# Vague Prototype-Oriented Diffusion Model for Multi-class Anomaly Detection

**Yuxin Li** [1 *] **Yaoxuan Feng** [1 *] **Bo Chen** [1] **Wenchao Chen** [1] **Yubiao Wang** [1] **Xinyue Hu** [1] **Baoling sun** [2]
**Chunhui Qu** [1 3] **Mingyuan Zhou** [4]

## Abstract

Multi-class unsupervised anomaly detection aims to create a unified model for identifying anomalies in objects from multiple classes when only normal data is available. In such a challenging setting, widely used reconstruction-based networks persistently grapple with the "identical shortcut" problem, wherein the infiltration of abnormal information from the condition biases the output towards an anomalous distribution. In response to this critical challenge, we introduce a Vague Prototype-Oriented Diffusion Model (VPDM) that extracts only fundamental information from the condition to prevent the occurrence of the "identical shortcut" problem from the input layer. This model leverages prototypes that contain only vague information about the target as the initial condition. Subsequently, a novel conditional diffusion model is introduced to incrementally enhance details based on vague conditions. Finally, a Vague Prototype-Oriented Optimal Transport (VPOT) method is proposed to provide more accurate information about conditions. All these components are seamlessly integrated into a unified optimization objective. The effectiveness of our approach is demonstrated across diverse datasets, including the MVTec, VisA, and MPDD benchmarks, achieving state-of-the-art results.

[*]Equal contribution [1]National Key Laboratory of Radar Signal Processing, Xidian University, Xi'an, 710071, China. [2]Xidian University, Xi'an, 710071, China. [3]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China [4]McCombs School of Business, The University of Texas at Austin, Austin, TX 78712. Correspondence to: Bo Chen <bchen@mail.xidian.edu.cn>, Wenchao Chen <wchen_xidian@163.com>.

*Proceedings of the $41^{st}$ International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. 

## 1. Introduction

Unsupervised anomaly detection aims to identify and localize anomalies when only normal data is available, garnering significant attention in recent years across diverse application scenarios such as medical image analysis (Fernando et al., 2021), video inspection (Ramachandra et al., 2020), and defect detection (Bergmann et al., 2019). Given the need to detect anomalies across multiple tasks, a common approach involves modeling the distribution of normal samples following the one-for-one scheme (Gong et al., 2019; Bergmann et al., 2020; Li et al., 2023). However, this scheme can be memory-intensive, especially with an increasing number of classes, and may not align well with scenarios characterized by substantial intra-class diversity among normal samples (You et al., 2022a; Lu et al., 2023). Recent advancements in multi-class unsupervised anomaly detection (You et al., 2022a; Lu et al., 2023; He et al., 2023) aim to address these challenges by developing unified models for multiple classes. Modeling the normal distribution remains an inherently challenging task, and the complexity escalates when endeavoring to accurately capture multi-class distributions within a unified model.

A prevalent approach to learning the distribution of normal data involves representation-based methods (Roth et al., 2022; Gudovskiy et al., 2022; Shi et al., 2021; Zaheer et al., 2020; Liang et al., 2023). These methods operate under the assumption that a well-trained model cannot effectively generate samples deviating from the normal distribution. Consequently, when guided by an anomaly sample, the model tends to produce normal samples, leading to noticeable reconstruction errors that can serve as indicators for detecting anomalies (You et al., 2022a; Mousakhan et al., 2023). However, this assumption may not always hold true, as sometimes the infiltration of abnormal information from the condition biases the output towards an anomalous distribution. This phenomenon, where abnormal inputs are well-reconstructed, is known as the "identical shortcut" problem (Gong et al., 2019; You et al., 2022b). Moreover, in multi-class scenarios, the complexity of the normal data distribution is heightened, exacerbating the effects of the "identical shortcut" problem (You et al., 2022a).

Recently, there has been a surge in interest in diffusion-

based generative models for their ability to generate high-dimensional data (Ho et al., 2020; Han et al., 2022). Current diffusion models designed for anomaly detection (He et al., 2023; Mousakhan et al., 2023; Yin et al., 2023) primarily emphasize the creation of influential conditional embeddings derived from abnormal inputs. These embeddings are subsequently fed into the denoising network, steering the reverse process within the diffusion model. For instance, DiAD (He et al., 2023) employs pixel-level semantics as conditions to guide the diffusion model, while LafitE (Yin et al., 2023) employs a condition-guided approach using feature editing, which is determined by input samples to guide the generation processes. However, while these methods aim to use information-rich conditions to generate high-quality results, the conditions still retain some anomaly information, potentially leading to the persistence of the "identical shortcut" problem.

Drawing inspiration from the human memory retrieval mechanism (Ratcliff, 1978), which starts recalling the appearance of a normal sample with vague elements such as shapes and colors, followed by the gradual recall of more detailed information, we propose a novel conditional diffusion model guided by vague conditions. We suggest that initiating the process with vague conditions, initially excluding anomalous information from the input layer, and then incrementally introducing details through a generative model will effectively mitigate the occurrence of the "identical shortcut" problem at its source. In summary, we present the Vague Prototype-Oriented Diffusion Model (VPDM), meticulously designed to counteract the infiltration of abnormal condition information at its origin. To achieve this, we utilize prototypes that contain only vague information about the target as the initial condition. By leveraging prototypes (Tanwisuth et al., 2021; Guo et al., 2022; Wang et al., 2022) with fundamental shape and color information, the model receives adequate guidance for generating corresponding tasks. Simultaneously, the exclusion of anomalous details helps avoid misleading the model. Additionally, we introduce a novel conditional diffusion model to incrementally enhance details based on vague conditions. Finally, we introduce the Vague Prototype-Oriented Optimal Transport (VPOT) model, leveraging Optimal Transport (OT) (Peyré et al., 2019) technology to offer more precise information about conditions. All these components are integrated into a unified optimization objective.

The main contributions of our work are summarized as follows:

- Drawing inspiration from the human memory retrieval mechanism and the intricacies of practical multi-task anomaly detection, we introduce Vague Prototype-Oriented Diffusion Model. This model leverages vague conditions to fundamentally address the challenge of "identical shortcut" problem.
- Given that the vague condition contains less information, it increases the difficulty of generating samples. We introduce a conditional diffusion model that considers the vague condition across both the forward and reverse processes within the diffusion model. This results in a diffusion model that generates samples by gradually adding details based on vague conditions.
- We introduce the VPOT model, which leverages the OT distance between distributions to guide the learning of prototypes, with the goal of summarizing the normal distribution across multiple classes.
- We present comprehensive experimental results and comparisons on MVTec-AD, VisA, and MPDD, showcasing that our method attains SOTA performance across these datasets.

## 2. Background

### 2.1. Multi-class Unsupervised Anomaly Detection

Multi-class unsupervised anomaly detection aims to create a unified model capable of identifying anomalies within objects spanning multiple classes when only normal data is available (You et al., 2022a). It relies on the hypothesis that reconstruction models trained solely on normal samples excel in normal regions but struggle in anomalous regions (Chen et al., 2022; You et al., 2022b; Zavrtanik et al., 2021; Li et al., 2021; Pirnay & Chai, 2022). However, these methods face the challenge of the "identical shortcut" problem, where the model may learn to restore anomalies effectively (Lu et al., 2023). In response, researchers adopt various strategies to address this issue, such as incorporating memory mechanisms (Hou et al., 2021; Yin et al., 2023), instructional information (Shi et al., 2021; Cao et al., 2022), iteration mechanisms (Dehaene et al., 2020), and pseudo-anomalies (Collin & De Vleeschouwer, 2021). Despite their emphasis on enhancing generative capacity through effective condition guidance, these models remain susceptible to anomalous information within the conditions, potentially leading to the "identical shortcut" problem. Unlike these approaches, we present VPDM, which utilizes vague conditions to fundamentally tackle the challenge of the "identical shortcut" problem arising from the intrusion of anomaly information.

### 2.2. Diffusion Model

Diffusion probabilistic models (Sohl-Dickstein et al., 2015) take the form $p_\theta(\boldsymbol{x}^0) := \int p_\theta(\boldsymbol{x}^{0:T}) d\boldsymbol{x}^{1:T}$, where $\boldsymbol{x}^1, ..., \boldsymbol{x}^T$ represent latent variables (Ho et al., 2020). One well-known diffusion model is the denoising diffusion probabilistic model (DDPM) (Ho et al., 2020), which consists

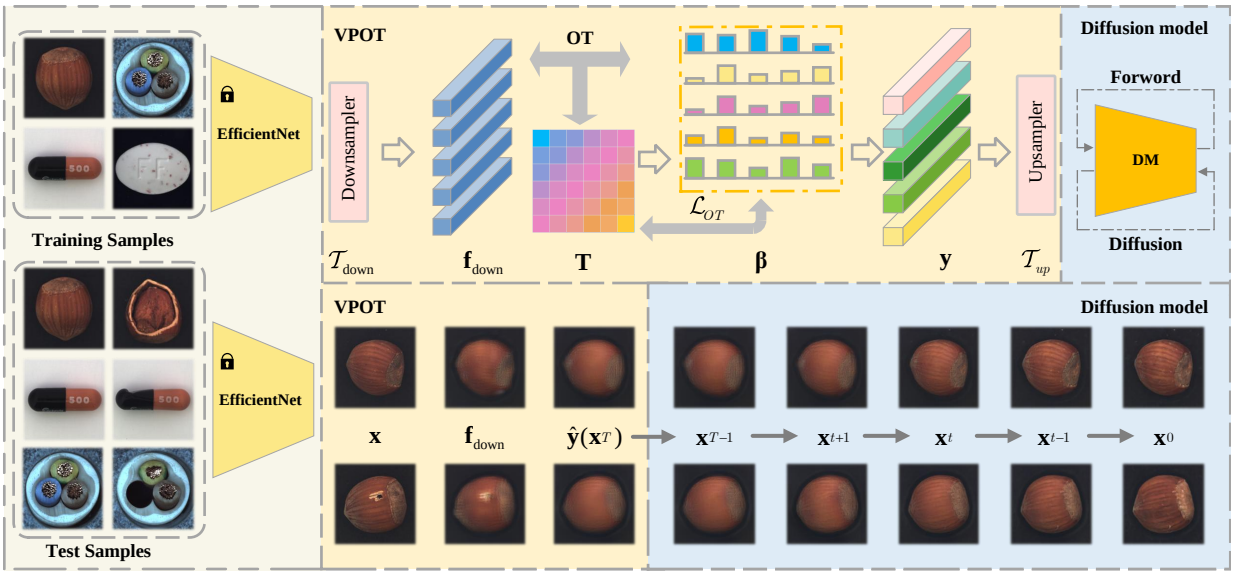


*Figure 1.* The overall framework of proposed Vague Prototype-Oriented Optimal Transport (VPOT), which consist of a pre-trained feature extractor (EfficientNe), a proposed VPOT model and a well designed diffusion model. The upper part provides insight into the workings of VPOT, while the lower half visualizes how VPDM generates normal samples for the corresponding class. The class-based information is initially extracted through low-pass filtering, then mapped to the vague normal pattern in the vague prototype, and finally, details are incrementally added through a specially designed diffusion model.

of two processes: the forward (diffusion) process and the reverse process. Following the Markov chain, the forward process gradually adds noise, transforming an input vector $\boldsymbol{x}^0$ into a Gaussian noise vector $\boldsymbol{x}^T$ over $T$ steps:

$$\begin{aligned} q(\boldsymbol{x}^{1:T} \mid \boldsymbol{x}^0) &:= \textstyle\prod_{t=1}^{T} q(\boldsymbol{x}^t \mid \boldsymbol{x}^{t-1}), \\ q(\boldsymbol{x}^t \mid \boldsymbol{x}^{t-1}) &:= \mathcal{N}(\sqrt{1-\beta^t}\boldsymbol{x}^{t-1}, \beta^t \boldsymbol{I}) \end{aligned} \tag{1}$$

where $\beta^t$ represents a small positive constant denoting the noise level. In practical applications, we directly sample $\boldsymbol{x}^t$ from $\boldsymbol{x}^0$ as the following: $q(\boldsymbol{x}^t \mid \boldsymbol{x}^0) = \mathcal{N}(\sqrt{\alpha^t}\boldsymbol{x}^0, (1-\alpha^t)\boldsymbol{I})$, where $\bar{\alpha}^t := 1-\beta^t$ and $\alpha^t := \prod_{t=1}^{T} \bar{\alpha}^t$. The reverse process involves denoising $\boldsymbol{x}^t$ back to $\boldsymbol{x}^0$ and is defined as a Markov chain with a learned Gaussian transition:

$$\begin{aligned} p_\theta(\boldsymbol{x}^{0:T}) &:= p(\boldsymbol{x}^T) \textstyle\prod_{t=1}^{T} p_\theta(\boldsymbol{x}^{t-1} \mid \boldsymbol{x}^t), \\ p_\theta(\boldsymbol{x}^{t-1} \mid \boldsymbol{x}^t) &:= \mathcal{N}(\boldsymbol{\mu}_\theta(\boldsymbol{x}^t, t), \boldsymbol{\sigma}_\theta(\boldsymbol{x}^t, t)) \end{aligned} \tag{2}$$

In DDPM (Ho et al., 2020), the parameterization of $p_\theta(\boldsymbol{x}^{t-1} \mid \boldsymbol{x}^t)$ is defined as:

$$\begin{aligned} \boldsymbol{\mu}_\theta(\boldsymbol{x}^t, t) &= \tfrac{1}{\alpha^t}(\boldsymbol{x}^t - \tfrac{\beta^t}{\sqrt{1-\alpha^t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}^t, t)), \\ \boldsymbol{\sigma}_\theta(\boldsymbol{x}^t, t) &= (\bar{\beta}^t)^{1/2}, \\ \text{if } t = 1: \ \bar{\beta}^t = \beta^1, \quad &\text{else}: \ \bar{\beta}^t = \tfrac{1-\alpha^{t-1}}{1-\alpha^t}\beta^t \end{aligned} \tag{3}$$

where the $\boldsymbol{\epsilon}_\theta$ is denoising function and which can be trained by solving the following optimization problem:

$$\min_\theta \mathcal{L}(\theta) := \mathbf{E}_{\boldsymbol{x}^0 \sim q(\boldsymbol{x}^0), \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), t} \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}^t, t) \right\|^2 \tag{4}$$

Using the trained denoising function $\boldsymbol{\epsilon}_\theta$, we can generate samples step by step from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ randomly. However, in the context of multi-class unsupervised anomaly detection, the objective is to generate the normal samples $\boldsymbol{x}$ conditioned on the no the normal or abnormal images. Several studies (He et al., 2023; Mousakhan et al., 2023; Yin et al., 2023) have explored adapting diffusion models for this task by injecting conditional information into the reverse process to guide the generative process.

### 2.3. Optimal Transport

OT is a widely used tool for quantifying the difference between two distributions (Peyré et al., 2019). Specifically, considering two discrete distributions as $p = \sum_{i=1}^{n} a_i \delta_{x_i}$ and $q = \sum_{j=1}^{m} b_j \delta_{y_j}$, where $x_i, y_j \in \mathbb{R}^d$ and $\delta_x$ is the Dirac function that places a unit point mass at $x$. The OT distance between $p$ and $q$ can be expressed as:

$$\mathrm{OT}(p, q) = \min_{\mathbf{T} \in \Pi(p,q)} \langle \mathbf{T}, \mathbf{C} \rangle \tag{5}$$

where $\langle \cdot, \cdot \rangle$ denotes the Frobenius dot-product, $\mathbf{C} \in \mathbb{R}_{\geq} 0^{n \times m}$ is the transport cost matrix. $\mathbf{T} \in \mathbb{R}_{>0}^{n \times m}$ refers to the doubly stochastic transport probability matrix that $\Pi(p, q) := \{\mathbf{T} \mid \sum_{i=1}^{n} T_{ij} = b_j, \sum_{j=1}^{m} T_{ij} = a_i\}$, which can be learned by minimizing $\mathrm{OT}(p, q)$. As the optimization of Eq. 5 often demands a high computational cost, the Sinkhorn algorithm for discrete OT, which is achieved by introducing the entropic regularization $H = -\sum_{ij} T_{ij} \ln T_{ij}$, is commonly used in practice to reduce the computation (Peyré et al., 2019).

## 3. Methodology

Taking inspiration from the human memory retrieval mechanism (Ratcliff, 1978), where recalling the appearance of a normal sample starts with vague elements such as shapes and colors, followed by the progressive recall of more detailed information, in this section, we introduce VPDM, a novel framework that employs prototypes containing only vague information about the target as the initial condition. By excluding anomalous information from the input layer, we alleviate concerns about anomaly information infiltration. As depicted in Fig. 1, VPDM consist of two main components, we first present the VPOT model. VPOT leverages OT (Peyré et al., 2019) technology to provide more precise information about conditions. In contrast to existing models (Yin et al., 2023; He et al., 2023) that use information-rich conditions to guide the model, the vague prototype eliminates anomaly information at the input level. However, this also leads to the loss of some normal information, making the generation process more challenging. To overcome this, the second part of VPDM is a novel conditional diffusion model that incrementally enhances details based on vague conditions. Finally, these two models are integrated into a unified framework, leveraging a hybrid optimization approach. From a conceptual standpoint, VPDM can be viewed as a Bayesian generative model (Tran et al., 2019), where the generative process can be expressed as:

$$p(\boldsymbol{x}^0) = \int_{\boldsymbol{x}^{1:T}} p(\boldsymbol{x}^T \mid \hat{\boldsymbol{y}}) \prod_{t=1}^{T} p(\boldsymbol{x}^{t-1} \mid \boldsymbol{x}^t, \hat{\boldsymbol{y}}) d\boldsymbol{x}^{1:T} \tag{6}$$

Where the $\boldsymbol{x}^{1:T}$ are latents of the same dimensionality as the data $x^0 \sim p(\boldsymbol{x}^0)$, $\hat{\boldsymbol{y}}$ is the vague condition generated by VPOT model.

### 3.1. Learning Vague Prototype with Optimal Transport

Existing works (Hou et al., 2021; Yin et al., 2023; Shi et al., 2021) predominantly focus on designing information-rich conditions to enhance the model's generative capability. In contrast, VOPT is proposed to characterizing the normal distribution using a set of vague prototypes containing only fundamental task information. This adjustment provides several advantages. Firstly, the occurrence of the "identical shortcut" is primarily caused by the infiltration of anomaly information from input samples, enabling the model to proficiently generate anomaly samples. The vague prototypes we propose contain minimal information, primarily vague elements like shapes and colors. By excluding anomalous information from the input layer, we alleviate concerns about anomaly information infiltration. Secondly, the prototypes encapsulate various normal dynamic patterns for different tasks (Wang et al., 2022; Li et al., 2023), allowing VPDM to cover multi-class with diverse characteristics through this group of prototypes. This enables the model to better capture the different patterns inherent in multi-class scenarios.

Finally, we introduce the OT algorithm to learn the vague prototypes and enhance the quality of the prototypes by optimizing the OT loss. This approach improves the model's ability to capture various patterns, beneficial for multi-class settings (Wang et al., 2022; Tanwisuth et al., 2021).

In the VPOT model, the initial step involves extracting fundamental information from the input image while filtering out undesired anomalies and intricate details. Drawing inspiration from recent works (Choi et al., 2021; Wang et al., 2023) that employ low-pass filtering for information segmentation in feature space, the retained low-frequency information primarily encompasses fundamental task elements, such as shape and color, offering a vague representation of the current sample. The discarded high-frequency information includes intricate details like edges and textures, contributing minimally to guiding the model in generating the corresponding task but potentially introducing significant anomalous information, leading to the generation of anomalous samples. The low-frequency segment from the input image is then extracted to learn vague prototypes. Leveraging a pre-trained EfficientNet ($\phi$) (Tan & Le, 2021) for visual token extraction, we obtain the feature $\boldsymbol{f} = \phi(\boldsymbol{x}) \in \mathbb{R}^{w \times h \times c}$, where $w$ and $h$ represent the width and height of the features extracted by $\phi$, and $c$ is the number of channels. To capture fundamental task information in $\boldsymbol{f}$, we employ a downsampler $\mathscr{T}_{\text{down}}(\cdot)$, resulting in $\boldsymbol{f}_{\text{down}} = \mathscr{T}_{\text{down}}(\boldsymbol{f}) \in \mathbb{R}^{(w/N) \times (h/N) \times c}$, where $N$ represents the number of downsamples. This operation is analogous to low-pass filtering (Choi et al., 2021).

Considering a set of randomly initialized vague prototypes $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_K] \in \mathbb{R}^{K \times d}$, where $K$ is the number of vague prototypes and $d = w \times h \times c/N^2$ represents the feature dimension. To model the normal distribution over training data, we can represent $K_f$ features as an empirical distribution over $K_f$ :

$$P_f = \sum_{i=1}^{K_f} \frac{1}{K_f} \delta_{\boldsymbol{f}^i_{\text{down}}}, \boldsymbol{f}^i_{\text{down}} \in \mathbb{R}^d \tag{7}$$

The vague prototypes represent distinct fundamental information from multiple classes by combining them with each other. Each prototype holds a similar level of importance and captures a foundational element. Consequently, the distribution over vague prototypes can be defined as an empirical distribution:

$$P_{\boldsymbol{\beta}} = \sum_{i=1}^{K} \frac{1}{K} \delta_{\boldsymbol{\beta^i}}, \ \boldsymbol{\beta^i} \in \mathbb{R}^d \tag{8}$$

where $\boldsymbol{\beta}$ is vague prototypes. In this way, we can get the transport probability matrix $\boldsymbol{T} \in \mathbb{R}^{K_f \times K}$ by pushing $P_{\boldsymbol{\beta}}$ to $P_f$:

$$\boldsymbol{T} = \mathbf{OT}(P_f, P_{\boldsymbol{\beta}}) = \min_{\boldsymbol{T}} \langle \boldsymbol{T}, \boldsymbol{C} \rangle \stackrel{\text{def.}}{=} \sum_{i}^{K_f} \sum_{j}^{K} T_{ij} C_{ij} \tag{9}$$

$\mathbf{C} \in \mathbb{R}_{\geq 0}^{K_f \times K}$ is the transport cost matrix, where we use the Euclidean distance between the embedding $\boldsymbol{f}_{\text{down}}$ and the prototype $\boldsymbol{\beta}$, denoted as $C_{ij} = \sqrt{(f_{\text{down}}^i - \beta^j)^2}$. The transport probability matrix $\boldsymbol{T}$ should satisfy $\Pi(\boldsymbol{g}, \boldsymbol{h}) := \left\{ \boldsymbol{T} \mid \boldsymbol{T}\mathbf{1}_{K_g} = \boldsymbol{g}, \boldsymbol{T}^\top \mathbf{1}_{N_j} = \boldsymbol{h} \right\}$, where $\boldsymbol{g} = [1/K]$ and $\boldsymbol{h} = [1/K_f]$ are two probability vectors defined in Eq. 7 and Eq. 8. OT provides an optimal transport plan from the embedding $P_f$ to the prototype $P_{\boldsymbol{\beta}}$ based on the cost matrix $\mathbf{C}$, allowing us to construct a vague condition $\boldsymbol{y}$ using the transport probability $\boldsymbol{T}$ and prototypes $\boldsymbol{\beta}$:

$$\boldsymbol{y} = \boldsymbol{T} \times \boldsymbol{\beta}, \boldsymbol{y} \in \mathbb{R}^{K_j \times d} \tag{10}$$

In this manner, the vague condition $\boldsymbol{y}$ has undergone refinement to minimize the influence of anomalous information. Simultaneously, the prototypes encapsulate diverse normal dynamic patterns for different tasks, enhancing the ability of $\boldsymbol{y}$ to effectively capture the distinct patterns inherent in multi-class scenarios. Drawing inspiration from existing OT-based prototype-oriented methods (Guo et al., 2022; Tanwisuth et al., 2021), we employ the entropic constraint (Cuturi, 2013) to learn the prototypes $\boldsymbol{\beta}$. The average OT loss for all training sets is defined as:

$$\begin{aligned} \mathcal{L}_{OT} &= \min_{\boldsymbol{\beta}} \mathbb{E}_{\boldsymbol{f}_{\text{down}} \sim \mathscr{T}_{\text{down}}(\phi(D_x))} \Big[ \sum_{i}^{K_f} \sum_{j}^{K} T_{ij} C_{ij} \\ &\quad + \sum_{i}^{K_f} \sum_{j}^{K} T_{ij} ln(T_{ij}) \Big] \\ &= \min_{\boldsymbol{\beta}} \mathbb{E}_{\boldsymbol{f}_{\text{down}} \sim \mathscr{T}_{\text{down}}(\phi(D_x))} \big[ \mathbf{OT}(P_f, P_{\boldsymbol{\beta}}) \big] \end{aligned} \tag{11}$$

$\mathscr{T}_{\text{down}}(\cdot)$ is the downsampler. $\phi(\cdot)$ represents the pre-trained EfficientNet, and $D_x$ is the training set consisting of normal samples. Finally, an upsampler $\mathscr{T}_{\text{up}}(\cdot)$ has been employed to process $\boldsymbol{y}$ and generate $\hat{\boldsymbol{y}}$ to guide the diffusion model, where $\hat{\boldsymbol{y}} = \mathscr{T}_{\text{up}}(\boldsymbol{y}) \in \mathbb{R}^{w \times h \times c}$.

### 3.2. Vague Prototype-Oriented Diffusion Model

Diverging from existing conditional diffusion models that use information-rich conditions to alleviate the difficulty of the generation process, VPDM employs a vague condition to guide the model, imposing higher demands on the generative model. To address this challenge, we modify the endpoint of the forward process in the feature space mapped by EfficientNet ($\phi$), denoted as $\boldsymbol{x}^T$, which typically follows a standard normal distribution $\mathcal{N}(0, 1)$. We introduce the vague condition $\hat{\boldsymbol{y}}$ into the endpoint $p(\boldsymbol{x}^T)$ as follows:

$$p(\boldsymbol{x}^T \mid \hat{\boldsymbol{y}}) = \mathcal{N}(\hat{\boldsymbol{y}}, \boldsymbol{I}) \tag{12}$$

With this configuration, the generation process commences at $\mathcal{N}(\hat{\boldsymbol{y}}, \boldsymbol{I})$. Subsequently, based on the fundamental task information in the vague condition $\hat{\boldsymbol{y}}$, details are gradually added, thereby generating normal samples. To better utilize the guidance of $\hat{\boldsymbol{y}}$, we also incorporate it into the forward process in VPDM. With the diffusion schedule ${\beta^t}_{t=1:T} \in (0, 1)$, the conditional distributions for the forward process at all other time steps can be defined as:

$$\begin{aligned} q\left(\boldsymbol{x}^t \mid \boldsymbol{x}^{t-1}, \hat{\boldsymbol{y}}\right) &\sim \mathcal{N}(\boldsymbol{x}^t \mid \mu_1, \beta^t \boldsymbol{I}) \\ \mu_1 &= \sqrt{1-\beta^t}\boldsymbol{x}^{t-1} + (1 - \sqrt{1-\beta^t})\hat{\boldsymbol{y}} \end{aligned} \tag{13}$$

Inspired by the DDPM (Ho et al., 2020), we sample $\boldsymbol{x}^t$ directly from $\boldsymbol{x}^0$ with an arbitrary timestep $t$:

$$\begin{aligned} q\left(\boldsymbol{x}^t \mid \boldsymbol{x}^0, \hat{\boldsymbol{y}}\right) &\sim \mathcal{N}(\boldsymbol{x}^t \mid \mu_2, (1 - \sqrt{\alpha^t})\boldsymbol{I}) \\ \mu_2 &= \sqrt{\alpha^t}\boldsymbol{x}^0 + (1 - \sqrt{\alpha^t})\hat{\boldsymbol{y}} \end{aligned} \tag{14}$$

where $\bar{\alpha}^t := 1 - \beta^t$ and $\alpha^t := \prod_{t=1}^{T} \bar{\alpha}^t$. In Eq. 13, the mean term $\mu_1$ in the forward process can be conceptualized as an interpolation between the true data $\boldsymbol{x}^0$ and the conditional representation $\hat{\boldsymbol{y}}$. This process is the reverse of the human recall mechanism, gradually reducing details from a complete image to a vague basic concept. As the VPOT model eliminates anomalous information in $\hat{\boldsymbol{y}}$ and the diffusion model solely focuses on adding details, VPDM minimizes the infiltration of anomalous information. Meanwhile, the diffusion model only needs to focus on adding details from $\hat{\boldsymbol{y}}$ to $\boldsymbol{x}^0$ rather than generating a sample from noise, reducing the difficulty of generation while enhancing the quality of the generated samples. Considering the forward process in Eq. 13, the corresponding manageable posterior for the forward process is:

$$\begin{aligned} q\left(\boldsymbol{x}^{t-1} \mid \boldsymbol{x}^0, \boldsymbol{x}^t, \hat{\boldsymbol{y}}\right) &\sim \mathcal{N}\left(\boldsymbol{x}^{t-1} \mid \gamma_0 \boldsymbol{x}^0 + \gamma_1 \boldsymbol{x}^t + \gamma_2 \hat{\boldsymbol{y}}, \tilde{\beta}^t \boldsymbol{I}\right) \\ \gamma_0 = \tfrac{\beta^t \sqrt{\alpha^{t-1}}}{1-\alpha^t}, \gamma_1 &= \tfrac{(1-\alpha^{t-1})\sqrt{\bar{\alpha}^t}}{1-\alpha^t}, \\ \gamma_2 = 1 + \tfrac{(\sqrt{\alpha^t}-1)(\sqrt{\bar{\alpha}^t}+\sqrt{\alpha^{t-1}})}{1-\alpha^t}, \tilde{\beta}^t &= \tfrac{(1-\alpha^{t-1})}{1-\alpha^t}\beta^t \end{aligned} \tag{15}$$

The derivation can be found in Appendix A.

### 3.3. Model Training

In this paper, we integrate the VPOT model and denoising model into a unified optimization objective. The pre-trained EfficientNet ($\phi$) does not participate in training, so our focus is solely on the vague prototypes $\boldsymbol{\beta}$ and the denoising network $\boldsymbol{\varepsilon}_\theta(\cdot)$. As shown in Eq. 6, the optimization objective of the diffusion model part is to maximize the evidence lower bound (ELBO) of the log marginal likelihood, formulated as:

$$\begin{aligned} &\log p\left(\boldsymbol{x}^0 \mid \hat{\boldsymbol{y}}\right) \geq \log \mathbb{E}_{q(\boldsymbol{x}^{1:T}, \mid \boldsymbol{x}^0, \hat{\boldsymbol{y}})} \left[ \tfrac{p(\boldsymbol{x}^{0:T} \mid \hat{\boldsymbol{y}})}{q(\boldsymbol{x}^{1:T}, \mid \boldsymbol{x}^0, \hat{\boldsymbol{y}})} \right] \\ &= \mathbb{E}_q[-\log p(\boldsymbol{x}^0 \mid \boldsymbol{x}^1, \hat{\boldsymbol{y}})] + \mathbf{D}_{KL}(q(\boldsymbol{x}^T \mid \boldsymbol{x}^0, \hat{\boldsymbol{y}}) \| p(\boldsymbol{x}^T \mid \hat{\boldsymbol{y}})) \\ &+ \textstyle\sum_{t=2}^{T} \mathbf{D}_{KL}(q\left(\boldsymbol{x}^{t-1} \mid \boldsymbol{x}^0, \boldsymbol{x}^t, \hat{\boldsymbol{y}}\right) \| p(\boldsymbol{x}^{t-1} \mid \boldsymbol{x}^t, \hat{\boldsymbol{y}})) \end{aligned} \tag{16}$$

*Table 1.* Anomaly detection/localization results with AUROC metric on MVTec-AD. All methods are evaluated under the multi-class settings. The learned model is applied to detect anomalies for all categories without fine-tuning. The best results are bold with black.

| | Category | US | PSVDD | PaDiM | MKD | DRAEM | RD4AD | UniAD | DiAD | HVQ-Trans | Ours |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Object | Bottle | 84.0 / 67.9 | 85.5 / 86.7 | 97.9 / 96.1 | 98.7 / 91.8 | 97.5 / 87.6 | 98.7 / 97.7 | 99.7 / 98.1 | 99.7 / 98.4 | **100** / 98.3 | **100** ±0.00 / **98.6** ±0.01 |
| | Cable | 60.0 / 78.3 | 64.4 / 62.2 | 70.9 / 81.0 | 78.2 / 89.3 | 57.8 / 71.3 | 85.0 / 83.1 | 95.2 / 97.3 | 94.8 / 96.8 | **99.0** / **98.1** | 97.8±0.19 / **98.1**±0.05 |
| | Capsule | 57.6 / 85.5 | 61.3 / 83.1 | 73.4 / 96.9 | 68.3 / 88.3 | 65.3 / 50.5 | 95.5 / 98.5 | 86.9 / 98.5 | 89.0 / 97.1 | 95.4 / **98.8** | **97.0**±0.21 / **98.8**±0.02 |
| | Hazelnut | 95.8 / 93.7 | 83.9 /97.4 | 85.5 / 96.3 | 97.1 / 91.2 | 93.7 / 96.9 | 87.1 / 98.7 | 99.8 / 98.1 | 99.5 / 98.3 | **100** / **98.8** | 99.9±0.01 / 98.7±0.04 |
| | Metal Nut | 62.7 / 76.6 | 80.9 / 96.0 | 88.0 / 84.8 | 64.9 / 64.2 | 72.8 / 62.2 | 99.4 / 94.1 | 99.2 / 94.8 | 99.1 / **97.3** | **99.9** / 96.3 | 98.9±0.03 / 96.0±0.01 |
| | Pill | 56.1 / 80.3 | 89.4 / 96.5 | 68.8 / 87.7 | 79.7 / 69.7 | 82.2 / 94.4 | 52.6 / 96.5 | 93.7 / 95.0 | 95.7 / 95.7 | 95.8 / **97.1** | **97.9**±0.23 / 96.4±0.06 |
| | Screw | 66.9 / 90.8 | 80.9 / 74.3 | 56.9 / 94.1 | 75.6 / 92.1 | 92.0 / 95.5 | **97.3** / **99.4** | 87.5 / 98.3 | 90.7 / 97.9 | 95.6 / 98.9 | 95.5±0.26 / 99.3±0.01 |
| | Toothbrush | 57.8 / 86.9 | 99.4 / 98.0 | 95.3 / 95.6 | 75.3 / 88.9 | 90.6 / 97.7 | 99.4 / **99.0** | 94.2 / 98.4 | **99.7** / **99.0** | 93.6 / 98.6 | 94.6±0.22 / 98.8±0.02 |
| | Transistor | 61.0 / 68.3 | 77.5 / 78.5 | 86.6 / 92.3 | 73.4 / 71.7 | 74.8 / 64.5 | 92.4 / 86.4 | **99.8** / 97.9 | **99.8** / 95.1 | 99.7 / 97.9 | 99.7±0.02 / **99.1**±0.01 |
| | Zipper | 78.6 / 84.2 | 77.8 / 95.1 | 79.7 / 94.8 | 87.4 / 86.1 | 98.8 / **98.3** | **99.6** / 98.1 | 95.8 / 96.8 | 95.1 / 96.2 | 97.9 / 97.5 | 99.0±0.06 / 98.0±0.09 |
| Texture | Carpet | 86.6 / 88.7 | 63.3 / 78.6 | 93.8 / 97.6 | 69.8 / 95.5 | 98.0 / 98.6 | 97.1 / **98.8** | 99.8 / 98.5 | 99.4 / 98.6 | 99.9 / 98.7 | **100**±0.00 / **98.8**±0.03 |
| | Grid | 69.2 / 64.5 | 66.0 / 70.8 | 73.9 / 71.0 | 83.8 / 82.3 | 99.3 / 98.7 | **99.7** / **99.2** | 98.2 / 96.5 | 98.5 / 96.6 | 97.0 / 97.0 | 98.6±0.07 / 98.0±0.01 |
| | Leather | 97.2 / 95.4 | 60.8 / 93.5 | 99.9 / 84.8 | 93.6 / 96.7 | 98.7 / 97.3 | **100** / **99.4** | **100** / 98.8 | 99.8 / 98.8 | **100** / 98.8 | **100**±0.00 / 99.2±0.01 |
| | Tile | 93.7 / 82.7 | 88.3 / 92.1 | 93.3 / 80.5 | 89.5 / 85.3 | 99.8 / **98.0** | 97.5 / 95.6 | 99.3 / 91.8 | 96.8 / 92.4 | 99.2 / 92.2 | **100**±0.00 / 94.5±0.12 |
| | Wood | 90.6 / 83.3 | 72.1 / 80.7 | 98.4 / 89.1 | 93.4 / 80.5 | **99.8** / **96.0** | 99.2 / **96.0** | 98.6 / 93.2 | 99.7 / 93.3 | 97.2 / 92.4 | 98.2±0.11 / 95.3±0.07 |
| | Mean | 74.5 / 81.8 | 76.8 / 85.6 | 84.2 / 89.5 | 81.9 / 84.9 | 88.1 / 87.2 | 93.4 / 96.0 | 96.5 / 96.8 | 97.2 / 96.8 | 98.0 / 97.3 | **98.4**±0.04 / **97.8**±0.02 |

**Algorithm 1** Training

1: Initialize the parameters;
2: **repeat**
3: Draw $\boldsymbol{x}^0 \sim q(\boldsymbol{x}^0)$
4: Draw $t \sim \text{Uniform}(\{1, 2, ..., T\})$
5: Draw $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$
6: Draw $\hat{\boldsymbol{y}} = \text{VPOT}(\boldsymbol{x}^0)$
7: Compute the loss in Eq. 17
8: Take numerical optimization step on: $\nabla\mathcal{L}$
9: **until** converged

**Algorithm 2** Inference

1: Draw $\hat{\boldsymbol{y}} = \text{VPOT}(\boldsymbol{x}^0)$
2: $\boldsymbol{x}^T \sim \mathcal{N}(\hat{\boldsymbol{y}}, \boldsymbol{I})$
3: **for** t = $T$ to 1 **do**
4: Calculate reparameterize: $\boldsymbol{X}^t = (1/\alpha^t)(\boldsymbol{x}^t - (1 - \sqrt{\alpha^t})\hat{\boldsymbol{y}} - \sqrt{1-\alpha^t}\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}^t, \hat{\boldsymbol{y}}, t))$
5: **if** $t > 1$: draw $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$
6: $\boldsymbol{x}^{t-1} = \gamma_0 \boldsymbol{X}^t + \gamma_1 \boldsymbol{x}^t + \gamma_2 \hat{\boldsymbol{y}} + \sqrt{\tilde{\boldsymbol{\beta}}^t}\boldsymbol{\varepsilon}$
7: **else**: $\boldsymbol{x}^{t-1} = \boldsymbol{X}^t$
8: **end for**

where $\mathbf{D}_{KL}(q\|p)$ represents the Kullback–Leibler (KL) divergence from distribution $p$ to distribution $q$. For the VPOT model, the OT loss presented in Eq. 11 plays a pivotal role in our comprehensive loss formulation. This loss function is instrumental in guiding the learning of prototypes, ensuring that the vague conditions $\hat{\boldsymbol{y}}$ are effectively aligned with the normal data distribution. The final loss function for VPDM can be expressed as follows:

$$\mathcal{L} = \mathcal{L}_{ELBO} + \mathcal{L}_{OT} \tag{17}$$

Here, $\mathcal{L}_{ELBO}$ denotes the ELBO as defined in Eq. 16. Leveraging Eq. 17, the vague prototypes are directed by $\mathcal{L}_{ELBO}$ to learn how to contribute to generating the final samples. Additionally, they are guided by $\mathcal{L}_{OT}$, providing a principled and unsupervised approach to encourage the vague prototypes to capture the diverse normal patterns within multi-class. We regard this as a unique advantage of hybrid optimization. Finally, the training and inference processes are outlined in Algorithms 1 and 2.

### 3.4. Anomaly Localization and Detection

The result of anomaly localization is an anomaly score map, denoted as $\boldsymbol{S}$, calculated as the L2 norm of the reconstruction differences, expressed as $\boldsymbol{S} = \|\boldsymbol{x}^0 - \boldsymbol{x}^0_{rec}\|_2 \in \mathbb{R}^{w \times h}$. Since the VPOT model and the proposed diffusion model operate on the feature space mapped by EfficientNet ($\phi$), $\boldsymbol{S}$ is up-sampled to the image size using bi-linear interpolation to obtain the localization results. Anomaly detection aims to identify whether an image contains anomalous regions. We transform the anomaly score map, denoted as $S$, to the anomaly score of the image by taking the maximum value of the average-pooled $S$.

## 4. Experimental Evaluation

### 4.1. Experiments Setup

**Dataset:** Three datasets are utilized in our paper: (1) MVTec-AD dataset (Bergmann et al., 2019) serves as a simulation of real-world industrial production scenarios, specifically designed for unsupervised anomaly detection. (2) VisA dataset (Zou et al., 2022) is a recently published large dataset, which consists of 9,621 normal and 1,200 anomalous high-resolution images. (3) MPDD (Jezek et al., 2021) contains 6 classes of metal parts, focusing on defect detection during the fabrication of painted metal parts. More details about the datasets can be find in Appendix B.

**Evaluation metrics:**We report the Area Under the Receiver Operator Curve (AUROC) on imagelevel anomaly detection

*Table 2.* Anomaly detection/localization results with AUROC metric on VisA. All methods are evaluated under the multi-class settings. The learned model is applied to detect anomalies for all categories without fine-tuning. The best results are bold with black.

| Category | | DRAEM | JNLD | OmniAL | UniAD | DiAD | HVQ-Trans | Ours |
|---|---|---|---|---|---|---|---|---|
| Complex structure | PCB1 | 83.9 / 94.0 | 82.9 / 98.0 | 77.7 / 97.6 | 95.4 / 99.3 | 88.1 / 98.7 | 96.7 / 99.4 | **98.2**±0.02 / **99.6**±0.03 |
| | PCB2 | 81.7 / 94.1 | 79.1 / 95.0 | 81.0 / 93.9 | 93.6 / 97.8 | 91.4 / 95.2 | 93.4 / 98.0 | **97.5**±0.03 / **98.8**±0.01 |
| | PCB3 | 87.7 / 94.1 | 90.1 / 98.5 | 88.1 / 94.7 | 88.6 / 98.3 | 86.2 / 96.7 | 92.0 / 98.3 | **94.5**±0.08 / **98.7**±0.01 |
| | PCB4 | 87.1 / 72.3 | 96.2 / 97.5 | 95.3 / 97.1 | 99.4 / **97.9** | 99.6 / 97.0 | 99.5 / 97.7 | **99.9**±0.01 / 97.8±0.06 |
| Multiple instances | Macaroni 1 | 68.6 / 89.8 | 90.5 / 93.3 | 92.6 / 98.6 | 92.2 / 99.3 | 85.7 / 94.1 | 93.1 / 99.4 | **97.5**±0.02 / **99.6**±0.01 |
| | Macaroni 2 | 60.3 / 83.2 | 71.3 / 92.1 | 75.2 / 97.9 | 85.9 / 98.0 | 62.5 / 93.6 | **86.2** / 98.5 | 85.7±0.12 / **99.0**±0.03 |
| | Capsules | 89.6 / 96.6 | **91.4 / 99.6** | 90.6 / 99.4 | 72.0 / 98.3 | 58.2 / 97.3 | 77.1 / 99.0 | 79.5±0.31 / 99.1±0.01 |
| | Candles | 70.2 / 82.6 | 85.4 / 94.5 | 86.8 / 95.8 | 96.8 / 99.2 | 92.8 / 97.3 | 96.8 / 99.2 | **97.2**±0.07 / **99.4**±0.01 |
| Single instance | Cashew | 67.3 / 68.5 | 82.5 / 94.1 | 88.6 / 95.0 | 92.4 / 98.7 | 91.5 / 90.9 | **94.9 / 99.2** | 90.0±0.13 / 98.0±0.02 |
| | Chewing gum | 90.0 / 92.7 | 96.0 / 98.9 | 96.4 / 99.0 | **99.4 / 99.2** | 99.1 / 94.7 | **99.4** / 98.8 | 99.0±0.01 / 98.6±0.02 |
| | Fryum | 86.2 / 83.2 | 91.9 / 90.0 | **94.6** / 92.1 | 89.8 / 97.7 | 89.8 / 97.6 | 90.4 / 97.7 | 92.0±0.03 / **98.6**±0.04 |
| | Pipe fryum | 87.1 / 72.3 | 87.5 / 92.5 | 86.1 / 98.2 | 97.4 / 99.2 | 96.2 / **99.4** | 98.5 / **99.4** | **98.8**±0.01 / **99.4**±0.01 |
| Mean | | 80.5 / 87.0 | 87.1 / 95.2 | 87.8 / 96.6 | 91.9 / 98.6 | 86.8 / 96.0 | 93.2 / 98.7 | **94.2**±0.09 / **98.9**±0.02 |

and pixel-wise anomaly localization following the previous works (He et al., 2023; You et al., 2022a; Lu et al., 2023).

**Implementation details:** In the VPOT model, the number of prototypes is set to 50, and the downsampler sampling multiplier is 4. For the diffusion model, the number of timesteps is configured as $T = 1000$, and a linear noise schedule is employed with $\beta^1 = 10^{-4}$ and $\beta^T = 0.02$, consistent with the setup in Ho et al. (2020). More details about the implementation can be find in Appendix D.

**Baselines:** We extensively compare our model with 14 baselines with different experiment settings. Such as a unified SOTA HVQ-Trans (Lu et al., 2023) method and the diffusion-based methods DiAD(He et al., 2023). More details about the baselines can be find in Appendix C.

### 4.2. Main Result

#### 4.2.1. Quantitative analysis

**Anomaly detection:** The results of anomaly detection on the MVTec-AD dataset are comprehensively presented in Table 1. HVQ-Trans, a state-of-the-art method known for its well-designed transformer structure, and DiAD, representing the latest diffusion-based anomaly detection model, serve as the primary baselines for comparison. In this evaluation, the proposed VPDM consistently demonstrates superior performance, outperforming all competitive baselines. Notably, VPDM showcases a significant performance boost, surpassing HVQ-Trans by 1.65% and 2.15% on *Capsule* and *Pill*, respectively. This notable improvement can be attributed to the efficacy of the proposed VPOT model and the underlying diffusion model in effectively addressing the challenges posed by the "identical shortcut" problem. The results underscore the robustness and effectiveness of VPDM in anomaly detection tasks.

In comparison to MVTec-AD, VisA presents greater challenges due to its more intricate structures and scenes featuring multiple misaligned instances. Table 2 showcases the superior performance of VPDM compared to other methods in the multi-class setting. Our proposed model outperforms the top-performing comparison methods, HVQ-Trans, and DiAD, by 1.06% and 7.86%, respectively. In conclusion, VPDM demonstrates effectiveness and efficiency in anomaly detection applications. For a detailed quantitative analysis on MPDD, please refer to Appendix E.

*Table 3.* Ablation studies were conducted to assess anomaly detection/localization results using the AUROC metric on MVTec-AD.

| Vague | Prototype | OT | OT-loss | $\mathcal{N}(0, \boldsymbol{I})$ | $\mathcal{N}(\hat{\boldsymbol{y}}, \boldsymbol{I})$ | Result |
|---|---|---|---|---|---|---|
| ✓ | - | - | - | - | ✓ | 68.4 / 68.8 |
| ✓ | ✓ | - | - | - | ✓ | 93.5 / 94.1 |
| ✓ | ✓ | ✓ | - | - | ✓ | 97.1 / 96.9 |
| ✓ | ✓ | ✓ | ✓ | ✓ | - | 76.7 / 80.1 |
| - | ✓ | ✓ | ✓ | - | ✓ | 96.5 / 96.8 |
| ✓ | ✓ | ✓ | ✓ | - | ✓ | **98.4 / 97.8** |

**Anomaly localization:** Anomaly localization aims to detect anomalous regions given an anomalous sample. The localization results on MVTec-AD are presented in Table 1. Our model consistently outperforms all competitive baselines on average. Notably, even against strong SOTA baselines such as HVQ-Trans and DiAD, our model exhibits superior performance, surpassing them by 0.51% and 1.02%, respectively. This achievement can be attributed to the innovative concepts introduced, such as the incorporation of vague conditions. While HVQ-Trans mitigates the infiltration of anomalous information through a specially designed transformer structure, it still experiences some leakage due to excessive anomalous information entering from the input layer. In contrast, our proposed VPDM effectively reduces anomalous information from the input layer, alleviating the subsequent challenges associated with troubleshooting anomalous information. The localization results on VisA are detailed in Table 2, where our model consistently outperforms all competitive baselines on average. For additional

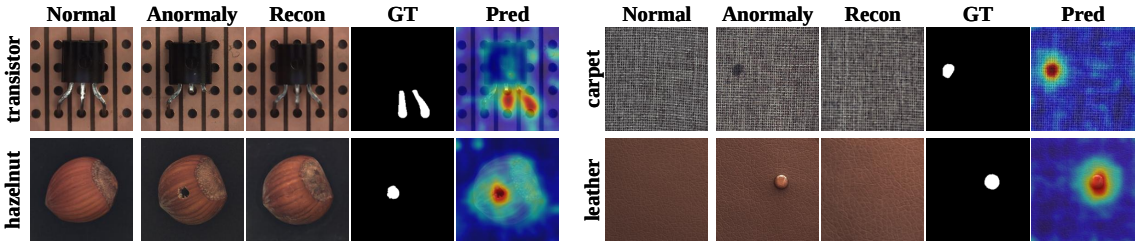


*Figure 2.* Qualitative results for anomaly localization on MVTec-AD. From left to right: normal sample as the reference, anomaly, our reconstruction, ground-truth, and our predicted anomaly map.

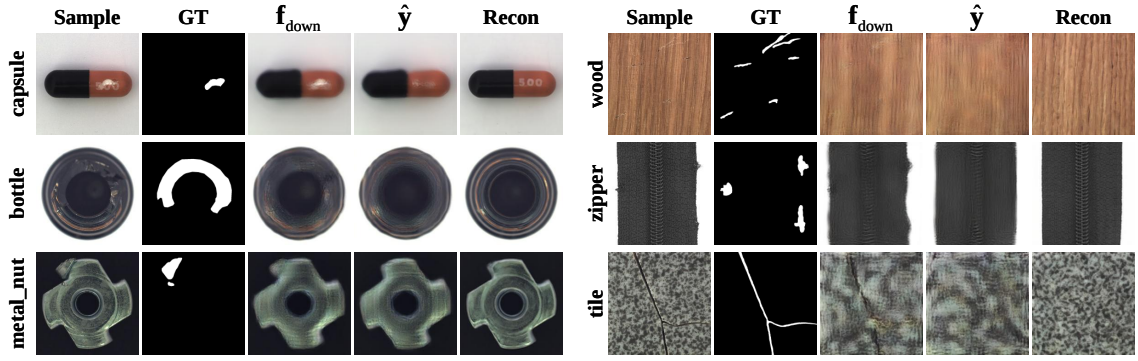


*Figure 3.* Visualization of features at various steps of VPDM, from left to right: normal sample, ground truth, downsampler feature ($\boldsymbol{f}_{\text{down}}$), VPOT output ($\hat{\boldsymbol{y}}$), and reconstruction.

quantitative analysis on MPDD, please refer to Appendix E.

**Ablation study:** To assess the effectiveness of the proposed modules, which include the vague (low-pass) operation, prototypes, OT, and various diffusion models, we conducted comprehensive ablation studies on MVTec-AD. The results, as presented in Table 3, reveal the following key observations: (i) The performance of VPDM without the vague process drops by 1.93% (from 98.2 to 96.5), indicating that while the other components of VPDM contribute to mitigating the "identical shortcut", it is still crucial to reduce anomalous information from the input layer through the vague process; (ii) When we replace the diffusion model with DDPM, the performance drops significantly by 22.05% (from 98.2 to 76.7), highlighting the pivotal role of the proposed diffusion model in VPDM. The vague condition provided by the VPOT model includes only fundamental information, making the generation process more challenging. It is difficult for the general diffusion model to generate the desired result directly from noise; (iii) The application of OT and OT loss demonstrates an improvement of 3.81% 3.71% (from 93.5 to 97.1) and 1.32% (from 97.1 to 98.2), respectively. This validates the advantage of using OT to index prototypes and indicates that optimizing OT loss is beneficial for prototypes; (iv) Without the VPOT model, using the result of the vague process to guide the proposed diffusion model results in a significant drop of 30.49% (from 68.4 to 98.2). This emphasizes the critical importance of the proposed VPOT module. For additional ablation studies, please refer to Appendix F.

#### 4.2.2. Qualitative Analysis

To showcase the capability of modeling normal distributions, we visualize the generated results. As depicted in Fig.2, VPDM successfully reconstructs anomalies into their corresponding normal samples, accurately localizing anomalous regions through reconstruction differences for both object anomalies (Left) and texture damages (Right). The distinctiveness of the proposed VPDM lies in excluding detailed information from the input images, retaining only the basic information input to the model. The images were low-pass filtered to index the normal distribution in the vague prototype, which was then used to guide the diffusion model in generating samples, we illustrated this process in Fig. 3. In Fig. 3, $\boldsymbol{f}_{\text{down}}$ denotes the low-pass filtered image where most anomalous components are eliminated. Using this vague feature to query the normal distribution in vague prototypes via OT, the result $\hat{\boldsymbol{y}}$ completely eliminates anomalous information, becoming a vaguely normal sample. Subsequently, the proposed well-designed diffusion model adds details progressively, allowing VPDM to successfully reconstruct anomalies into their corresponding normal samples. For additional qualitative analysis, please refer to Appendix G.

## 5. Conclusion

In this paper, we introduce the VPDM for multi-class anomaly detection, specifically designed to counteract the infiltration of abnormal condition information at its origin, thus avoiding the "identical shortcut" problem. VPDM utilizes prototypes containing only vague information about the target as the initial condition. A carefully designed diffusion model is subsequently employed to progressively enrich these vague prototypes with finer details. By leveraging prototypes with fundamental shape and color information, the model receives sufficient guidance for generating corresponding tasks. Simultaneously, the exclusion of anomalous details helps prevent the model from being misled. Finally, we introduce the VPOT model, leveraging OT technology to offer more precise information about conditions. Experimental results and comparisons on the MVTec-AD, VisA, and MPDD datasets demonstrate that VPDM achieves state-of-the-art performance.

## Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant U21B2006; in part by Shaanxi Youth Innovation Team Project; in part by the Fundamental Research Funds for the Central Universities QTZX24003 and QTZX22160; in part by the 111 Project under Grant B18039; The work of Wenchao Chen acknowledges the support of the stabilization support of National Radar Signal Processing Laboratory under Grant (JKW202X0X) and National Natural Science Foundation of China (NSFC) (6220010437).

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

## References

Bergmann, P., Fauser, M., Sattlegger, D., and Steger, C. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9592–9600, 2019.

Bergmann, P., Fauser, M., Sattlegger, D., and Steger, C. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4183–4192, 2020.

Cao, Y., Wan, Q., Shen, W., and Gao, L. Informative knowledge distillation for image anomaly segmentation. *Knowledge-Based Systems*, 248:108846, 2022.

Chen, L., You, Z., Zhang, N., Xi, J., and Le, X. Utrad: Anomaly detection and localization with u-transformer. *Neural Networks*, 147:53–62, 2022.

Choi, J., Kim, S., Jeong, Y., Gwon, Y., and Yoon, S. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14347–14356. IEEE Computer Society, 2021.

Collin, A.-S. and De Vleeschouwer, C. Improved anomaly detection by training an autoencoder with skip connections on images corrupted with stain-shaped noise. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 7915–7922. IEEE, 2021.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Defard, T., Setkov, A., Loesch, A., and Audigier, R. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pp. 475–489. Springer, 2021.

Dehaene, D., Frigo, O., Combrexelle, S., and Eline, P. Iterative energy-based projection on a normal data manifold for anomaly localization. *arXiv preprint arXiv:2002.03734*, 2020.

Deng, H. and Li, X. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9737–9746, 2022.

Fernando, T., Gammulle, H., Denman, S., Sridharan, S., and Fookes, C. Deep learning for medical anomaly detection–a survey. *ACM Computing Surveys (CSUR)*, 54(7):1–37, 2021.

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S., and Hengel, A. v. d. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1705–1714, 2019.

Gudovskiy, D., Ishizaka, S., and Kozuka, K. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 98–107, 2022.

Guo, D., Tian, L., Zhang, M., Zhou, M., and Zha, H. Learning prototype-oriented set representations for meta-learning. In *International Conference on Learning Representations*, 2022.

Han, X., Zheng, H., and Zhou, M. Card: Classification and regression diffusion models. *Advances in Neural Information Processing Systems*, 35:18100–18115, 2022.

He, H., Zhang, J., Chen, H., Chen, X., Li, Z., Chen, X., Wang, Y., Wang, C., and Xie, L. Diad: A diffusion-based framework for multi-class anomaly detection. *arXiv preprint arXiv:2312.06607*, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hou, J., Zhang, Y., Zhong, Q., Xie, D., Pu, S., and Zhou, H. Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8791–8800, 2021.

Jezek, S., Jonak, M., Burget, R., Dvorak, P., and Skotak, M. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT)*, pp. 66–71. IEEE, 2021.

Li, C.-L., Sohn, K., Yoon, J., and Pfister, T. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9664–9674, 2021.

Li, Y., Chen, W., Chen, B., Wang, D., Tian, L., and Zhou, M. Prototype-oriented unsupervised anomaly detection for multivariate time series. In *International Conference on Machine Learning*, pp. 19407–19424. PMLR, 2023.

Liang, Y., Zhang, J., Zhao, S., Wu, R., Liu, Y., and Pan, S. Omni-frequency channel-selection representations for unsupervised anomaly detection. *IEEE Transactions on Image Processing*, 2023.

Lu, R., Wu, Y., Tian, L., Wang, D., Chen, B., Liu, X., and Hu, R. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Mousakhan, A., Brox, T., and Tayyub, J. Anomaly detection with conditioned denoising diffusion models. *arXiv preprint arXiv:2305.15956*, 2023.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *In NeurIPS*, 32, 2019.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Pirnay, J. and Chai, K. Inpainting transformer for anomaly detection. In *International Conference on Image Analysis and Processing*, pp. 394–406. Springer, 2022.

Ramachandra, B., Jones, M. J., and Vatsavai, R. R. A survey of single-scene video anomaly detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2293–2312, 2020.

Ratcliff, R. A theory of memory retrieval. *Psychological review*, 85(2):59, 1978.

Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., and Gehler, P. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2022.

Salehi, M., Sadjadi, N., Baselizadeh, S., Rohban, M. H., and Rabiee, H. R. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14902–14912, 2021.

Shi, Y., Yang, J., and Qi, Z. Unsupervised anomaly segmentation via deep feature reconstruction. *Neurocomputing*, 424:9–22, 2021.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. *International conference on machine learning*, pp. 2256–2265, 2015.

Tan, M. and Le, Q. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pp. 10096–10106. PMLR, 2021.

Tanwisuth, K., Fan, X., Zheng, H., Zhang, S., Zhang, H., Chen, B., and Zhou, M. A prototype-oriented framework for unsupervised domain adaptation. *Advances in Neural Information Processing Systems*, 34:17194–17208, 2021.

Tran, T., Do, T.-T., Reid, I., and Carneiro, G. Bayesian generative active deep learning. *International Conference on Machine Learning*, pp. 6295–6304, 2019.

Wang, D., Guo, D., Zhao, H., Zheng, H., Tanwisuth, K., Chen, B., and Zhou, M. Representing mixtures of word embeddings with mixtures of topic embeddings. In *International Conference on Learning Representations*, 2022.

Wang, Z., Zhang, Z., Zhang, X., Zheng, H., Zhou, M., Zhang, Y., and Wang, Y. Dr2: Diffusion-based robust degradation remover for blind face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1704–1713, 2023.

Yi, J. and Yoon, S. Patch svdd: Patch-level svdd for anomaly detection and segmentation. In *Proceedings of the Asian conference on computer vision*, 2020.

Yin, H., Jiao, G., Wu, Q., Karlsson, B. F., Huang, B., and Lin, C. Y. Lafite: Latent diffusion model with feature editing for unsupervised multi-class anomaly detection. *arXiv preprint arXiv:2307.08059*, 2023.

You, Z., Cui, L., Shen, Y., Yang, K., Lu, X., Zheng, Y., and Le, X. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35: 4571–4584, 2022a.

You, Z., Yang, K., Luo, W., Cui, L., Zheng, Y., and Le, X. Adtr: Anomaly detection transformer with feature reconstruction. In *International Conference on Neural Information Processing*, pp. 298–310. Springer, 2022b.

Yu, J., Zheng, Y., Wang, X., Li, W., Wu, Y., Zhao, R., and Wu, L. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677*, 2021.

Zaheer, M. Z., Lee, J.-h., Astrid, M., and Lee, S.-I. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14183–14193, 2020.

Zavrtanik, V., Kristan, M., and Skočaj, D. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8330–8339, 2021.

Zhao, Y. Just noticeable learning for unsupervised anomaly localization and detection. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 01–06. IEEE, 2022.

Zhao, Y. Omnial: A unified cnn framework for unsupervised anomaly localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3924–3933, 2023.

Zou, Y., Jeong, J., Pemula, L., Zhang, D., and Dabeer, O. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pp. 392–408. Springer, 2022.

## A. Derivation for Forward Process Posteriors:

In this section, we derive the mean and variance of the forward process posteriors $q\left(\boldsymbol{x}^{t-1} \mid \boldsymbol{x}^{t-1}, \boldsymbol{x}^0, \hat{\boldsymbol{y}}\right)$ in Eq. 15:

$$\begin{aligned}
q\left(\boldsymbol{x}^{t-1} \mid \boldsymbol{x}^{t-1}, \boldsymbol{x}^0, \hat{\boldsymbol{y}}\right) &\propto q\left(\boldsymbol{x}^t \mid \boldsymbol{x}^{t-1}, \hat{\boldsymbol{y}}\right) q\left(\boldsymbol{x}^{t-1} \mid \boldsymbol{x}^0, \hat{\boldsymbol{y}}\right) \\
&\propto \exp(-\frac{1}{2}(\frac{(\boldsymbol{x}^t - (1-\sqrt{\bar{\alpha}^t})\hat{\boldsymbol{y}} - \sqrt{\bar{\alpha}^t}\boldsymbol{x}^{t-1})^2}{\beta^t} \\
&+ \frac{(\boldsymbol{x}^{t-1} - \sqrt{\alpha^{t-1}}\boldsymbol{x}^0 - (1-\sqrt{\alpha^{t-1}})\hat{\boldsymbol{y}})^2}{1-\alpha^{t-1}})) \\
&\propto \exp(-\frac{1}{2}(\frac{\bar{\alpha}^t(\boldsymbol{x}^{t-1})^2 - 2\sqrt{\bar{\alpha}^t}(\boldsymbol{x}^t - (1-\sqrt{\bar{\alpha}^t})\hat{\boldsymbol{y}})\boldsymbol{x}^{t-1}}{\beta^t} \\
&+ \frac{(\boldsymbol{x}^{t-1})^2 - 2(\sqrt{\alpha^{t-1}}\boldsymbol{x}^0 + (1-\sqrt{\alpha^{t-1}})\hat{\boldsymbol{y}})\boldsymbol{x}^{t-1}}{1-\alpha^{t-1}})) \\
&= \exp(-\frac{1}{2}(\mathscr{B}_1(\boldsymbol{x}^{t-1})^2 - 2\mathscr{B}_2\boldsymbol{x}^{t-1}))
\end{aligned}$$

where

$$\mathscr{B}_1 = \frac{\bar{\alpha}^t(1-\alpha^{t-1}) + \beta^t}{\beta^t(1-\alpha^{t-1})} = \frac{1-\alpha^t}{\beta^t(1-\alpha^{t-1})}$$

$$\mathscr{B}_2 = \frac{\sqrt{\alpha^{t-1}}}{1-\alpha^{t-1}}\boldsymbol{x}^0 + \frac{\sqrt{\bar{\alpha}^t}}{\beta^t}\boldsymbol{x}^t + (\frac{\sqrt{\bar{\alpha}^t}(\sqrt{\bar{\alpha}^t}-1)}{\beta^t} + \frac{1-\sqrt{\alpha^{t-1}}}{1-\alpha^{t-1}})\hat{\boldsymbol{y}}$$

and we have the posterior variance:

$$\tilde{\beta}^t = 1/\mathscr{B}_1 = \frac{(1-\alpha^{t-1})}{1-\alpha^t}\beta^t$$

Meanwhile, the following coefficients of the terms in the posterior mean through dividing each coefficient in $\mathscr{B}_2$ by $\mathscr{B}_1$

$$\gamma_0 = \frac{\sqrt{\alpha^{t-1}}}{1-\alpha^{t-1}}/\mathscr{B}_1 = \frac{\beta^t\sqrt{\alpha^{t-1}}}{1-\alpha^t}$$

$$\gamma_1 = \frac{\sqrt{\bar{\alpha}^t}}{\beta^t}/\mathscr{B}_1 = \frac{(1-\alpha^{t-1})\sqrt{\bar{\alpha}^t}}{1-\alpha^t}$$

$$\begin{aligned}
\gamma_2 &= (\frac{\sqrt{\bar{\alpha}^t}(\sqrt{\bar{\alpha}^t}-1)}{\beta^t} + \frac{1-\sqrt{\alpha^{t-1}}}{1-\alpha^{t-1}})/\mathscr{B}_1 \\
&= \frac{\bar{\alpha}^t - \alpha^t - \sqrt{\bar{\alpha}^t}(1-\alpha^{t-1}) + \beta^t - \beta^t\sqrt{\alpha^{t-1}}}{1-\alpha^{t-1}} \\
&= 1 + \frac{(\sqrt{\alpha^t}-1)(\sqrt{\bar{\alpha}^t} + \sqrt{\alpha^{t-1}})}{1-\alpha^t}
\end{aligned}$$

which together give us the posterior mean

$$\boldsymbol{\mu}(\boldsymbol{x}^0, \boldsymbol{x}^t, \hat{\boldsymbol{y}}) = \gamma_0\boldsymbol{x}^0 + \gamma_1\boldsymbol{x}^t + \gamma_2\hat{\boldsymbol{y}}$$

*Table 4.* Anomaly detection results with AUROC metric on MPDD. All methods are evaluated under the multi-class settings. The best results are bold with black.

| Normal Indices | PatchSVDD | PaDiM | DRAEM | RevDistill | PatchCore | FastFlow | UniAD | HVQ-Trans | Ours |
|---|---|---|---|---|---|---|---|---|---|
| Bracket Black | 85.8 / 67.9 | 71.1 / 93.1 | 81.2 / 97.9 | 81.0 / 97.3 | 77.3 / 96.9 | 81.4 / 82.4 | 95.9 / 94.3 | 91.6 / 96.1 | **97.8**±0.05 / **98.2**±0.01 |
| Bracket Brown | 97.3 / 63.2 | 75.0 / 95.0 | 85.0 / 53.8 | 86.0 / 97.2 | 83.1 / 95.3 | 97.5 / 80.3 | 94.2 / **98.7** | 90.3 / 98.2 | **97.9**±0.23 / 98.6±0.01 |
| Bracket White | 87.2 / 55.8 | 73.0 / 97.2 | 78.8 / 95.7 | 83.6 / 98.8 | 75.8 / **99.6** | 72.3 / 98.1 | 84.8 / 95.0 | 89.7 / 94.5 | **95.2**±0.31 / 99.2±0.02 |
| Connector | **99.8** / 90.2 | 83.8 / 97.2 | 88.8 / 85.1 | 99.5 / **99.5** | 96.4 / 98.4 | 94.0 / 94.0 | 89.8 / 97.9 | 88.3 / 97.9 | 97.5±0.11 / 98.9 ±0.01 |
| Metal Plate | 84.6 / 91.0 | 51.1 / 90.2 | **100** / **99.2** | **100** / **99.2** | **100** / 98.6 | 99.7 / 97.9 | 77.6 / 93.3 | 94.4 / 96.4 | **100**±0.00 / 98.9 ±0.03 |
| Tubes | 79.1 / 41.7 | 75.6 / 88.7 | **96.2** / 98.2 | 95.5 / **99.1** | 68.5 / 97.3 | 77.1 / 96.9 | 74.8 / 92.1 | 78.9 / 97.1 | 93.4±0.19 / 98.2±0.01 |
| Mean | 89.0 / 68.3 | 71.6 / 93.6 | 88.3 / 88.3 | 90.9 / 98.5 | 83.5 / 97.7 | 87.0 / 91.6 | 86.2 / 95.2 | 88.9 / 96.7 | **96.9**±0.11 / **98.6**±0.01 |

## B. Dataset:

**MVTec-AD dataset:** MVTec-AD dataset (Bergmann et al., 2019) serves as a simulation of real-world industrial production scenarios, specifically designed for unsupervised anomaly detection. It features a diverse range of 5 texture types and 10 object types, totaling 5,354 high-resolution images across different domains. The training set is composed of 3,629 images containing solely anomaly-free samples, while the test set comprises 1,725 images, including both normal and abnormal instances. Detailed pixel-level annotations are provided for precise evaluation of anomaly localization.

**VisA dataset:** VisA dataset (Zou et al., 2022) includes 10,821 high-resolution images, consisting of 9,621 normal images and 1,200 anomaly images featuring 78 distinct anomaly types. Organized into 12 subsets, each corresponds to a specific object type, with the 12 objects categorized into three types: Complex structure, Multiple instances, and Single instance.

**MPDD dataset:** MPDD (Jezek et al., 2021) contains 6 classes of metal parts, focusing on defect detection during the fabrication of painted metal parts. Its training set is composed of 888 normal samples without defects, and the test set is composed of 458 samples either normal or anomalous. In particular, samples in MPDD have non-homogeneous backgrounds with diverse spatial orientations, different positions, and various light intensities, leading to greater challenges in anomaly detection.

## C. Baselines:

We conduct and analyze a variety of qualitative and quantitative comparison experiments on MVTec-AD, VisA, and MPDD. We choose the basic method US (Bergmann et al., 2020), RevDistill (Deng & Li, 2022),PatchCore (Roth et al., 2022), FastFlow (Yu et al., 2021) and PSVDD (Yi & Yoon, 2020), a synthesizing-based method DRAEM (Zavrtanik et al., 2021), three embedding-based methods MKD(Salehi et al., 2021), PaDiM (Defard et al., 2021) and RD4AD (Deng & Li, 2022), the reconstruction-based methods JNLD (Zhao, 2022), OmniAL(Zhao, 2023), UniAD(You et al., 2022a), a unified SOTA HVQ-Trans (Lu et al., 2023) method and the diffusion-based methods DiAD(He et al., 2023).

## D. Implementation details:

The image size is chosen as 224 × 224, and the size for resizing feature maps is set to 32 × 32. The feature maps from stage-1 to stage-4 of EfficientNet-b4 (Tan & Le, 2021) are resized and concatenated to form a 272-channel feature map. In the VPOT model, the number of prototypes is set to 50, and the downsampler sampling multiplier is 4. For the diffusion model, the number of timesteps is configured as $T = 1000$, and a linear noise schedule is employed with $\beta^1 = 10^{-4}$ and $\beta^T = 0.02$, consistent with the setup in Ho et al. (2020). We used the Adam optimizer with a learning rate of 0.001 and a batch size of 32. All experiments were implemented in PyTorch (Paszke et al., 2019) and conducted on an NVIDIA RTX 3090 24GB GPU.

## E. More Quantitative Results:

**Anomaly detection:**The anomaly detection results on MPDD, meticulously outlined in Table 4, provide a thorough examination of the effectiveness of our proposed VPDM. In a noteworthy display of superior performance, our model consistently outperforms all competitive baselines, showcasing its robust capabilities in handling anomaly detection challenges. Specifically, when benchmarked against HVQ-Trans and UniAD, two formidable methods renowned for their transformer structure and diffusion-based approaches, respectively, our VPDM surpasses them by a substantial 8.26% and

11.04%on average. This impressive margin underscores the prowess of our innovative VPOT model and diffusion model in effectively mitigating the intricate "identical shortcut" problem. The deliberate design choices within VPDM, emphasizing the use of vague conditions, set it apart from existing conditional diffusion models. The integration of the proposed VPOT model and diffusion model emerges as a powerful solution, achieving state-of-the-art performance on the MPDD dataset.

**Anomaly localization:**Anomaly localization, a pivotal aspect of our study, is geared towards identifying anomalous regions within a given anomalous sample. The comprehensive localization results on MPDD, meticulously presented in Table 4, affirm the consistent superiority of our proposed VPDM over all competitive baselines. In a striking demonstration of its efficacy, our model outperforms strong SOTA baselines, including HVQ-Trans and UniAD, by a remarkable 1.93% and 3.45%, respectively, on average. This notable accomplishment can be attributed to the innovative concepts integrated into VPDM, particularly the strategic incorporation of vague conditions. While HVQ-Trans employs a specially designed transformer structure to mitigate the infiltration of anomalous information, it is susceptible to some leakage due to the excessive entry of anomalous information from the input layer. In stark contrast, our VPDM adeptly addresses this challenge by effectively reducing anomalous information from the input layer. This reduction proves pivotal in alleviating subsequent challenges associated with troubleshooting anomalous information, establishing VPDM as a leading solution for anomaly localization on the MPDD dataset.

## F. More Ablation Studies:

A visual ablation study on VPDM is presented in Fig. 4. Each block consists of four columns: the first column displays the real sample along with its labeling, the second column depicts the direct use of the original input($\boldsymbol{x}$) as a condition, and the subsequent images showcase the corresponding generated results. The third column displays the image that has undergone low-pass filtering, followed by its generated result as a condition. In the fourth column, the (VPOT) condition is reconstructed by the prototype after applying the low-pass, along with its corresponding generated result. In Fig. 4, $\boldsymbol{f}_{\text{down}}$ denotes the low-pass filtered image where most anomalous components are eliminated. Using this vague feature to query the normal distribution in vague prototypes via OT, the result $\hat{\boldsymbol{y}}$ completely eliminates anomalous information, becoming a vaguely normal sample.

The distinctive feature of the proposed VPDM lies in the exclusion of detailed information from the input images, retaining only the fundamental information input to the model. The images underwent low-pass filtering to characterize the normal distribution in the vague prototype, which was subsequently employed to guide the diffusion model in generating samples. Figure 4 illustrates the impact of the proposed VPDM. When using the input image directly as a condition, the "identical shortcut" problem becomes severe, leading to the reconstruction of anomalous. Although this issue is somewhat alleviated after low-pass filtering, the vague image still contains a certain amount of anomalous information, resulting in misleading results. Ultimately, the problem is effectively addressed by using VPOT to reconstruct the entire condition before generating samples, demonstrating the efficacy of the proposed model.

## G. More Qualitative Results:

To delve deeper into the comparative analysis between VPDM and existing methodologies, we present a visual exploration of anomaly detection results involving VPDM, HVQ-Trans, DiAD, and UniAD, as depicted in Fig.5 for MVTec-AD and Fig. 6 for MPDD. HVQ-Trans, leveraging a specialized transformer structure, endeavors to mitigate the infiltration of anomalous information; however, it grapples with potential leakage arising from an excess of anomalous data at the input layer. In contrast, VPDM adeptly tackles this challenge by efficiently reducing anomalous information from the input layer.

Furthermore, we provide detailed visualizations of the generated results. The results, showcased in Fig.7 for MVTec-AD, Fig.8 for VisA and Fig. 9 for MPDD underscore the remarkable performance of the VPDM. In these visual representations, VPDM consistently and successfully transforming anomalous into their corresponding normal samples. The visualizations not only illustrate the accuracy of the model in localizing anomalous regions but also showcase its versatility in handling different types of anomalies, including object anomalies and texture damages. By meticulously capturing reconstruction differences, VPDM effectively discerns and highlights the specific areas that deviate from the normal distribution. These visual insights serve as a testament to the model's efficacy in addressing the challenges posed by anomalies, further solidifying VPDM as a powerful solution for anomaly detection and localization across various datasets and scenarios.

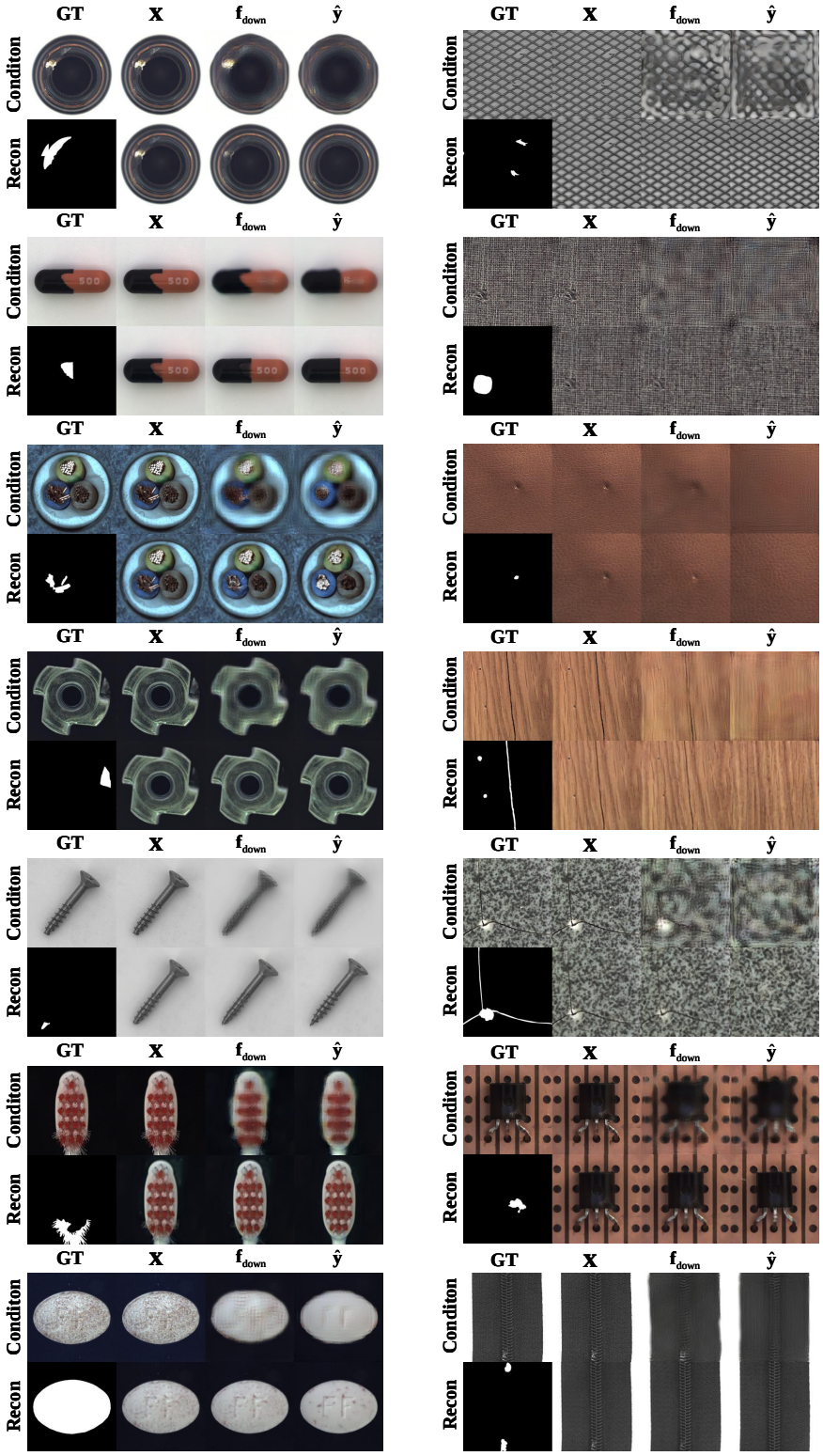


*Figure 4.* A visual ablation study on VPDM is presented. Each block consists of four columns: the first column displays the real sample along with its labeling, the second column depicts the direct use of the original input($\boldsymbol{x}$) as a condition, and the subsequent images showcase the corresponding generated results. The third column displays the image that has undergone low-pass filtering, followed by its generated result as a condition. In the fourth column, the (VPOT) condition is reconstructed by the prototype after applying the low-pass, along with its corresponding generated result.

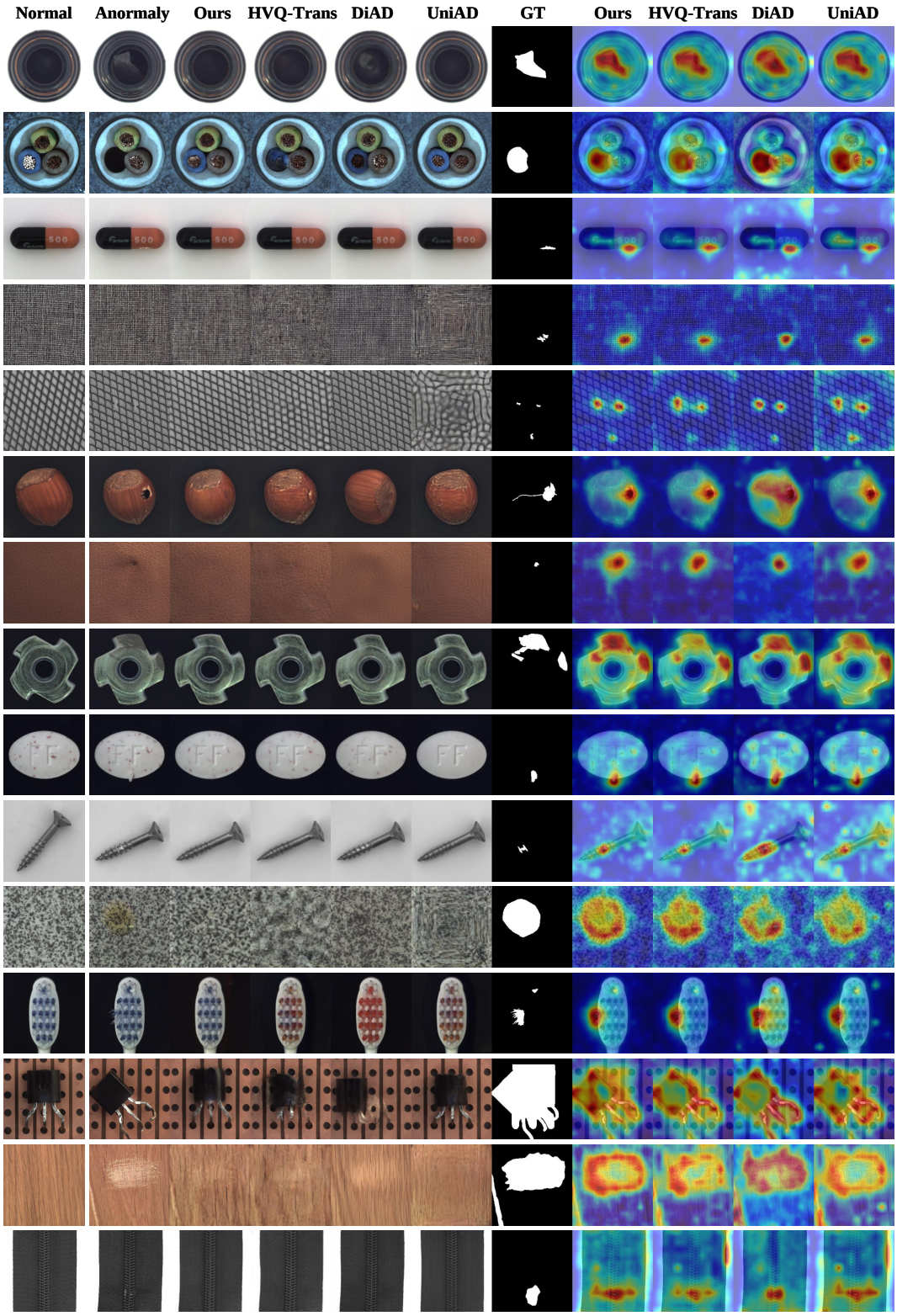


*Figure 5.* Qualitative results for anomaly localization on MVTec-AD. From left to right: normal sample as the reference, anomaly, our reconstruction, HVQ-Trans reconstruction, DiAD reconstruction, UniAD reconstruction, ground-truth, our predicted anomaly map, HVQ-Trans predicted anomaly map, DiAD predicted anomaly map and UniAD predicted anomaly map.

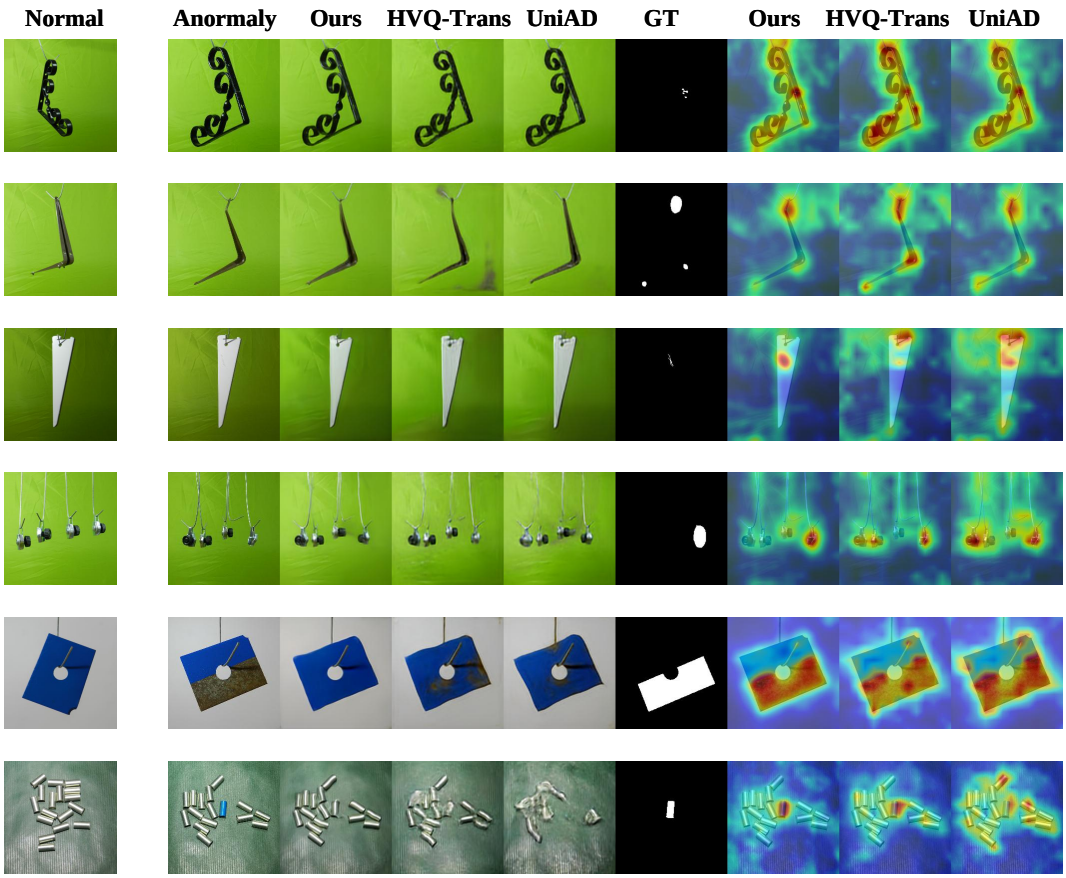


*Figure 6.* Qualitative results for anomaly localization on MPDD. From left to right: normal sample as the reference, anomaly, our reconstruction, HVQ-Trans reconstruction, UniAD reconstruction, ground-truth, our predicted anomaly map, HVQ-Trans predicted anomaly map and UniAD predicted anomaly map.

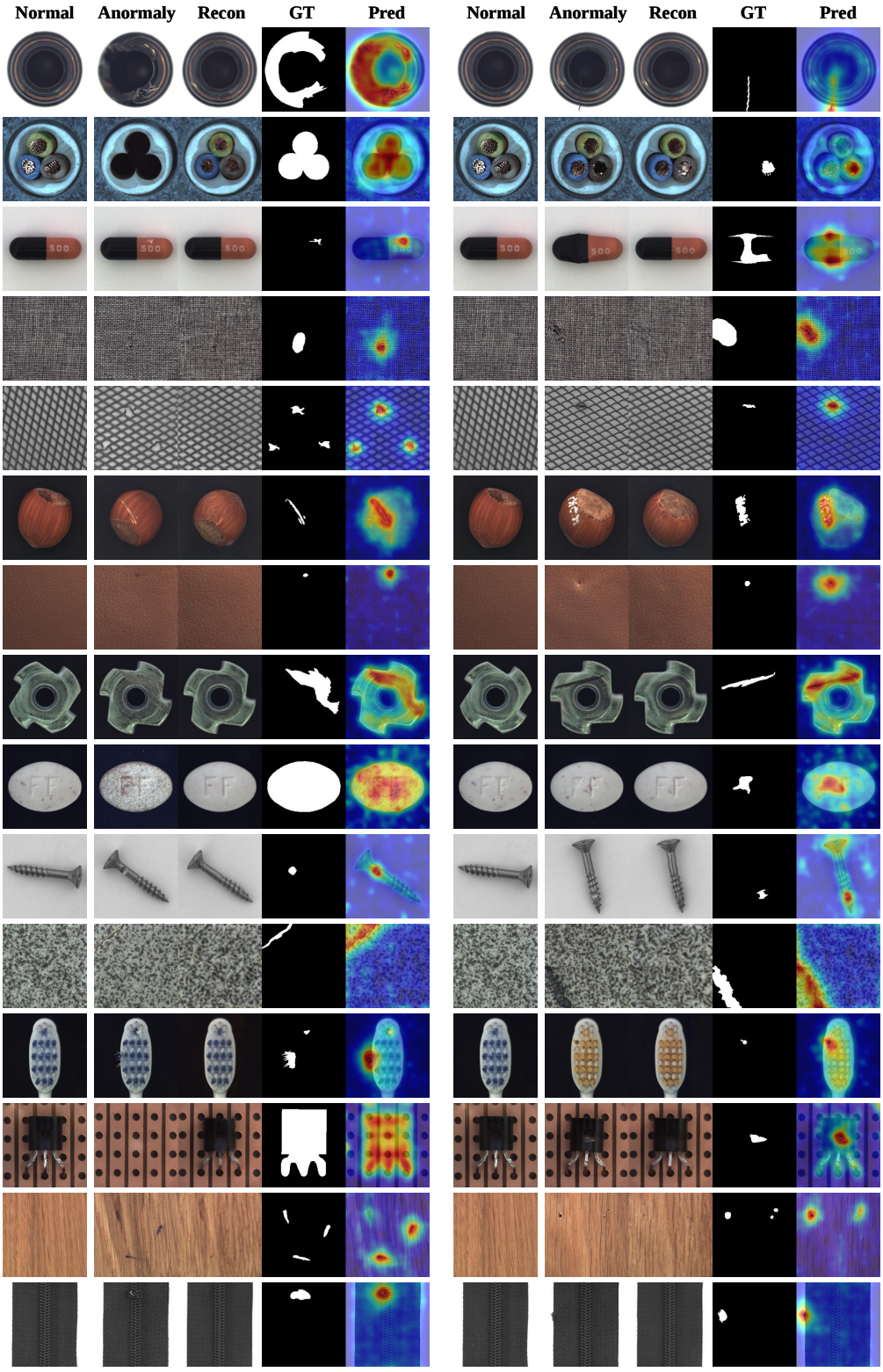


*Figure 7.* Qualitative results for anomaly localization on MVTec-AD. From left to right: normal sample as the reference, anomaly, our reconstruction, ground-truth, and our predicted anomaly map.

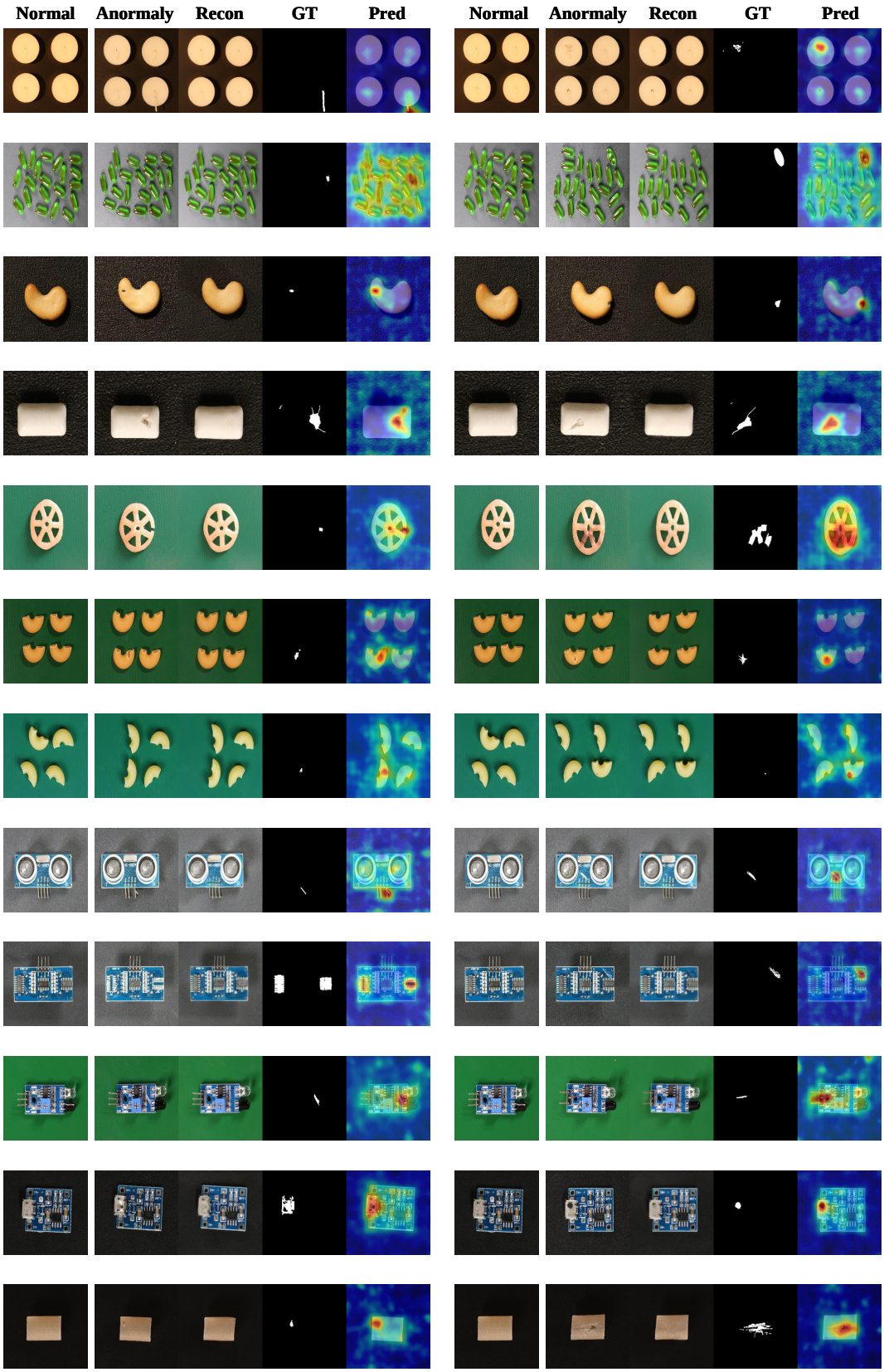


*Figure 8.* Qualitative results for anomaly localization on VisA. From left to right: normal sample as the reference, anomaly, our reconstruction, ground-truth, and our predicted anomaly map.

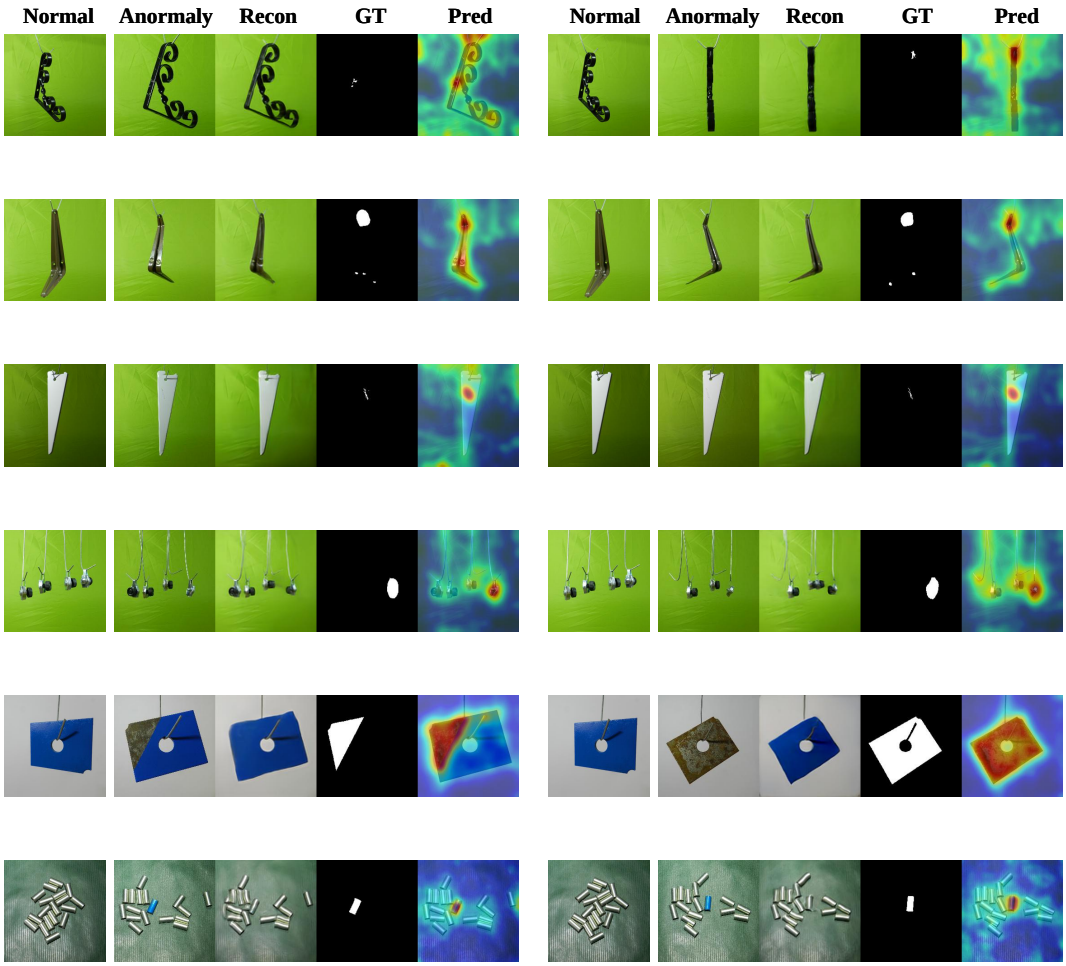


*Figure 9.* Qualitative results for anomaly localization on MPDD. From left to right: normal sample as the reference, anomaly, our reconstruction, ground-truth, and our predicted anomaly map.