

FailureMem: A Failure-Aware Multimodal Framework for Autonomous Software Repair

Ruize Ma^{1,2*}, Yilei Jiang^{3*}, Shilin Zhang^{1,2*}, Zheng Ma^{2†}, Yi Feng^{1‡}

Zhi Wang¹, Xiangyu Yue³, Chuanyi Li¹, Vincent Ng⁴, Lewei Lu²

¹State Key Laboratory for Novel Software Technology, Nanjing University

²Sensetime Research

³Multimedia Laboratory, The Chinese University of Hong Kong

⁴Human Language Technology Research Institute, University of Texas at Dallas

ruizema@smail.nju.edu.cn, fy@nju.edu.cn

Abstract

Multimodal Automated Program Repair (MAPR) extends traditional program repair by requiring models to jointly reason over source code, textual issue descriptions, and visual artifacts such as GUI screenshots. While recent LLM-based repair systems have shown promising results, existing approaches face several limitations: rigid workflow pipelines restrict exploration during debugging, visual reasoning is often performed over full-page screenshots without localized grounding, and failed repair attempts are rarely transformed into reusable knowledge. To address these challenges, we propose FailureMem, a multimodal repair framework that integrates three key mechanisms: a hybrid workflow-agent architecture that balances structured localization with flexible reasoning, active perception tools that enable region-level visual grounding, and a Failure Memory Bank that converts past repair attempts into reusable guidance. Experiments on SWE-bench Multimodal demonstrate FailureMem improves the resolved rate over GUIRepair by 3.7%. Our code is available at <https://github.com/Ruize-Ma/FailureMem>.

1 Introduction

Software defects are an inevitable part of software development. Automated Program Repair (APR) aims to automatically generate patches that fix buggy programs while preserving intended behavior, typically using signals such as source code and test cases to synthesize code modifications (Huang et al., 2023, 2025a; Jiang et al., 2023; Wu et al., 2023; Huang et al., 2025b).

Recent advances in large language models (LLMs) have significantly expanded the capability of APR systems. Leveraging their strong

code understanding and generation abilities, LLM-based approaches have progressed from function-level repair via fine-tuning and prompting (Huang et al., 2023; Jiang et al., 2023; Xia and Zhang, 2022, 2024) to repository-scale issue resolution, where autonomous agents analyze entire codebases and generate patches for real-world software issues (Yang et al., 2024b; Xia et al., 2025; Zhang et al., 2024). Benchmarks such as SWE-bench have been proposed to evaluate this capability by requiring models to resolve GitHub issues through patch generation that passes repository test suites.

However, real-world debugging often relies on information beyond code and text. Many issue reports include visual artifacts, such as screenshots of graphical user interfaces, diagrams, or UI mockups, which convey important clues about the intended system behavior. To capture this richer context, recent work introduced SWE-bench Multimodal, which augments issue instances with visual information, giving rise to the task of Multimodal Automated Program Repair (MAPR). Compared with text-only settings, solving such tasks requires models to jointly reason over code, natural language descriptions, and visual observations, making the repair problem substantially more challenging.

Formally, MAPR can be defined as follows: given a repository, a problem description (e.g., a GitHub issue), and associated visual artifacts (e.g., screenshots), the goal is to generate a patch that modifies the repository to resolve the reported issue. Figure 1 illustrates a representative instance. The issue report describes a rendering bug in a Nav component and includes a screenshot showing the faulty sidebar; the task is to generate a patch that fixes the visual defect reflected in the interface. Existing methods, such as GUIRepair (Huang et al., 2025c), have taken an important step toward addressing this problem by incorporating visual inputs into the repair process. However, current approaches still face several practical challenges.

*Equal contribution.

†Project leader.

‡Corresponding author.

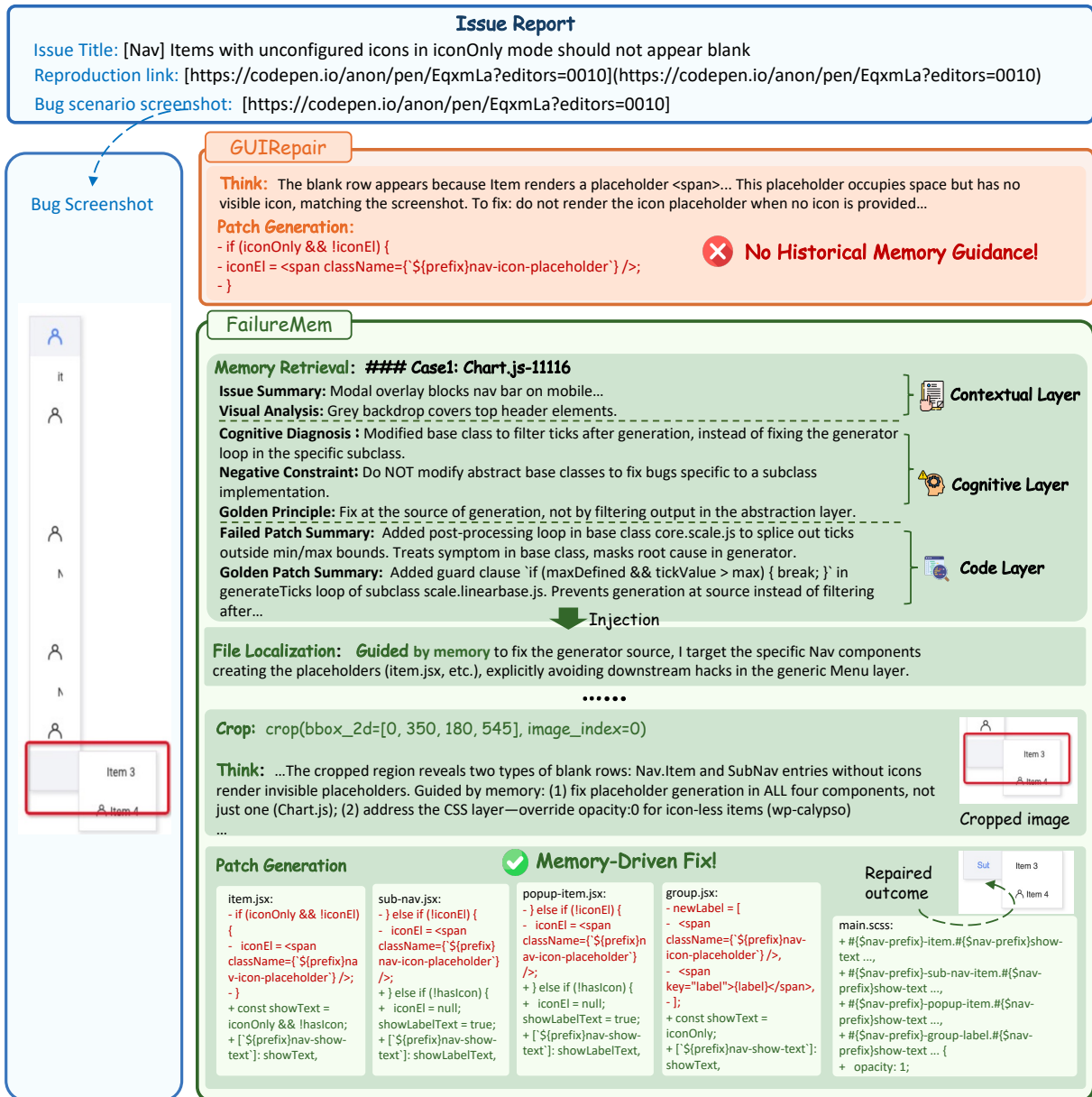


Figure 1: GUIRepair vs. FailureMem on alibaba-fusion/next-895. FailureMem retrieves historical repair guidance, crops the screenshot to isolate the bug region, and produces a comprehensive five-file patch.

Challenge 1. Existing methods struggle to balance structured reasoning and exploratory search. Workflow-based approaches follow predefined pipelines (e.g., issue understanding, file localization, patch generation), which efficiently constrain the search space but may overlook alternative hypotheses. In contrast, agentic systems enable flexible exploration via autonomous reasoning, yet they may lose direction during multi-step exploration and repeatedly inspect irrelevant components.

Challenge 2. Visual reasoning is often performed over full-page screenshots, which distribute attention across many irrelevant interface regions. In

Figure 1, the screenshot contains multiple UI elements, including the main content area, navigation links, and surrounding layout components, while the actual defect is localized to the sidebar rendering of the Nav component. Processing the entire screenshot equally can dilute the model’s attention and make it harder to isolate the precise visual anomaly that should guide the repair process.

Challenge 3. Current repair agents rarely transform previous repair attempts into reusable knowledge. For instance, when attempting to fix the sidebar rendering issue, a repair agent may generate several candidate patches that modify layout prop-

erties or component structure. However, these attempts are typically treated as isolated trials rather than structured experiences. As a result, the agent cannot effectively retrieve analogous repair cases, analyze why earlier fixes failed (e.g., modifying the wrong layout container), or apply concrete repair pattern, such as adjusting component-level layout constraints, to guide subsequent repair attempts.

To address these challenges, we propose **FailureMem**, a multimodal repair framework for GUI issue resolution. FailureMem combines three key designs. First, it introduces a **hybrid workflow-agent architecture** that keeps file and element localization within a structured workflow while reserving an agentic loop for patch generation, enabling controlled search with flexible exploration. Second, it improves visual grounding through **active perception tools**, including *Crop* and *Grounding*, together with an interactive *Bash* environment that allows the agent to inspect repositories and verify assumptions before editing code. Third, it incorporates a **Failure Memory Bank** that transforms historical repair trajectories into reusable guidance via contextual, cognitive, and code-level representations. We evaluate FailureMem on SWE-bench Multimodal. With GPT-5.1, it improves the resolved rate over GUIRepair (a SOTA MAPR model) by 3.7%.

Our contributions are four-fold: (1) We propose a hybrid workflow-agent architecture for multimodal program repair that balances structured localization with flexible patch exploration; (2) We introduce active perception tools for localized visual grounding together with an interactive repository exploration environment; (3) We design a Failure Memory Bank that converts historical repair trajectories into reusable guidance for both localization and patch generation; (4) Experiments on SWE-bench Multimodal demonstrate consistent improvements over strong baselines.

2 A Motivating Example

We use an example from SWE-bench Multimodal (Figure 1) to illustrate the challenges of MAPR and preview how FailureMem addresses them.

Issue `alibaba-fusion/next-895` reports that in the `Nav` component’s `iconOnly` mode, items without an `icon` prop render as blank rows in the vertical sidebar. The screenshot shows icon entries interspersed with empty horizontal bars.

How GUIRepair Fails. GUIRepair localizes two files (`item.jsx` and `nav.jsx`) but misses other affected components and the `Nav` stylesheet. Its patch modifies only `item.jsx`, removing the placeholder logic for `Nav.Item`. However, the bug spans two layers: four components generate invisible placeholders (JS layer), and a global CSS rule hides labels under `iconOnly` mode (CSS layer). The patch therefore fixes one component in one layer and fails regression tests.

This failure reflects the three challenges from Section 1. First, GUIRepair follows a fixed localization workflow that restricts exploration, causing it to miss the broader set of affected components and style rules. Second, visual reasoning operates on the full-page screenshot without region-level focus, making it difficult to associate the visual symptom (empty rows between icons) with placeholder elements generated by multiple `Nav` components. Third, repair attempts are treated as isolated trials, preventing the system from leveraging historical repair experience or reusable repair patterns.

How FailureMem Works. FailureMem first retrieves historical cases from a Failure Memory Bank (Challenge 3). Two cases provide useful guidance. A `Chart.js` axis-tick bug emphasizes fixing errors at the source of generation rather than filtering outputs in abstraction layers. A `wp-calypso` layout bug advises targeted CSS adjustments instead of modifying reusable component logic.

Guided by these principles, the model localizes all four affected components and the `Nav` stylesheet, going beyond the two files identified by GUIRepair. Unlike GUIRepair’s rigid workflow—where early localization errors propagate through the pipeline, FailureMem combines structured workflow with agentic reasoning, allowing the agent to revisit earlier steps and expand the search when inconsistencies arise (Challenge 1).

During patch generation, the agent uses visual cropping to isolate the sidebar region, confirming that blank rows correspond to icon-less entries rendering invisible placeholders (Challenge 2). The resulting patch modifies five files: it removes placeholder generation from four components (`Item`, `SubNav`, `PopupItem`, `Group`), introduces a `nav-show-text` CSS class for icon-less items, and adds an `opacity: 1` override in `main.scss`. The previously blank rows now display text labels, and the patch passes all tests.

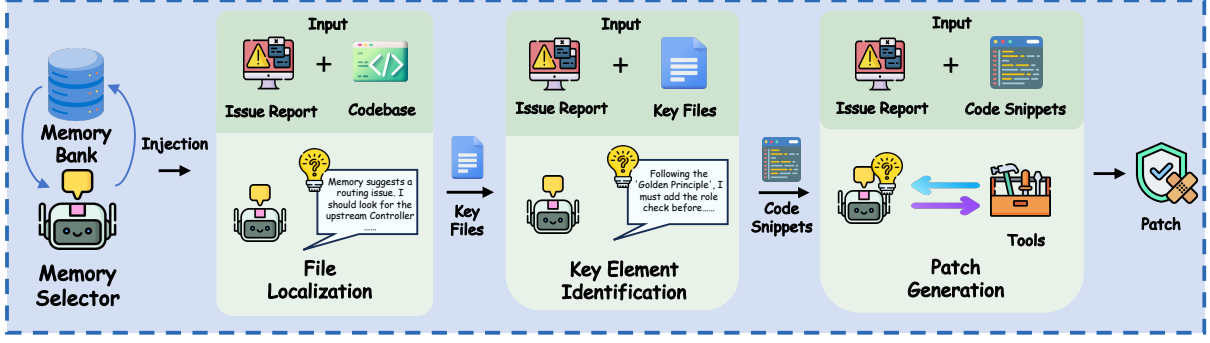


Figure 2: Overview of FailureMem.

3 Method

We propose FailureMem, a MAPR framework combining a hybrid workflow-agent architecture, active perception tools, and a failure memory bank. Figure 2 shows the overall architecture.

3.1 Problem Formulation

Given a software repository with codebase $\mathcal{C}$ and a multimodal issue report comprising a textual description $\mathcal{D}$ and visual screenshots $\mathcal{I}$, MAPR aims at generating a code patch Δ such that the patched codebase $\mathcal{C}' = \mathcal{C} \oplus \Delta$ passes a held-out test suite:

$$\mathcal{F}(\mathcal{C}') \models T_{spec} \quad (1)$$

where $\oplus$ is the patch application operator, $\mathcal{F}(\cdot)$ denotes program execution, and T_{spec} includes both fail-to-pass tests (verifying the issue is fixed) and pass-to-pass tests (verifying no regression).

3.2 Failure-Aware Retrieval Module

LLM-based repair agents often produce plausible but incorrect patches and may repeat errors from prior attempts. To address this limitation, we introduce a failure memory bank, which stores historical repair trajectories. We further design a failure-aware retrieval module that contrasts failed patches with ground-truth fixes to extract reusable repair patterns. By retrieving and reusing prior experiences, the model reduces recurring mistakes and improves transfer to new issues.

Hierarchical Memory Structure Design. The failure memory bank is defined as a collection of N entries: $\mathcal{B} = \{\mathcal{M}_i\}_{i=1}^N$. As illustrated in Figure 1, we design three complementary layers to capture different levels of repair knowledge in each entry.

$$\mathcal{M}_i = \langle \mathcal{L}_{ctx}^{(i)}, \mathcal{L}_{cog}^{(i)}, \mathcal{L}_{code}^{(i)} \rangle$$

Contextual Layer ($\mathcal{L}_{ctx}$) provides retrieval keys for matching similar cases. It contains two text-based fields: the *Issue Summary*, which abstracts the bug scenario, and the *Visual Analysis*, which describes observable visual symptoms (e.g., “The modal overlay obscures the navigation bar”). Instead of storing raw screenshots, we encode visual information as text. This decision reduces token cost during retrieval and avoids background noise in images, ensuring that the selector focuses on bug-relevant signals.

Cognitive Layer ($\mathcal{L}_{cog}$) provides high-level reasoning guidance. It includes a *Cognitive Diagnosis* that explains the causal mechanism of failure, a *Negative Constraint* that prohibits incorrect repair strategies (e.g., “Do not modify downstream rendering logic for upstream data errors”), and a *Golden Principle* that captures transferable design patterns (e.g., fixing errors at the source of generation rather than applying downstream filtering). This layer focuses on abstract repair knowledge to improve generalization across issues.

In contrast, Code Layer ($\mathcal{L}_{code}$) supplies implementation-level evidence. It stores *Failed* and *Golden Patch Summaries*, explicitly highlighting structural divergences between incorrect and correct solutions. By grounding abstract principles in concrete code changes, this layer supports actionable and executable repairs. Together, the Cognitive Layer shapes repair strategy at the reasoning level, while the Code Layer ensures alignment with precise implementation patterns.

Failure Memory Bank Construction. To construct the memory bank, we process historical failure cases offline (see Appendix A for detailed data statistics and the distillation pipeline). For each case, we collect the agent’s failed patch P_{fail} , the developer’s ground-truth patch P_{gold} , and the original issue report, including the textual description $\mathcal{D}$

and screenshots $\mathcal{V}$. Using Gemini 3 Pro as the backbone model, we analyze these multimodal inputs and distill them into hierarchical memory entries. Formally, for each historical failure case i , the construction process is:

$$\mathcal{M}_i = f_{\text{distill}}(P_{\text{fail}}^{(i)}, P_{\text{gold}}^{(i)}, \mathcal{D}^{(i)}, \mathcal{V}^{(i)})$$

where f_{distill} denotes the distillation model that takes the failed patch, ground-truth patch, issue description, and visual screenshots as inputs, and outputs the structured three-layer memory entry.

The model is prompted to generate structured content for all three memory layers. For the Contextual Layer, it abstracts the issue report into a concise *Issue Summary* and converts screenshots $\mathcal{V}$ into a textual *Visual Analysis*, preserving visual symptoms for efficient retrieval. For the Cognitive Layer, it contrasts P_{fail} and P_{gold} to identify the root cause, extracting the *Cognitive Diagnosis*, *Negative Constraints*, and *Golden Principles*. For the Code Layer, it summarizes the implementation differences into *Failed* and *Golden Patch Summaries*, capturing key divergences while filtering out project-specific noise.

Data Boundaries. The Failure Memory Bank is constructed exclusively from the development split of SWE-bench Multimodal, which contains 102 instances from 5 repositories. We use only the 84 failed trajectories along with their developer gold patches as source material. No data from the test split is accessed during memory construction or inference. Importantly, the development and test splits share zero repository overlap. The development repositories are Automattic/wp-calypso, chartjs/Chart.js, markedjs/marked, diegomura/react-pdf, and processing/p5.js, while the 12 test repositories (alibaba-fusion/next, bpmn-io/bpmn-js, carbon-design-system/carbon, eslint, GoogleChrome/lighthouse, grommet, highlightjs/highlight.js, openlayers, prettier, PrismJS/prism, quarto-dev/quarto-cli, scratchfoundation/scratch-gui) are entirely disjoint. As a result, the memory bank cannot contain repository-specific patterns such as file paths, API names, or component structures that would directly transfer to the test set. The observed gains on the test split therefore reflect cross-repository transfer of abstract repair knowledge rather than in-distribution memorization.

Retrieving Memory Entries. Before file localization, we retrieve relevant historical memory en-

tries. We employ a Selector Agent to identify the top- k most relevant cases. Given a new issue with description $\mathcal{D}_q$ and screenshots $\mathcal{V}_q$, the Selector Agent receives all candidate Contextual Layers and directly selects the top- k most relevant entries in a single pass:

$$\mathcal{R} = \text{Selector}(\mathcal{D}_q, \mathcal{V}_q, \{\mathcal{L}_{\text{ctx}}^{(i)}\}_{i=1}^N, k)$$

where $\mathcal{R} \subset \mathcal{B}$ and $|\mathcal{R}| = k$.

The retrieval process is modeled as a semantic selection task. The Selector Agent is presented with the current issue report, which contains both the textual description $\mathcal{D}$ and the associated visual screenshots $\mathcal{V}$. It compares these inputs against the *Contextual Layer* ($\mathcal{L}_{\text{ctx}}$) of the candidate memories, specifically aligning the current problem with the historical *Issue Summary* and *Visual Analysis*. This approach allows the agent to leverage its semantic understanding to find relevant precedents based on symptom similarity, rather than relying on superficial keyword overlap.

Once $\mathcal{R}$ is obtained, the guidance context is constructed by extracting the Cognitive and Code layers from each retrieved entry:

$$\mathcal{G} = \{(\mathcal{L}_{\text{cog}}^{(i)}, \mathcal{L}_{\text{code}}^{(i)}) \mid \mathcal{M}_i \in \mathcal{R}\}$$

This guidance $\mathcal{G}$ is then injected into the model’s context at each subsequent phase.

3.3 Repair with a Hybrid Agent Framework

After retrieving the top- k memory entries, we adopt a hybrid agent–workflow architecture that balances structured fault localization with flexible patch generation. To reduce computational overhead in repository-scale contexts, we confine the agentic loop to the final patch generation stage, while using a deterministic workflow for earlier localization steps. This design narrows the search space and improves token efficiency. At each phase, the injected Cognitive and Code layers further align the model’s reasoning with historical constraints, guiding both localization and repair decisions.

Phase 1: File Localization. The process begins by scanning the repository’s directory tree to identify a candidate set of files relevant to the issue. The model receives the issue description, the file tree structure, and the retrieved memory entries. At this stage, the injected *Cognitive Layer* helps the model filter out irrelevant domains (e.g., distinguishing

between test files and source code) based on historical failure patterns. This phase operates in a single-pass inference mode to efficiently narrow the search space without invoking external tools.

Phase 2: Key Element Identification. Once the candidate files are determined, the system identifies the specific classes or functions requiring modification. To handle the token limit, we utilize a *Skeleton Compression* strategy (detailed in Appendix B.1). We parse the candidate files to retain only class signatures, function headers, and docstrings, while abstracting implementation bodies. The model analyzes these skeletons alongside the injected memory to pinpoint the key elements. The memory injection here is critical for preventing architectural misalignments, such as guiding the model to focus on the controller logic rather than view definitions when the memory suggests a backend root cause. Like the previous phase, this step does not utilize an agentic loop.

Phase 3: Agentic Patch Generation. This final stage is the only phase where the model operates as an autonomous agent with a multi-turn execution loop. The input comprises the *full raw code* of the identified key elements, the issue description, and the retrieved memory. We equip the agent with a specific suite of tools to verify logic and resolve multimodal ambiguities:

(1) *Active Visual Perception.* To address the resolution limitations of standard multimodal models, the agent utilizes a *Crop* tool to zoom into specific regions for detailed inspection. Additionally, a *Grounding* tool enables the agent to draw bounding boxes around bug-related UI elements. This action explicitly focuses the model’s attention on relevant visual contexts, mitigating the noise from unrelated interface components. Implementation details are provided in Appendix B.2.

(2) *Interactive Environment.* A *Bash* tool (operating within a strictly sandboxed execution environment as described in Appendix B.3) allows the agent to execute commands. This enables the agent to actively explore the repository structure and verify code logic dynamically, such as checking dependency versions or running reproduction scripts, before committing to specific modifications.

During this loop, the model iteratively refines its plan using these tools, while continuously receiving guidance from the injected *Cognitive* and *Code* layers. These memory layers help the agent apply proven repair principles, reference success-

ful implementation patterns, and avoid failure modes identified in past repair attempts. This dual-sided memory guidance significantly improves the agent’s ability to make reliable repair decisions.

4 Evaluation

4.1 Experimental Setup

Benchmark. We evaluate FailureMem on SWE-bench Multimodal (SWE-bench M) (Yang et al., 2025), which consists of 619 real-world GitHub issues spanning 17 popular JavaScript repositories. The benchmark is partitioned into a development split (102 instances from 5 repositories) used for system development and memory bank construction, and a test split (517 instances from 12 repositories) used for evaluation. Visual information is strictly necessary for resolving over 83% of the tasks (Yang et al., 2025).

Baselines. We primarily benchmark against GUIRepair (Huang et al., 2025c), the current SOTA open-source multimodal workflow-based approach. We also report results from text-centric workflows like Agentless (Ma et al., 2025) and generalist agents like SWE-agent (Yang et al., 2024b), alongside commercial systems (e.g., Globant Code Fixer) as upper-bound references.

Implementation Details. We reproduce the official implementation of GUIRepair. All experiments use a sampling temperature of 0 for deterministic generation and follow the Pass@1 evaluation protocol, allowing only a single predicted patch without re-ranking. Experiments are conducted on four NVIDIA A100 (80GB) GPUs.

4.2 Results and Discussion

Table 1 shows FailureMem consistently outperforms the baseline GUIRepair across all tested foundation models. Specifically, when equipped with GPT-5.1, our framework achieves a resolved rate of 33.1%, yielding a significant absolute improvement of 3.7% over the baseline. We observe similar consistent gains with GPT-4.1 (+2.3%) and Claude 4.5 (+2.3%), validating that the benefits of our failure-aware mechanism are model-agnostic. Furthermore, FailureMem surpasses all other reference baselines, including both agent-based frameworks such as SWE-agent and workflow-based systems like Agentless Lite.

Ablations. To analyze the contributions of each component, we perform an ablation study on SWE-

Method	LLM	(Resolved Rate %)	Repository Breakdown (Resolved Rate %)											
			next	bpmn-js	carbon	eslint	lighthouse	grommet	highlight.js	openlayers	prettier	prism	quarto-cli	scratch-gui
Other Baselines														
SWE-agent Multimodal	GPT-4o	12.4	0.0	27.8	1.5	0.0	5.6	0.0	2.6	51.9	7.7	0.0	0.0	0.0
Computer-Use Agents	GPT-4o	20.1	-	-	-	-	-	-	-	-	-	-	-	-
Agentless Lite	GPT-5.1	28.4	15.4	53.7	14.9	9.1	9.3	4.8	10.3	94.9	7.7	2.6	0.0	0.0
Zencoder	NA	27.5	15.4	51.9	11.2	18.2	13.0	19.0	10.3	81.0	7.7	18.4	8.3	0.0
GUIRepair vs. Ours (GPT-4.1)														
GUIRepair	GPT-4.1	28.8	17.9	66.7	12.7	36.4	7.4	4.8	2.6	96.2	7.7	2.6	4.2	0.0
FailureMem (Ours)	GPT-4.1	31.1 (+2.3)	23.1	68.5	12.7	18.2	13.0	9.5	5.1	97.5	15.4	13.2	4.2	0.0
GUIRepair vs. Ours (GPT-5.1)														
GUIRepair	GPT-5.1	29.4	23.1	57.4	11.9	36.4	13.0	4.8	5.1	97.5	15.4	5.3	4.2	0.0
FailureMem (Ours)	GPT-5.1	33.1 (+3.7)	25.6	63.0	17.9	36.4	18.5	14.3	7.7	97.5	7.7	10.5	4.2	0.0
GUIRepair vs. Ours (GPT-5.2)														
GUIRepair	GPT-5.2	29.8	28.2	61.1	11.9	27.3	13.0	19.0	0.0	97.5	15.4	0.0	4.2	0.0
FailureMem (Ours)	GPT-5.2	33.3 (+3.5)	28.2	66.7	15.7	27.3	20.4	9.5	10.3	97.5	15.4	10.5	4.2	0.0
GUIRepair vs. Ours (Claude 4 Sonnet)														
GUIRepair	Claude 4	28.6	20.5	61.1	11.9	45.5	11.1	4.8	5.1	96.2	7.7	0.0	0.0	0.0
FailureMem (Ours)	Claude 4	32.5 (+3.9)	23.1	68.5	14.9	36.4	16.7	9.5	7.7	97.5	15.4	13.2	4.2	0.0
GUIRepair vs. Ours (Claude 4.5 Opus thinking)														
GUIRepair	Claude 4.5	31.5	23.1	64.8	13.4	54.5	16.7	9.5	5.1	97.5	15.4	5.3	4.2	0.0
FailureMem (Ours)	Claude 4.5	33.8 (+2.3)	23.1	68.5	17.2	36.4	18.5	19.0	7.7	97.5	15.4	13.2	4.2	0.0

Table 1: **Comprehensive Performance on SWE-bench Multimodal.** We compare FailureMem against state-of-the-art baselines. Columns under *Repository Breakdown* show the resolved rate (%) for each specific repository.

bench Multimodal with the GPT-5.1 backend (See Table 2). We define a **Base** configuration that removes all proposed components (Active Perception, Bash Environment, and FailureMem) from our framework, resulting in a standard multimodal workflow without dynamic exploration or historical memory. This configuration is functionally equivalent to GUIRepair without its RAG-based file retrieval module, which accounts for the 0.8% difference between GUIRepair (29.4%) in Table 1 and Base (28.6%) in Table 2.

The **+Active Perception** variant yields a 1.4% improvement over the baseline (28.6% $\rightarrow$ 30.0%). Our error analysis indicates that the Base agent frequently hallucinates DOM element attributes when processing downsampled full-page screenshots. The inclusion of the Crop and Grounding tools effectively mitigates this by allowing the model to request high-resolution re-sampling of specific regions, thereby converting vague visual signals into precise coordinate-level constraints.

The **+Bash** environment contributes a 1.6% gain (30.2%). We observed that purely generative agents often propose theoretically valid but environmentally incompatible patches, such as referencing

non-existent relative paths or deprecated API versions. The interactive shell shifts the paradigm from “write-and-pray” to “verify-then-commit”, enabling the agent to validate file structures and dependencies before attempting a fix.

Notably, **+FailureMem** delivers the largest individual performance boost (+2.3%, reaching 30.9%). This suggests that a significant portion of repair failures stems not from a lack of capability, but from *cognitive recurrence*—the tendency to repeatedly attempt plausible but incorrect solutions (e.g., modifying the view layer for a state logic bug). By injecting negative constraints, FailureMem effectively prunes these high-probability failure branches early in the reasoning process.

The **Full Framework** achieves 33.1%, a cumulative improvement of +4.5%. This performance exceeds any single component, confirming that the modules address orthogonal failure modes: Active Perception ensures the *input* is accurate, FailureMem ensures the *plan* is sound, and Bash ensures the *execution* is valid.

Memory Ablations. We investigate the distinct contributions of the cognitive and code layers

Configuration	Resolved Rate (%)	Improvement
Base (Standard Workflow)	28.6	-
+ Active Perception	30.0	(+1.4)
+ Bash Environment	30.2	(+1.6)
+ FailureMem (Memory Only)	30.9	(+2.3)
Full Framework	33.1	(+4.5)

Table 2: **Ablation Analysis on GPT-5.1.** We compare the Full Framework against the Base and single-component variants. “Active Perception” enables Crop/Grounding tools; “Bash” provides an interactive shell; “FailureMem” injects negative memory constraints.

Configuration	Memory Fields			Performance	
	Cognitive	Code	Failure	Resolved	Rate (%)
(A) Cognitive Only	✓	-	✓	154 / 517	29.8
(B) Code Only (RAG)	-	✓	-	158 / 517	30.6
(C) Positive Only	✓	✓	-	158 / 517	30.6
(D) Raw Patch	✓	Diff	✓	159 / 517	30.8
(E) Full Summary (Ours)	✓	✓	✓	171 / 517	33.1

Table 3: **Component Analysis of FailureMem.** We evaluate the impact of different memory compositions. The *Failure* column indicates the inclusion of Failed Patch Summaries and Negative Constraints. Configuration (E) represents the proposed method.

within FailureMem. To rigorously isolate these components, we conducted experiments using the GPT-5.1 backbone. In all experimental settings, the retrieval mechanism remains constant: the Selector Agent identifies the top-3 relevant memory entries based on the Contextual Layer. We vary only the specific information fields injected into the repair agent’s context. Table 3 presents the results of five configurations. Variant (A) retains only the cognitive components (Diagnosis and Negative Constraints). Variant (B) simulates a standard RAG baseline by providing only the Golden Patch Summary. Variant (C) represents a positive-reinforcement setup, providing both the Golden Principle and Golden Patch Summary. Variant (D) replaces the natural language summaries in the Code Layer with raw git diffs. Finally, Configuration (E) represents our full framework.

Variant (A) relies on high-level reasoning, providing the Cognitive Diagnosis and Negative Constraints without concrete code examples. This configuration yields the lowest performance at 29.8%. While the model correctly identifies architectural constraints, it struggles to translate abstract directives into valid syntax specific to the target repository. In multimodal repair tasks involving complex DOM APIs, abstract reasoning must be supported by explicit code references provided by the Code

Layer to facilitate correct implementation.

We further analyzed whether the form of the Code Layer influences performance. Variant (D) replaces the distilled patch summaries with raw code diffs, resulting in a performance drop to 30.8%. Raw patches frequently contain project-specific artifacts, such as variable naming conventions or unrelated context lines, which introduce noise into the context window. Natural language summaries extract the core repair logic from these implementation details. This filtering process allows the agent to transfer the underlying fix pattern to the current issue more effectively than direct code copying.

The comparison between Variant (C) and our proposed framework (E) highlights the core contribution of this work. Variant (C) provides comprehensive positive guidance, including both Golden Principles and Golden Patch Summaries, yet its performance plateaus at 30.6%, identical to the Code Only baseline. This result suggests that providing correct examples alone is insufficient to prevent the model from repeating common mistakes. The inclusion of failure components in Configuration (E) improves the resolved rate to 33.1%. By explicitly contrasting the Failed Patch Summary with the Golden Patch Summary, the framework provides a discriminative signal. This signal enables the agent to distinguish between the correct solution and plausible but incorrect alternatives that positive examples alone cannot identify.

Memory Size Impacts. We evaluate different numbers of retrieved memory entries, $k \in \{1, 3, 5, 10\}$, using GPT-5.1 and Claude 4. In Figure 3, both models follow an inverted U-shaped trend: performance improves from $k = 1$ to $k = 3$, peaking at 33.1% (GPT-5.1) and 32.5% (Claude 4), then declines as k increases further. At $k = 10$, resolved rates drop to 31.4% and 30.8%, likely due to context dilution from excessive retrieved code. We therefore set $k = 3$ as the default to balance information coverage and reasoning focus.

Cross-Repository Generalization. Since the development and test splits of SWE-bench Multimodal share zero repository overlap (Section 3.2), the standard evaluation already measures cross-repository transfer. To validate that the memory bank does not rely on any single source repository, we conduct a leave-one-repository-out experiment. We construct 5 variants of the memory bank, each excluding all entries from one development repository, and evaluate on the full test set. Table 4

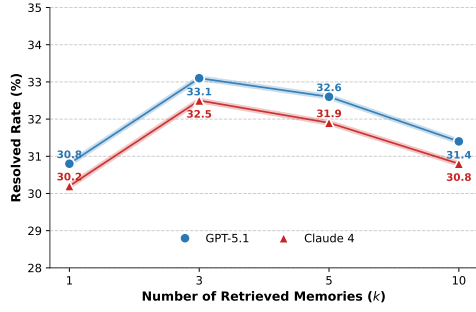


Figure 3: **Impact of Retrieval Size k .** Both models exhibit an inverted U-shape trend, peaking at $k = 3$.

Excluded Dev Repository	Removed	GPT-5.1 (%)	Claude 4 (%)
None (Full Memory)	0	33.1	32.5
Automatic/wp-calypso	37	31.3	30.6
chartjs/Chart.js	16	31.9	31.3
markedjs/marked	11	31.5	31.1
diegomura/react-pdf	10	32.3	31.7
processing/p5.js	10	31.7	31.7

Table 4: Leave-one-repository-out evaluation.

reports results with GPT-5.1 and Claude 4 Sonnet.

Performance remains stable across all ablations. Even removing wp-calypso, which contributes 37 of the 84 entries (44%), causes only a 1.8% drop for GPT-5.1, still substantially outperforming GUIRepair (29.4%). No single development repository dominates the transfer, confirming that the memory bank provides diverse repair knowledge rather than overfitting to any specific codebase.

Active Perception Usage. We report tool invocation statistics from the Full Framework (GPT-5.1) across 509 test instances that reached the agentic patch generation phase. The Crop tool was invoked in 145 instances (28.5%) and the Grounding tool in 56 instances (11.0%), with at least one perception tool used in 201 instances (39.5%). Usage varies across repositories: UI-intensive repositories such as carbon-design-system (51.1%) and prettier (53.8%) show higher invocation rates, while repositories with fewer visual elements such as openlayers (22.8%) see lower usage. This pattern confirms that the agent selectively invokes visual tools when screenshots contain fine-grained details requiring focused inspection.

To further isolate each tool’s contribution, we ran two additional ablations starting from the Base configuration: Base + Crop only achieves 29.6%, Base + Grounding only achieves 29.8%, and Base + Both (Active Perception) achieves 30.0%. Both tools individually contribute comparable gains over the 28.6% baseline. Through manual inspection of

resolved cases, the two tools address complementary scenarios: Crop is more effective for single-component issues where magnifying a region removes surrounding noise, while Grounding is more effective for layout bugs where spatial relationships between elements must be preserved in context. The near-additive behavior when combining both tools is consistent with this complementarity, as on a given instance typically one tool is appropriate while the other provides no additional benefit.

5 Related Work

Research on LLM-based Automated Program Repair (APR) has progressed from early fine-tuning approaches (Wang et al., 2023; Xia et al., 2023a; Yang et al., 2026; Silva et al., 2025) and prompting-based methods (Fan et al., 2023; Xia et al., 2023b; Xia and Zhang, 2022; Zhang et al., 2023a; Xia and Zhang, 2024; Zhao et al., 2024; Xu et al., 2025; Yang et al., 2024a; Xiao et al.; Peng et al., 2024; Xiang et al., 2024; Jiang et al., 2025; Fan et al., 2026), to autonomous agents for repository-level issue resolution (Jimenez et al., 2024; Antoniadou et al., 2025; Liu et al., 2024; Meng et al., 2024). While systems such as SWE-agent (Yang et al., 2024b) and Agentless (Ma et al., 2025) improve general debugging, they lack the multimodal capabilities required for GUI-based tasks. Specialized tools exist for limited visual domains, such as UI design (Yuan et al., 2025) and accessibility (Zhang et al., 2023b), but general-purpose multimodal repair remains underexplored. The current SOTA, GUIRepair (Huang et al., 2025c), extends agentless workflows with cross-modal reasoning to incorporate visual information. However, existing approaches suffer from passive perception, relying only on static inputs without active visual exploration, and statelessness, treating each repair attempt independently without leveraging historical experience. These limitations highlight the need for tool-using, memory-enabled agents for robust multimodal reasoning.

6 Conclusion

We proposed FailureMem, an experience-driven framework for Multimodal Automated Program Repair that integrates a hybrid workflow-agent architecture, active visual grounding, and a hierarchical failure memory bank. Experiments on SWE-bench Multimodal showed consistent improvements over strong baselines.

Limitations

Despite FailureMem’s performance gains, two limitations remain. The iterative agentic loop and detailed memory contexts increase inference costs to \$0.33 per issue, a 13% rise compared to GUIRepair’s \$0.29. We consider this tradeoff acceptable given the 3.7% improvement in resolution rate and the relatively low cost of autonomous compute versus human effort. Additionally, FailureMem depends on the diversity of its offline memory bank. Rare or unprecedented failure modes without historical analogues bypass the contrastive distillation process, causing the system to revert to standard agentic behavior.

Ethics Statement

This work studies Multimodal Automated Program Repair and proposes FailureMem to improve the reliability of automated software debugging. Our goal is to assist developers by identifying and repairing software defects more effectively. The system is intended as a decision-support tool, and generated patches should be reviewed and validated by developers before deployment in production environments. Automated repair systems may occasionally generate incorrect or incomplete patches, and human oversight remains essential for ensuring software safety and correctness.

The memory bank used in FailureMem is constructed from publicly available software repositories and issue reports, which may include textual descriptions and screenshots of user interfaces. These artifacts are processed to extract abstract repair knowledge rather than storing raw sensitive content. We do not intentionally collect personal or confidential data, and the framework is designed to store summarized repair experiences rather than proprietary code whenever possible. Nevertheless, practitioners should ensure that memory construction follows appropriate data governance policies when applied to private repositories.

Automated program repair technologies may also be misused if deployed without sufficient safeguards. For example, blindly applying automatically generated patches could introduce regressions or security vulnerabilities. Our framework therefore emphasizes learning from failed repair attempts and negative constraints, aiming to reduce recurring mistakes rather than encouraging fully autonomous code modification.

Finally, the use of large language models and

multimodal reasoning systems requires substantial computational resources. Our experiments are conducted on GPU infrastructure, which has environmental implications. Future work will explore more efficient memory retrieval and reasoning strategies to reduce computational costs while maintaining repair effectiveness.

Acknowledgments

We thank the reviewers for their valuable comments on an earlier draft of this paper. This work was supported by National Natural Science Foundation of China (No. 62406139), State Key Laboratory for Novel Software Technology at Nanjing University (KFKT2026A09, ZZKT2026A43).

References

- Antonis Antoniadis, Albert Örwall, Kexun Zhang, Yuxi Xie, Anirudh Goyal, and William Wang. 2025. SWE-Search: Enhancing software agents with Monte Carlo tree search and iterative refinement. In *Proceedings of the 13th International Conference on Learning Representations*, pages 64485–64515.
- Kaixuan Fan, Kaituo Feng, Manyuan Zhang, Tianshuo Peng, Zhixun Li, Yilei Jiang, Shuang Chen, and Xiangyu Yue. 2026. Exploring reasoning reward model for agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 1975–1991.
- Zhiyu Fan, Xiang Gao, Martin Mirchev, Abhik Roychoudhury, and Shin Hwei Tan. 2023. Automated repair of programs from large language models. In *Proceedings of the 45th International Conference on Software Engineering (ICSE)*, pages 1469–1481.
- Kai Huang, Xiangxin Meng, Jian Zhang, Yang Liu, Wenjie Wang, Shuhao Li, and Yuqing Zhang. 2023. An empirical study on fine-tuning large language models of code for automated program repair. In *Proceedings of the 38th International Conference on Automated Software Engineering (ASE)*, pages 1162–1174.
- Kai Huang, Jian Zhang, Xinlei Bao, Xu Wang, and Yang Liu. 2025a. Comprehensive fine-tuning large language models of code for automated program repair. *IEEE Transactions on Software Engineering*, 51(4):904–928.
- Kai Huang, Jian Zhang, Xiangxin Meng, and Yang Liu. 2025b. Template-guided program repair in the era of large language models. In *Proceedings of the 47th International Conference on Software Engineering (ICSE)*, pages 1895–1907.
- Kai Huang, Jian Zhang, Xiaofei Xie, and Chunyang Chen. 2025c. Seeing is fixing: Cross-modal reasoning with multimodal llms for visual software issue

- repair. In *Proceedings of the 40th International Conference on Automated Software Engineering (ASE)*, pages 1156–1168.
- Nan Jiang, Kevin Liu, Thibaud Lutellier, and Lin Tan. 2023. Impact of code language models on automated program repair. In *Proceedings of the 45th International Conference on Software Engineering (ICSE)*, pages 1430–1442.
- Yilei Jiang, Yaozhi Zheng, Yuxuan Wan, Jiaming Han, Qunzhong Wang, Michael R. Lyu, and Xiangyu Yue. 2025. [ScreenCoder: Advancing visual-to-code generation for front-end automation via modular multi-modal agents](#). *Preprint*, arXiv:2507.22827.
- Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. 2024. SWE-bench: Can language models resolve real-world Github issues? In *Proceedings of the 12th International Conference on Learning Representations*.
- Yizhou Liu, Pengfei Gao, Xincheng Wang, Jie Liu, Yexuan Shi, Zhao Zhang, and Chao Peng. 2024. [MarsCode agent: AI-native automated bug fixing](#). *Preprint*, arXiv:2409.00899.
- Yingwei Ma, Qingping Yang, Rongyu Cao, Binhua Li, Fei Huang, and Yongbin Li. 2025. Alibaba lingmaagent: Improving automated issue resolution via comprehensive repository exploration. In *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering*, pages 238–249.
- Xiangxin Meng, Zexiong Ma, Pengfei Gao, and Chao Peng. 2024. [An empirical study on LLM-based agents for automated bug fixing](#). *Preprint*, arXiv:2411.10213.
- Yun Peng, Shuzheng Gao, Cuiyun Gao, Yintong Huo, and Michael Lyu. 2024. Domain knowledge matters: Improving prompts with fix templates for repairing Python type errors. In *Proceedings of the 46th International Conference on Software Engineering (ICSE)*, pages 1–13.
- André Silva, Sen Fang, and Martin Monperrus. 2025. RepairLLaMA: Efficient representations and fine-tuned adapters for program repair. *IEEE Transactions on Software Engineering*, 51(8):2366–2380.
- Weishi Wang, Yue Wang, Shafiq Joty, and Steven CH Hoi. 2023. RAP-Gen: Retrieval-augmented patch generation with CodeT5 for automatic program repair. In *31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (FSE)*, pages 146–158.
- Yi Wu, Nan Jiang, Hung Viet Pham, Thibaud Lutellier, Jordan Davis, Lin Tan, Petr Babkin, and Sameena Shah. 2023. How effective are neural networks for fixing security vulnerabilities. In *Proceedings of the 32nd International Symposium on Software Testing and Analysis (ISSTA)*, pages 1282–1294.
- Chunqiu Steven Xia, Yinlin Deng, Soren Dunn, and Lingming Zhang. 2025. Demystifying LLM-based software engineering agents. *Proceedings of the ACM on Software Engineering*, 2(FSE):801–824.
- Chunqiu Steven Xia, Yifeng Ding, and Lingming Zhang. 2023a. The plastic surgery hypothesis in the era of large language models. In *Proceedings of the 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 522–534.
- Chunqiu Steven Xia, Yuxiang Wei, and Lingming Zhang. 2023b. Automated program repair in the era of large pre-trained language models. In *Proceedings of the 45th International Conference on Software Engineering (ICSE)*, pages 1482–1494.
- Chunqiu Steven Xia and Lingming Zhang. 2022. Less training, more repairing please: revisiting automated program repair via zero-shot learning. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*, pages 959–971.
- Chunqiu Steven Xia and Lingming Zhang. 2024. Automated program repair via conversation: Fixing 162 out of 337 bugs for \$0.42 each using ChatGPT. In *33rd International Symposium on Software Testing and Analysis (ISSTA)*, pages 819–831.
- Jiahong Xiang, Xiaoyang Xu, Fanchu Kong, Mingyuan Wu, Zizheng Zhang, Haotian Zhang, and Yuqun Zhang. 2024. [How far can we go with practical function-level program repair?](#) *Preprint*, arXiv:2404.12833.
- Yuan-An Xiao, Weixuan Wang, Dong Liu, Junwei Zhou, Shengyu Cheng, and Yingfei Xiong. PredicateFix: Repairing static analysis alerts with bridging predicates. In *Proceedings of the 48th International Conference on Software Engineering (ICSE)*.
- Junjielong Xu, Ying Fu, Shin Hwei Tan, and Pinjia He. 2025. Aligning the objective of LLM-based program repair. In *Proceedings of the 47th International Conference on Software Engineering (ICSE)*, pages 2548–2560.
- Boyang Yang, Haoye Tian, Weiguo Pian, Haoran Yu, Haitao Wang, Jacques Klein, Tegawendé F Bissyandé, and Shunfu Jin. 2024a. CREF: An LLM-based conversational software repair framework for programming tutors. In *Proceedings of the 33rd International Symposium on Software Testing and Analysis (ISSTA)*, pages 882–894.
- Boyang Yang, Haoye Tian, Jiadong Ren, Hongyu Zhang, Jacques Klein, Tegawendé F Bissyandé, Claire Le Goues, and Shunfu Jin. 2026. MORepair: Teaching LLMs to repair code via multi-objective fine-tuning. *ACM Transactions on Software Engineering and Methodology*, 35(2):1–38.
- John Yang, Carlos Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir

Press. 2024b. SWE-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems*, 37:50528–50652.

John Yang, Carlos E. Jimenez, Alex L. Zhang, Kilian Lieret, Joyce Yang, Xindi Wu, Ori Press, Niklas Muennighoff, Gabriel Synnaeve, Karthik R. Narasimhan, Diyi Yang, Sida I. Wang, and Ofir Press. 2025. SWE-bench Multimodal: Do AI systems generalize to visual software domains? In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*.

Mingyue Yuan, Jieshan Chen, Zhenchang Xing, Aaron Quigley, Yuyu Luo, Tianqi Luo, Gelareh Mohammadi, Qinghua Lu, and Liming Zhu. 2025. Design-repair: Dual-stream design guideline-aware frontend repair with large language models. In *Proceedings of the 47th International Conference on Software Engineering (ICSE)*, pages 2483–2494.

Quanjun Zhang, Chunrong Fang, Tongke Zhang, Bowen Yu, Weisong Sun, and Zhenyu Chen. 2023a. Gamma: Revisiting template-based automated program repair via mask prediction. In *Proceedings of the 38th International Conference on Automated Software Engineering (ASE)*, pages 535–547.

Yuntong Zhang, Haifeng Ruan, Zhiyu Fan, and Abhik Roychoudhury. 2024. Autocoderover: Autonomous program improvement. In *Proceedings of the 33rd International Symposium on Software Testing and Analysis (ISSTA)*, pages 1592–1604.

Yuxin Zhang, Sen Chen, Lingling Fan, Chunyang Chen, and Xiaohong Li. 2023b. Automated and context-aware repair of color-related accessibility issues for android apps. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*, pages 1255–1267.

Jiuang Zhao, Donghao Yang, Li Zhang, Xiaoli Lian, Zitian Yang, and Fang Liu. 2024. Enhancing automated program repair with solution design. In *Proceedings of the 39th International Conference on Automated Software Engineering (ASE)*, pages 1706–1718.

A Data Construction for FailureMem

We detail the offline construction process of the Failure Memory Bank, specifically focusing on how empirical failure trajectories are collected and distilled into structured memory entries.

A.1 Data Source and Trajectory Collection

Our memory bank is strictly derived from the development set of the SWE-bench Multimodal benchmark, which comprises 102 real-world visual software issues. To capture natural, high-probability failure patterns, we first ran FailureMem without

the Failure Memory Bank (i.e., using only the hybrid workflow-agent architecture and active perception tools, powered by GPT-5.1) to attempt all 102 instances.

The evaluation yielded a clear behavioral split: the memory-free variant correctly resolved 18 issues but failed to produce valid patches for the remaining 84 instances. We exclusively utilize these 84 failed trajectories as our source material for memory construction. The 18 successful instances were intentionally excluded to ensure the memory bank remains focused on learning from failures rather than redundant positive imitation. For each of the 84 failed instances, we collected the incorrect patch output (P_{fail}) from this variant to serve as the contrastive counterpart to the developer-verified ground truth (P_{gold}).

A.2 Offline Distillation Pipeline

To transform raw failed trajectories into the hierarchical memory entries defined in Section 3.2, we implemented an automated offline distillation pipeline. We utilized Gemini 3 Pro as the reasoning engine to conduct a contrastive root cause analysis.

For each of the 84 failed instances, the pipeline constructs a multimodal context tuple $\langle \mathcal{D}, \mathcal{I}, P_{fail}, P_{gold} \rangle$, where:

- $\mathcal{D}$: The original natural language issue description.
- $\mathcal{I}$: Up to three visual symptom screenshots (encoded as base64 images).
- P_{fail} : The incorrect patch generated by the memory-free variant.
- P_{gold} : The correct developer patch.

The distillation model is prompted to compare P_{fail} against P_{gold} and abstract the multimodal inputs into our three-layer memory architecture:

Contextual Layer ($\mathcal{L}_{ctx}$): The model synthesizes the raw inputs into two text-based fields: the Issue Summary, which abstracts the bug scenario, and the Visual Analysis, which textually describes the visual symptoms (e.g., “The modal overlay obscures the navigation bar”). This deliberate translation from raw screenshots to text minimizes token consumption and prevents visual background noise from distracting the selector agent during the retrieval phase.

Cognitive Layer ($\mathcal{L}_{cog}$): The model extracts high-level reasoning guidance consisting of three

components. It generates a Cognitive Diagnosis that explains the causality behind the failure; formulates a Negative Constraint that explicitly forbids the specific incorrect strategies (e.g., “Do not modify downstream rendering logic for upstream data errors”); and outlines a Golden Principle that defines the correct design pattern.

Code Layer ($\mathcal{L}_{code}$): To provide concrete references, the model generates a Failed Patch Summary and a Golden Patch Summary. These summaries highlight the specific implementation divergence between the incorrect and correct solutions, offering the agent an empirical reference for the repair strategy while filtering out project-specific noise found in raw diffs.

To ensure the reliability of the memory bank, the pipeline includes a strict field-level validation step. Any generated entry missing the required structural layers, containing empty strings for critical reasoning steps, or exhibiting hallucinated syntax is automatically rejected and retried with exponential backoff. This rigorous filtering resulted in a high-fidelity memory bank containing 84 well-structured memory entries ready for retrieval.

A.3 Memory Bank Statistics and Case Example

The final memory bank comprises 84 distinct memory entries derived from the development set failures. Table 5 details the distribution of these empirical failure trajectories across the repositories included in the SWE-bench Multimodal benchmark. This distribution ensures the retrieved memory entries cover a wide spectrum of visual structures, rendering frameworks (e.g., PDF generation, Canvas drawing), and architectural patterns.

Table 5: **Repository Distribution of the Memory Bank.** The 84 failed trajectories used for extracting negative constraints span 5 distinct repositories.

Repository	Count of Failed Trajectories
Automattic/wp-calypso	37
chartjs/Chart.js	16
markedjs/marked	11
diegomura/react-pdf	10
processing/p5.js	10
Total Memory Entries	84

To illustrate the exact payload injected into the agent’s context during Phase 3, we present a complete, distilled memory entry for the instance `Automattic__wp-calypso-21964` in Figure 4. Extraneous metadata utilized exclusively for

offline pipeline routing has been omitted for clarity.

B Extended Methodology Details

This appendix provides detailed engineering descriptions of the core components discussed in the main text: the skeleton compression strategy used for context management (§B.1), the implementation of Active Perception tools (§B.2), and the sandboxed execution environment for codebase exploration (§B.3).

B.1 Skeleton Compression Strategy

Providing the complete source code of all candidate files to the model is often impractical: a single file in a large front-end repository can span thousands of lines, and the full context of multiple files easily exceeds the effective context window of current language models. To address this, we design a *skeleton compression* strategy that retains the structural outline of each file while aggressively removing implementation bodies, yielding a concise yet informative representation.

AST-Based Structure Extraction. We leverage an Abstract Syntax Tree (AST) parser to identify the structural elements of each source file. Concretely, we employ the Babel parser—a widely-used, plugin-extensible parser for JavaScript and TypeScript—invoked as an external process. The parser is configured with a comprehensive set of syntax plugins (including JSX, TypeScript, class properties, optional chaining, nullish coalescing, decorators, pipeline operators, and others) to ensure broad compatibility across diverse repository coding conventions. For repositories that use non-standard or unconventional coding patterns, we maintain repository-specific parser configurations that handle edge cases such as recursive node traversal with cycle detection or null-safe property access during AST walking.

The parser performs a recursive traversal of the AST and extracts two categories of structural elements:

- **Classes:** Each class declaration is recorded with its name, start line, end line, and a list of its methods (each with name, start line, and end line). Both named and anonymous class declarations are captured, including those assigned via `module.exports`.
- **Functions:** Named function declarations, arrow function expressions, and function expressions

assigned to variables or object properties are recorded with their names and line spans.

Skeleton Construction. Given the extracted structure, we construct the skeleton representation through the following procedure:

1. **Initialize an empty template** of the same length as the original file (one empty string per line).
2. **Retain class boundaries:** For each class, copy the declaration header (start line) and the closing brace (end line) into the template. For each method within the class, similarly retain the method signature line and closing line.
3. **Retain function signatures:** For each function, copy the start line and the end line. Additionally, if the function signature spans multiple lines (e.g., functions with long parameter lists), we expand downward from the start line for up to 20 additional lines, stopping when we encounter an empty line or a line ending with a block delimiter (`{` or `}`). This ensures that multi-line function signatures are preserved in their entirety.
4. **Preserve comments, imports, and exports:** All comment lines (single-line `//`, multi-line `/* */`, and JSDoc `*` lines), `import` statements, and `export` statements are unconditionally retained. These provide semantic context about the file’s dependencies and public interface.
5. **Collapse consecutive blank lines:** Sequences of more than two consecutive blank lines are collapsed to at most two, preventing large gaps where function bodies were removed.

The compression is applied conditionally: files shorter than a configurable threshold (in lines) are provided in their entirety, since the overhead of full inclusion is minimal. Files exceeding this threshold are compressed. In rare cases where the compressed skeleton itself exceeds 5,000 lines (e.g., for exceptionally large auto-generated files), we truncate to the first 5,000 lines. If compression yields an empty result (e.g., for declarative configuration files with no class or function structure), the full file content is used as a fallback.

This skeleton format typically achieves a 5–20× compression ratio compared to the original file, making it feasible to present dozens of candidate files simultaneously within a single prompt while preserving the structural information necessary for fault localization.

B.2 Implementation of Active Perception Tools

Our framework equips the agent with two complementary visual analysis tools—CROP and GROUNDING—that enable it to actively manipulate and inspect bug scenario screenshots during the repair process. Both tools operate on raw pixel coordinates specified directly by the agent.

Coordinate Specification. When the agent invokes a visual tool, it outputs a bounding box in the form $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$, where coordinates are specified in absolute pixel units relative to the top-left corner of the image. The agent determines these coordinates by reasoning about the spatial layout of the screenshot based on its visual understanding. No external grounding module (e.g., Set-of-Mark prompting or DOM tree parsing) is employed; the agent directly estimates pixel regions from the rendered screenshot. An image index parameter allows the agent to select which screenshot to operate on when multiple bug scenario images are available.

Crop Tool. The CROP tool extracts a sub-region from a screenshot to enable detailed inspection of fine-grained visual artifacts. Upon receiving the agent’s bounding box, the system:

1. Decodes the base64-encoded screenshot into a pixel buffer.
2. Validates and clamps the bounding box coordinates to the image boundaries to prevent out-of-bounds errors.
3. Extracts the specified rectangular region.
4. Re-encodes the cropped region as a new image and injects it into the subsequent conversation turn as a user message containing the cropped image.

This tool is designed for scenarios requiring pixel-level examination, such as verifying whether a border is 1 px too thick, whether an icon is rendered at incorrect resolution, or whether text overflow is occurring at a specific breakpoint.

Grounding Tool. The GROUNDING tool annotates the original screenshot with a bounding box overlay to highlight the bug-affected region while preserving the full page context. Given the agent’s bounding box and an optional text label, the system:

1. Decodes the screenshot into a pixel buffer.
2. Draws a colored rectangle (red, 3 px width) at the specified coordinates on a copy of the image.

3. If a text label is provided, renders it above the bounding box with a contrasting background for readability.
4. Re-encodes the annotated image and injects it into the conversation as a new visual input.

This tool is particularly useful for layout and positioning bugs, where the spatial relationship between the highlighted region and the surrounding page elements is critical for diagnosis.

Tool Usage Protocol. Both tools follow a strict single-tool-per-turn protocol: the agent issues exactly one tool call per response, then waits for the system to return the actual result (the processed image) before proceeding. This prevents hallucinated tool outputs and ensures that subsequent reasoning is grounded in real visual evidence. The agent is encouraged to use visual tools as its *first* action when bug scenario images are available, establishing a concrete visual understanding before proceeding to code-level analysis. The processed images (both originals and tool outputs) are persisted to disk for post-hoc analysis.

B.3 Bash Environment Constraints

To support codebase exploration during patch generation, we provide the agent with a sandboxed shell execution environment. This allows the agent to inspect files, search for patterns, and understand the repository structure when the provided code snippets are insufficient for generating a correct fix.

Execution Model. Each shell command is executed as an independent subprocess with its working directory set to the repository root. The `cd` command is explicitly blocked, as directory changes do not persist across invocations; instead, the agent uses relative paths from the project root (e.g., `cat src/components/Button.js`). This stateless execution model simplifies security enforcement and prevents path-related confusion.

Security Isolation. We enforce a multi-layered security policy to ensure that the agent’s shell access is strictly read-only:

- **Command blacklist:** Destructive file system operations (`rm`, `mv`, `cp`, `mkdir`, `touch`, `chmod`), version control mutations (`git reset`, `git checkout`, `git clean`, `git merge`, `git rebase`, `git push`), network utilities (`curl`, `wget`, `ssh`, `scp`), package managers (`npm`, `pip`, `apt`), and process control commands (`kill`, `sudo`) are rejected

before execution.

- **Redirect blocking:** Output redirection operators (`>`, `>>`, `2>`, `&>`) are forbidden to prevent any file writes. Pipe operators (`|`) are allowed for command chaining (e.g., `grep | head`) but are validated to ensure the downstream command is not a write-capable utility.
- **Injection prevention:** Command substitution (`$(...)` and backticks) and background execution (`&`) are blocked to prevent privilege escalation and uncontrolled process spawning. In-place file editing via `sed -i` is specifically detected and rejected while allowing read-only `sed` usage for text extraction.
- **Command chaining:** Semicolon-separated command chains are permitted, but each sub-command is individually validated against the blacklist before execution.

Resource Limits. To prevent runaway processes and excessive context injection, we impose the following constraints:

- **Timeout:** Each command is subject to a 120-second execution timeout. Commands exceeding this limit are terminated, and the agent receives a timeout notification.
- **Output truncation:** Command output is capped at 300 lines or 50 KB (whichever is reached first). When truncation occurs, the agent is informed and advised to narrow its query (e.g., by using more specific patterns or limiting the search scope to a subdirectory).
- **Directory exclusion:** Common non-source directories (`node_modules`, `.git`, `dist`, `build`, `coverage`, etc.) are excluded from search and traversal operations to reduce noise and improve response time.

Permitted Operations. Within these constraints, the agent can freely execute read-only operations including file reading (`cat`, `head`, `tail`), pattern searching (`grep` with recursive, case-insensitive, and context-line options), file discovery (`find`, `ls`), and text processing (`wc`, `sort`, `awk` in read-only mode). This set of operations is sufficient for the agent to navigate unfamiliar codebases, trace import chains, locate related files, and gather the contextual information necessary for patch generation.

Memory Entry Example: Automattic__wp-calypso-21964

Contextual Layer ($\mathcal{L}_{ctx}$)

- **Issue Summary:** OAuth `client_id` parameter is dropped when navigating from Signup back to Login form, causing loss of custom branding/styling.
- **Visual Analysis:** The login page reverts to generic styling instead of custom branded styling after a user clicks 'Already have an account?'. The expected behavior is that the login page should retain custom branding (driven by `client_id`) when returning from the Signup form.

Cognitive Layer ($\mathcal{L}_{cog}$)

- **Cognitive Diagnosis:** The agent attempted to fix a View-layer link generation issue by modifying Controller-layer route initialization, failing to propagate data to the actual UI component. It calculated `initialUrl` in `client/signup/controller.js` to include query strings, assuming this would implicitly preserve the parameters during navigation. However, the 'Back to Login' link is explicitly rendered by a React component using a helper function, so the controller-level variable had no effect on the rendered UI. The Golden Patch demonstrates the 'Selector Pattern': retrieving the persisted state from the store within the View component and passing it explicitly to the URL generation utility.
- **Negative Constraints:**
 - Do NOT attempt to fix UI link generation issues by modifying controller initialization variables that are not passed to the view.
 - Do NOT assume URL parameters persist automatically across route changes; explicitly inject them into link generators.
- **Golden Principle:** Reconstruct navigation state explicitly in View components using Store Selectors, rather than relying on implicit Controller context preservation.

Code Layer ($\mathcal{L}_{code}$)

- **Failed Patch Summary:** Modified `client/signup/controller.js` (Controller). Added logic to reconstruct `initialUrl` from context including query strings. The bug wasn't about the controller losing context, but about the View component (SignupForm) generating a link that lacked the parameter. Modifying a local variable in the controller without passing it to the view has no effect on the rendered HTML.
- **Golden Patch Summary:** The fix involves updating the URL generation utility to support the `client_id` parameter, and connecting the View component to the Redux store to retrieve the current client ID.
 - `lib/paths/login/index.js` (Utility): Updated the login function to append `?client_id=...` if `oauth2ClientId` is provided, centralizing URL logic to ensure consistency.
 - `components/signup-form/index.jsx` (View): Used the connect wrapper to retrieve `getCurrentOAuth2Client(state)` from Redux and explicitly injected `oauth2ClientId` into the `getLoginLink` method, ensuring the component has access to global context data.

Figure 4: An instantiated Memory Entry utilized during the final repair phase. The structured fields align with the defined hierarchical memory architecture ($\mathcal{L}_{ctx}$, $\mathcal{L}_{cog}$, $\mathcal{L}_{code}$). It is presented across both columns for readability.