UNIFIED ANOMALY UNDERSTANDING AND GENERATION THROUGH CAUSAL DIFFUSION INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing anomaly understanding and anomaly generation methods are usually treated as isolated tasks and rely on separate models, which limits knowledge sharing and scalability. In this paper, we investigate how to unify anomaly understanding and anomaly generation in a single latent diffusion model. Motivated by causal effects, we propose a novel factual-counterfactual generation framework, which enables us to capture the pure anomaly-related response as the direct causal effect of the defect reference on the generation process and suppress background interference by subtracting the counterfactual normal-reference effect from the total attention response. Extensive experiments validate our method's superiority in both anomaly generation and downstream detection tasks. Compared with specialized models, it slashes the FID generation score by roughly 50 points and boosts downstream detection performance by around 20% in PRO and 10% in max-F1. The codes will be public after acceptance.

1 INTRODUCTION

Anomalies are rare, meaningful deviations from normal patterns. In general, as shown in Figure 1(a), there are two principal anomaly-centric objectives: anomaly understanding (or detection) aims to localize regions deviating from normal distributions (Bergmann et al., 2019a; Iqbal et al., 2023), while anomaly generation focuses on synthesizing scarce or expensive abnormal data for augmentation and training (Zhang et al., 2021). To date, these two perspectives have been commonly treated as isolated problems: detection models focus on discriminating or reconstructing normal patterns (Roth et al., 2022; Wyatt et al., 2022), while generation models pay more attention to authenticity and diversity (Zhang et al., 2021), with little shared representation, training objective, or parameter transfer between them.

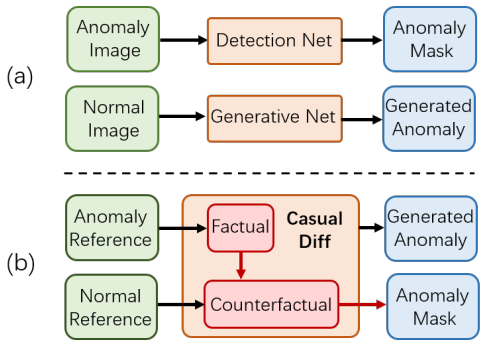


Figure 1: Comparison of (a) traditional isolated models for anomaly detection and generation, and (b) our unified Causal-Diff, which unifies both tasks into a single model with shared parameters through the causal inference mechanism (Axel Sauer, 2021).

In this work, we make the first attempt to unify anomaly detection and generation within a single model via fully shared parameters. However, this unification is highly non-trivial due to severe task conflicts: detection maps pixel-level anomalies from a test image for precision, whereas generation synthesizes complete defect images from multimodal references for photorealism. This mismatch in modalities, objectives, and parameter optimization makes naive unifying prone to suboptimal performance. Note that, this fundamentally differs from prior reconstruction-based detection methods (Wyatt et al., 2022), which employ generative models but still rely on two independent models with separate parameters and objectives.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address this challenge, we propose a unified diffusion-based framework: given a defect reference image depicting the target anomaly, it synthesizes a completed anomaly image where the background image is filled with the target abnormal content. Crucially, the defect reference branch that injects target anomaly information into the generation process naturally exposes attention patterns between the reference tokens and the generation tokens. Since these patterns reveal which reference regions drive the generation, and thus its highlighted regions on the defect reference image can be directly interpreted as its mask result for the detection task. In this way, the same model that generates anomalies can simultaneously understand anomalies, bridging such two tasks through shared computation.

To further suppress background interference, such as normal textures resembling anomalies, we apply the causal inference principle of contrasting factual and counterfactual observations (Axel Sauer, 2021). As shown in Figure 1(b), we replace the single reference branch with two complementary ones: a factual branch driven by a defect reference image and a counterfactual branch driven by a normal reference image of the same scene. The factual branch captures the total effect of the defect reference (including both genuine anomaly and correlated background responses), while the counterfactual branch serves as a baseline to isolate the normal background influence. By computing the causal effect as the difference between these two attention maps, Causal-Diff precisely pinpoints anomalous regions while filtering out background noise, common textures, and lighting variations, ensuring more accurate and interpretable detection without sacrificing generative quality.

The proposed architecture is named as Causal-Diff, compatible with mainstream diffusion backbones, such as Stable Diffusion (Rombach et al., 2022) and FLUX (Black Forest Labs, 2024). This allows it to inherit powerful generative priors from large-scale models. We hope Causal-Diff serves as a new unified paradigm: for common defects with sufficient data, it deploys directly as a general detector; for rare anomalies, the same model acts as a generator to synthesize training data for specialized downstream models. Main contributions of the proposed work can be summarized as follows:

- We propose Causal-Diff, the first unified anomaly detection and generation framework with fully shared parameters. By treating detection as a by-product of the same causal attention computation that drives generation, the model bridges the two tasks without requiring task-specific heads or separate training stages.
- We introduce a causal-effect attention mechanism that explicitly contrasts factual and counterfactual attention responses over defect and normal references. This design isolates the pure anomaly-related response and effectively suppresses interference from noisy or visually similar normal regions on the reference image.
- Extensive experiments demonstrate the superiority of our method in both anomaly generation and detection. Compared with SOTA specialized anomaly generation models, Causal-Diff dramatically improves image fidelity, reducing the FID score by around 50 points. Concurrently, it enhances downstream anomaly detection performance, outperforming specialized detectors by around 20% in PRO and 10% in max-F1.

2 Related Work

2.1 Anomaly Detection

Anomaly detection has evolved rapidly from handcrafted-feature methods to deep learning-based approaches. Early deep methods such as PatchCore (Roth et al., 2022) build a memory bank of normal features and localize anomalies via nearest-neighbor distance. With the rise of generative models, diffusion-based detectors have attracted increasing attention; for example, AnoDDPM (Wyatt et al., 2022) exploits the diffusion prior for unsupervised anomaly scoring. Recently, the community has shifted toward unified multi-class detectors. Dinomaly (Guo et al., 2025) advocates a "less is more" philosophy for multi-class unsupervised anomaly detection, while MambaAD (He et al., 2024) explores state-space models for efficient multi-class feature modeling. UniNet (Wei et al., 2025) further proposes a contrastive learning-guided unified framework with feature selection to suppress irrelevant cues. Beyond the standard one-class setting, few-shot and generalist detectors have also emerged. InCTRL (Zhu & Pang, 2024) performs in-context residual learning with few-shot normal prompts, enabling cross-dataset generalization. Meanwhile, GLAD (Yao et al., 2024) improves reconstruction-based detection via global and local adaptive diffusion models. Collectively,

these advances highlight a trend toward unified, scalable, and foundation-model-powered anomaly detectors.

2.2 Anomaly Generation

Anomaly generation aims to synthesize realistic defective samples to augment scarce anomaly data. Early GAN-based methods such as Defect-GAN (Zhang et al., 2021) generate defect appearances by disentangling defect and background factors. More recently, diffusion models have dominated this area due to their high-quality synthesis and controllability. Qiu et al. (Qiu et al., 2025) propose a noise-consistent siamese-diffusion framework for medical image synthesis and segmentation, ensuring consistent noise supervision between synthesized images and masks. Li et al. (Li et al., 2025) introduce an image-mask alignment regularization to jointly generate anomaly images and their segmentation labels, mitigating mask-label drift. Yao et al. (Yao et al., 2025c) further address drift challenges via hybrid energy optimizations, refining attention activation and anomaly patterns. Beyond generic synthesis, recent works focus on hard or crucial samples. Wan et al. (Wan et al., 2025) revisit hard training sample generation for medical image segmentation, while Yao et al. (Yao et al., 2025a) propose Crucial-Diff, a unified diffusion model that synthesizes crucial images and annotations for data-scarce scenarios. These methods collectively move toward controllable, high-fidelity, and task-aware anomaly generation.

3 Methodology

3.1 Preliminaries on Causal Inference

Necessary concepts of causal inference that underpin our model is first reviewed. A causal graph encodes the cause-and-effect relationships among a set of random variables as a directed acyclic graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}_\mathcal{G}\}$, where each node in $\mathcal{V}$ denotes a random variable and each edge in $\mathcal{E}_\mathcal{G}$ indicates a direct causal influence. Following causal inference convention, we use uppercase letters (*e.g.*, X, M, Y) to denote random variables and lowercase letters (*e.g.*, x, m, y) to denote their observed values. In the simplest setting with a treatment X, a mediator M, and an outcome Y, the graph contains two causal pathways from X to Y: a direct path $X \to Y$ and an indirect path $X \to M \to Y$ through the mediator. Counterfactual notation allows us to express what value the outcome Y would take under hypothetical interventions on the treatment and mediator; for example, $Y_{x,m} = Y(X = x, M = m)$ denotes the outcome when X is set to x and M is set to m.

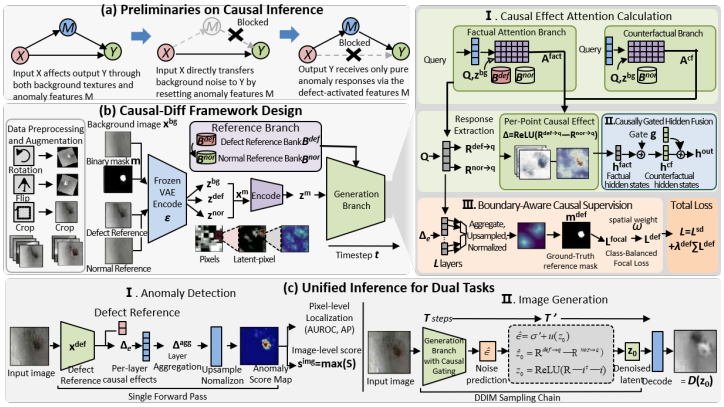


Figure 2: Illustration of our framework. Guided by (a) causal inference priors, (b) Causal-Diff processes VAE-encoded latents through dual reference banks. It applies factual/counterfactual attention to calculate causal effects with boundary-aware supervision, enabling (c) unified anomaly detection and image generation.

Factually, the mediator takes its natural value under treatment x, $M_x = M(X = x)$, yielding outcome Y_{x,M_x}. Counterfactually, it is held at its baseline value under treatment x^*, denoted as M_{x^*}. The total effect (TE) of shifting treatment from x^* to x decomposes into the natural direct effect (NDE) and total indirect effect (TIE). TE measures the overall outcome change, NDE captures the direct effect with the mediator blocked at baseline, and TIE isolates the effect mediated through M.

These three effects are summarized as follows:

$$\begin{aligned} \mathrm{TE} &= Y_{x,M_x} - Y_{x^*,M_{x^*}}, \\ \mathrm{NDE} &= Y_{x,M_{x^*}} - Y_{x^*,M_{x^*}}, \\ \mathrm{TIE} &= \mathrm{TE} - \mathrm{NDE} \;=\; Y_{x,M_x} - Y_{x,M_{x^*}}. \end{aligned} \tag{1}$$

Intuitively, NDE represents the direct influence on the outcome, while TIE captures the indirect influence mediated by M.

3.2 Causal-Diff Framework Design

Causal-Diff is a unified latent diffusion framework for both anomaly understanding and generation. As shown in Figure 2, given a background image $\mathbf{x}^{\text{bg}}$, a binary mask $\mathbf{m}$ indicating the region to inpaint, and dual references (defect $\mathbf{x}^{\text{def}}$ and normal $\mathbf{x}^{\text{nor}}$), the model extracts latent features into two reference banks. During denoising, a dual-branch causal attention mechanism contrasts responses from both references to isolate pure anomaly-related patterns. This shared computation serves dual inference modes using a single set of parameters: in generation mode, it guides the synthesis of a completed image $\hat{\mathbf{x}}$ with realistic anomalies in the masked region; in detection mode, the same attention responses are aggregated into an anomaly score map $\mathbf{S}$ to localize defects on the defect reference image.

The correspondence between Causal-Diff and the causal preliminaries in the previous section is: the defect reference $\mathbf{x}^{\text{def}}$ acts as treatment X, the reference banks' multi-modal knowledge acts as mediator M, and the generated image acts as outcome Y. The factual and counterfactual branches compute outcomes under active and baseline treatments, respectively. Contrasting them isolates the causal effect of $\mathbf{x}^{\text{def}}$ to guide anomalous inpainting.

3.2.1 Causal Effect Attention Calculation

The proposed Causal-Diff first compresses the input and the two references into latent codes via a frozen VAE:

$$\mathbf{z}^{\text{bg}} = \mathcal{E}(\mathbf{x}^{\text{bg}})s, \quad \mathbf{z}^{\text{def}} = \mathcal{E}(\mathbf{x}^{\text{def}})s, \quad \mathbf{z}^{\text{nor}} = \mathcal{E}(\mathbf{x}^{\text{nor}})s, \tag{2}$$

where $\mathcal{E}$ denotes the VAE encoder and s is a fixed scaling factor. The masked input is built as $\mathbf{x}^{\text{m}} = \mathbf{x}^{\text{bg}} \odot (1 - \mathbf{m}) + \boldsymbol{\xi} \odot \mathbf{m}$, where $\boldsymbol{\xi}$ denotes standard Gaussian noise injected into the inpainting region, and encoded to $\mathbf{z}^{\text{m}} = \mathcal{E}(\mathbf{x}^{\text{m}}) \cdot s$, after which the generation branch receives the concatenated conditional input $\mathbf{u}_t = \mathrm{Concat}(\mathbf{z}^{\text{bg}}, \mathbf{m}_{\downarrow}, \mathbf{z}^{\text{m}})$ at each timestep t, where $\mathbf{u}_t$ encodes the current noisy latent, the masked background, and the spatial region to inpaint, and $\mathbf{m}_{\downarrow}$ denotes the binary mask $\mathbf{m}$ downsampled to the latent spatial resolution.

Then, references are processed in two passes: first, the defect latent $\mathbf{z}^{\text{def}}$ is processed in write mode to store intermediate features into the defect bank $\mathcal{B}^{\text{def}}$; second, the normal latent $\mathbf{z}^{\text{nor}}$ is processed similarly to save features into the normal bank $\mathcal{B}^{\text{nor}}$. This sequential population can be written as follows:

$$\mathcal{B}^{\mathsf{r}} = \{\mathbf{b}_\ell^{\mathsf{r}}\}_{\ell=1}^{L}, \quad \mathbf{b}_\ell^{\mathsf{r}} = \mathcal{R}_\phi^{(\ell)}(\mathbf{z}^{\mathsf{r}}, t, \mathbf{c}), \quad \mathsf{r} \in \{\text{def}, \text{nor}\}, \tag{3}$$

where r indexes the reference type (defect or normal), $\mathcal{R}_\phi^{(\ell)}$ denotes the reference branch at layer ℓ, L is the number of hooked layers, t is the diffusion timestep, $\mathbf{c}$ is the conditional context derived from a CLIP visual embedding of the reference image, $\mathcal{B}^{\mathsf{r}}$ is the corresponding reference bank, and $\mathbf{b}_\ell^{\mathsf{r}}$ is the layer-wise feature tensor stored in that bank.

Consider a single transformer block inside the generation branch at layer ℓ. Let $\mathbf{q}_\ell \in \mathbb{R}^{B \times N_\ell^q \times d_\ell}$ denote the normalized query hidden states, where B is the batch size, N_ℓ^q is the number of query tokens, and d_ℓ is the hidden dimension. The features stored in the two banks at this layer are denoted by $\mathbf{b}_\ell^{\text{def}} \in \mathbb{R}^{B \times N_\ell^{\text{def}} \times d_\ell}$ and $\mathbf{b}_\ell^{\text{nor}} \in \mathbb{R}^{B \times N_\ell^{\text{nor}} \times d_\ell}$. In the factual reference branch, the query attends to the concatenation of the query itself, the defect reference bank, and the normal reference bank. After computing the query and key projections $\mathbf{Q}_\ell = \mathbf{q}_\ell(\mathbf{W}_\ell^Q)^\top$ and $\mathbf{K}_\ell^{\text{fact}} = \mathrm{Concat}(\mathbf{q}_\ell, \mathbf{b}_\ell^{\text{def}}, \mathbf{b}_\ell^{\text{nor}})(\mathbf{W}_\ell^K)^\top$, where $\mathbf{W}_\ell^Q$ and $\mathbf{W}_\ell^K$ are the learnable query and key projection matrices, the factual attention distribution is obtained by applying the softmax function across the

key dimension, as follows:

$$\mathbf{A}_\ell^{\text{fact}} \;=\; \mathrm{Softmax}\left(\frac{\mathbf{Q}_\ell \mathbf{K}_\ell^{\text{fact}\top}}{\sqrt{d_\ell}}\right). \tag{4}$$

This distribution captures how much each query token attends to the query itself, the defect reference, and the normal reference when all three sources of information are available.

The counterfactual reference branch blocks the defect reference's influence to isolate the baseline attention distribution observed in its absence. The counterfactual key matrix is $\mathbf{K}_\ell^{\text{cf}} = \mathrm{Concat}(\mathbf{q}_\ell, \mathbf{b}_\ell^{\text{nor}})(\mathbf{W}_\ell^K)^\top$, and the corresponding counterfactual attention is calculated as follows:

$$\mathbf{A}_\ell^{\text{cf}} \;=\; \mathrm{Softmax}\left(\frac{\mathbf{Q}_\ell \mathbf{K}_\ell^{\text{cf}\top}}{\sqrt{d_\ell}}\right). \tag{5}$$

From these two distributions, we extract the per-point responses from the reference tokens to the query tokens. After averaging the attention probabilities over heads, the query-to-defect response and the query-to-normal response are given by:

$$\begin{aligned}\mathbf{R}_\ell^{\text{def}\to q} &= (\bar{\mathbf{A}}_\ell^{\text{fact}})_{:,:,N_\ell^q+1:\,N_\ell^q+N_\ell^{\text{def}}},\\ \mathbf{R}_\ell^{\text{nor}\to q} &= (\bar{\mathbf{A}}_\ell^{\text{cf}})_{:,:,N_\ell^q+1:\,N_\ell^q+N_\ell^{\text{nor}}},\end{aligned} \tag{6}$$

where $\bar{\mathbf{A}}$ denotes the head-averaged attention map and the last axis of each tensor indexes the reference tokens. Thus the response values are aligned with the spatial coordinates of the defect or normal reference image. Following Eq.1, the per-point causal effect at layer ℓ is then defined as the rectified difference between these two response tensors, as follows:

$$\mathbf{\Delta}_\ell \;=\; \mathrm{ReLU}\Big(\mathbf{R}_\ell^{\text{def}\to q} \;-\; \mathbf{R}_\ell^{\text{nor}\to q}\Big). \tag{7}$$

The ReLU operation clips negative values to zero, preserving only regions with stronger attention to the defect reference than the normal reference to indicate anomalies.

3.2.2 Causally Gated Hidden Fusion

To obtain a query-level gate, we aggregate the causal effect in Eq. (7) across the reference-token dimension and apply a sigmoid function, producing a soft gate that selectively emphasizes query tokens with strong defect-reference influence. The gate vector is computed as follows:

$$\boldsymbol{g}_\ell \;=\; \sigma\left(\frac{1}{N_\ell^{\text{def}}}\sum_{j=1}^{N_\ell^{\text{def}}}(\mathbf{\Delta}_\ell)_{:,:,j}\right). \tag{8}$$

Here $\sigma(\cdot)$ denotes the element-wise sigmoid function. Meanwhile, we compute the hidden states from both the factual and counterfactual attention paths, restricted to the query positions, denoted by $\mathbf{h}_\ell^{\text{fact}}$ and $\mathbf{h}_\ell^{\text{cf}}$. The final hidden state at layer ℓ is obtained by blending these two hidden states according to the causal gate, as follows:

$$\mathbf{h}_\ell^{\text{out}} \;=\; \boldsymbol{g}_\ell \odot \mathbf{h}_\ell^{\text{fact}} \;+\; (1-\boldsymbol{g}_\ell)\odot \mathbf{h}_\ell^{\text{cf}}, \tag{9}$$

where broadcasting across the channel dimension is implicit. Performed at every hooked layer, this gated fusion adaptively balances normal structure preservation and anomalous content synthesis on a per-token basis.

3.2.3 Boundary-Aware Causal Supervision

To align the causal effect with the true anomaly, we introduce a boundary-aware causal supervision loss comparing the causal effect map with the ground-truth mask $\mathbf{m}^{\text{def}}$. For each layer ℓ, we aggregate the absolute causal effect over query tokens, reshape it into a pseudo-spatial map, and upsample and normalize it to $[0,1]$. Let $s_{p,\ell}$ be the normalized score at pixel p and layer ℓ, and $r_p = \mathbf{m}^{\text{def}}_{\downarrow,p}$ be the downsampled mask label. Due to small anomalous regions, this reference mask is highly imbalanced. We therefore adopt a class-balanced focal loss formulation:

$$\ell^{\text{focal}}(s,r) = -\alpha^{\text{bal}} r(1-s)^\gamma \log s - (1-\alpha^{\text{bal}})(1-r)s^\gamma \log(1-s), \tag{10}$$

where $\alpha^{\text{bal}} \in [0,1]$ balances the positive and negative classes and $\gamma \geq 0$ is the focusing parameter. The per-layer reference loss is calculated as follows:

$$\mathcal{L}_\ell^{\text{def}} \;=\; \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \omega_p \, \ell^{\text{focal}}(s_{p,\ell}, r_p), \tag{11}$$

where $\mathcal{P}$ is the set of all pixels in the supervision map and ω_p is a spatial importance weight that emphasizes pixels near the anomaly boundary. The boundary weight is computed from the signed distance transform $\mathcal{T}(\cdot)$ of the reference mask as:

$$\omega_p \;=\; 1 \;+\; \lambda^{\text{bdry}} \exp\left(-\frac{\left(\mathcal{T}(\mathbf{m}_\downarrow^{\text{def}})_p\right)^2}{2(\varsigma^{\text{bdry}})^2} \right), \tag{12}$$

where λ^{bdry} and ς^{bdry} control the boundary emphasis. The total training objective combines the standard latent diffusion loss $\mathcal{L}^{\text{sd}}$ with the averaged reference supervision loss, calculated as follows:

$$\mathcal{L} \;=\; \mathcal{L}^{\text{sd}} \;+\; \lambda^{\text{def}} \cdot \frac{1}{L} \sum_{\ell=1}^{L} \mathcal{L}_\ell^{\text{def}}, \tag{13}$$

where L is the number of hooked layers and λ^{def} is a balancing hyperparameter.

3.3 Unified Inference for Dual Tasks

The Causal-Diff model seamlessly switches between tasks. In detection mode, a single forward pass yields layer-wise causal tensors $\mathbf{\Delta}_\ell$. Averaged over queries, these are reshaped and upsampled via $\psi_\ell(\cdot)$ into a unified 2D anomaly map $\mathbf{S}^{\text{agg}} \in \mathbb{R}^{H \times W}$ across L layers, calculated as follows:

$$\mathbf{S}^{\text{agg}} = \frac{1}{L} \sum_{\ell=1}^{L} \psi_\ell \left(\frac{1}{N_\ell^q} \sum_{k=1}^{N_\ell^q} \mathbf{\Delta}_\ell[k,:] \right). \tag{14}$$

For image-level detection, this aggregated spatial anomaly map is further reduced to a scalar score s^{img} by extracting its maximum value across all spatial coordinates (i,j), as follows:

$$s^{\text{img}} = \max_{(i,j)} \mathbf{S}^{\text{agg}}(i,j). \tag{15}$$

Conversely, in image generation mode, the model receives a masked background and dual references to iteratively execute the full DDIM sampling chain. Using the same causal mechanism at each step to predict noise, it synthesizes a final anomaly-inpainted image via the VAE decoder.

4 Experiments

4.1 Datasets and Metrics

A variety of datasets encompassing three diverse scenarios: industrial domain (MVTec (Bergmann et al., 2019b) and VisA (Zou et al., 2022)), medical domain (Polyp (Zhang et al., 2023)) and indoor domain (Floor Dirty (Yao et al., 2025b)), are utilized to comprehensively evaluate the proposed work.

Table 1: Quantitative comparison of generated image quality on the Floor Dirty and Polyp datasets.

Dataset	Category	Method	FID ↓	KID ↓	CLIP ↑
Floor Dirty	Faeces	Anomagic	170.9486	0.06632	0.4486
		Ours	**164.9159**	**0.06253**	**0.4669**
	Stains	Anomagic	139.5493	0.01843	0.3354
		Ours	**138.5543**	**0.01719**	**0.3989**
	Avg	Anomagic	155.2490	0.04238	0.3920
		Ours	**151.7351**	**0.03986**	**0.4329**
Polyp	Avg	Anomagic	113.9588	0.07604	0.4683
		Ours	**95.3700**	**0.04475**	**0.6183**

4.2 Comparisons with SOTA Methods

4.2.1 Comparison on Generation Quality

Table 2: Quantitative comparison of the generated image quality on the MVTec and VisA datasets.

Category	Anomagic(Jiang et al., 2026) FID↓	KID↓	IS↑	CLIP↑	Ours FID↓	KID↓	IS↑	CLIP↑
MVTec Dataset								
Bottle	48.3819	0.29779	1.6301	0.6595	**7.9699**	**-0.01941**	**1.6491**	**0.7655**
Cable	49.8407	0.32646	2.0711	0.6885	**19.6171**	**0.14276**	**2.0748**	**0.7484**
Capsule	260.7583	1.66469	**1.3390**	0.5345	**94.2125**	**0.32158**	1.3036	**0.6421**
Carpet	85.4861	0.32702	1.3176	0.6090	**14.1246**	**0.00336**	**1.3306**	**0.7584**
Grid	184.2018	0.47260	1.8447	0.5643	**21.6004**	**0.01716**	**2.0524**	**0.5891**
Hazelnut	211.7136	1.83808	**2.0032**	0.5171	**69.6026**	**0.55475**	1.9906	**0.6735**
Leather	442.1049	4.67403	1.5531	0.3889	**117.4312**	**0.90616**	**1.6007**	**0.5736**
Metal Nut	68.9035	0.49276	**1.5669**	0.6438	**18.8440**	**0.02659**	1.4017	**0.7938**
Pill	82.7668	0.93140	1.3932	0.7657	**14.3568**	**0.03021**	**1.4715**	**0.8485**
Screw	34.6817	**0.05285**	**1.1289**	0.4880	**31.7602**	0.05833	1.1275	**0.5037**
Tile	730.7167	4.98737	2.9016	0.4608	**35.9057**	**0.10191**	**3.0128**	**0.8224**
Toothbrush	244.6243	**1.30057**	1.2024	**0.5365**	**209.7146**	1.53260	**1.2187**	0.4973
Transistor	**36.4273**	**0.12619**	1.5189	0.7108	64.7540	0.56885	**1.5522**	**0.7373**
Wood	228.2822	1.37187	**2.5976**	0.4471	**55.3913**	**0.24703**	2.5672	**0.7974**
Zipper	95.0384	0.38628	**1.2890**	0.4479	**21.3648**	**0.07709**	1.2887	**0.5972**
MVTec Average	186.9285	1.28333	1.6905	0.5642	**53.1100**	**0.32774**	**1.7095**	**0.6899**
VisA Dataset								
Candle	139.9629	1.09533	1.3756	0.5343	**71.9285**	**0.67713**	**1.3776**	**0.5928**
Capsules	**6.1141**	**0.01005**	**1.5319**	0.8420	22.7845	0.22414	1.4950	**0.8544**
Cashew	122.7646	0.71554	**1.2944**	0.4679	**46.6641**	**0.18286**	1.2580	**0.5490**
Chewinggum	166.7569	0.86495	**1.4231**	0.4340	**13.4749**	**0.02692**	1.4090	**0.6046**
Fryum	97.7398	0.90976	**1.2979**	0.6278	**48.5206**	**-0.00829**	1.2221	**0.7227**
Macaroni1	195.5129	0.71639	1.0310	0.5856	**110.2438**	**-0.11236**	**1.1418**	**0.7062**
Macaroni2	108.0578	0.27364	**1.3040**	**0.5753**	**43.3873**	**-0.05137**	1.2934	0.5138
PCB1	56.0963	0.36600	**1.2659**	0.7289	**39.5929**	**0.20395**	1.2141	**0.7716**
PCB2	**58.6363**	**0.23830**	**1.1616**	0.7201	87.7791	0.25399	1.0368	**0.7572**
PCB3	**22.7012**	**0.14129**	**1.1993**	**0.6298**	37.1037	0.22083	1.1795	0.6241
PCB4	79.3507	**0.35301**	**1.2900**	0.3963	**73.1510**	0.35417	1.2313	**0.4878**
Pipe Fryum	147.7758	1.03932	**1.5118**	0.4152	**100.8456**	**-0.07945**	1.3473	**0.7198**
VisA Average	100.1224	0.56030	**1.3072**	0.5798	**57.9563**	**0.26800**	1.2671	**0.6587**

We compare the generation quality with Anomagic (Jiang et al., 2026), a zero-shot anomaly generation model. For the industrial scenario, as shown in Table 2, both models yield exceptionally high FIDs on complex structural categories such as Toothbrush and Leather, highlighting severe generation difficulties. However, while performance converges closely on Screw (around 3-point gap), ours dramatically slashes the MVTec average FID from 186.93 to 53.11, proving its robust superiority in conquering complex background noise. Similar performance gaps can also be observed in Table 1 for the medical and indoor scenarios. For the qualitative comparison, as shown in Figure 3, ours outperforms Anomagic by blending bottle defects seamlessly with realistic lighting, preserving intricate polyp tissue textures, and successfully synthesize the target black stain—whereas Anomagic completely fails to generate the defect and distorts the entire scene.

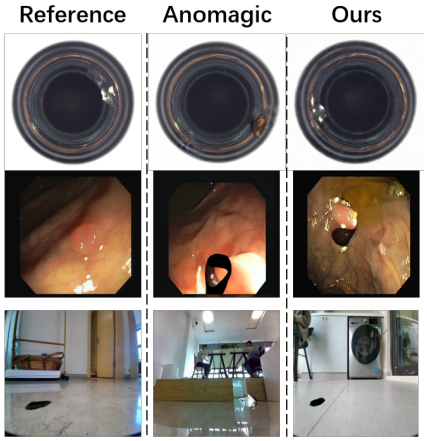


Figure 3: Qualitative comparisons of synthetic anomaly images generated by different models across industrial, medical and indoor scenarios.

4.2.2 COMPARISON ON DETECTION ABILITY

To evaluate the detection mode of Causal-Diff, we compare it with WinClip (Jeong et al., 2023), a zero-shot anomaly detector, on the MVTec and VisA datasets. As shown in Table 3, WinCLIP struggled heavily with pixel-level localization on the structurally complex VisA dataset. Conversely, ours bridges the task conflict seamlessly, shattering SOTA baselines by boosting the VisA max-F1 by nearly 10% and lifting the PRO score by over five points.

Table 3: Anomaly detection performance comparison through the detection mode of Causal-Diff.

Method	MVTec			VisA		
	AUROC↑	PRO↑	max-F1↑	AUROC↑	PRO↑	max-F1↑
Trans-MM(Chefer et al., 2021)	57.5	21.9	12.1	49.4	10.2	3.1
MaskCLIP(Zhou et al., 2022)	63.7	40.5	18.5	60.9	27.3	7.3
WinCLIP(Jeong et al., 2023)	85.1	64.6	31.7	79.6	56.8	14.8
Ours	**86.6**	**83.3**	**40.2**	**79.8**	**62.3**	**23.5**

Visualization of the causal effect calculation can be seen in Figure 4. The causal mechanism flawlessly isolates anomalies across all cases by filtering common textures and reflections: it strips away the bottle's reflective inner rim highlighted by factual attention, subtracts the complex structural noise of the toothbrush bristles, eliminates sprawling gloss variations on the tile surface, and suppresses edge lighting gradients on the capsule to cleanly expose the localized crack or scratch.

4.2.3 EFFECTIVENESS OF GENERATED TRAINING DATA

Beyond its competitive performance in detection mode as demonstrated in the precious section, Causal-Diff can also operate in generation mode to synthesize high-quality training data for downstream tasks in data-scarce domains, such as medical and indoor scenarios, based on SegFormer (Xie et al., 2021). As shown in Table 4, specialized methods severely degrade on structurally intricate categories, particularly on the Faeces and Polyp datasets. Conversely, Ours overcomes these bottlenecks through its causal mechanism to generate highly effective training samples: it nearly doubles performance on Faeces by surging over 25 points in IoU and 36 points in Dice, while consistently boosting the average Polyp performance by around 20 points across both metrics.

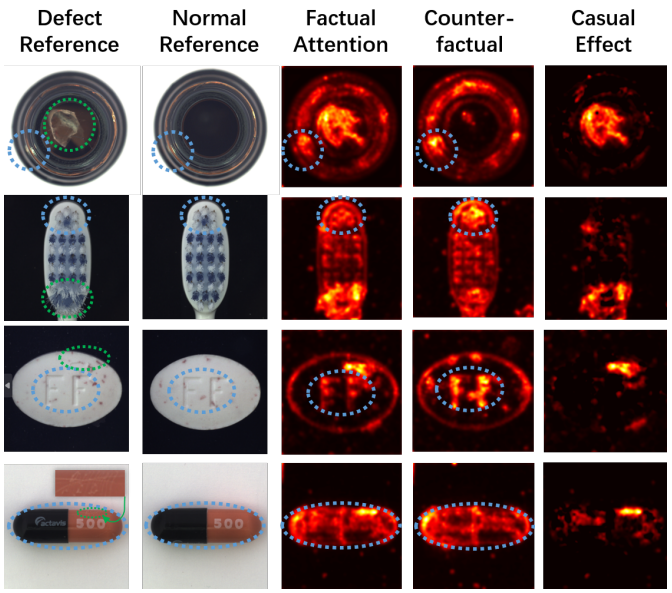


Figure 4: Visualizing the causal mechanism. By subtracting counterfactual attention (normal noise) from factual attention, our causal effect successfully suppresses background interference to isolate precise anomaly responses across multiple categories.

4.3 ABLATION STUDY

4.3.1 EFFECT OF GATING STRATEGIES

This ablation study explores gate fusion strategies: "No Gate" uses no fusion, "Gate for Attn" fuses at the attention map level, and "Ours" fuses at the feature level.

Table 4: Comparison on detectors trained on synthetic data from the generation mode of Causal-Diff.

Dataset	Category	Method	AUROC ↑	AP ↑	max-F1 ↑	PRO ↑	IOU ↑	Dice ↑
Floor Dirty	Stains	Anomagic	**95.90**	**59.04**	**57.89**	**70.79**	41.48	**52.99**
		Ours	95.24	57.51	56.90	68.74	**41.52**	52.38
	Faeces	Anomagic	94.80	62.58	58.22	69.80	36.90	45.42
		Ours	**98.46**	**88.84**	**83.33**	**87.76**	**74.48**	**82.63**
	Avg	Anomagic	95.35	60.81	58.06	70.30	39.19	49.21
		Ours	**96.85**	**73.18**	**70.12**	**78.25**	**58.00**	**67.51**
Polyp	CVC-300	Anomagic	72.61	16.06	27.65	20.87	13.50	19.41
		Ours	**90.46**	**33.51**	**44.18**	**51.35**	**30.80**	**41.77**
	CVC-ClinicDB	Anomagic	79.67	44.16	45.94	11.58	14.59	20.07
		Ours	**82.04**	**57.50**	**56.21**	**26.56**	**37.80**	**45.10**
	CVC-ColonDB	Anomagic	71.23	22.04	28.63	16.90	10.44	15.22
		Ours	**74.92**	**27.17**	**30.61**	**43.05**	**24.32**	**32.33**
	ETIS	Anomagic	73.64	22.43	26.78	29.30	14.17	19.91
		Ours	**82.89**	**40.87**	**40.66**	**44.08**	**24.68**	**33.60**
	Kvasir	Anomagic	75.46	47.97	44.30	20.84	19.09	25.97
		Ours	**80.18**	**57.29**	**55.47**	**42.21**	**38.28**	**50.10**
	Avg	Anomagic	74.52	30.53	34.66	19.90	14.36	20.11
		Ours	**82.10**	**43.27**	**45.43**	**41.45**	**31.18**	**40.58**

As shown in Table 5, The data reveals that feature-level fusion significantly outperforms "Gate for Attn", which suffers from performance degradation due to information loss in the attention maps. Crucially, by dynamically exploiting counterfactual guidance, our feature-level gating mechanism achieves the best performance across all evaluation metrics.

Table 5: Ablation study on gating strategies.

Setting	Generation		Detection		
	FID ↓	KID ↓	AUROC ↑	PRO ↑	max-F1 ↑
No Gate	57.8288	0.33200	85.7	82.1	38.7
Gate for Attn	63.5386	0.33171	85.3	81.7	39.1
Ours	**53.1100**	**0.32774**	**86.6**	**83.3**	**40.2**

4.3.2 Effect of Causal Optimization Strategies

This ablation study explores different optimizations for causal effect. As shown in Table 6, the proposed boundary-aware causal supervision strategy outperforms all alternative reference losses, achieving a minimum FID of 53.11 and top downstream detection scores. In contrast, standard CE and Dice losses perform exceptionally poorly because they lack spatial regression capabilities, fail to penalize pixel-level deviations and thus generate severe structural artifacts that mislead downstream models.

Table 6: Ablation study on causal optimization.

Setting	Generation		Detection		
	FID ↓	KID ↓	AUROC ↑	PRO ↑	max-F1 ↑
CE Loss	206.3718	1.36157	80.11	79.62	38.03
Dice Loss	313.0130	2.57345	80.19	79.98	38.12
MSE Loss	58.5815	0.48851	84.82	81.28	39.49
Ours	**53.1100**	**0.32774**	**86.6**	**83.3**	**40.2**

5 Conclusion

This paper introduces a novel causal inference framework for anomaly understanding and generation. Future work will focus on scaling this causal gating mechanism to multi-class industrial anomaly scenarios and real-time generation pipelines.

AI USE STATEMENT

We used generative AI tools to polish the language and improve the readability of the manuscript. The authors take full responsibility for the final content.

REFERENCES

Andreas Geiger Axel Sauer. Counterfactual generative networks. In *International Conference on Learning Representations (ICLR)*, 2021.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD – a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9584–9592, 2019a.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. 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, 2019b.

Black Forest Labs. FLUX. https://github.com/black-forest-labs/flux, 2024.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20405–20415, 2025.

Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. Mambaad: Exploring state space models for multi-class unsupervised anomaly detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 71162–71187, 2024.

Hasan Iqbal, Umar Khalid, Chen Chen, and Jing Hua. Unsupervised anomaly detection in medical images using masked diffusion model. In *International Workshop on Machine Learning in Medical Imaging*, pp. 372–381. Springer, 2023.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

Yuxin Jiang, Wei Luo, Hui Zhang, Qiyu Chen, Haiming Yao, Weiming Shen, and Yunkang Cao. Anomagic: Crossmodal prompt-driven zero-shot anomaly generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 5485–5493, 2026.

Xiangyue Li, Xiaoyang Wang, Zhibin Wan, Quan Zhang, Yupei Wu, Tao Deng, and Mingjie Sun. Free lunch of image-mask alignment for anomaly image generation and segmentation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1449–1457, 2025.

Kunpeng Qiu, Zhiqiang Gao, Zhiying Zhou, Mingjie Sun, and Yongxin Guo. Noise-consistent siamese-diffusion for medical image synthesis and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15672–15681, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14318–14328, 2022.

Zhibin Wan, Zhiqiang Gao, Mingjie Sun, Yang Yang, Cao Min, Hongliang He, and Guohong Fu. Rethinking hard training sample generation for medical image segmentation. *Pattern Recognition*, 172:112533, 2025.

Shun Wei, Jielin Jiang, and Xiaolong Xu. Uninet: A contrastive learning-guided unified framework with feature selection for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9994–10003, 2025.

Julian Wyatt, Adam Leach, Sebastian M. Schmon, and Chris G. Willcocks. AnoDDPM: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 650–656, 2022.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, pp. 12077–12090, 2021.

Haiming Yao, Yunkang Cao, Wei Luo, Weihang Zhang, Wenyong Yu, and Weiming Shen. Glad: Towards better reconstruction with global and local adaptive diffusion models for unsupervised anomaly detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 1–17, 2024.

Siyue Yao, Mingjie Sun, Eng Gee Lim, Ran Yi, Baojiang Zhong, and Moncef Gabbouj. Crucial-diff: A unified diffusion model for crucial image and annotation synthesis in data-scarce scenarios. *IEEE Transactions on Image Processing*, 34:7024–7036, 2025a.

Siyue Yao, Haotian Xu, Mingjie Sun, Siyue Yu, Jimin Xiao, and Eng Gee Lim. Driftremover: Hybrid energy optimizations for anomaly images synthesis and segmentation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 2251–2259, 2025b.

Siyue Yao, Haotian Xu, Mingjie Sun, Siyue Yu, Jimin Xiao, and Eng Gee Lim. Driftremover: Hybrid energy optimizations for anomaly images synthesis and segmentation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, 2025c.

Gongjie Zhang, Kaiwen Cui, Tzu-Yi Hung, and Shijian Lu. Defect-GAN: High-fidelity defect synthesis for automated defect inspection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2524–2534, 2021.

Hui Zhang, Zuxuan Wu, Zheng Wang, Zhineng Chen, and Yu-Gang Jiang. Prototypical residual networks for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16281–16291, 2023.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European conference on computer vision*. Springer, 2022.

Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17826–17836, 2024.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pp. 392–408, 2022.

APPENDIX

A IMPLEMENTATION DETAILS

In terms of the baseline model selection. For the encoder, we use the fine-tuned MSE VAE: stabilityai/sd-vae-ft-mse, which compresses a 512×512 RGB image into a $64 \times 64 \times 4$ latent tensor with scaling factor $s = 0.18215$. The generation backbone is the inpainting-specific UNet from sd-legacy/stable-diffusion-inpainting. Its input is a 9-channel latent tensor obtained by concatenating (i) the noisy latent $\mathbf{z}_t$, (ii) the downsampled binary mask $\mathbf{m}$, and (iii) the masked-image latent $\mathbf{z}^{\text{bg}}$. In addition, we initialize the reference branch with the standard UNet weights from sd-legacy/stable-diffusion-v1-5 and load only the `unet` subfolder. ReferenceNet runs in "write" mode to build the factual (defect) and counterfactual (normal) token banks.

In terms the training procedure, the focusing parameter is set to $\gamma = 2.0$, which is the standard focal-loss choice and down-weights easy background pixels so that training concentrates on hard anomaly pixels. The class-balancing weight is set to $\alpha^{\text{bal}} = 0.75$, up-weighting the positive (anomaly) class because the anomaly region is typically much smaller than the background in industrial inspection data. The boundary emphasis multiplier is $\lambda^{\text{bdry}} = 2.0$, and the boundary bandwidth is $\varsigma^{\text{bdry}} = 2.0$ pixels on the 32×32 supervision map; this combination puts moderate extra weight on a narrow band around the defect boundary, helping the model localize anomaly edges without overwhelming the interior or background pixels.

At inference time, we replace the training DDPM scheduler with a DDIM scheduler ($1{,}000$ training steps, same β range and scaled-linear schedule). The default generation setting uses 50 DDIM steps and a classifier-free guidance scale of 7.5 with a fixed random seed of 42. For unconditional guidance, we pass an all-zero CLIP image embedding through the same projection network. The defect reference image is also fed as the detection reference; the aggregated causal-effect response is interpreted as the anomaly score map on the reference image, as described in the Method chapter.

B ADDITIONAL EXPERIMENT

We further demostrate the Causal-Diff's ability for generation mode to synthesize high-quality training data for downstream tasks in industrial domain datasets, such as MVTec and VisA. This experiment treats the problem as a segmentation task. Consequently, we adopt SegFormer (Xie et al., 2021) as the training architecture to moving away from training on real normal images with synthetic additions to a framework trained purely on synthetic anomalies. Full evaluation results across diverse datasets are provided in Table 7. Evaluating downstream segmentation models trained purely on synthetic data reveals that the performance gap between the proposed Casual-Diff and other methods is dictated largely by the structural complexity of the target anomalies. For instance, while both methods perform adequately on simpler industrial backgrounds, our auto-focusing approach achieves a significantly wider margin on the structurally complex VisA dataset (AUROC: 91.41% *vs.* 79.69%) because manual masks struggle to blend anomalies naturally into intricate geometries.

Table 7: Performance comparison on the baseline SegFormer trained using different synthetic-only training data on MVTec and VisA datasets. The best results are highlighted in **bold**.

Dataset	Method	AUROC	AP	max-F1	PRO	IOU	Dice
MVTec	DRAEM	87.11	45.81	46.63	57.98	35.96	41.93
	Anomagic	90.16	47.88	48.70	52.60	42.71	48.14
	Ours	**93.14**	**61.49**	**62.43**	**64.84**	**48.55**	**57.02**
VisA	DRAEM	77.45	15.82	22.16	21.66	40.49	42.48
	Anomagic	79.69	29.85	37.34	16.40	49.89	52.15
	Ours	**91.41**	**39.10**	**45.19**	**30.26**	**52.29**	**56.48**