RISA: RECURSIVE IMPROVEMENT FOR SCENE-CONSISTENT VISUAL ANOMALY SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual anomaly synthesis provides a practical way to supplement scarce real-world anomaly data for industrial inspection and safety monitoring. Although instruction-guided image editing enables controllable anomaly generation from natural-language descriptions, existing methods typically rely on one-shot generation or loosely constrained iterative refinement, making it difficult to correct synthesis errors without disturbing the spatial structure and unaffected content of the original scene. We introduce RISA, a multimodal agent framework for Recursive Improvement for Scene-Consistent Anomaly Synthesis. RISA formulates anomaly synthesis as a progressive process of observation, diagnosis, and repair. At each iteration, the agent evaluates the synthesized image at both global and local scales, identifies discrepancies in anomaly appearance, scale, placement, and integration with the surrounding scene, and adaptively revises its editing strategy based on feedback from previous attempts. To maintain scene consistency throughout refinement, RISA combines multimodal scene understanding with spatially constrained image manipulation, including cropping, localized editing, and compositing, enabling the agent to jointly determine what anomaly to generate, where to place it, and how to integrate it into the original scene while minimizing unintended changes. This recursive inference-time improvement process requires no parameter updates to the underlying models. Experiments across industrial inspection and safety monitoring scenarios demonstrate that RISA consistently improves anomaly fidelity while better preserving the structural and visual consistency of the original scene.

1 INTRODUCTION

Industrial inspection and safety monitoring require the recognition of diverse anomalies, including fine cracks on component surfaces, metallic debris on airport runways, road potholes, and abnormal human states in public spaces. Collecting real examples of these anomalies is essential for developing and evaluating visual recognition systems. However, anomalies occur unpredictably and vary considerably in appearance, location, and environmental conditions, making comprehensive coverage through data collection difficult. Synthesizing specified anomalies in existing real-world images therefore offers a practical means of supplementing scarce samples and constructing diverse test scenarios (Hu et al., 2024; 2026).

Instruction-guided image editing provides a promising approach to controllable anomaly synthesis. Using natural-language descriptions of the desired changes, users can guide models to insert foreign objects, modify local surfaces, or alter human poses and states in real images. Multimodal language models further strengthen the joint understanding of visual content and editing intent, enabling editing instructions to incorporate scene context. These advances provide a foundation for synthesizing diverse anomalies from task descriptions.

Nevertheless, existing approaches mainly follow two paradigms. One line of work (Brooks et al., 2022; Fu et al., 2024) improves instruction following through stronger alignment between language and visual content, but largely relies on guidance established before generation, without explicitly inspecting the output and revising the editing plan. Another line of work (Madaan et al., 2023) uses feedback and iterative refinement to improve initial outputs, allowing subsequent attempts to build on earlier results. When applied to anomaly synthesis, however, iterative refinement without

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.

appropriate spatial constraints may alter unrelated content, disrupting existing spatial relationships and scene structure.

Therefore, an effective anomaly synthesis agent should not only understand the task and execute edits, but also identify discrepancies in previous attempts and use feedback to revise its strategy. The central challenge is to translate visual observations into actionable decisions: determining whether a problem arises from the anomaly itself, its spatial configuration, or its integration into the image, and selecting an appropriate operation accordingly. Refinement must also account for both local details and the global scene, avoiding damage to already satisfactory content. This calls for an integrated process of observation, diagnosis, and repair, in which subsequent attempts can exploit information from previous results.

To this end, we introduce RISA, a multimodal agent framework for controllable anomaly synthesis through recursive self-improvement. Rather than a single execution of an editing instruction, RISA formulates anomaly synthesis as a process of progressive refinement guided by visual feedback. The agent examines generated results at both global and local scales, identifies discrepancies between the synthesized anomaly and the task requirements, and adapts subsequent editing strategies using previous attempts and their feedback. Refinement addresses not only the semantic and morphological accuracy of the anomaly, but also its scale, placement, and integration with the surrounding scene. Through repeated observation, diagnosis, and repair, RISA leverages within-task trial and error to improve anomaly fidelity and visual consistency at inference time, without updating the parameters of the underlying models.

RISA combines multimodal scene understanding with image manipulation tools. The agent jointly interprets the input image and anomaly description to determine a semantically appropriate editing location, spatial extent, and relevant contextual constraints. It then executes the edits through cropping, localized editing, and compositing tools. Local views support fine-grained anomaly editing, while the full image provides a reference for object scale and spatial relationships. This combination of semantic understanding and spatially constrained image manipulation allows the agent to determine both what to generate and where to generate it, while executing and refining edits within explicit spatial constraints. Experiments demonstrate that RISA improves target anomaly fidelity and scene consistency across industrial inspection and safety monitoring scenarios.

Our main contributions are as follows:

- We develop a tool-assisted generation mechanism that combines multimodal scene understanding with localized image editing.
- We propose RISA, a recursive-improvement framework for controllable visual anomaly synthesis that diagnoses visual discrepancies and uses feedback from previous attempts to adjust editing and repair strategies.
- We introduce SQAE-300, a benchmark spanning industrial inspection and safety monitoring, and demonstrate that RISA improves anomaly synthesis quality over baseline methods.

2 Related Work

2.1 Image Generation and Editing

Diffusion models (Rombach et al., 2021; Ho et al., 2020) have become a widely adopted foundation for image generation and editing (Meng et al., 2021), enabling diverse visual transformations through textual conditioning. Instruction-based editing extends these capabilities by allowing users to specify desired changes through natural-language commands. Representative approaches learn from paired editing examples or incorporate multimodal language models to better interpret visual content and editing intent (Brooks et al., 2022; Fu et al., 2024; Huang et al., 2024). Beyond understanding what to change, accurate editing requires determining where a modification belongs and how it should relate to the surrounding scene (Zhang et al., 2023; Li et al., 2023). Attention-controlled and inversion-based editing (Hertz et al., 2022; Mokady et al., 2023), together with mask-guided generation (Avrahami et al., 2023; Ju et al., 2024), improves spatial control and content preservation. Context-aware editing (Kim et al., 2025) further considers the compatibility between instructions and image content, identifying relevant regions while suppressing inappropriate modifications.

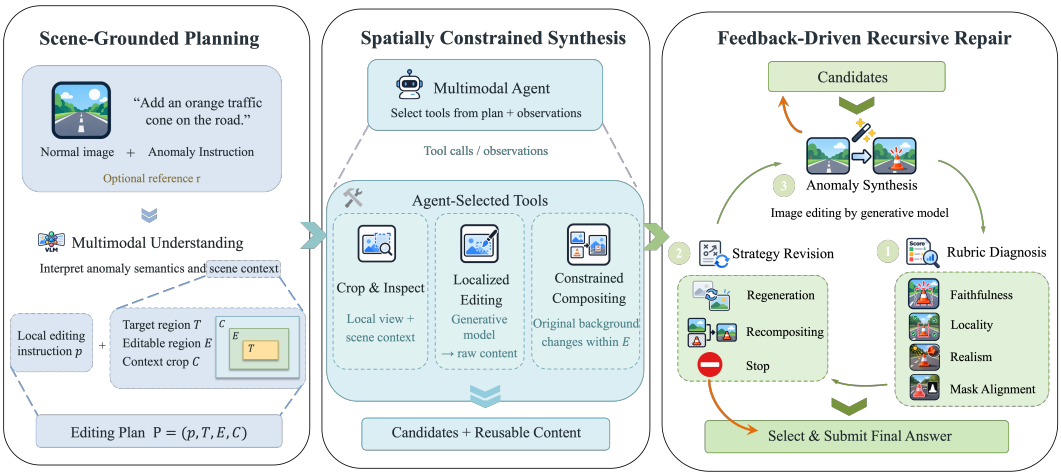


Figure 1: Overview of RISA. **Left:** The agent formulates a scene-grounded editing plan. **Center:** The agent selects image editing tools to synthesize anomalies while preserving non-target content. **Right:** The agent iteratively refines candidates using visual feedback and selects a final result.

This progression from instruction following to contextual understanding motivates agent-based editing, where scene interpretation, tool use, and result inspection are integrated into a sequential decision-making process (Wu et al., 2023; Yao et al., 2022). Multimodal agents can use generated images as feedback to revise subsequent decisions, as demonstrated by iterative approaches to image design (Yang et al., 2023). For localized editing, this perspective connects semantic reasoning with concrete operations such as region selection, cropping, editing, and compositing. It provides a basis for adapting not only the editing instruction but also the spatial scope and sequence of operations, with the aim of achieving the intended change while preserving consistency with the original scene.

2.2 Recursive Self-Improvement

Feedback-driven self-improvement enables models to refine their outputs and adapt their behavior through successive attempts (Gou et al., 2024). Self-Refine (Madaan et al., 2023) alternates between feedback generation and output revision, while Reflexion (Shinn et al., 2023) converts task outcomes into verbal reflections that inform subsequent decisions. These approaches demonstrate how previous attempts can provide useful information for inference-time adaptation without updating model parameters.

In visual generation, iterative refinement extends this principle by incorporating direct inspection of generated images. Idea2Img (Yang et al., 2023) and MIRA (Zeng et al., 2026) use visual feedback to guide successive generation or editing decisions. Such approaches connect multimodal understanding with sequential decision-making: generated results serve not only as outputs but also as evidence for guiding subsequent actions.

Building on this perspective, we consider recursive self-improvement within a single task, where feedback informs successive revisions of the generation strategy rather than changes to model parameters or persistent capabilities. The central challenge is to translate observed discrepancies into actionable decisions. In image editing, this requires determining whether to revise the instruction, adjust the editing region, or modify the compositing operation, while preserving already satisfactory content. This formulation extends feedback-driven refinement from prompt revision to the coordinated adjustment of editing actions under explicit spatial and contextual constraints.

3 METHODS

3.1 PROBLEM SETUP

Given a normal image I, an anomaly description d, and an optional reference image r, AgentAD aims to synthesize the specified anomaly while preserving unrelated image content. A multimodal agent serves as the decision-maker, interpreting the task, inspecting images, and selecting operations through a shared set of image editing tools. The tools perform cropping, localized generation, and compositing, while a visual evaluator provides feedback on the resulting candidates.

The process begins with scene-grounded planning, in which the agent determines the intended change and its spatial scope. Spatially constrained synthesis then translates these decisions into an initial candidate. Finally, feedback-driven recursive repair allows the agent to inspect the result and selectively revise the editing instruction, spatial configuration, or composition. Previous candidates and their feedback remain available throughout the task, enabling subsequent decisions to build on earlier attempts without updating model parameters, as illustrated in Figure 1.

3.2 SCENE-GROUNDED PLANNING

Introducing an anomaly into an existing image requires determining not only what to generate, but also where the modification belongs and which surrounding content should remain unchanged. The first stage translates the anomaly specification into an executable editing plan grounded in the input scene. The multimodal agent identifies the relevant object or surface, determines the intended modification and its spatial extent, and identifies the surrounding content to be preserved. This tool-assisted planning process is summarized as

$$P_0 = \mathrm{MLLM}_\theta(I, d, r), \qquad P_0 = (p_0, T_0, E_0, C_0),$$

where θ denotes the fixed model parameters, p_0 is the local editing instruction, and T_0, E_0, and C_0 specify the target region, editable region, and context crop, respectively. The planning process combines the agent's semantic decisions with tool-assisted spatial preparation.

The three regions serve distinct but complementary purposes:

$$T_0 \subseteq E_0 \subseteq C_0.$$

The target region T_0 describes the intended spatial extent of the anomaly. The editable region E_0 defines where modifications are permitted and may extend beyond T_0 to accommodate necessary local transitions, such as contact shadows or changes along a damaged surface. The context crop C_0 includes surrounding visual information that supports generation, such as nearby structures, surface appearance, and lighting conditions. Importantly, providing this additional context does not expand the region that may be modified in the final image.

The agent specifies the target and editable regions in the original image coordinates and selects a context strategy, while the cropping tool constructs the corresponding model input and handles coordinate transformations. The agent can inspect the prepared crop before generation and revise the instruction or spatial configuration if necessary.

3.3 SPATIALLY CONSTRAINED SYNTHESIS

The agent executes its editing plan through a set of image manipulation tools, including crop preview, localized editing, and recompositing. These tools allow the agent to inspect the prepared input, request new content, or reuse an existing local result. Then, the agent selects the operation and its parameters, while the tools handle coordinate transformations, image model invocation, and spatially constrained composition.

Before requesting an edit, the agent can invoke the crop-preview tool to inspect the context region C_0. The tool extracts the region from the original image and applies resizing and, when necessary, padding to prepare the model input. This preview allows the agent to assess whether the selected region provides sufficient context and to revise the spatial plan before generation.

When the agent invokes the localized editing tool, the tool calls the generative model with the prepared crop and editing instruction:

$$J_0 = \mathcal{G}\big(\mathcal{T}_{C_0}(I), p_0\big),$$

where $\mathcal{G}$ denotes the generative model and $\mathcal{T}_{C_0}$ maps the context region to the model input. The generated result is then mapped back to the original image coordinates, yielding $\widetilde{J}_0$, and composited as

$$I_0 = (1 - A_0) \odot I + A_0 \odot \widetilde{J}_0, \qquad \mathrm{supp}(A_0) \subseteq E_0.$$

Here, I_0 is the initial candidate and A_0 is the compositing weight map, and $\odot$ denotes element-wise multiplication. Restricting its support to E_0 preserves the original pixels outside the permitted editing region, regardless of changes introduced elsewhere in the generated crop.

The tools retain the raw generated content and its spatial mapping alongside each candidate. The agent can subsequently invoke the recompositing tool to reuse a historical result and adjust its composition within the compatible source crop, without requesting another generation. All candidates are composited against the original image, avoiding the accumulation of background changes across iterations. This separation between agent-selected operations and tool-enforced spatial constraints supports flexible editing decisions while preserving non-target content.

3.4 Feedback-Driven Recursive Repair

The initial candidate may satisfy the spatial constraints while still deviating from the requested anomaly in appearance, placement, or visual integration. RISA addresses these discrepancies through recursive self-improvement: the agent inspects generated results, interprets visual feedback, and revises subsequent editing decisions. This process operates within a single task without updating model parameters.

Rubric-Guided Visual Diagnosis. Effective repair requires identifying why a candidate falls short of the task requirements, rather than relying on an overall quality score alone. We employ GPT-5.5 as the visual evaluator to assess each candidate under a predefined rubric:

$$f_t = (\mathbf{q}_t, \mathcal{D}_t) = \mathrm{MLLM}(I, I_t, M_t, d, \mathcal{L}_t),$$

where I is the original image, I_t and M_t are the candidate and its generated mask, d is the anomaly description, and $\mathcal{L}_t$ denotes supplementary inspection views, including aligned crops, mask overlays, and pixel differences. The feedback f_t consists of two complementary components: numerical scores $\mathbf{q}_t$ and corresponding textual diagnoses $\mathcal{D}_t$.

The numerical assessment comprises four scores in $[0, 1]$:

$$\mathbf{q}_t = (F_t, L_t, R_t, M_t^{\mathrm{align}}),$$

corresponding to the following criteria:

- **Faithfulness (F_t).** Agreement between the generated anomaly and the task description in appearance, location, scale, and physical placement.
- **Locality (L_t).** Preservation of the original image content outside the generated mask.
- **Realism (R_t).** Visual consistency with the surrounding scene in lighting, perspective, geometry, material, occlusion, and shadows.
- **Mask alignment (M_t^{align}).** Agreement between the generated mask and the actual image changes, penalizing both excessive coverage and omitted changes. This criterion measures change-mask consistency rather than accuracy against a ground-truth semantic anomaly mask.

The textual diagnoses $\mathcal{D}_t$ complement these scores by explaining the visual evidence underlying the assessment. Each finding identifies the evaluation criterion, the degree to which the corresponding requirement is satisfied, and the observation supporting that judgment. These fields provide an explicit basis for repair, such as incorrect placement, object truncation, or an unnatural compositing boundary, without prescribing a particular editing action.

For example, an image may faithfully depict the requested object but exhibit an unnatural boundary, indicating that its composition may require adjustment rather than its content being regenerated. Each finding receives a program-assigned identifier that subsequent repair actions can reference. These references establish a traceable link between an observed discrepancy and the intended modification; they do not themselves establish that the problem has been resolved. The resulting candidate is assessed again under the same task requirements and scoring criteria.

Feedback-Conditioned Strategy Revision. The agent uses diagnostic feedback to determine whether further inspection, regeneration, or recompositing is needed. At iteration t, the available history consists of previous editing plans, operations, generated results, and their evaluations:

$$H_t = \big[(P_i, J_i, I_i, f_i)\big]_{i=0}^{t}.$$

Conditioned on this history and the remaining budget, the agent constructs an updated plan when further repair is needed:

$$P_{t+1} = \mathrm{MLLM}_\theta(I, d, r, H_t).$$

The updated plan adapts the editing instruction and spatial configuration based on previous outcomes, selecting between two forms of repair:

- **Regeneration.** When the generated content does not adequately satisfy the task, the agent revises the editing instruction or spatial configuration and invokes the localized editing tool. The updated plan P_{t+1} produces a new local result:

  $$J_{t+1} = \mathcal{G}\big(\mathcal{T}_{C_{t+1}}(I), p_{t+1}\big).$$

  The result is mapped to the original image coordinates and composited within E_{t+1}.
- **Recompositing.** When existing local content can be retained, the agent selects a historical candidate $k \le t$ and adjusts its composition:

  $$I_{t+1} = (1 - A_{t+1}) \odot I + A_{t+1} \odot \widetilde{J}_k, \qquad \mathrm{supp}(A_{t+1}) \subseteq E_{t+1} \subseteq C_k.$$

  Here, $\widetilde{J}_k$ is recovered using the stored spatial mapping of candidate k. This operation reuses the generated content without another image-model call, with $J_{t+1} = J_k$ and its original spatial mapping retained.

For each targeted repair, the agent can reference the candidate and diagnostic findings that motivate the modification and describe the intended change. The tools validate these references and separately record the source of reused image content, since the candidate informing a decision need not be the one used for recompositing.

After execution, the new candidate is evaluated under the same rubric, and the resulting record is appended to the history:

$$H_{t+1} = H_t \,\|\, (P_{t+1}, J_{t+1}, I_{t+1}, f_{t+1}),$$

where $\|$ denotes sequence append. This recursive process allows subsequent decisions to incorporate the outcomes of earlier repairs, without assuming that every modification improves the result.

Candidate Selection and Termination. All candidates are retained because additional edits may degrade the result. The agent may stop early when a candidate adequately satisfies the task requirements; once the editing budget is exhausted, it must select an existing valid candidate. The final output need not be the most recent result. The tools validate the selection and preserve its operation history.

4 Experiments

4.1 Experimental Setup

Datasets. To address the need for instruction-guided anomaly editing benchmarks that explicitly account for surveillance viewpoints, we introduce **SQAE-300** (Safety and Quality Anomaly Editing), combining safety-monitoring scenarios with industrial inspection. It comprises 300 editing tasks constructed from 66 source images: 180 safety-related tasks from PRW (Zheng et al., 2017), Mall (Chen et al., 2012), SCUT-HEAD Part A (Peng et al., 2018), and ShanghaiTech Part B (Zhang et al., 2016), and 120 industrial defect tasks from MVTec AD (Bergmann et al., 2019). Each task specifies an anomaly and constraints on its location, scale, and preservation of unrelated content. The benchmark contains 244 text-guided and 56 reference-guided tasks, covering safety hazards and localized industrial defects.

Table 1: **Visual quality metrics on SQAE-300**, across reasoning models and agent harnesses. Higher scores indicate better performance.

Method	Reasoning Model	Harness	SQAE-300					
			F (%) ↑	L (%) ↑	R (%) ↑	M (%) ↑	P-SSIM (%) ↑	Local IS ↑
Direct Editing	–	–	73.24	48.27	72.08	9.84	59.54	6.71
Best-of-K	–	–	75.34	55.27	74.97	13.48	62.86	6.86
Tool-Augmented Agent	GPT-5.5	Codex	73.86	94.32	73.88	81.65	65.61	7.55
		Kimi	73.93	93.11	72.81	79.29	62.59	7.46
		Pi	74.30	94.08	73.66	81.30	64.26	7.28
	Kimi-K3	Codex	73.33	95.39	72.78	83.34	65.60	7.63
		Kimi	75.04	94.71	73.30	82.20	64.69	7.78
		Pi	73.70	94.57	73.52	82.23	65.16	7.50
	DeepSeek-v4-flash	Codex	73.86	95.40	71.98	83.16	61.95	6.75
		Kimi	75.12	95.32	73.07	83.35	66.05	7.71
		Pi	71.11	94.49	70.99	82.06	60.62	7.42
	Average		73.81	94.60	72.89	82.06	64.06	7.44
RISA (Ours)	GPT-5.5	Codex	80.31	95.02	74.17	80.74	82.22	6.75
		Kimi	79.93	95.20	75.54	82.91	87.63	7.99
		Pi	80.66	95.61	76.49	84.00	86.95	7.86
	Kimi-K3	Codex	82.73	96.17	77.83	84.64	88.07	7.27
		Kimi	79.33	94.99	75.79	83.06	77.29	7.26
		Pi	78.90	95.04	75.52	83.15	75.89	7.82
	DeepSeek-v4-flash	Codex	82.92	96.32	77.64	85.36	89.08	7.94
		Kimi	81.60	96.14	76.79	85.23	87.26	7.00
		Pi	81.20	95.63	76.16	83.66	82.68	7.80
	Average		80.84	95.57	76.21	83.64	84.12	7.53

Comparison Methods. We compare four methods: (1) **Direct Editing**, which performs a single full-image edit without agent intervention; (2) **Best-of-K**, which independently generates K candidates and selects one using a fixed selection rule, with the same maximum image-generation allowance as RISA;; (3) **Tool-Augmented Agent**, which retains the editing tools and spatial pipeline without feedback-driven recursive repair; and (4) **RISA (Ours)**, which uses the complete agent-driven editing and refinement framework. Comparisons share the same task inputs and editing backbone.

Evaluation Metrics. We report the four GPT-5.5 evaluation scores defined in Section 3.4: **Faithfulness**, **Locality**, **Realism**, and **Mask Alignment**. Each score lies in $[0, 1]$ and is reported on a 0–100 scale; higher is better. We additionally report two offline computational metrics.

- **P-SSIM** (Wang et al., 2004) measures structural preservation in protected regions:

$$\text{P-SSIM} = \rho \cdot \frac{1}{|\mathcal{W}_P|} \sum_{w \in \mathcal{W}_P} \text{SSIM}(X_w, Y_w), \tag{1}$$

  where ρ is the pass rate defined in Appendix B, X and Y are the source and edited images, and $\mathcal{W}_P$ contains windows entirely outside a shared, task-defined editing region.

- **Local IS**, following TF-IDG (Xu et al., 2025), applies Inception Score (Salimans et al., 2016) to local anomaly regions. We use shared task-region crops z_i for paired comparisons:

$$\text{Local IS} = \exp\left(\frac{1}{N}\sum_{i=1}^{N} D_{\text{KL}}\big(p(y \mid z_i) \,\|\, \bar{p}(y)\big)\right), \tag{2}$$

  where $\bar{p}(y) = N^{-1}\sum_i p(y \mid z_i)$. All methods use identical evaluation regions and preprocessing. Local IS reflects image quality via both semantic correctness and visual sharpness.

Models and Implementation Details. We use GPT Image 2 as the image editing model and GPT-5.5 as the visual evaluator, examining different reasoning models and agent harnesses under a shared task interface.

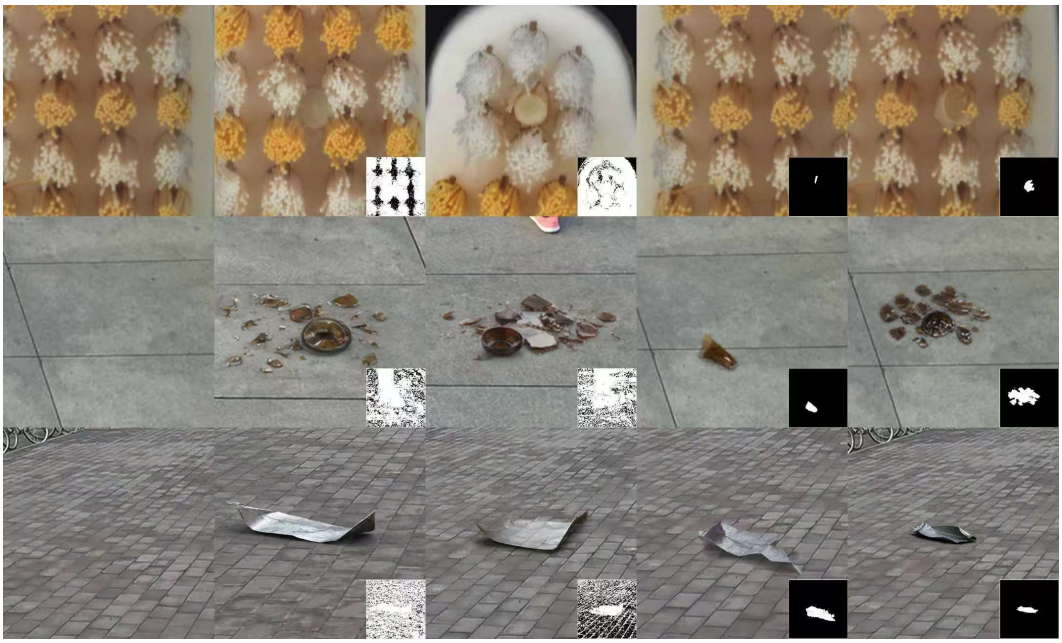


Figure 2: **Qualitative comparison of anomaly synthesis.** Columns show the input, Direct Editing, Best-of-K, Tool-Augmented Agent, and RISA. Rows show missing bristles, broken glass, and metal debris. Each generated image includes its output mask as a bottom-right inset.

A CLI-based harness manages the interaction context, including tool calls and observations. Each task runs in an isolated Docker container, with filesystem permissions separating the agent's writable workspace from authoritative experiment records. The agent accesses task inputs, candidate images, and diagnostic feedback through shared task tools, which also execute editing operations and validate result submission. Candidate outputs and operation histories are persisted independently of the agent session, supporting traceable execution and task-state reconstruction after interruption.

RQ1: Does RISA improve scene-consistent anomaly synthesis? Table 1 highlights the complementary benefits of controlled tool use and recursive improvement. For agent-based methods, we report averages over the same nine model–harness configurations.

Direct Editing achieves faithfulness and realism scores of 73.24 and 72.08, respectively, indicating that the editing model can generate relevant anomalies with plausible visual appearance. However, its locality of 48.27 reveals substantial limitations in preserving unrelated content, while its mask alignment of 9.84 indicates poor correspondence between the output masks and actual image changes. The Tool-Augmented Agent increases locality to 94.60 and mask alignment to 82.06. In contrast, faithfulness and realism improve only modestly, to 73.81 and 72.89. These results suggest that spatially controlled tool execution substantially improves content preservation but is insufficient to fully address deviations in anomaly semantics and appearance.

RISA further improves anomaly fidelity and scene compatibility through feedback-driven recursive repair. Compared with the Tool-Augmented Agent, it increases average faithfulness from 73.81 to 80.84 and realism from 72.89 to 76.21, yielding gains of 7.03 and 3.32 points, respectively. Locality and mask alignment also increase to 95.57 and 83.64. The computational metrics complement these evaluator-based results: P-SSIM increases from 64.06 to 84.12, supporting improved structural preservation in protected regions, while Local IS increases modestly from 7.44 to 7.53. Together, these results support the complementary roles of tool-mediated spatial control and recursive strategy revision in achieving scene-consistent anomaly synthesis.

Under the same maximum image-generation budget, RISA exceeds Best-of-K by 5.50 points in faithfulness and 1.24 points in realism, while achieving substantially higher locality (95.57 versus 55.27) and mask alignment (83.64 versus 13.48). It also achieves higher P-SSIM (84.12 ver-

Table 2: **Downstream anomaly segmentation**. Image-level AUROC and pixel-level AUROC for a U-Net trained on synthetic data from each generation method. Best results in bold.

Method	Image AUROC ↑	Pixel AUROC ↑
Direct Editing	48.44	50.07
Best-of-K	59.50	66.10
Tool-Augmented Agent	68.92	59.01
RISA	**70.31**	**81.00**

sus 62.86) and Local IS (7.53 versus 6.86). This comparison shows that the complete framework offers benefits beyond independent sampling and selection, while the comparison with the Tool-Augmented Agent more directly demonstrates the additional value of recursive repair.

RQ2: Are the benefits consistent across reasoning models and agent harnesses? RISA improves faithfulness, locality, realism, and P-SSIM over the Tool-Augmented Agent in all nine model–harness configurations. With GPT-5.5 fixed, faithfulness increases by 6.45, 6.00, and 6.36 points under Codex, Kimi, and Pi, respectively. Realism also improves under all three harnesses, while locality remains above 95. These consistent gains indicate that the benefits are not confined to a particular agent execution environment.

The improvements also extend across reasoning models. Averaged over the three harnesses, the faithfulness gains are 6.27 points for GPT-5.5, 6.30 for Kimi-K3, and 8.54 for DeepSeek-v4-flash. DeepSeek-v4-flash with Codex achieves the highest faithfulness score of 82.92, whereas Kimi-K3 with Codex achieves the highest realism score of 77.83. Although mask alignment and Local IS do not improve uniformly across configurations, the consistent gains in task fidelity, scene compatibility, and structural preservation support the applicability of recursive refinement across different reasoning models and harnesses.

RQ3: How do the methods differ visually? Figure 2 compares representative results using GPT-5.5 with the Codex harness for the agent-based methods. Direct Editing and Best of K produce recognizable anomalies but also alter surrounding content, whereas the Tool-Augmented Agent better preserves the scene but can underexpress the requested change: the missing bristle tuft is barely visible, and the requested glass pile appears as a single fragment. RISA produces more explicit tuft removal, a recognizable pile of broken glass, and metal debris with a more appropriate scale relative to the paving. These examples illustrate how recursive refinement can improve anomaly fidelity and scene compatibility while retaining localized edits.

4.2 Downstream Anomaly Segmentation

To assess the downstream utility of the synthesized anomalies (Zavrtanik et al., 2021), we train a binary U-Net segmentation model (Ronneberger et al., 2015) using synthetic anomalous images and their corresponding generated masks, together with an equal number of defect-free images from the corresponding dataset, assigned all-zero masks. For the agent-based methods, we report results using GPT-5.5 with the Codex harness. We train the segmentation model for 100 epochs with a learning rate of 10^{-4}, keeping the training protocol identical across generation methods. Evaluation is performed on a subset of the official real-image test splits of MVTec AD, selecting for our target application scenarios. Table 2 reports the segmentation results, with RISA achieving the highest image- and pixel-level AUROC among the compared methods.

5 Conclusion

We presented RISA for scene-consistent visual anomaly synthesis. Multimodal planning, spatially constrained editing, and feedback-driven repair enable recursive strategy revision without model parameter updates. Experiments on SQAE-300 demonstrate improved anomaly fidelity and scene consistency across reasoning models and agent harnesses. Future work will improve feedback reliability and explore transferring repair experience across tasks.

AI USE STATEMENT

In our experiments, we used large language models, including GPT-5.5, DeepSeek-v4-flash, and Kimi-K3, as well as image generation models such as GPT Image 2, as components of the evaluated systems and anomaly synthesis pipeline, as described in the paper.

Separately, generative AI tools were used only to assist with language polishing and manuscript revision. All technical content, experimental design, implementation, analysis, and reported results were reviewed and verified by the authors. We take full responsibility for the final manuscript, code, data, and reported results.

REFERENCES

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Trans. Graph.*, 42 (4), jul 2023. ISSN 0730-0301. doi: 10.1145/3592450. URL https://doi.org/10.1145/3592450.

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)*, June 2019.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *Bmvc*, volume 1, pp. 3, 2012.

Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding Instruction-based Image Editing via Multimodal Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2024.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Nan Duan, Weizhu Chen, et al. Critic: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, volume 2024, pp. 57734–57811, 2024.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jinwei Hu, Francesco Borsatti, Arianna Stropeni, Davide Dalle Pezze, Manuel Barusco, and Gian Antonio Susto. Mirage: Model-agnostic industrial realistic anomaly generation and evaluation for visual anomaly detection. *arXiv preprint arXiv:2603.13507*, 2026.

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.

Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8362–8371, 2024.

Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion, 2024.

Hyunseung Kim, Chiho Choi, Srikanth Malla, Sai Padmanabhan, Saurabh Bagchi, and Joon Hee Choi. Camila: Context-aware masking for image editing with language alignment. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 20998–21017. Curran Associates, Inc., 2025. doi: 10.52202/

085713-0707. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/1df27e31a0c0b916797a49198cb28d6a-Paper-Conference.pdf.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22511–22521. IEEE, 2023.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6038–6047. IEEE, 2023.

Dezhi Peng, Zikai Sun, Zirong Chen, Zirui Cai, Lele Xie, and Lianwen Jin. Detecting heads using feature refine net and cascaded multi-scale architecture. *arXiv preprint arXiv:1803.09256*, 2018.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.

Ruyi Xu, Yen-Tzu Chiu, Tai-I Chen, Oscar Chew, Yung-Yu Chuang, and Wen-Huang Cheng. Training-free industrial defect generation with diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24214–24223, 2025. doi: 10.1109/ICCV51701.2025.02244.

Zhengyuan Yang, Jianfeng Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. Idea2img: Iterative self-refinement with gpt-4v(ision) for automatic image design and generation. *arXiv preprint arXiv:2310.08541*, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

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.

Ziyun Zeng, Hang Hua, and Jiebo Luo. Mira: Multimodal iterative reasoning agent for image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9563–9573, 2026.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824. IEEE, 2023.

Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1367–1376, 2017.