# FINER UNDERSTANDING, FINER CONTROL: TOWARDS CONTROLLABLE AND GENERALIZABLE ANOMALY GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Industrial anomaly generation has emerged as an effective approach to alleviate defect-data scarcity, yet existing methods, while generalizing to unseen object categories, remain limited in precisely controlling fine-grained anomaly attributes. In this work, we study zero-shot controllable anomaly generation, which requires both accurate attribute control and strong cross-category generalization. To avoid disturbing the pre-trained generalizable generative prior through direct generator training, our key insight is to cast controllable generation as a visual understanding problem with the generator kept frozen, and this can be inherently supported by leveraging a unified multimodal understanding–generation model (UMM). Based on it, we present the first Controllable and Generalizable Anomaly Generation framework, comprising grounded attribute understanding, on-policy generation alignment, and inference-time corrective steering. In particular, in the UMM, we first establish reliable anomaly attribute understanding and then use the grounded attribute space to align condition signal for the frozen generator via reinforcement learning. Our method successfully enables precise control over position, size, shape, and orientation on zero-shot object categories. Extensive cross-dataset experiments demonstrate state-of-the-art controllability and visual quality. We further show that stronger attribute understanding leads to more faithful generation and improves targeted augmentation for downstream anomaly detection.

## 1 INTRODUCTION

Industrial anomaly generation has become a increasingly important task as to seek to alleviate defect-data scarcity through synthetic images Sun et al. (2025); Jin et al. (2025). For targeted data augmentation, however, increasing the number of synthetic samples alone may provide limited benefit when difficult defect configurations remain underrepresented in the observed data Baitieva et al. (2024). This motivates us to focus on *controllable* anomaly generation, which deliberately expands coverage toward desired defect configurations while maintaining visual plausibility.

Existing approaches have greatly improved anomaly synthesis for diversity and generalizability across anomaly objects. Training-based methods Gui et al. (2024); Hu et al. (2024); Dai et al. (2025) learn anomaly-specific representations or adapt generative models using available defect examples, whereas training-free methods Sun et al. (2025); Rao et al. exploit pretrained generative priors through attention manipulation, prompt refinement, or reference-guided optimization, enabling the capability of zero-shot generation on unseen objects. Anomagic Jiang et al. (2026) extends the training-based route by learning cross-modal conditioning and anomaly inpainting from a cross-dataset corpus and also supports zero-shot generation.

However, we find that it remains very challenging for these approaches to achieve precise control for anomaly attributes at the same time, for example, requiring a tiny, elongated defect near the upper-left boundary of an unseen object from training data, oriented diagonally. As in Fig. 1, these approaches either lack attribute controllability, or heavily rely on precise condition to achieve it. We formulate this challenge as *zero-shot controllable anomaly generation* problem, and therefore raise a central question: **How can we improve attribute controllability while preserving generalization to unseen data and anomaly types?**



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.

We argue that the generator-centric paradigm of prior anomaly generation methods limits its controllability and generalizability: The pre-trained generator itself already possesses the capability to synthesize visually plausible anomalies, yet training a pre-trained generator for attribute-specific control can distort its generative prior and weaken generalization to unseen object categories. Inspired by recent advances in unified multimodal understanding–generation models (UMMs) Shi et al. (2026); Li et al. (2026); Tian et al. (2026), we instead explore a new perspective: *leveraging the inherent visual understanding capability for precise control of anomaly generation within a single UMM, while keeping the generation module frozen to retain the generalizability in the pre-trained generative prior*. In this paradigm, we cast controllable generation as a visual understanding problem and only need to optimize the shared semantic backbone. By establishing reliable anomaly understanding in a grounded attribute space, the attribute control problem becomes using understanding as a proxy to align requested and generated attributes within this shared space.

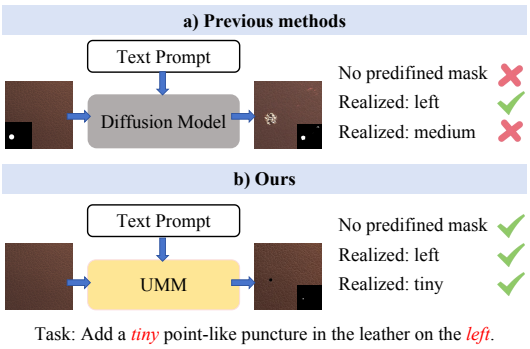



Figure 1: **Controllability for zero-shot anomaly generation methods.** Existing methods Rao et al.; Sun et al. (2025) emphasize plausibility on unseen categories but offer limited attribute control through text. We enable precise anomaly control by leveraging a unified understanding and generation model while preserving the pre-trained generalizability.

Building on the above motivation, we present the first **Controllable and Generalizable Anomaly Generation framework (CGAG)** for fine-grained geometric attributes. Given a normal image, an anomaly category, and a request of attributes over e.g., position, size, shape, and orientation, we can synthesize an anomaly that satisfies the specified attributes on arbitrary unseen object categories. Specifically, we train a single unified understanding-generation model with progressive reinforcement learning. We first ground attribute understanding on real anomalies, and then leverage this capability to align generated anomalies with the requested attributes. At inference time, a differentiable attribute discrepancy further steers the latent trajectory toward the requested configuration.

Our experiments are conducted on the cross-dataset setting, i.e., VisA-to-MVTec AD transfer, for evaluating zero-shot anomaly generation. The result yields a key finding that stronger geometric attribute understanding is associated with more faithful generation result, validating our central thesis that understanding is a strong learner for controllable anomaly generation. We achieve state-of-the-art performance in both attribute controllability and visual quality compared to baseline methods, and further show that targeted augmentation with our method consistently improves downstream anomaly classification and segmentation accuracy. In summary, our contributions are threefold:

- We formulate the problem of zero-shot controllable anomaly generation, aiming to jointly achieve precise attribute control and generalization to unseen object categories and anomaly types. We further introduce the first controllable and generalizable framework for fine-grained anomaly synthesis building on a unified understanding and generation model.
- We propose an understanding–generation coupling mechanism leveraging visual understanding and a frozen generator for zero-shot controllable generation, comprising grounded attribute understanding, on-policy generation alignment, and inference-time steering.
- Our method achieves state-of-the-art performance in both attribute controllability and visual quality. We further validate that stronger attribute understanding leads to more faithful controllable generation and demonstrate its effectiveness for targeted data augmentation.

## 2 Related Work

### 2.1 Industrial Anomaly Generation

Recent research has progressively expanded from anomaly detection and understanding Zhang et al. (2026); Jin et al. (2026); Xu et al. (2025); Zhang et al. (2025), which aim to characterize anomaly

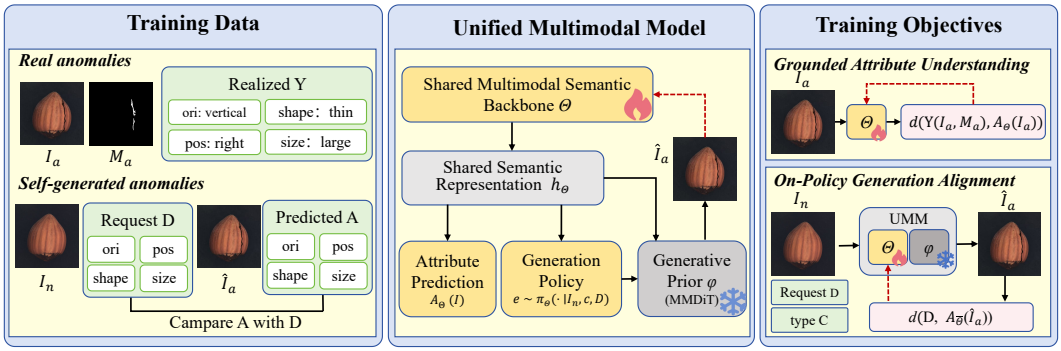



Figure 2: **Overview of Generalizable and Controllable Anomaly Generation**. We optimize a shared multimodal semantic model $\Theta$ while keeping the generative prior $\Phi$ frozen. The shared representation simultaneously supports attribute understanding and generation conditioning: Stage I aligns predicted attributes $A$ with mask-derived realized attributes $Y$ on real anomalies ($A \approx Y$), and Stage II aligns $A$ with the request $D$ on self-generated anomalies ($A \approx D$), leading to the desired control ($Y \approx D$). The same attribute feedback is further used for attribute steering at inference.

semantics and attributes, to more recent anomaly generation, which seeks to synthesize diverse and realistic defects for downstream applications. Early anomaly generation methods construct synthetic perturbations for detector training, as exemplified by CutPaste Li et al. (2021) and DRAEM Zavrtanik et al. (2021). To capture richer defect appearances, AnomalyDiffusion Hu et al. (2024), AnoGen Gui et al. (2024), and DualAnoDiff Jin et al. (2025) learn anomaly representations or adapt generators from limited examples, while SeaS Dai et al. (2025) recombines learned anomaly attributes. Alternatively, training-free methods exploit pretrained priors through text-driven guidance in AnomalyAny Sun et al. (2025) and reference-guided optimization in O2MAG Rao et al.. Cross-dataset training further enables zero-shot generation in Anomagic Jiang et al. (2026), while outcome feedback and reinforcement learning support iterative synthesis refinement in AnomalyAgent Su et al. (2026). Our method emphasizes precise attribute control without target-specific parameter updates, while retaining the cross-dataset zero-shot generalization ability of prior methods.

### 2.2 Unified Models and Understanding-Guided Generation

Unified multimodal modeling brings visual understanding and synthesis into a common framework, as explored by EMU Sun et al. (2024), DreamLLM Dong et al. (2024), BAGEL Deng et al. (2025), Muddit Shi et al. (2026), OneCAT Li et al. (2026), and InternVL-U Tian et al. (2026). Recent UMMs typically improve understanding–generation unification by designing shared architectures that preserve task-specific processing while aligning both capabilities in a common representation space. Representative methods achieve this through decoupled visual pathways Wu et al. (2025), unified autoregressive–flow or diffusion objectives Ma et al. (2025), specialized vision experts Li et al. (2025), and shared tokenization mechanisms Qu et al. (2025), enabling tighter coordination between visual understanding and generation. Complementary approaches improve semantic alignment through prompt learning or sampling guidance: Promptist Hao et al. (2023) optimizes prompt policies using generation rewards, while Attend-and-excite Chefer et al. (2023) steers sampling through attention-based guidance. Our framework connects these two optimization routes through reliable attribute understanding grounded in real anomalies, using it both to assess the generated results and to enable differentiable guidance for the generator.

## 3 Method

### 3.1 Problem Formulation: Controlling Anomaly Attributes

We study zero-shot controllable anomaly generation, where a model is expected to synthesize anomalous images with explicitly specified attributes on unseen objects without target-specific fine-

tuning. Given a normal image $I_n$, an anomaly category $c$, and an attribute request $D$, our goal is to synthesize an anomalous image $\hat{I}_a$ whose realized attributes match the requested configuration.

We characterize each generated anomaly by a set of attributes $Y$, including *position, size, shape, and orientation*, which can be deterministically derived from its anomaly mask through a set of geometric rules $R$ (See supplementary B for details). The generation objective is therefore to produce a plausible anomaly and at the same time to minimize the discrepancy $d$ between the requested attributes $D$ and their actual attribute $Y$.

### 3.2 Overview: Recasting Controlled Generation as Visual Understanding

To allow for both controllability and generalizability as discussed in the introduction, we tackle the above generation objective by leveraging a unified understanding and generation model, keeping its generator frozen, and turning its visual understanding into generation feedback. In the unified model, a semantic model parameterized by $\Theta$ encodes multimodal inputs into a shared semantic representation $h_\Theta$. This representation supports two task interfaces: attribute understanding predicts the properties of an anomalous image, while generation conditioning autoregressively produces a semantic instruction $e$ for the pre-trained generator $\mathcal{G}$ with parameters $\Phi$ to yield the generation $\hat{I}_a$:

$$\hat{I}_a = \mathcal{G}_\Phi(I_n, e), \quad e \sim \pi_\Theta(\cdot \mid I_n, c, D). \tag{1}$$

So formally, our training goal on the source dataset is to keep the generative prior $\Phi$ frozen and optimize the semantic backbone $\Theta$ so that:

$$\Theta = \arg\min_\Theta \mathbb{E}\big[d(Y(\hat{I}_a), D)\big]. \tag{2}$$

The requested attributes $D$, predicted attributes $A$, and realized attributes $Y$ lie in the same discrete attribute space, enabling explicit, attribute-wise comparison. We recast controllable generation as visual understanding by using anomaly attribute understanding as an intermediate proxy between requested configurations and generated results. Specifically, Grounded Attribute Understanding (§3.3) first aligns $A$ with $Y$ on real anomalies. On-Policy Generation Alignment (§3.4) then uses the discrepancy between $A$ and $D$ on self-generated anomalies as feedback to improve generation conditioning. With reliable attribute predictions, the overall pipeline can be summarized as:

$$\underbrace{A \approx Y}_{\text{ground understanding}}, \qquad \underbrace{A \approx D}_{\text{align generation}} \implies \underbrace{Y \approx D}_{\text{desired control.}} \tag{3}$$

Finally, Attribute Steering (§3.5) extends the same attribute feedback to inference, guiding the generation trajectory toward the requested configuration without updating model parameters.

### 3.3 Grounded Attribute Understanding

To ground the shared semantic representation correctly in the anomaly attribute space, we first optimize the semantic policy on real anomalous images. Given an anomalous image $I_a$, we obtain its anomaly region $M_a$ following Anomagic Jiang et al. (2026), and derive its attributes $Y(I_a, M_a)$ using deterministic geometric rules (Supplementary B) over the anomaly region and object foreground, covering position, size, shape, and orientation. For the same image, the current semantic policy samples $K$ structured attribute responses from $\pi_\Theta(\cdot \mid I_a)$, each containing the predicted attribute values $A^{(k)} = \{a_t^{(k)}\}$. Each sampled response is evaluated against the mask-derived actual attributes using:

$$r_{\text{attr}}^{(k)} = \frac{1}{|\mathcal{A}(Y)|} \sum_{t \in \mathcal{A}(Y)} \left[1 - C_t\Big(y_t, a_t^{(k)}\Big)\right], \tag{4}$$

where $C_t(y_t, a_t^{(k)})$ is an attribute-specific cost measuring the degree of mismatch between the realized value $y_t$ (See Supplementary C) and the predicted value $a_t^{(k)}$, and $\mathcal{A}(\cdot)$ is the set of applicable attributes. We optimize $\pi_\Theta$ with GRPO to favor predictions that better match the correct attributes of real anomalies. This stage grounds the model's attribute understanding toward:

$$A_\Theta(I_a) \approx Y(I_a, M_a), \tag{5}$$

making its own attribute prediction a grounded semantic proxy for the realized anomaly properties used in subsequent generation alignment.

### 3.4 On-Policy Generation Alignment

After grounding attribute understanding on real anomalies, we turn it into a feedback signal for generation using self-generated samples, without requiring additional annotated anomaly data. Given a normal image $I_n$, anomaly category $c$, and attribute request $D$, the current semantic policy samples $K$ generation instructions from $\pi_\Theta(\cdot \mid I_n, c, D)$. Each sampled instruction $e^{(k)}$ conditions the frozen generative prior to synthesize:

$$\hat{I}_a^{(k)} = \mathcal{G}_\Phi\left(I_n, e^{(k)}; \xi\right). \tag{6}$$

Here, $\xi$ is the auxiliary visual guidance used during early training, comprising the reference anomaly image, its corresponding mask, and the target mask, to facilitate stable and high-quality anomaly synthesis. Each sampled instruction $e^{(k)}$ conditions the frozen generative prior to synthesize Each generated image is then fed back into an EMA copy of the semantic model, $\bar{\Theta}$, to obtain its predicted attributes $A_{\bar{\Theta}}(\hat{I}_a^{(k)}) = \{a_t^{(k)}\}$. Since Sec. 3.3 grounds the model prediction toward the real attributes of the generation, its discrepancy from the requested attributes provides a semantic feedback signal:

$$r_{\text{attr}}^{(k)} = \frac{1}{|\mathcal{A}(D)|} \sum_{t \in \mathcal{A}(D)} \left[1 - C_t\left(d_t, a_t^{(k)}\right)\right], \tag{7}$$

where $d_t$ is the required value for an attribute $t \in \mathcal{A}(D)$. To ensure that attribute feedback is applied only to meaningful generations, we additionally introduce a visual-validity reward $r_{\text{tex}}^{(k)}$. We compare visual features between the generated and original normal images and measure whether the induced feature shift is concentrated in the spatial region implied by the request while the remaining content stays largely unchanged (See Supplementary F). The final reward for each sampled instruction is:

$$r^{(k)} = r_{\text{tex}}^{(k)} + \mathbb{1}\left[r_{\text{tex}}^{(k)} \geq \tau_{\text{tex}}\right] r_{\text{attr}}^{(k)}. \tag{8}$$

We optimize $\pi_\Theta$ with GRPO while keeping the pre-trained generative prior $\Phi$ fixed. Because the training samples are produced by the current policy and subsequently evaluated by the same grounded semantic model, this forms an on-policy self-feedback loop. This stage can be characterized as aligning $A$ with $D$. Together with the grounding $A \approx Y$, the realized attributes $Y$ on the generated image are progressively driven toward the requested configuration $D$.

### 3.5 Test-Time Attribute Steering

On-policy alignment improves generation behavior at the policy level, but individual samples may still deviate from the requested configuration. We therefore reuse the same attribute discrepancy as a differentiable objective to steer the generation trajectory at inference time. While the rewards in Secs. 3.3 and 3.4 operate on discrete attribute predictions, trajectory steering retains the full prediction distribution of the trained semantic model and defines a soft differentiable form:

$$\mathcal{L}_{\text{attr}}(I, D; \Theta_\star) = \frac{1}{|\mathcal{A}(D)|} \sum_{t \in \mathcal{A}(D)} \mathbb{E}_{v \sim p_{\Theta_\star, t}(\cdot \mid I)}\left[C_t(d_t, v)\right]. \tag{9}$$

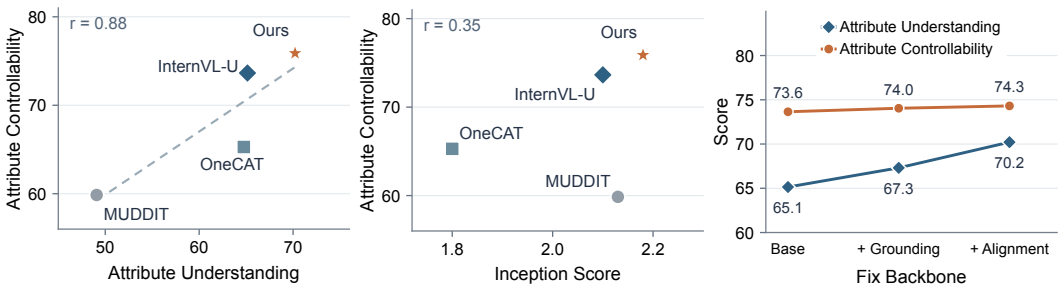
Here, $p_{\Theta_\star, t}(v \mid I)$ denotes the model-predicted probability of candidate value $v$ for attribute $t$, and $C_t(d_t, v)$ measures its discrepancy from the requested value $d_t$. The expectation therefore computes a probability-weighted attribute discrepancy to the request. In contrast to the discrete reward based on a sampled or selected attribute value, this soft formulation preserves gradients through the full prediction distribution. Within the generator $\Phi$, at selected sampling steps, let $z_s$ denote the current latent state and $\hat{I}_s$ the corresponding decoded clean-image estimate. We backpropagate the attribute discrepancy to the latent state:

$$g_s = \nabla_{z_s} \mathcal{L}_{\text{attr}}\left(\hat{I}_s, D; \Theta_\star\right), \tag{10}$$

and apply an RMS-normalized update before the next generation transition:

$$z_s^+ = z_s - \eta_s \frac{g_s}{\text{RMS}(g_s) + \epsilon}, \tag{11}$$

where $\eta_s$ controls the steering strength and $\epsilon$ ensures numerical stability. At this stage, both the trained semantic model $\Theta_\star$ and the pretrained generative prior $\Phi$ remain fixed; gradients are used only to adjust the current generation trajectory. Attribute Steering thus extends the same grounded attribute criterion from policy-level alignment during training to sample-specific correction at inference, without requiring anomaly masks or additional model adaptation.




(a) (b) (c)

Figure 3: Relationship between attribute understanding, image quality, and controllability. (a) Attribute understanding is strongly correlated with controllability across unified models ($r = 0.88$). (b) Inception Score shows a much weaker correlation with controllability ($r = 0.35$). (c) Within the same backbone, improved understanding is accompanied by consistent gains in controllability.

# 4 Experiments

## 4.1 Experimental Setup

**Datasets.** We use VisA Zou et al. (2022) for attribute-understanding initialization and outcome-driven semantic policy learning. Evaluation covers all 15 MVTec AD Bergmann et al. (2019) categories, assessing attribute understanding, controllable generation, and downstream anomaly detection. Our generation model is directly transferred without target-specific parameter updates and requires neither anomaly masks nor anomaly references at inference.

**Compared methods.** We compare with two groups of baselines. The first includes representative industrial anomaly generation methods, including DRAEM Zavrtanik et al. (2021), AnomalyDiffusion Hu et al. (2024), AnoGen Gui et al. (2024), SeaS Dai et al. (2025), Anomagic Jiang et al. (2026), and O2MAG Rao et al.. These methods differ in how they adapt to a target domain. AnomalyDiffusion, AnoGen, and SeaS perform task-specific training-time adaptation using a few real anomaly images, while O2MAG is training-free but conditions generation on a real anomaly reference at inference time. Anomagic is closer to our cross-dataset setting, where a model trained on source-domain data is directly evaluated on unseen target categories without target-specific fine-tuning. We further compare with unified multimodal understanding–generation models, including Muddit Shi et al. (2026), OneCAT Li et al. (2026), and InternVL-U Tian et al. (2026), to assess whether general-purpose unified models can provide comparable fine-grained controllability for industrial anomalies.

**Evaluation metrics.** Attribute understanding is evaluated on real images against mask-derived labels. Controllability measures agreement between requested attributes and those derived from generated anomaly masks, reporting position, size, shape, orientation, and their unweighted average (Avg.). Generation quality is measured by Inception Score (IS) and IC-L. Downstream performance is evaluated using image-level accuracy (I-ACC) and pixel-level PRO (P-PRO) on both the full test set and the predefined hard-case subset.

**Downstream refinement protocol.** We use the discriminative network of DRAEM Zavrtanik et al. (2021) as the downstream detector for all augmentation settings. Generic and targeted augmentation use the same downstream architecture, normal-image pool, defect types, number of synthetic samples, and training schedule; the only difference is the distribution of attribute requests. Generic augmentation samples from the default attribute distribution, whereas targeted augmentation follows attribute configurations diagnosed from the hard-case pool.

**Implementation details.** We implement our framework in PyTorch with InternVL-U Tian et al. (2026) and run experiments on three NVIDIA A6000 GPUs. GRPO updates understanding-side LoRA adapters with rank 16, a learning rate of $5 \times 10^{-6}$, and a rollout group size of 6, while the visual generator remains frozen. Full training and inference configurations are in the supplementary.

Table 1: **Comparison of attribute controllability and visual quality on MVTec AD**. Controllability measures how faithfully generated anomalies realize the requested position, size, shape, and orientation, with Avg. denoting their unweighted mean. IS and IC-L measure visual quality. Our method achieves the best alignment across all four attributes and the highest average controllability, while attaining the highest IS and matching the best IC-L. Industrial generators are included as visual-quality references; N/A indicates that a method does not provide the evaluated mask-free fine-grained attribute-control interface. † denotes methods requiring adaptation with real anomalies.

| Method | Params. | Controllability Alignment | | | | | Visual Quality | |
|---|---|---|---|---|---|---|---|---|
| | | Pos.↑ | Size↑ | Shape↑ | Ori.↑ | Avg.↑ | IS↑ | IC-L↑ |
| **Industrial Generators** | | | | | | | | |
| AnomalyDiffusion† [AAAI'24] | – | N/A | N/A | N/A | N/A | N/A | 1.70 | 0.28 |
| AnoGen† [ECCV'24] | – | N/A | N/A | N/A | N/A | N/A | 1.67 | 0.30 |
| SeaS† [ICCV'25] | – | N/A | N/A | N/A | N/A | N/A | 1.88 | 0.34 |
| Anomagic [AAAI'26] | – | N/A | N/A | N/A | N/A | N/A | 1.92 | 0.27 |
| O2MAG [CVPR'26] | – | N/A | N/A | N/A | N/A | N/A | 1.65 | 0.30 |
| **Unified Models** | | | | | | | | |
| Muddit [ICLR'26] | 2B | 65.51 | 56.25 | 58.58 | 59.09 | 59.86 | <u>2.13</u> | 0.36 |
| OneCAT [CVPR'26] | 3B | 72.66 | 66.19 | 60.85 | 61.43 | 65.28 | 1.80 | 0.30 |
| InternVL-U [arXiv'26] | 4B | <u>88.10</u> | <u>66.68</u> | <u>62.07</u> | <u>77.73</u> | <u>73.65</u> | 2.10 | **0.38** |
| **Ours** | 4B | **90.08** | **70.58** | **63.45** | **79.42** | **75.88** | **2.18** | **0.38** |

### 4.2 Finer Understanding Leads to Finer Control

We further investigate the relationship between anomaly understanding and generation controllability. As shown in Fig. 3(a), attribute understanding is strongly correlated with attribute controllability across different unified models ($r = 0.88$): models that better recognize fine-grained anomaly attributes also more faithfully realize the requested configurations during generation. In contrast, Fig. 3(b) shows a much weaker correlation between Inception Score and controllability ($r = 0.35$). This suggests that improving generic visual quality does not necessarily translate into more precise attribute control; generating plausible anomalies and generating the requested anomalies are distinct capabilities. More importantly, the same trend holds within a fixed backbone. In Fig. 3(c), progressively strengthening attribute understanding from Base to +Grounding and +Alignment increases the understanding score from 65.1 to 70.2, while controllability rises consistently from 73.6 to 74.3. Since the underlying backbone and generation prior remain unchanged, this progression directly supports our central design principle: finer understanding provides a stronger semantic basis for finer generation control. Together, these results motivate shifting controllable anomaly generation from solely improving the generator toward explicitly strengthening its understanding of the attributes to be controlled.

### 4.3 Zero-Shot Generation and Downstream Application

**Zero-shot controllable generation.** Tab. 1 evaluates whether generated anomalies not only appear realistic, but also realize the geometric configuration explicitly requested by the user. Among unified models, our method achieves the strongest attribute alignment across all four dimensions, reaching 90.08, 70.58, 63.45, and 79.42 for position, size, shape, and orientation, respectively, and improving the average controllability from 73.65 to 75.88 over the 4B InternVL-U backbone. The improvement is consistent across attributes, demonstrating that the proposed attribute control strengthens compositional realization of the requested anomaly configuration rather than optimizing a particular geometric factor. Crucially, this gain in controllability preserves the visual generation prior. Compared with InternVL-U, our method increases IS from 2.10 to 2.18 while retaining an IC-L of 0.38, yielding the highest IS among all evaluated methods and the joint-best IC-L. These results support our central premise: fine-grained anomaly control can be substantially strengthened from the understanding side, without sacrificing the visual quality provided by the pretrained generator. Existing industrial generators are therefore included primarily as visual-quality references, whereas unified models enable a direct evaluation of mask-free requested-attribute control.

Table 2: Failure-driven detector refinement with generic and targeted augmentation. We report I-ACC and P-PRO on the full evaluation set and the fixed hard-case subset. All augmentation settings use the same DRAEM discriminative network, synthetic-data budget, normal-image pool, defect types, and training schedule; only the attribute-request distribution differs.

| | | AD Baselines | | | Generic Augmentation | | Targeted Augmentation | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Subset** | **Metric** | PatchCore | WinCLIP | DRAEM | Anomagic | Ours | Muddit | OneCAT | InternVL-U | **Ours** |
| **Overall** | I-ACC ↑ | 86.84 | 75.98 | 87.25 | 84.34 | <u>87.76</u> | 86.76 | 75.59 | 86.66 | **89.58** |
| | P-PRO ↑ | 83.38 | 22.45 | 89.58 | 86.80 | <u>89.84</u> | 87.46 | 78.91 | 89.61 | **91.31** |
| **Hard Cases** | I-ACC ↑ | 70.62 | 69.95 | 73.16 | 73.30 | <u>79.13</u> | 76.49 | 67.61 | 75.76 | **82.39** |
| | P-PRO ↑ | <u>87.36</u> | 18.32 | 84.38 | 84.66 | 84.66 | 85.02 | 74.14 | 84.03 | **88.17** |

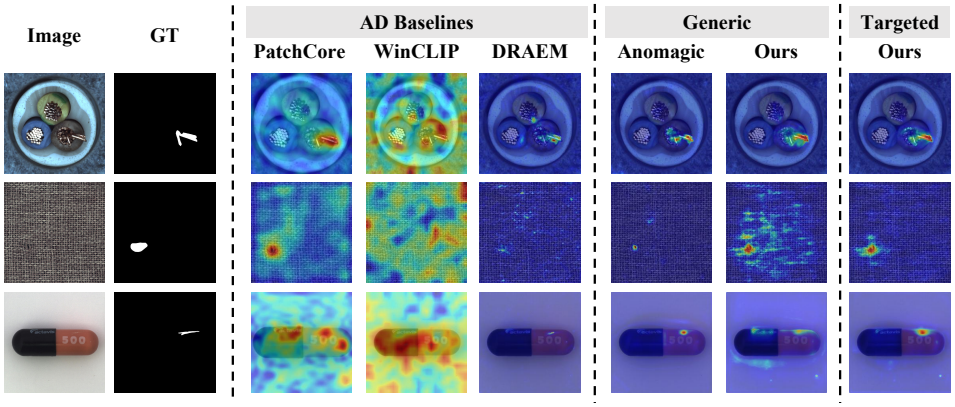



Figure 4: Qualitative comparison on representative hard cases.

**Downstream application on targeted augmentation.** We further study whether fine-grained controllability can be used to repair diagnosed detector failure modes. A fixed hard-case pool is constructed by aggregating the predictions of PatchCore, WinCLIP, and DRAEM and selecting the lowest-performing samples. We analyze the compositional attribute configurations concentrated in these failures and use them to define the request distribution for targeted synthesis. The resulting samples are then used to refine the same DRAEM discriminative network used throughout the downstream comparison. As shown in Tab. 2, targeted synthesis produces stronger refinement gains than generic augmentation under an identical data budget and training protocol. Overall I-ACC/P-PRO improve from 87.76/89.84 to 89.58/91.31, while the gains are amplified on the diagnosed hard cases, from 79.13/84.66 to 82.39/88.17. The substantially larger hard-case recovery demonstrates the practical value of controllability: synthetic data can be deliberately allocated to known coverage gaps in the detector's failure space, rather than simply increasing data volume. Fig. 4 provides a localization-level view of the same effect. The selected examples come directly from the detector-identified hard-case subset. Across these samples, the baseline detectors exhibit either diffuse responses over normal structures or weak localization of the actual defect. Generic augmentation improves anomaly sensitivity, whereas hard-case-targeted augmentation further concentrates the response around the ground-truth region and suppresses unrelated activations. This qualitative behavior is consistent with the +3.51-point P-PRO improvement observed on hard cases.

### 4.4 Ablation Studies on Attribute Control

Tab. 3 examines the contributions of attribute grounding, policy alignment, and semantic steering. Starting from the unified baseline, grounding and policy alignment strengthen training-side attribute control, increasing average requested-attribute alignment from 73.65 to 74.31 while improving IS from 2.10 to 2.17. Semantic steering provides a stronger direct control signal at inference time: when applied alone, it raises the average alignment to 75.10, demonstrating that attribute-level feedback can effectively guide the generation trajectory toward the requested configuration. The three components are complementary. Combining grounding, policy alignment, and semantic steering

Table 3: Ablation of the semantic control components. We evaluate attribute grounding, policy alignment, and semantic steering in terms of requested-attribute alignment and visual quality.

| Components | | | Controllability Alignment | | | | | Quality | |
|---|---|---|---|---|---|---|---|---|---|
| Grounding | Alignment | Steering | Pos. | Size | Shape | Ori. | Avg. | IS | IC-L |
| – | – | – | 88.10 | 66.68 | 62.07 | 77.73 | 73.65 | 2.10 | 0.38 |
| ✓ | – | – | 88.80 | 68.07 | 61.18 | 78.12 | 74.04 | 2.15 | 0.38 |
| ✓ | ✓ | – | 88.58 | 68.16 | 61.85 | 78.64 | 74.31 | 2.17 | **0.39** |
| – | – | ✓ | 89.84 | 70.22 | 62.18 | 78.17 | 75.10 | 2.16 | 0.38 |
| ✓ | ✓ | ✓ | **90.08** | **70.58** | **63.45** | **79.42** | **75.88** | **2.18** | 0.38 |

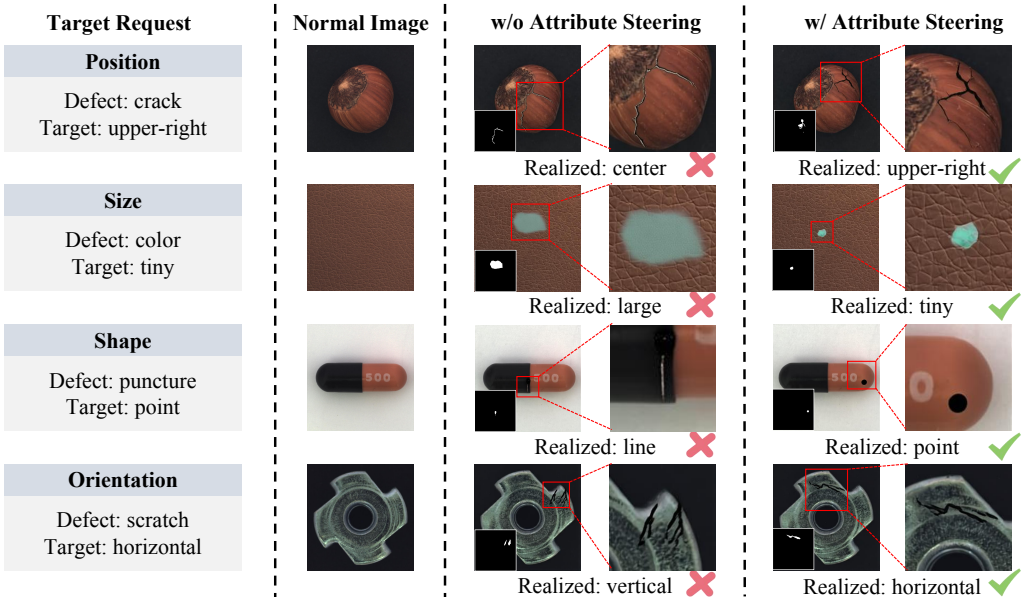



Figure 5: Qualitative examples of semantic steering. Given the same normal image and target request, semantic steering corrects mismatched generated attributes.

achieves the best results on all four attributes, reaching 90.08, 70.58, 63.45, and 79.42 for position, size, shape, and orientation, respectively, and improves the average alignment to 75.88 (+2.23 points). Meanwhile, IS increases to 2.18 and IC-L remains at 0.38, showing that the controllability gain is obtained while maintaining visual quality.

Fig. 5 visualizes how semantic steering acts on individual generations. Without steering, the synthesized defects can remain visually plausible while realizing an incorrect configuration, such as center instead of upper-right, large instead of tiny, or vertical instead of horizontal. Semantic steering consistently moves these outcomes toward the requested attributes across position, size, shape, and orientation, providing a direct qualitative explanation for the controllability gains in Tab. 3.

## 5 Conclusion

In this work, we study zero-shot controllable anomaly generation and present a unified understanding–generation framework for precise attribute control without adapting the pretrained generator. By grounding anomaly attribute understanding, aligning generation through on-policy feedback, and applying inference-time attribute steering, our method achieves strong controllability while preserving cross-dataset generalization and visual quality. Extensive experiments demonstrate the promise of leveraging visual understanding as a reusable control interface for generation. In the future, we will explore richer anomaly attributes and more complex compositional controls on anomaly generation.

## AI USE DISCLOSURE

We used generative AI tools to assist with code debugging, refactoring, and optimization, and with drafting and revising preliminary versions of the manuscript. AI-assisted writing was used primarily to improve clarity, structure, and readability. All AI-generated code was manually inspected and tested by the authors, and all AI-assisted text was reviewed, edited, and verified against the underlying methodology, implementation, and experimental results. We did not rely on generative AI to independently generate experimental results or determine the scientific conclusions of this work. We take full responsibility for the final content of the paper and all AI-assisted artifacts included in it.

## REFERENCES

Aimira Baitieva, David Hurych, Victor Besnier, and Olivier Bernard. Supervised anomaly detection for complex industrial images. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17754–17762. IEEE, 2024.

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, 2019.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023.

Zhewei Dai, Shilei Zeng, Haotian Liu, Xurui Li, Feng Xue, and Yu Zhou. Seas: few-shot industrial anomaly image generation with separation and sharing fine-tuning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23135–23144. IEEE, 2025.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683, 2025. URL https://arxiv.org/abs/2505.14683.

Runpei Dong, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, et al. Dreamllm: Synergistic multimodal comprehension and creation. In *International Conference on Learning Representations*, volume 2024, pp. 6666–6702, 2024.

Guan Gui, Bin-Bin Gao, Jun Liu, Chengjie Wang, and Yunsheng Wu. Few-shot anomaly-driven generation for anomaly classification and segmentation. In *Computer Vision – ECCV 2024*, volume 15141, pp. 210–226, 2024. doi: 10.1007/978-3-031-73010-8_13. URL https://link.springer.com/chapter/10.1007/978-3-031-73010-8_13.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:66923–66939, 2023.

Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, Yabiao Wang, and Chengjie Wang. Anomalydiffusion: Few-shot anomaly image generation with diffusion model. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 8526–8534, 2024.

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*, volume 40, pp. 5485–5493, 2026. doi: 10.1609/aaai.v40i7.37466. URL https://ojs.aaai.org/index.php/AAAI/article/view/37466.

Ying Jin, Jinlong Peng, Qingdong He, Teng Hu, Jiafu Wu, Hao Chen, Haoxuan Wang, Wenbing Zhu, Mingmin Chi, Jun Liu, et al. Dual-interrelated diffusion model for few-shot anomaly image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30420–30429. IEEE, 2025.

Yizhou Jin, Yuezhu Feng, Jinjin Zhang, Peng Wang, Qingjie Liu, and Yunhong Wang. Reasoning-driven anomaly detection and localization with image-level supervision. *arXiv preprint arXiv:2603.27179*, 2026.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9659–9669. IEEE, 2021.

Han Li, Xinyu Peng, Yaoming Wang, Zelin Peng, Xin Chen, Rongxiang Weng, Jingang Wang, Xunliang Cai, Wenrui Dai, and Hongkai Xiong. Onecat: Decoder-only auto-regressive model for unified understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30235–30245, 2026.

Hao Li, Changyao Tian, Jie Shao, Xizhou Zhu, Zhaokai Wang, Jinguo Zhu, Wenhan Dou, Xiaogang Wang, Hongsheng Li, Lewei Lu, et al. Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29767–29779. IEEE, 2025.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7739–7751. IEEE, 2025.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2545–2555. IEEE, 2025.

Haoxiang Rao, Zhao Wang, Chenyang Si, Yan Lyu, Yuanyi Duan, Fang Zhao, and Caifeng Shan. One-to-More: High-fidelity training-free anomaly generation with attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Qingyu Shi, Jinbin Bai, Zhuoran Zhao, Wenhao Chai, Kaidong Yu, Jianzong Wu, Shuangyong Song, Yunhai Tong, Xiangtai Li, Xuelong Li, and Shuicheng YAN. Beyond text-to-image: Liberating generation with a unified discrete diffusion model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=pG0WTde3pR.

Jiaming Su, Tengchao Yang, Ruikang Zhang, Zhengan Yan, Haoyu Sun, and Linfeng Zhang. Anomalyagent: Agentic industrial anomaly synthesis via tool-augmented reinforcement learning. *arXiv preprint arXiv:2604.07900*, 2026.

Han Sun, Yunkang Cao, Hao Dong, and Olga Fink. Unseen visual anomaly generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25508–25517, 2025.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *International conference on learning representations*, volume 2024, pp. 12352–12380, 2024.

Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, Penghao Yin, Sitao Chen, Ganlin Yang, Mingxin Liu, Zirun Zhu, Ziqian Fan, Leyao Gu, Haomin Wang, Qi Wei, Jinhui Yin, Xue Yang, Zhihang Zhong, Qi Qin, Yi Xin, Bin Fu, Yihao Liu, Jiaye Ge, Qipeng Guo, Gen Luo, Hongsheng Li, Yu Qiao, Kai Chen, and Hongjie Zhang. InternVL-U: Democratizing unified multimodal models for understanding, reasoning, generation and editing. arXiv preprint arXiv:2603.09877, 2026. URL https://arxiv.org/abs/2603.09877.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12966–12977. IEEE, 2025.

Jiacong Xu, Shao-Yuan Lo, Bardia Safaei, Vishal M Patel, and Isht Dwivedi. Towards zero-shot anomaly detection and reasoning with multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20370–20382. IEEE, 2025.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Dræm–a discriminatively trained reconstruction embedding for surface anomaly detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8310–8319. IEEE, 2021.

Jinjin Zhang, Guodong Wang, Yizhou Jin, and Di Huang. Towards training-free anomaly detection with vision and language foundation models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15204–15213. IEEE, 2025.

Kai Zhang, Zekai Zhang, Xihe Sun, Anpeng Wang, Jingmeng Nie, Qinghui Chen, Han Hao, Jianyuan Guo, and Jinglin Zhang. Adseeker: A knowledge-grounded reasoning framework for industry anomaly detection and reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21379–21388, 2026.

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. Springer, 2022.

# A Dataset Construction

We use VisA for both training stages and MVTec-AD to evaluate anomaly understanding. Anomaly attributes are derived from real anomalous images and their pixel-level masks.

## A.1 Training data

Each anomalous image selected for the two VisA training sets contains a single anomaly. Grounded attribute understanding (Stage I) uses 1,163 images across 11 categories: 660 normal and 503 anomalous images. Each image is paired with a question about anomaly presence. Anomalous images also receive questions about position, size, and shape. For 129 anomalies with band or thin shapes, orientation is queried as both a categorical label and a numerical angle. Ground-truth answers follow the attribute extraction rules in Section B. Table 4 reports the number of question–answer (QA) pairs for each question type.

For on-policy generation alignment (Stage II), we pair each VisA anomalous image with its most similar normal image to construct conditional editing examples. We apply geometric transformations, including translation, scaling, and stretching, to the reference anomaly mask. The resulting target mask lies within the foreground object region. Target attributes are determined from the reference anomaly and the transformed mask. We use Qwen2.5-VL-7B to generate an editing prompt describing these attributes. The dataset contains 2,683 examples from 11 categories, using 1,879 distinct normal images and 600 distinct anomalous reference images. Table 5 reports the number of examples per category.

## A.2 Anomaly-understanding test data

The anomaly-understanding test set contains 1,420 MVTec-AD images across 15 categories: 467 normal and 953 anomalous images. The anomalous images cover 69 defect types. Using the same question types as Stage I, we construct 4,759 QA pairs. For 240 anomalies with band or thin shapes, orientation is queried as both a categorical label and a numerical angle (Table 6).

# B Attribute Extraction Rules

Let $M$ be a binary anomaly mask and $F$ the corresponding foreground mask. For texture categories, $F$ covers the entire image. We assign anomaly attributes using the following rules.

**Position.** The bounding box of $F$ defines a $3 \times 3$ grid relative to the object. We assign *center* if the mask pixel farthest from the foreground center lies in the central cell. Otherwise, we divide the image into upper, lower, left, and right half-planes through the foreground center. Each direction is selected if its half-plane contains more than half of the mask pixels. The selected directions are combined into one label. The eight noncentral labels are *upper-left*, *upper*, *upper-right*, *left*, *right*, *lower-left*, *lower*, and *lower-right*.

| Question type | Items |
|---|---|
| Presence | 1,163 |
| Position | 503 |
| Size | 503 |
| Shape | 503 |
| Orientation | 258 |
| Total | 2,930 |

Table 4: VisA QA distribution for Stage I.

| Category | Examples | Category | Examples |
|---|---|---|---|
| Candle | 313 | PCB 1 | 223 |
| Cashew | 329 | PCB 2 | 168 |
| Chewing gum | 270 | PCB 3 | 215 |
| Fryum | 202 | PCB 4 | 137 |
| Macaroni 1 | 241 | Pipe fryum | 291 |
| Macaroni 2 | 294 | Total | 2,683 |

Table 5: VisA category distribution for Stage II.

| Question type | Items |
|---|---|
| Presence | 1,420 |
| Position | 953 |
| Size | 953 |
| Shape | 953 |
| Orientation | 480 |
| Total | 4,759 |

Table 6: MVTec-AD test QA distribution.

| Attribute | Prompts |
|---|---|
| Position | 14,264 |
| Shape | 2,518 |
| Size | 14,600 |
| Orientation | 5,118 |
| Total | 36,500 |

Table 7: MVTec-AD generation prompt distribution.

**Size.** Relative size is the mask-to-foreground area ratio $\rho = |M|/|F|$. The labels *tiny*, *small*, *medium*, and *large* have upper bounds $0.005$, $0.02$, $0.06$, and $0.12$, respectively; larger masks are *extensive*.

**Shape.** Let $P$ and $A_{\text{hull}}$ denote the perimeter and convex-hull area of $M$. We characterize shape using the principal-axis ratio $r = \sqrt{\lambda_1/\lambda_2}$, circularity $c = 4\pi|M|/P^2$, solidity $s = |M|/A_{\text{hull}}$, and relative medial thickness $t$. We apply the following rules in order. A mask is *thin* if $r \geq 5$ and $t < 0.20$. Otherwise, it is *irregular* if $s < 0.65$ or $c < 0.20$. Among the remaining masks, we assign *band* if $r \geq 2.5$, then *point* if $c \geq 0.50$, and *blob* otherwise.

**Orientation.** Only band and thin masks receive orientation labels. Their principal-axis angle $\theta$ is normalized to $[0°, 180°)$. Angles below $22.5°$ or at least $157.5°$ are *horizontal*; those in $[67.5°, 112.5°)$ are *vertical*; all others are *diagonal*. For numerical orientation questions, the answer is the integer part of $\theta$.

## C Reward and Cost Matrices

We use the same attribute reward definition for training and evaluation. For attribute $a$, let $y$ denote the target and $\hat{y}$ the prediction. The cost $C_a[y, \hat{y}] \in [0, 1]$ measures their discrepancy. For discrete attributes, targets index matrix rows and predictions index columns. For orientation expressed as a numerical angle, the same notation denotes the continuous cost defined below. The reward is

$$R_a(y, \hat{y}) = 1 - C_a[y, \hat{y}].$$

**Position.** Let $g(p) = (x_p, y_p) \in \{0, 1, 2\}^2$ map each position label to its column and row in the $3 \times 3$ grid. The position cost is

$$\begin{aligned} C_{\text{sq}}(\hat{p}, p) &= \frac{(x_{\hat{p}} - x_p)^2 + (y_{\hat{p}} - y_p)^2}{8} \\ &= \left( \frac{\|g(\hat{p}) - g(p)\|_2}{2\sqrt{2}} \right)^2 . \end{aligned}$$

We override the distance-based cost with 1 when a non-center target is predicted as center:

$$C_{\text{pos}}[p, \hat{p}] = \begin{cases} 1, & \hat{p} = \text{center},\ p \neq \text{center}, \\ C_{\text{sq}}(\hat{p}, p), & \text{otherwise}. \end{cases}$$

The denominator, 8, is the squared distance between opposite corners of the grid. For a center target, we retain the distance-based cost, which is zero for a correct prediction.

**Size.** The ordered labels *tiny*, *small*, *medium*, *large*, and *extensive* have indices $0, \ldots, 4$, giving

$$C_{\text{size}}[s, \hat{s}] = \frac{|i(s) - i(\hat{s})|}{4}.$$

**Shape.** Let $i(s)$ index the class order $(\text{point}, \text{blob}, \text{band}, \text{thin}, \text{irregular})$. With target classes as rows and predicted classes as columns, the symmetric semantic cost matrix is

$$\mathbf{C}_{\text{shape}} = \begin{bmatrix} 0 & 0.25 & 1 & 1 & 0.50 \\ 0.25 & 0 & 0.25 & 1 & 0.50 \\ 1 & 0.25 & 0 & 0.25 & 0.50 \\ 1 & 1 & 0.25 & 0 & 0.50 \\ 0.50 & 0.50 & 0.50 & 0.50 & 0 \end{bmatrix}.$$

The class cost is $C_{\text{shape}}[s, \hat{s}] = (\mathbf{C}_{\text{shape}})_{i(s), i(\hat{s})}$, which assigns partial credit to geometrically related classes.

**Orientation.** Let $j(\text{horizontal}) = 0$, $j(\text{diagonal}) = 1$, and $j(\text{vertical}) = 2$. The orientation cost is

$$C_{\text{dir}}[d, \hat{d}] = \frac{|j(d) - j(\hat{d})|}{2}.$$

For numerical orientation, the axis angle has period $180^\circ$. We use

$$C_{\text{angle}}[\theta, \hat{\theta}] = \frac{\min\{|\hat{\theta} - \theta|,\ 180^\circ - |\hat{\theta} - \theta|\}}{90^\circ}.$$

## D Training and Inference Settings

**Training settings.** We apply LoRA adapters to the language model's attention and feed-forward projections, with default scaling factor $\alpha = 32$ and dropout probability $p = 0.05$. For Stage II, we initialize the adapter from Stage I and continue optimizing it. We provide the reference anomaly as auxiliary conditioning for the first 800 steps of Stage II, then gradually phase it out until step 1,200. No reference anomaly is provided thereafter. For reward evaluation, we use an exponential moving average (EMA) of the Stage II adapter, updated throughout training. We use AdamW, with default weight decay $\lambda_{\text{wd}} = 0$ and gradient clipping at $\|g\|_2 \leq 1$. The GRPO clipping parameter is $\epsilon = 0.2$, and the KL weight is $\beta_{\text{KL}} = 0.02$. We train both stages for 1,500 steps. The best checkpoints are obtained at step 1,400 for Stage I and step 1,250 for Stage II.

**Inference settings.** We generate a $512 \times 512$ image from a normal image and an editing prompt using 20 denoising steps. We apply trajectory steering at steps $8, 10, 12, 14, 16$, indexed from zero. Each selected step uses a latent update step size of 0.01 and at most 10 update rounds. We introduce guidance for the specified attributes once the language model detects an anomaly in the image.

## E Attribute-Directed Generation Prompts

The MVTec-AD generation test set contains 36,500 prompt records, with 500 per anomaly class. Each record specifies one controlled attribute. We summarize the prompt counts for position, shape, size, and orientation in Table 7. The orientation group includes both categorical orientations and numerical angles. We provide example prompts for controllable generation in Table 8.

Table 8: Representative MVTec-AD generation prompts. Defect types use dataset labels with underscores replaced by spaces. Prompt wording is lightly edited for clarity; bold text indicates the controlled attribute.

| Attribute | Class | Defect type | Prompt |
|---|---|---|---|
| Position | Bottle | broken small | Add a small chip to the **upper-right** of the bottle. |
| | | contamination | Add a contamination mark to the **left** of the bottle. |
| | Wood | hole | Add a hole to the **lower-left** of the wood. |
| | | liquid | Add a liquid stain to the **upper** area of the wood. |
| Shape | Leather | color | Add a **blob-shaped** discolored patch to the leather. |
| | | fold | Add a **band-shaped** fold to the leather. |
| | Tile | glue strip | Add a **thin-shaped** glue strip to the tile. |
| | | oil | Add an **irregular** oil stain to the tile. |
| Size | Hazelnut | cut | Add a **tiny** cut to the hazelnut. |
| | | print | Add a **medium** unwanted print to the hazelnut. |
| | Metal nut | bent | Add a **large** bent section to the metal nut. |
| | | color | Add a **small** abnormal-color region to the metal nut. |
| Orientation | Capsule | crack | Add a **horizontally oriented** crack to the capsule. |
| | | scratch | Add a **vertically oriented** scratch to the capsule. |
| | Transistor | bent lead | Add a bent lead at $\mathbf{30}^\circ$ to the transistor. |
| | | cut lead | Add a **diagonal** cut lead to the transistor. |

# F Visualization of Generated Images

## F.1 Single-attribute generation

We present single-attribute generation examples for position, shape, size, and orientation in Figure 6. Within each group, we keep the normal input image and defect type fixed and vary the target value of one attribute in the prompt. These examples illustrate how the generated anomalies change with the specified attribute.

## F.2 Multi-attribute generation on hard cases

We present multi-attribute generation examples on hard cases in Figure 7. Each prompt specifies multiple attribute constraints, such as position and shape or size and shape. We show the generated images alongside the normal inputs and corresponding hard-case images to compare the requested attributes with the generated anomalies.

## F.3 Generation examples during training

Figure 8 shows representative training examples, including the defect type, target attribute values, normal input, reference anomaly image and mask, target mask, and generated candidates. We contrast candidates that pass the visual-quality gate, $r_{\mathrm{tex}} \geq \tau_{\mathrm{tex}}$, with those that fail it.

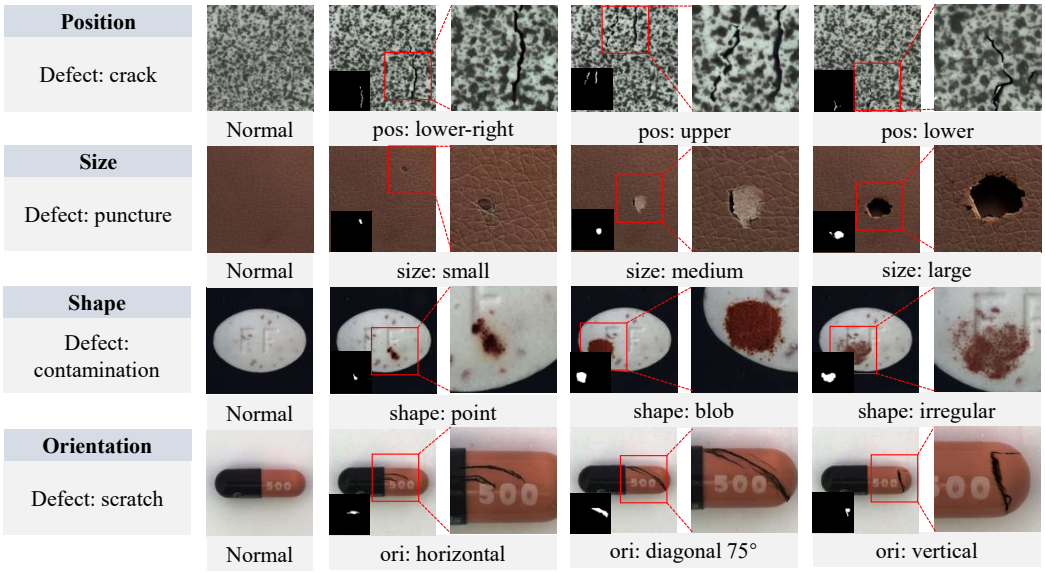



Figure 6: Single-attribute generation examples for position, size, shape, and orientation (top to bottom). Generated images are paired with local close-ups and labeled with target attribute values. Insets show anomaly masks, with anomalies marked in white.

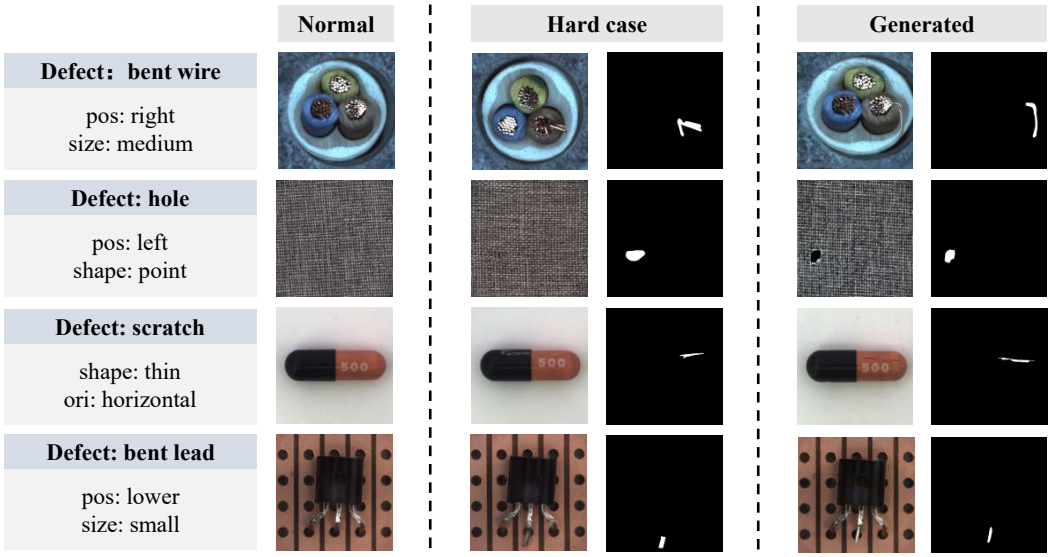



Figure 7: Multi-attribute generation on hard cases. Labels specify defect types and target attribute values. White mask regions indicate anomalies.

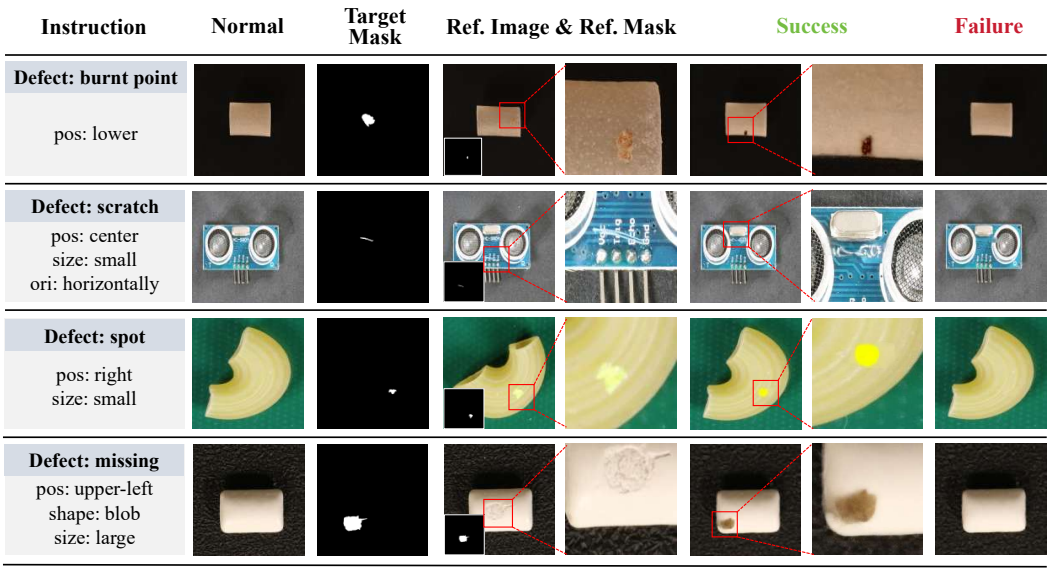



Figure 8: Training examples with defect types, target attribute values, normal inputs, reference images and masks, target masks, and generated candidates.