
ATCS	ReContrast	37.0	59.3	2.0	34.6	14.7	75.0	2.7	8.9	20.8	42.1	2.4	21.2
HiMT	ReContrast	**59.7**	71.1	2.7	52.7	**47.0**	59.3	0.8	40.8	**59.8**	61.2	**1.8**	**59.2**
EDM	CFA	42.9	81.5	6.1	35.8	25.5	68.6	2.5	16.5	22.9	98.5	16.6	14.0
ATT	CFA	10.0	**99.9**	58.6	5.4	5.5	**97.9**	20.8	2.9	8.9	**99.7**	56.5	4.8
ATH	CFA	23.3	19.6	**0.4**	**72.9**	13.3	10.3	**0.1**	**64.8**	49.0	53.7	**2.3**	**56.1**
ATCS	CFA	35.2	56.9	2.0	33.2	12.9	66.7	2.7	7.9	20.8	39.3	2.4	21.3
HiMT	CFA	**54.8**	63.3	2.4	50.7	**38.7**	46.4	0.8	37.0	**55.6**	63.0	2.9	49.8
EDM	PatchCore	43.4	83.4	5.9	34.4	17.4	87.7	4.4	10.5	21.8	98.3	18.6	13.3

Continued on next page

Method	Score Func	MVTec				VisA				BTAD			
		F1↑	Rec↑	FPR↓	Pr↑	**F1↑**	Rec↑	FPR↓	Pr↑	**F1↑**	Rec↑	FPR↓	Pr↑
ATT	PatchCore	6.3	**100**	99.9	3.3	1.2	**100**	99.9	0.6	5.8	**100**	99.7	3.1
ATH	PatchCore	34.6	30.0	**0.5**	**72.7**	31.0	29.8	**0.2**	**58.7**	30.2	37.1	**0.4**	**53.0**
ATCS	PatchCore	36.6	59.0	2.0	34.4	13.9	70.2	2.7	8.5	17.9	33.7	2.5	18.3
HiMT	PatchCore	**56.4**	65.1	2.2	52.3	**43.6**	48.4	0.6	42.7	**51.0**	60.4	3.4	44.2
EDM	RRD	18.3	98.0	30.0	10.6	7.0	**98.7**	15.9	3.8	14.8	99.5	29.7	8.3
ATT	RRD	13.0	**99.6**	53.4	7.3	5.5	97.9	39.5	2.9	11.4	**99.7**	38.2	6.2
ATH	RRD	49.1	65.8	2.1	**49.6**	41.9	52.1	0.5	40.7	42.8	91.6	7.4	28.1
ATCS	RRD	36.3	58.8	**2.0**	34.0	14.6	74.8	2.7	8.9	21.3	41.9	**2.4**	21.8
HiMT	RRD	**52.7**	65.6	2.8	47.4	**47.4**	56.0	**0.4**	**41.9**	**44.2**	68.0	8.1	**33.5**
EDM	AnomalyGPT	30.8	**87.2**	11.0	20.4	11.9	**47.6**	4.7	7.5	31.8	68.8	16.9	21.3
ATT	AnomalyGPT	40.2	34.3	**0.3**	**73.8**	6.3	3.7	**0.0**	45.2	11.4	**100**	100	6.1
ATH	AnomalyGPT	51.2	57.7	1.1	57.3	4.1	2.2	**0.0**	**64.8**	7.8	4.4	**0.1**	**68.7**
ATCS	AnomalyGPT	39.7	57.2	2.0	39.9	10.0	43.0	16.4	7.0	16.5	12.2	2.0	27.1
HiMT	AnomalyGPT	**53.3**	49.5	0.7	64.2	**16.4**	12.1	0.2	32.0	**33.8**	45.3	11.3	29.8

Table 2: Experimental results (%) of decision-making performance at **pixel level**. The ↑ indicates that larger values are better and vice versa. F1 is the primary metric, and Rec, Pr, and FPR are reported to further explain the source of F1.

Qualitative Results. The visualization results are shown in Figure 4. As shown in Figure 4, compared to EDM (the second row), our method (the third row) delivers more accurate anomaly locations. Moreover, our method avoids the conflicting decisions observed in the fourth and eighth columns in Figure 4.

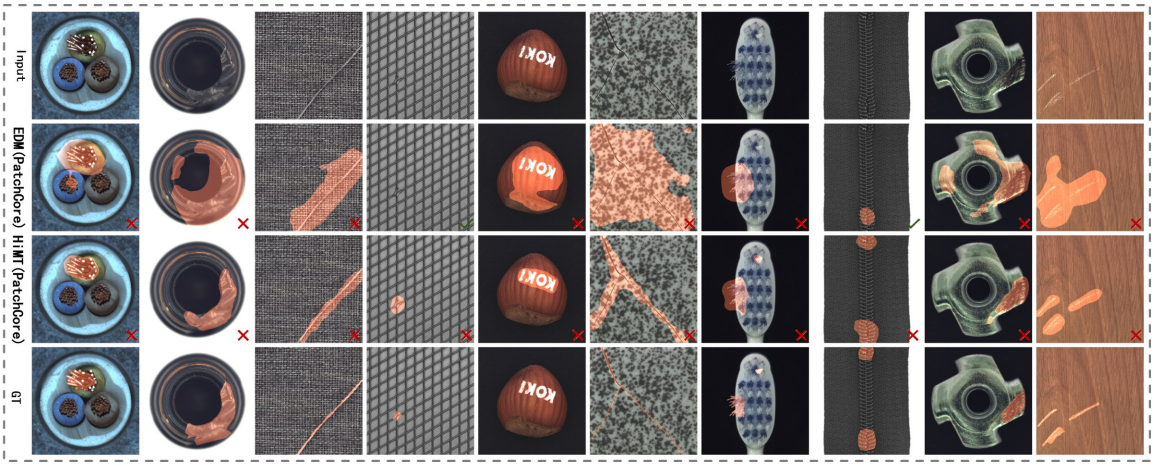
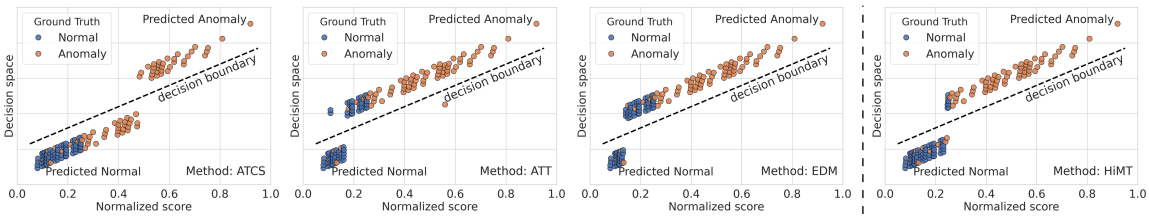


Figure 4: Qualitative results on the MVTec data set. Markers in the bottom-right indicate image-level decisions: green ✓for normal, red ✕ for anomalous. Red-highlighted pixels mark anomalous regions.

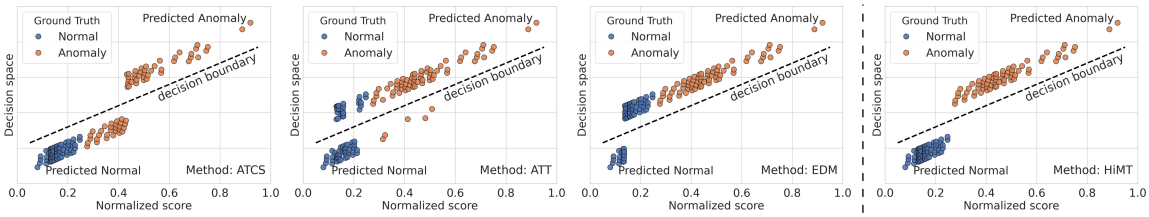
Decision Boundaries. To further explain the decision-making behavior of different methods, we visualize their decision boundaries and report the corresponding metrics on category-specific data. The results are shown in Figures 5 and 6. Experiments are conducted on the MVTec data set, using RRD as the score function. From Figures 5 and 6, we observe that:

1) Baselines exhibit bias toward classifying samples as either normal or anomalous, which results in suboptimal decisions. Aggressive decision boundaries classify more samples as anomalous. For example, in the Carpet category shown in Figure 5, EDM incorrectly identifies many normal samples as anomalous. Such a decision leads to higher recall but comes at the cost of reduced precision, as shown in the Carpet category of Figure 6. While conservative decision boundaries classify more samples as normal. For example, ATCS sets a conservative decision boundary in the Tile category of Figure 5. This improves precision and reduces FPR, but it misses many true anomalies and therefore reduces recall, as shown in the Tile category of Figure 6.

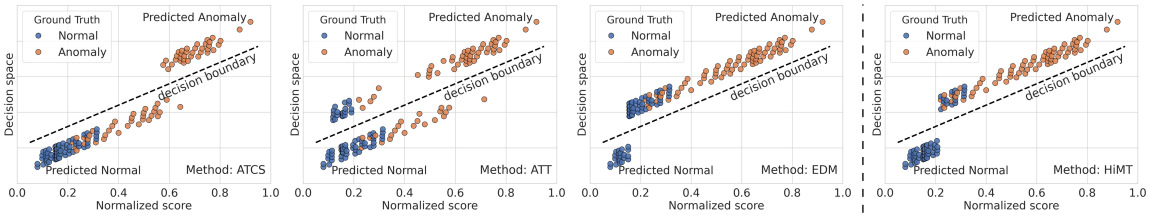
2) In contrast, our method leverages martingale theory and hypothesis testing framework to control the FDR, achieving a better trade-off in decision bias and producing more accurate decision boundaries (i.e., higher F1-score).



(a) Carpet

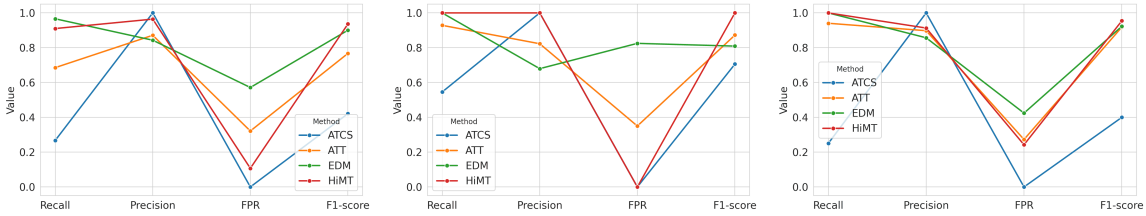


(b) Hazelnut



(c) Tile

Figure 5: Comparison of decision boundaries between HiMT and baseline methods.



(a) Carpet (b) Hazelnut (c) Tile

Figure 6: Metric comparison for the decision boundaries in Figure 5.

5.3 Separability Evaluation

To comprehensively evaluate our proposed HiMT framework, we further report AUROC and AUPRO metrics to evaluate the threshold-independent separability. In our framework, the p-values produced by MPC are used as decision statistics, and we convert them into anomaly scores by $(1 - p\text{-value})$, since smaller p-values indicate stronger evidence against the null hypothesis. The results are shown in Tables 3 and 4.

As shown in Tables 3 and 4, our method consistently improves both AUROC and AUPRO. These results demonstrate that HiMT not only achieves superior decision performance by effectively balancing recall and precision/FPR, but also enhances the separability between normal and anomalous samples.

5.4 Efficiency Comparisons

We evaluate the efficiency of our HiMT framework in terms of runtime and peak memory cost. All experiments are conducted on a workstation equipped with an NVIDIA RTX 4090 GPU and a 2.2 GHz Intel CPU.

We present runtime comparisons in Table 5. Each result is averaged over five runs, with standard deviations reported. As shown in Table 5, our HiMT framework achieves faster inference than most baselines, demonstrating its feasibility for real-time applications.

We report peak memory cost in Table 6. For each data set, we compute the peak memory cost required to process one category in parallel, and average the values over all categories. The results show that our method incurs only modest additional memory overhead.

5.5 Ablation Studies on FDR Control Levels

Theorem 10 establishes FDR control at specified levels α_{il} and α_{pl} at image level and pixel level, respectively. In this section, we perform an ablation study on these parameters to assess their impact on the decision-making performance of the proposed HiMT method, with results presented in Figure 7.

The results show that: (1) The F1-score is relatively insensitive to small changes around $\alpha_{il} = \alpha_{pl} = 0.05$; (2) When α_{il} and α_{pl} move beyond the locally insensitive range, all metrics change noticeably.

Model	MVTec		VisA		BTAD	
	AUROC↑	AUPRO↑	AUROC↑	AUPRO↑	AUROC↑	AUPRO↑
CFA	93.73	97.94	87.97	89.34	92.97	97.53
CFA + HiMT	**95.24**	**98.54**	**91.50**	**94.54**	**93.88**	**97.75**
CFLOW	89.16	94.77	80.28	84.93	91.76	97.57
CFLOW + HiMT	**92.22**	**96.62**	**85.51**	**91.04**	**92.84**	**97.83**
DMAD	98.37	99.17	96.25	96.80	92.55	96.83
DMAD + HiMT	**98.74**	**99.46**	**97.00**	**97.68**	**93.49**	**97.39**
DRAEM	98.19	98.97	90.59	92.58	87.75	88.43
DRAEM + HiMT	**98.38**	**99.10**	**92.36**	**94.86**	**89.98**	**92.60**
PaDiM	93.05	96.80	84.63	85.02	94.05	96.18
PaDiM + HiMT	**94.80**	**98.00**	**89.50**	**92.80**	**95.27**	**97.03**
PatchCore	97.08	98.99	89.41	90.83	92.38	98.50
PatchCore + HiMT	**97.46**	**99.15**	**92.14**	**94.73**	**93.81**	**98.85**
ReContrast	99.27	99.62	96.40	97.11	94.35	98.08
ReContrast + HiMT	**99.30**	**99.63**	**96.86**	**97.49**	**94.73**	**98.33**
RRD	96.55	97.76	95.79	96.65	92.21	97.44
RRD + HiMT	**97.41**	**98.55**	**96.26**	**97.18**	**96.44**	**98.28**

Table 3: Experimental results (%) of separability at image level. The ↑ indicates that larger values are better.

Model	MVTec		VisA		BTAD	
	AUROC↑	AUPRO↑	AUROC↑	AUPRO↑	AUROC↑	AUPRO↑
CFA	97.53	53.37	97.94	33.67	97.49	57.66
CFA + HiMT	**98.34**	**74.18**	**98.35**	**53.22**	**97.96**	**71.81**
CFLOW	94.44	41.13	96.19	22.18	94.13	32.84
CFLOW + HiMT	**97.09**	**78.23**	**97.53**	**49.08**	**96.24**	**73.02**
DMAD	97.28	51.81	98.94	38.86	97.95	56.39
DMAD + HiMT	**98.41**	**79.76**	**99.26**	**71.33**	**98.41**	**77.04**
DRAEM	97.35	69.02	92.95	32.71	79.84	14.93
DRAEM + HiMT	**98.59**	**80.28**	**97.10**	**52.75**	**89.64**	**43.01**
PaDiM	97.00	48.62	98.01	28.66	97.56	54.51
PaDiM + HiMT	**97.62**	**53.34**	**98.31**	**32.87**	**97.74**	**59.78**
PatchCore	97.64	55.88	98.04	39.75	96.98	50.95
PatchCore + HiMT	**98.47**	**76.58**	**98.57**	**55.72**	**97.68**	**68.68**
ReContrast	97.72	56.94	98.51	42.19	98.00	62.29
ReContrast + HiMT	**98.68**	**80.72**	**99.07**	**64.72**	**98.35**	**75.52**
RRD	97.25	54.39	98.82	42.37	98.01	60.03
RRD + HiMT	**98.40**	**79.33**	**99.16**	**64.10**	**98.81**	**84.71**

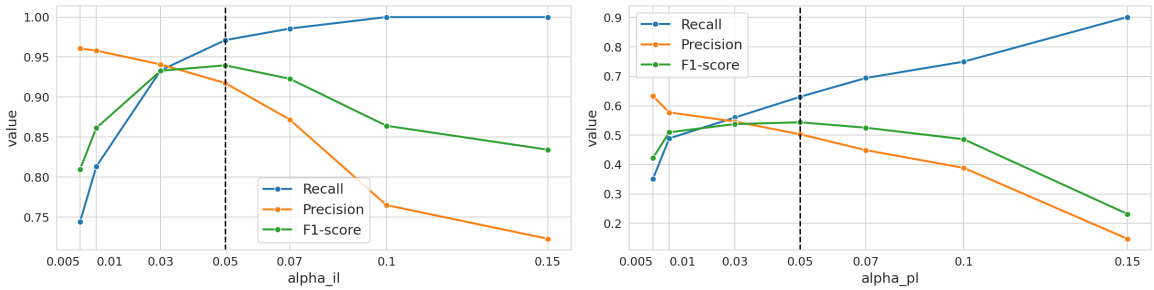
Table 4: Experimental results (%) of separability at pixel level. The ↑ indicates that larger values are better.

Method	datasize=100	datasize=500	datasize=1000
EDM	<1	<1	<1
ATT	2953.5±3.1	3000.9±4.3	3098.5±2.6
ATH	2973.5±5.3	2956.7±3.2	2949.6±2.9
ATCS	2350.4±7.4	2427.6±8.0	2524±7.3
HiMT	177.6±2.2	178.9±1.0	180.0±1.4

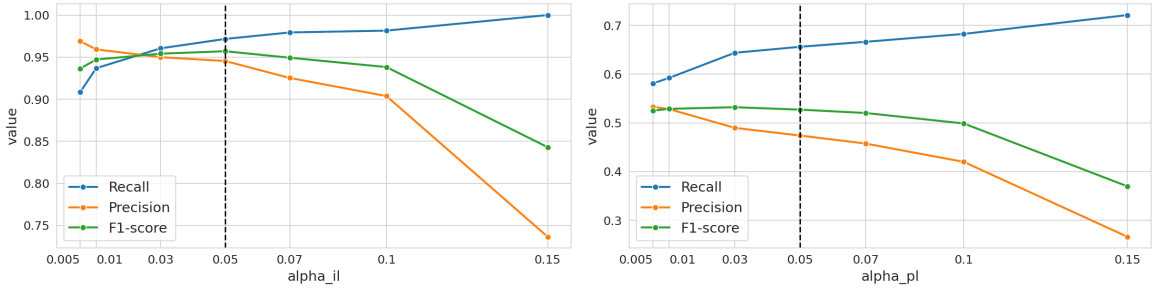
Table 5: Runtime comparison (ms).

Method	MVTec	VisA	BTAD
EDM	730.10	2177.20	1809.01
ATT	238.04	542.97	548.84
ATH	367.08	1094.42	909.51
ATCS	192.25	301.07	413.04
HiMT	36.03	56.30	77.19

Table 6: Peak memory comparison (MB).



(a) CFA



(b) RRD

Figure 7: Ablation results of α_{il} and α_{pl} on MVTec.

5.6 Ablation Studies on Patch Size

To construct a valid supermartingale process, the score map is divided into a sequence of disjoint patches, $s(Z_1), s(Z_2), \ldots, s(Z_k)$, where $k \in \{k_{il}, k_{pl}\}$. Here, k_{il}, k_{pl} denote the number of patches used for image-level and pixel-level testing, respectively. In this section, we perform ablation experiments on k_{il}, k_{pl} to evaluate their impact on the decision-making performance of the proposed HiMT method. The results are reported in Table 7.

The ablation results in Table 7 show that: 1) The patch sizes of 4 and 8 generally yield the best F1-scores at both image and pixel levels. 2) Very coarse partitions may hurt the statistical power of the martingale test for some score functions (e.g., DMAD with $k_{il} = 2^2$ at image level). Therefore, we set $k = 4^2$ for image-level computation and $k = 8^2$ for pixel-level experiments.

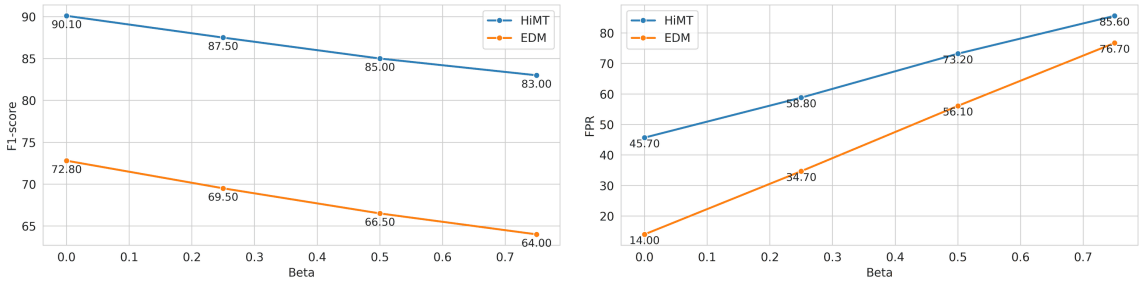
Score Function	$k_{il}=2^2$	$k_{il}=4^2$	$k_{il}=8^2$	$k_{il}=16^2$	$k_{pl}=2^2$	$k_{pl}=4^2$	$k_{pl}=8^2$	$k_{pl}=16^2$
DMAD	90.1	95.1	**95.9**	95.8	51.2	**55.0**	54.3	54.2
CFLOW	**90.1**	**90.1**	**90.1**	89.3	39.4	**41.0**	40.2	40.3
PatchCore	95.2	**95.3**	95.5	95.2	56.3	56.0	**56.6**	56.5
ReContrast	**97.6**	**97.6**	**97.6**	**97.6**	59.2	**59.7**	59.3	59.1

Table 7: Ablation results (F1-score) on patch size k_{il} (image level) and k_{pl} (pixel level) on MVTec data set.

5.7 Impact of Score Function Quality on Decision-making Performance

Our HiMT method is a decision-making framework that converts output scores into decisions with rigorous theoretical guarantees and strong empirical performance. It is score function-agnostic and can be applied to any score function to enhance decision-making performance. The quality of score function influences the performance of decision-making methods. In this section, we conduct experiments to study how the quality of the score function affects the performance of the decision methods.

Specifically, we use CFLOW and simulate a low quality score function by randomly selecting β proportion of normal samples and adding a perturbation γ to their scores. In the experiments, γ is set to the mean score of the test set. Experimental results are presented in Figure 8. From Figure 8, we observe that: 1) An increase in FPs leads to a higher FPR, which in turn affects the F1-score. On average, for every 25% increase in FPs, the F1-score drops by 2.37% for the HiMT method and by 2.93% for the EDM method; 2) Regardless of the quality of the score function used, our method consistently achieves superior decision performance.



(a) F1-score (b) FPR

Figure 8: The impact of the number of false positives (FPs) on decision performance. β denotes the proportion of FPs, with $\beta = 0$ indicating no perturbation.

6. Conclusion

In this paper, we identify two critical problems with existing decision-making approaches: distribution shift and theoretical guarantee. To address these challenges, we propose a novel

HiMT framework and provide theoretical guarantee for FDR control. Extensive experiments demonstrate the superior decision-making capabilities of our proposed HiMT framework.

Future directions of this work include optimizing score function quality; while this work provides a statistical foundation for any black-box score function, the reliability of its outputs depends on the score function's strength. Also, the martingale construction relies on assumptions. Although these assumptions are empirically supported in experiments, further relaxing them would make the framework more broadly applicable. Ultimately, by providing a general statistical framework for decision-making in VAD, we hope to help practitioners achieve reliable anomaly detection without sacrificing statistical validity.

Acknowledgments

We thank the action editor and anonymous reviewers for their constructive comments. This work is supported by the Key R&D Program of Hubei Province under Grant 2024BAB038, National Natural Science Foundation of China under Grant 624B2106, and National Key R&D Program of China under Grant 2023YFC3604702.

Appendix A. Omitted Proofs

A.1 Proof of Lemma 4

Proof Let $\{Z_1, Z_2, \ldots, Z_k\}$ denote the patches of image I, and let $Z \in \{Z_1, Z_2, \ldots, Z_k\}$ be an arbitrary patch in I. Let

$$N_0 = \sum_{j=1}^{r_0} \mathbb{1}_U\{s(x_j) \preceq S\} \quad \text{and} \quad N_1 = \sum_{j=1}^{r_1} \mathbb{1}_U\{s(\hat{x}_j) \preceq S\}. \tag{19}$$

The empirical randomized cumulative distribution function of normal scores in patch Z is

$$\hat{F}^{\text{rand}}_{z,s}(S) = \frac{\sum_{j=1}^{r_0} \mathbb{1}_U\{s(x_j) \preceq S\}}{r_0} = \frac{N_0}{r_0}. \tag{20}$$

If image I is normal, then $r_1 = N_1 = 0$. Thus, we have

$$\begin{aligned}
\mathbb{E}_Z[Y \mid S] &= \mathbb{E}_Z\left[\frac{\sum_{j=1}^{r_0} \mathbb{1}_U\{s(x_j) \preceq S\} + \sum_{j=1}^{r_1} \mathbb{1}_U\{s(\hat{x}_j) \preceq S\}}{r_0 + r_1} \mid S\right] \\
&= \mathbb{E}_Z\left[\frac{N_0 + N_1}{r_0 + r_1} \mid S\right] \\
&= \mathbb{E}_Z\left[\frac{N_0}{r_0} \mid S\right] \\
&= \mathbb{E}_Z[\hat{F}^{\text{rand}}_{z,s}(S) \mid S].
\end{aligned} \tag{21}$$

Since $\hat{F}^{\text{rand}}_{z,s}(S)$ is an unbiased estimator of $F^{\text{rand}}_{z,s}(S)$, we obtain

$$\mathbb{E}_Z[Y \mid S] = \mathbb{E}_Z[\hat{F}^{\text{rand}}_{z,s}(S) \mid S] = F^{\text{rand}}_{z,s}(S) = F^{\text{rand}}_{s}(S). \tag{22}$$

If I is an anomalous image, we assume that the patches in I satisfy

$$\mathbb{E}_Z\left[N_1 \mid S\right] \leq \mathbb{E}_Z\left[\frac{r_1}{r_0} N_0 \mid S\right]. \tag{23}$$

Equation (23) only constrains the population-level behavior of patches.

We empirically examine the validity of the assumption in Equation (23). We conduct experiments on the MVTec data set with DMAD and PatchCore as score functions. For each category, we split each anomalous image into 64 patches. We then randomly select 10 normal scores from the current category. For each selected normal score S, we compute $\mathbb{E}\left[N_1 \mid S\right]$ and $\mathbb{E}\left[\frac{r_1}{r_0} N_0 \mid S\right]$. The results are shown in Figures 9 and 10. Figures 9 and 10 show that Equation (23) holds across all categories and different score functions. These results provide empirical support for the relaxed assumption in Equation (23) in VAD task.

For an anomalous image I, we have

$$\begin{aligned}
\mathbb{E}_Z[Y \mid S] &= \mathbb{E}_Z\left[\frac{\sum_{j=1}^{r_0} \mathbb{1}_U\{s(x_j) \preceq S\} + \sum_{j=1}^{r_1} \mathbb{1}_U\{s(\hat{x}_j) \preceq S\}}{r_0 + r_1} \mid S\right] \\
&= \mathbb{E}_Z\left[\frac{N_0 + N_1}{r_0 + r_1} \mid S\right] \\
&= \mathbb{E}_Z\left[\frac{N_0}{r_0 + r_1} \mid S\right] + \mathbb{E}_Z\left[\frac{N_1}{r_0 + r_1} \mid S\right].
\end{aligned} \tag{24}$$

Since $r_0 + r_1$ is fixed within each patch, we have

$$\mathbb{E}_Z\left[\frac{N_0}{r_0+r_1} \mid S\right] + \mathbb{E}_Z\left[\frac{N_1}{r_0+r_1} \mid S\right] = \mathbb{E}_Z\left[\frac{N_0}{r_0+r_1} \mid S\right] + \frac{1}{r_0+r_1}\mathbb{E}_Z\left[N_1 \mid S\right]. \tag{25}$$

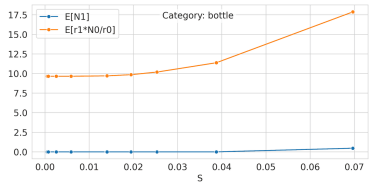
By Equation (23), we further have

$$\begin{aligned}\mathbb{E}_Z\left[\frac{N_0}{r_0+r_1} \mid S\right] + \frac{1}{r_0+r_1}\mathbb{E}_Z\left[N_1 \mid S\right] &\leq \mathbb{E}_Z\left[\frac{N_0}{r_0+r_1} \mid S\right] + \frac{1}{r_0+r_1}\mathbb{E}_Z\left[\frac{r_1}{r_0}N_0 \mid S\right] \\ &= \mathbb{E}_Z\left[\frac{N_0}{r_0} \mid S\right] \\ &= \mathbb{E}_Z[\hat{F}^{\text{rand}}_{z,s}(S) \mid S] \\ &= F^{\text{rand}}_s(S).\end{aligned} \tag{26}$$

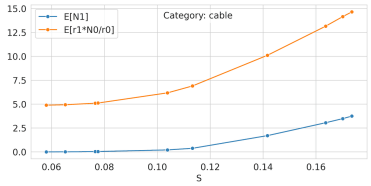
Therefore,

$$\mathbb{E}_Z[Y \mid S] \leq F^{\text{rand}}_s(S). \tag{27}$$

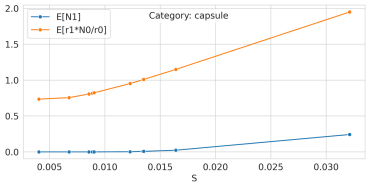
■



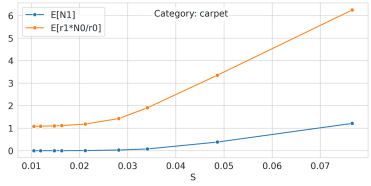
(a) Bottle



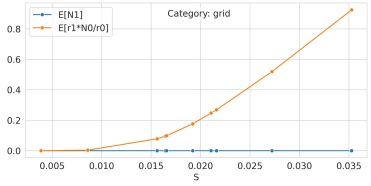
(b) Cable



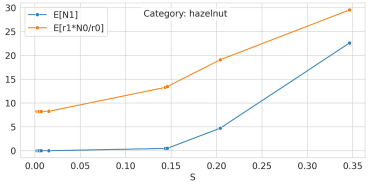
(c) Capsule



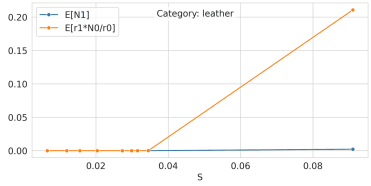
(d) Carpet



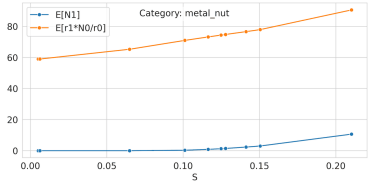
(e) Grid



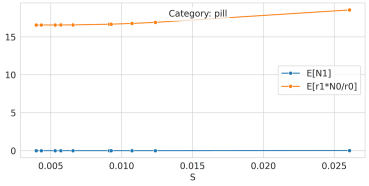
(f) Hazelnut



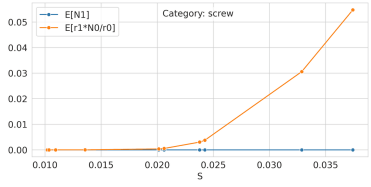
(g) Leather



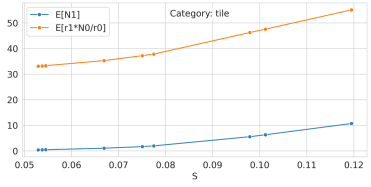
(h) Metal nut



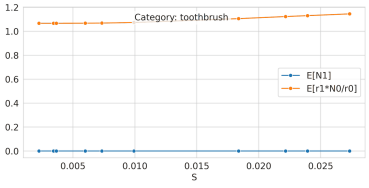
(i) Pill



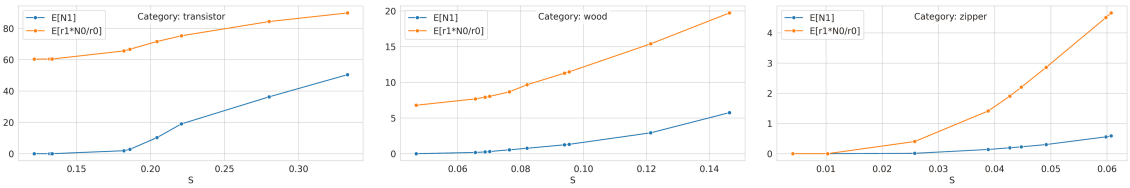
(j) Screw



(k) Tile

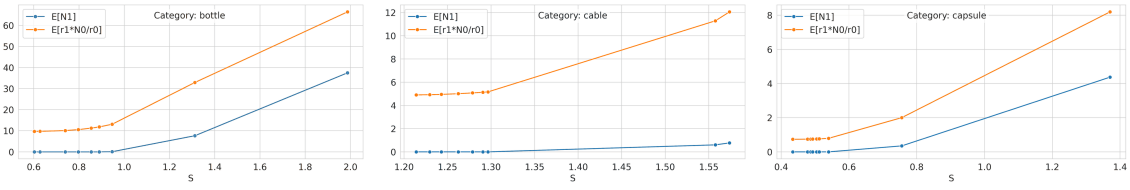


(l) Toothbrush

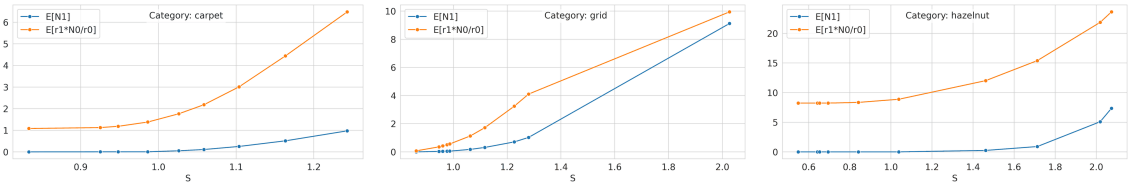


(m) Transistor (n) Wood (o) Zipper

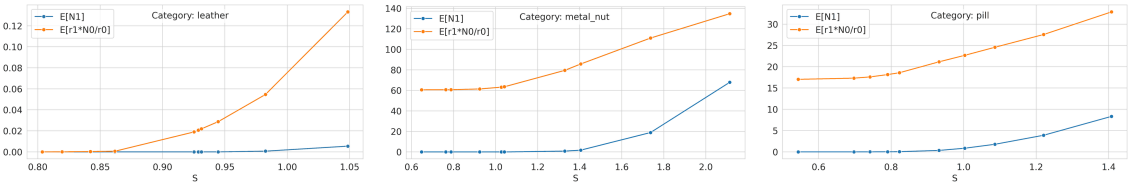
Figure 9: Comparison of $\mathbb{E}[N_1 \mid S]$ and $\mathbb{E}\left[\frac{r_1}{r_0} N_0 \mid S\right]$ on MVTec with DMAD.



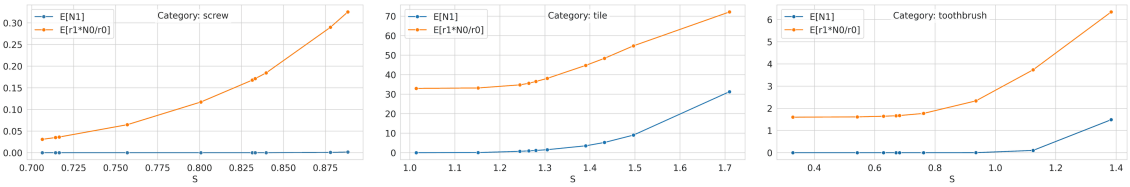
(a) Bottle (b) Cable (c) Capsule



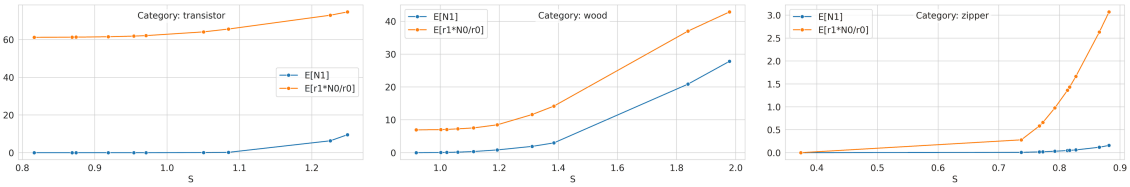
(d) Carpet (e) Grid (f) Hazelnut



(g) Leather (h) Metal nut (i) Pill



(j) Screw (k) Tile (l) Toothbrush



(m) Transistor (n) Wood (o) Zipper

Figure 10: Comparison of $\mathbb{E}[N_1 \mid S]$ and $\mathbb{E}\left[\frac{r_1}{r_0} N_0 \mid S\right]$ on MVTec with PatchCore.

A.2 Proof of Theorem 5

Proof According to Definition 1, it suffices to prove that $\{\Lambda_n, 0 \le n \le k\}$ satisfies the following three conditions: i) $\Lambda_n \ge 0$; ii) $\Lambda_0 = 1$; iii) $\mathbb{E}[\Lambda_n \mid \Lambda_0, \Lambda_1, \ldots, \Lambda_{n-1}] \le \Lambda_{n-1}$ for any integer $n \in [1, k]$.

By the definition of Λ_n in Theorem 5, we have $\Lambda_0 = 1$. Since $f(\cdot)$ is positive, we also have $\Lambda_n > 0$. It remains to prove that

$$\mathbb{E}[\Lambda_n \mid \Lambda_0, \Lambda_1, \ldots, \Lambda_{n-1}] \le \Lambda_{n-1}, \quad \forall n \in [1, k].$$

For $2 \le n \le k$, we have

$$\begin{aligned} \mathbb{E}[\Lambda_n \mid \Lambda_0, \Lambda_1, \ldots, \Lambda_{n-1}] &= \mathbb{E}\left[\prod_{i=1}^{n} f(Y_i) \mid Y_1, \ldots, Y_{n-1}\right] \\ &= \mathbb{E}\left[\Lambda_{n-1} f(Y_n) \mid Y_1, \ldots, Y_{n-1}\right] \\ &= \Lambda_{n-1}\, \mathbb{E}\left[f(Y_n) \mid Y_1, \ldots, Y_{n-1}\right]. \end{aligned} \tag{28}$$

In VAD, under the null hypothesis, we assume that the transformed statistics $\{Y_i, 1 \le i \le k\}$ are jointly independent. We empirically examine the dependence among the $\{Y_i, 1 \le i \le k\}$ using the Hilbert–Schmidt independence criterion (HSIC) (Gretton et al., 2007).

Specifically, we divide each score map into non-overlapping 4×4 patches and compute $\{Y_i, 1 \le i \le 16\}$ according to Equation (10). We then evaluate their pairwise independence using nine score functions on the MVTec data set. The results are shown in Figure 11. As shown in Figure 11, all reported p-values exceed the significance level of 0.01, and thus pairwise independence is not rejected at this significance level. These results provide empirical support for the independence assumption used in Theorem 5.

Therefore, we have

$$\begin{aligned} \mathbb{E}\left[f(Y_n) \mid Y_1, \ldots, Y_{n-1}\right] &= \mathbb{E}[f(Y_n)] \\ &= \mathbb{E}_S\left[\mathbb{E}_Z[f(Y_n) \mid S]\right]. \end{aligned} \tag{29}$$

Since $f(\cdot)$ is concave, by Jensen's inequality, we have

$$\mathbb{E}_Z[f(Y_n) \mid S] \le f\left(\mathbb{E}_Z[Y_n \mid S]\right). \tag{30}$$

According to Lemma 4, we have

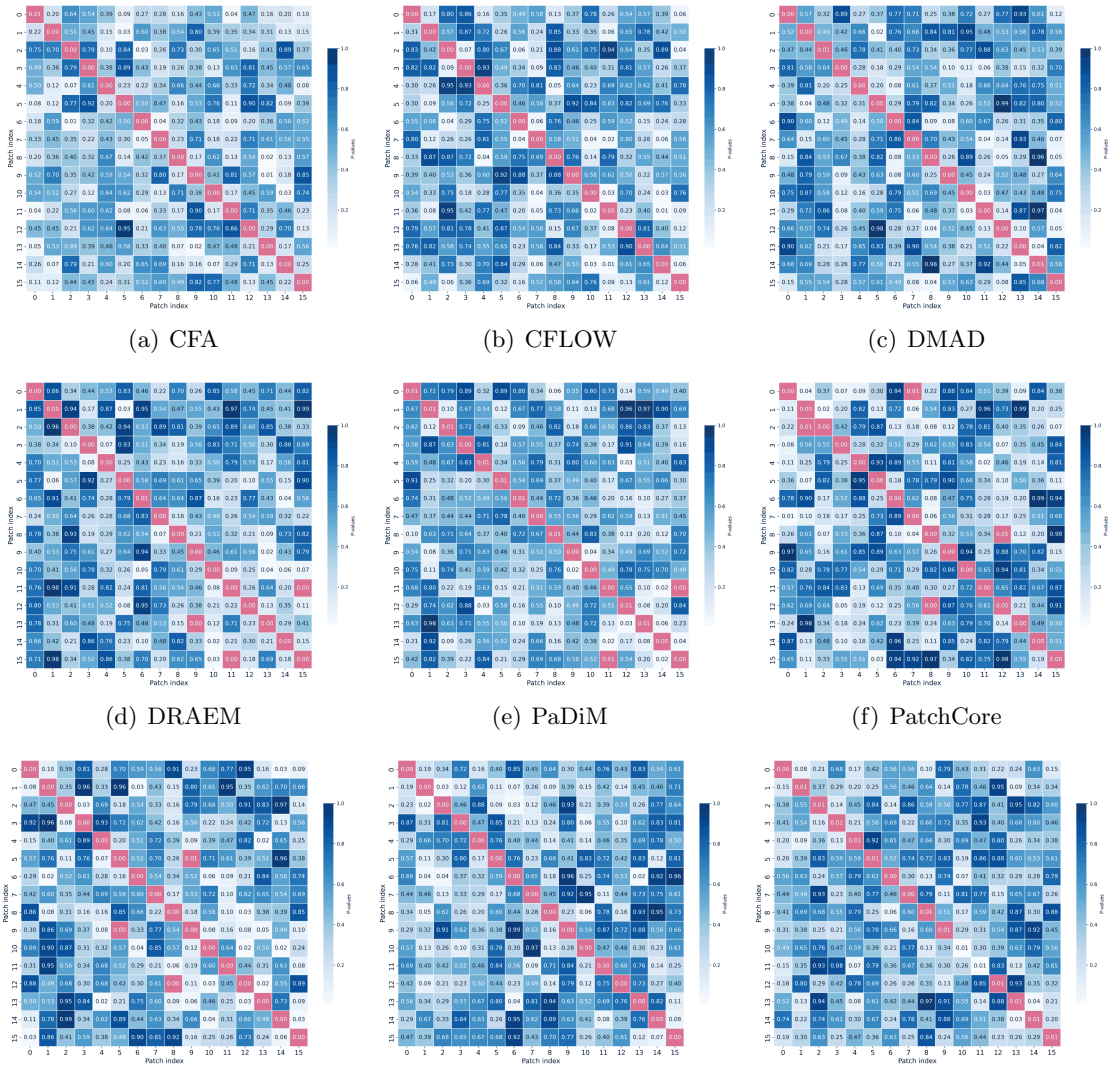
$$f\left(\mathbb{E}_Z[Y_n \mid S]\right) \le f\left(F_s^{\text{rand}}(S)\right), \tag{31}$$

where F_s^{rand} denotes the randomized CDF. By the randomized version of the probability integral transform (randomized PIT; Brockwell, 2007), $F_s^{\text{rand}}(S)$ follows the uniform distribution on $[0, 1]$. Let

$$V = F_s^{\text{rand}}(S).$$

Then $V \sim U(0, 1)$, and thus

$$\mathbb{E}_S\left[f\left(F_s^{\text{rand}}(S)\right)\right] = \mathbb{E}[f(V)] = \int_0^1 f(v)\, dv = 1. \tag{32}$$



(a) CFA (b) CFLOW (c) DMAD

(d) DRAEM (e) PaDiM (f) PatchCore

(g) ReContrast (h) RRD (i) Dinomaly

Figure 11: HSIC results for $\{Y_i, 1 \leq i \leq 16\}$ on MVTec data set. The entry at row i, column j reports the p-value of the independence test between the i-th and j-th patches. Values greater than 0.01 (blue and white regions) indicate that pairwise independence is not rejected at the 0.01 significance level. This provides empirical support for the independence assumption in Theorem 5.

Combining the above results, we obtain

$$\begin{aligned}\mathbb{E}[\Lambda_n \mid \Lambda_0, \Lambda_1, \ldots, \Lambda_{n-1}] &= \Lambda_{n-1}\, \mathbb{E}_S\left[\mathbb{E}_V[f(Y_n) \mid S]\right] \\ &\leq \Lambda_{n-1}\, \mathbb{E}_S\left[f\left(F_s^{\text{rand}}(S)\right)\right] \\ &= \Lambda_{n-1}.\end{aligned} \tag{33}$$

For $n = 1$, we have

$$\mathbb{E}[\Lambda_1 \mid \Lambda_0] = \mathbb{E}[f(Y_1)] \leq 1 = \Lambda_0. \tag{34}$$

Therefore, we have

$$\mathbb{E}[\Lambda_n \mid \Lambda_0, \Lambda_1, \ldots, \Lambda_{n-1}] \leq \Lambda_{n-1}, \quad \forall n \in [1, k]. \tag{35}$$

Therefore, $\{\Lambda_n, 0 \leq n \leq k\}$ is a test supermartingale according to Definition 1. ■

A.3 Proof of Lemma 9

Proof Since p-values $\{p_i\}_{i\in\mathcal{N}_0}$ are calculated from different images, $\{p_i\}_{i\in\mathcal{N}_0}$ are independent.

Next, we prove that p-values $\{p_i\}_{i\in\mathcal{N}_1}$ are PRDS, we first give the definition of PRDS and a useful proposition.

Definition 12 (Benjamini and Yekutieli, 2001) *Let $X = \{X_1, \ldots, X_n\}$ be a set of variables. X is PRDS on a subset $I_0 \subset \{1, \ldots, n\}$, if for any increasing set D, and for each $i \in I_0$, $\mathbb{P}[X \in D | X_i = x]$ is nondecreasing in x.*

Proposition 13 (Ma et al., 2024) *Let f be a strictly increasing or decreasing function. If variable set $\{X_1, \ldots, X_n\}$ is PRDS on subset I_0, then $\{f(X_1), \ldots, f(X_n)\}$ is PRDS on I_0 as well.*

Given an image I with patches $\{Z_1, Z_2, \ldots, Z_k\}$, let $\{S_1, S_2, \ldots, S_m\}$ be the pixel scores in I where $\{S_i, i \in I_0\}$ is the normal scores. Let $Y_i^j = Y(Z_i, S_j)$. According to the literature (Yu et al., 2023),

$$\{Y_i^1, Y_i^2, \ldots, Y_i^m\} \tag{36}$$

is PRDS on I_0 for all integer $i \in [1, k]$. Then, according to Proposition 13, we have

$$\{f(Y_i^1), f(Y_i^2), \ldots, f(Y_i^m)\} \tag{37}$$

are PRDS on I_0 where f is the betting function in Theorem 5.

Let $A_j^K = \prod_{i=1}^K f(Y_i^j)$. Our goal is to prove that $\{A_1^K, A_2^K, \ldots, A_m^K\}$ are PRDS for $K = k$. We first prove the case $K = 2$. Let $g(A_j^1) = A_j^2$, and $D' = \{g^{-1}(x) : x \in D\}$ for any increasing set D. Given A_a^1, A_b^1 satisfying $A_a^1 \leq A_b^1$ and $A_a^1 \in D'$, we have $A_a^2 \in D$ and $s_a \leq s_b$. Then we have $f(Y_2^a) \leq f(Y_2^b)$. Then we can conclude $A_a^2 \leq A_b^2$. Since D is increasing, then $A_b^2 \in D$ and $A_b^1 \in D'$. Thus, D' is increasing. If $\{A_1^1, A_2^1, \ldots, A_m^1\}$ is PRDS on I_0, it follows that

$$\mathbb{P}[\{A_1^2, A_2^2, \ldots, A_m^2\} \in D | A_j^2 = x] = \mathbb{P}[\{A_1^1, A_2^1, \ldots, A_m^1\} \in D' | A_j^1 = x] \tag{38}$$

is nondecreasing in x. Then $\{A_1^2, A_2^2, \ldots, A_m^2\}$ is PRDS on I_0.

For $K = 3, 4, \ldots, k$, repeat the above proof. Then, we conclude

$$\{A_1^k, A_2^k, \ldots, A_m^k\} \tag{39}$$

are PRDS on I_0. Applying Proposition 13 again, we conclude that

$$\{p_1, p_2, \ldots, p_m\} = \left\{ \frac{1}{A_1^k}, \frac{1}{A_2^k}, \ldots, \frac{1}{A_m^k} \right\} \tag{40}$$

are PRDS on I_0. ■

A.4 Proof of Theorem 10

Proof For the theoretical results at image level, according to Lemma 9, $p_1, p_2, \ldots, p_n$ are independent. We can conclude $FDR_{HiMT} \leq \alpha_{il}$ by applying the theoretical results of classical BH procedure in Theorem 8.

For the theoretical results at pixel level, according to Lemma 9, $p_1^i, p_2^i, \ldots, p_m^i$ are PRDS. We apply the theoretical results of classical BH procedure in Theorem 8 and conclude

$$FDR_{HiMT} = \mathbb{E}[FDP_{HiMT} | p_i \leq \alpha_{il}] \leq \alpha_{pl}. \tag{41}$$

According to Algorithm 2, the HiBH procedure performs pixel-level testing only if its parent image-level hypothesis has been rejected. Thus, we do not reject pixel-level hypothesis when $p_i > \alpha_{il}$. Then we conclude

$$\begin{aligned} FDR_{HiMT} &= \mathbb{E}[FDP_{HiMT}] \\ &= \mathbb{E}[FDP_{HiMT} | p_i \leq \alpha_{il}] \times \mathbb{P}[p_i \leq \alpha_{il}] + \mathbb{E}[FDP_{HiMT} | p_i > \alpha_{il}] \times \mathbb{P}[p_i > \alpha_{il}] \\ &= \mathbb{E}[FDP_{HiMT} | p_i \leq \alpha_{il}] \times \mathbb{P}[p_i \leq \alpha_{il}] \\ &\leq \mathbb{E}[FDP_{HiMT} | p_i \leq \alpha_{il}] \\ &\leq \alpha_{pl}. \end{aligned} \tag{42}$$

■

A.5 Proof of Proposition 11

Proof Let $\mathrm{FDP}_{i,\mathrm{BH}}^P$ denote the pixel-level FDP obtained by applying BH within image I_i. Since HiMT performs pixel-level BH only when $G_i = 1$, the pixel-level FDP of HiMT within image I_i is

$$\mathrm{FDP}_i^P = G_i \mathrm{FDP}_{i,\mathrm{BH}}^P. \tag{43}$$

By the linearity of expectation,

$$\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \mathrm{FDP}_i^P \right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[G_i \mathrm{FDP}_{i,\mathrm{BH}}^P \right]. \tag{44}$$

Using the law of total expectation, we have

$$\mathbb{E}\left[G_i \mathrm{FDP}_{i,\mathrm{BH}}^P \right] = \mathbb{P}(G_i = 1) \mathbb{E}\left[\mathrm{FDP}_{i,\mathrm{BH}}^P \mid G_i = 1 \right]. \tag{45}$$

Since the pixel-level BH step controls the FDR within each selected image at level α_{pl},

$$\mathbb{E}\left[\mathrm{FDP}^{P}_{i,\mathrm{BH}} \mid G_i = 1\right] \leq \alpha_{pl}. \tag{46}$$

Therefore,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{FDP}^{P}_{i}\right] \leq \frac{\alpha_{pl}}{n}\sum_{i=1}^{n}\mathbb{P}(G_i = 1). \tag{47}$$

Because $K = \sum_{i=1}^{n} G_i$, we have

$$\sum_{i=1}^{n}\mathbb{P}(G_i = 1) = \mathbb{E}\left[\sum_{i=1}^{n} G_i\right] = \mathbb{E}[K]. \tag{48}$$

Thus,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{FDP}^{P}_{i}\right] \leq \alpha_{pl}\frac{\mathbb{E}[K]}{n}. \tag{49}$$

Since $K \leq n$, it follows that

$$\alpha_{pl}\frac{\mathbb{E}[K]}{n} \leq \alpha_{pl}. \tag{50}$$

■

References

Muhammad Qasim Ali, Ehab Al-Shaer, Hassan Khan, and Syed Ali Khayam. Automated anomaly detector adaptation using adaptive threshold tuning. *ACM Trans. Inf. Syst. Secur.*, 15(4):1–30, 2013.

Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

Pallavi Basu, T Tony Cai, Kiranmoy Das, and Wenguang Sun. Weighted false discovery rate control in large-scale multiple testing. *Journal of the American Statistical Association*, 113(523):1172–1183, 2018.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 29(4):1165–1188, 2001.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD - A comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, pages 9592–9600, 2019.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *CVPR*, pages 4183–4192, 2020.

Gilles Blanchard and Etienne Roquain. Two simple sufficient conditions for FDR control. *Electronic Journal of Statistics*, 2:963–992, 2008.

AE Brockwell. Universal residuals: A multivariate transformation. *Statistics & Probability Letters*, 77(14):1473–1478, 2007.

Yunkang Cao, Xiaohao Xu, Yuqi Cheng, Chen Sun, Zongwei Du, Liang Gao, and Weiming Shen. Personalizing vision-language models with hybrid prompts for zero-shot anomaly detection. *IEEE Transactions on Cybernetics*, 55(4):1917–1929, 2025.

George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Pacific Grove, CA, 2002.

Agnes Coquio. The optional stopping theorem for quantum martingales. *Journal of Functional Analysis*, 238(1):149–180, 2006.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDim: A patch distribution modeling framework for anomaly detection and localization. In *ICPR*, pages 475–489, 2021.

Claude Dellacherie and Paul André Meyer. *Probabilities and Potential B: Theory of Martingales*. Hermann, 1988.

Luc Demortier. P-values and nuisance parameters. In *PHYSTAT-LHC Workshop*, pages 23–33, 2008.

Sandrine Dudoit and Mark J Van Der Laan. *Multiple Testing Procedures with Applications to Genomics*. Springer, 2008.

Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for medical anomaly detection–a survey. *ACM Computing Surveys (CSUR)*, 54:1–37, 2021.

Lasse Fischer, Timothy Barry, and Aaditya Ramdas. Multiple testing with anytime-valid Monte Carlo p-values. *Electronic Journal of Statistics*, 20(1):1324–1347, 2026.

Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. In *NeurIPS*, pages 585–592, 2007.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. AnomalyGPT: Detecting industrial anomalies using large vision-language models. In *AAAI*, pages 1932–1940, 2024.

Denis A. Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. CFLOW-AD: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *WACV*, pages 98–107, 2022.

Jia Guo, Shuai Lu, Lize Jia, Weihang Zhang, and Huiqi Li. ReContrast: Domain-specific anomaly detection via contrastive reconstruction. In *NeurIPS*, pages 10721–10740, 2023.

Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *CVPR*, pages 20405–20415, 2025.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *ICML*, pages 8759–8773, 2022.

Samet Hicsonmez, Abd El Rahman Shabayek, and Djamila Aouada. VLMDiff: leveraging vision-language models for multi-class anomaly detection with diffusion. In *WACV*, pages 6309–6319, 2026.

Ebenezer RHP Isaac and Akshat Sharma. Adaptive thresholding heuristic for KPI anomaly detection. In *COMSNETS*, pages 737–741, 2024.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. WinCLIP: Zero-/few-shot anomaly classification and segmentation. In *CVPR*, pages 19606–19616, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023.

Sungwook Lee, Seunghyun Lee, and Byung Cheol Song. CFA: Coupled-hypersphere-based feature adaptation for target-oriented anomaly localization. *IEEE Access*, 10:78446–78454, 2022.

Aodong Li, Chen Qiu, Marius Kloft, Padhraic Smyth, Maja Rudolph, and Stephan Mandt. Zero-shot anomaly detection via batch normalization. In *NeurIPS*, pages 40963–40993, 2023.

Wanhua Li, Xiaoke Huang, Zheng Zhu, Yansong Tang, Xiu Li, Jie Zhou, and Jiwen Lu. Ordinalclip: Learning rank prompts for language-guided ordinal regression. In *NeurIPS*, pages 35313–35325, 2022.

Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. PromptAD: Learning prompts with only normal samples for few-shot anomaly detection. In *CVPR*, pages 16838–16848, 2024a.

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 *ICLR*, 2024b.

Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. In *CVPR*, pages 8290–8299, 2018.

Robert Liptser and Albert Nikolaevich Shiryayev. *Theory of Martingales*. Springer Netherlands, 1989.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, pages 38–55, 2024.

Wenrui Liu, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Diversity-measurable anomaly detection. In *CVPR*, pages 12147–12156, 2023.

Xinyue Liu, Jianyuan Wang, Biao Leng, and Shuo Zhang. Unlocking the potential of reverse distillation for anomaly detection. In *AAAI*, pages 5640–5648, 2025.

Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenyong Yu. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *CVPR*, pages 9974–9983, 2025.

Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, and Hanwang Zhang. Unbiased multiple instance learning for weakly supervised video anomaly detection. In *CVPR*, pages 8022–8031, 2023.

Xinsong Ma, Xin Zou, and Weiwei Liu. A provable decision rule for out-of-distribution detection. In *ICML*, pages 34094–34114, 2024.

Xinsong Ma, Jie Wu, and Weiwei Liu. A closer look at generalized BH algorithm for out-of-distribution detection. In *ICML*, pages 42193–42208, 2025a.

Xinsong Ma, Jie Wu, and Weiwei Liu. SAC-BL: A hypothesis testing framework for unsupervised visual anomaly detection and location. *Neural Networks*, 185:107147, 2025b.

Xinsong Ma, Jie Wu, and Weiwei Liu. On the SAC-BL algorithm for anomaly detection. In *NeurIPS*, pages 43632–43650, 2025c.

Xinsong Ma, Jie Wu, Xin Zou, and Weiwei Liu. A unified decision rule for generalized out-of-distribution detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):3544–3555, 2026.

Alan Malek, Sumeet Katariya, Yinlam Chow, and Mohammad Ghavamzadeh. Sequential multiple hypothesis testing with type I error control. In *AISTATS*, pages 1468–1476, 2017.

Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. VT-ADL: A vision transformer network for image anomaly detection and localization. In *ISIE*, pages 01–06, 2021.

Eunju Park, Taekyung Kim, Minju Kim, Hojun Lee, and Gil-Jun Lee. SK-RD4AD: Skip-connected reverse distillation for robust one-class anomaly detection. In *CVPR*, pages 3945–3953, 2025.

Belinda Phipson and Gordon K Smyth. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), 2010.

Aaditya Ramdas, Rina Foygel Barber, Emmanuel J Candès, and Ryan J Tibshirani. Permutation tests using arbitrary permutation distributions. *Sankhya A*, 85(2):1156–1177, 2023a.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023b.

Joseph P Romano, Azeem M Shaikh, and Michael Wolf. Multiple testing. In *The new Palgrave dictionary of economics*, pages 9185–9189. Springer, 2018.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter V Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, pages 14318–14328, 2022.

Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt. Fully convolutional cross-scale-flows for image-based defect detection. In *WACV*, pages 1088–1097, 2022.

Claude Sammut and Geoffrey I Webb. *Encyclopedia of Machine Learning and Data Mining*. Springer, 2017.

Sanat K Sarkar. Some results on false discovery rate in stepwise multiple testing procedures. *The Annals of Statistics*, 30(1):239–257, 2002.

Eli Schwartz, Assaf Arbelle, Leonid Karlinsky, Sivan Harary, Florian Scheidegger, Sivan Doveh, and Raja Giryes. MAEDAY: MAE for few-and zero-shot anomaly-detection. *Computer Vision and Image Understanding*, 241:103958, 2024.

Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1):84–101, 2011.

Juliet Popper Shaffer. Multiple hypothesis testing. *Annual review of psychology*, 46(1): 561–584, 1995.

Sophia Sun, Abishek Sankararaman, and Balakrishnan Murali Narayanaswamy. Online adaptive anomaly thresholding with confidence sequences. In *ICML*, pages 47105–47132, 2024.

Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan Duong, Chanh D. Tr. Nguyen, and Steven Q. H. Truong. Revisiting reverse distillation for anomaly detection. In *CVPR*, pages 24511–24520, 2023.

Jean Ville. *Etude Critique de la Notion de Collectif*, volume 3. Gauthier-Villars Paris, 1939.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *CVPR*, pages 4921–4930, 2022.

Anderson M Winkler, Gerard R Ridgway, Matthew A Webster, Stephen M Smith, and Thomas E Nichols. Permutation inference for the general linear model. *Neuroimage*, 92: 381–397, 2014.

Rongcheng Wu, Hao Zhu, Shiying Zhang, Mingzhe Wang, Zhidong Li, Hui Li, Jianlong Zhou, Jiangtao Cui, Fang Chen, Pingyang Sun, et al. RcAE: Recursive reconstruction framework for unsupervised industrial anomaly detection. In *AAAI*, pages 10736–10744, 2026.

Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *ICCV*, pages 8301–8309, 2021.

Xue Yang, Enda Howley, and Michael Schukat. Agent-based dynamic thresholding for adaptive anomaly detection using reinforcement learning. *Neural Computing and Applications*, 37(23):18775–18791, 2025.

Xincheng Yao, Ruoqi Li, Zefeng Qian, Yan Luo, and Chongyang Zhang. Focus the discrepancy: Intra-and inter-correlation learning for image anomaly detection. In *ICCV*, pages 6803–6813, 2023.

Chenglin Yu, Xinsong Ma, and Weiwei Liu. Delving into noisy label detection with clean data. In *ICML*, pages 40290–40305, 2023.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem - A discriminatively trained reconstruction embedding for surface anomaly detection. In *ICCV*, pages 8330–8339, 2021.

Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. RealNet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *CVPR*, pages 16699–16708, 2024.

Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. Destseg: Segmentation guided denoising student-teacher for anomaly detection. In *CVPR*, pages 3914–3923, 2023.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *ECCV*, pages 392–408, 2022.