

Automatically Generating Web Applications from Requirements Via Multi-Agent Test-Driven Development

YUXUAN WAN*, The Chinese University of Hong Kong, China
TINGSHUO LIANG*, The Chinese University of Hong Kong, China
JIAKAI XU, Columbia University in the City of New York, USA
JINGYU XIAO, The Chinese University of Hong Kong, China
YINTONG HUO[†], Singapore Management University, Singapore
MICHAEL LYU, The Chinese University of Hong Kong, China

Developing full-stack web applications is complex and time-intensive, demanding proficiency across diverse technologies and frameworks. Although recent advances in multimodal large language models (MLLMs) enable automated webpage generation from visual inputs, current solutions remain limited to front-end tasks and fail to deliver fully functional applications. In this work, we introduce TDDev, the first test-driven development (TDD)-enabled LLM-agent framework for end-to-end full-stack web application generation. Given a natural language description or design image, TDDev automatically derives executable test cases, generates front-end and back-end code, simulates user interactions, and iteratively refines the implementation until all requirements are satisfied. Our framework addresses key challenges in full-stack automation, including underspecified user requirements, complex interdependencies among multiple files, and the need for both functional correctness and visual fidelity. Through extensive experiments on diverse application scenarios, TDDev achieves a 14.4% improvement on overall accuracy compared to state-of-the-art baselines, demonstrating its effectiveness in producing reliable, high-quality web applications without requiring manual intervention. The code of TDDev is available at <https://github.com/yxwan123/TDDev>.

CCS Concepts: • **Software and its engineering** → **Automatic programming**; • **Computing methodologies** → *Artificial intelligence*.

Additional Key Words and Phrases: Multi-modal Large Language Model, Code Generation, User Interface, Web Development

ACM Reference Format:

Yuxuan Wan, Tingshuo Liang, Jiakai Xu, Jingyu Xiao, Yintong Huo, and Michael Lyu. 2018. Automatically Generating Web Applications from Requirements Via Multi-Agent Test-Driven Development. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 22 pages. <https://doi.org/XXXXXXX.XXXXXXX>

*Both authors contributed equally to this research.

[†]Yintong Huo is the corresponding author.

Authors' Contact Information: Yuxuan Wan, The Chinese University of Hong Kong, Hong Kong, China, yxwan9@cse.cuhk.edu.hk; Tingshuo Liang, The Chinese University of Hong Kong, Hong Kong, China, 1155210994@link.cuhk.edu.hk; Jiakai Xu, Columbia University in the City of New York, New York, USA, ax2155@columbia.edu; Jingyu Xiao, The Chinese University of Hong Kong, Hong Kong, China, whalexiao99@gmail.com; Yintong Huo, Singapore Management University, Singapore, Singapore, ythuo@smu.edu.sg; Michael Lyu, The Chinese University of Hong Kong, Hong Kong, China, lyu@cse.cuhk.edu.hk.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, Woodstock, NY

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM

<https://doi.org/XXXXXXX.XXXXXXX>

Aspect	Front-End Development	Full-Stack Development
Goal	Show information and visuals in the browser	Front-End + read/write data and enforce rules
Output	Static webpage (a single HTML file)	Functional application (a project with multiple files)
Platform	Browser only	Browser + server
Data	No persistent data (or mock data embedded)	Real data stored
Examples	Marketing landing page; Personal homepage	Booking system; forum with posts/replies
Related Work	[4, 9, 11, 22, 26, 38, 41, 44, 46, 47, 50, 52, 54, 59, 60]	[34](benchmark study), Ours

Table 1. Differences between front-end and full-stack development.

1 Introduction

In the modern digital era, web applications play a pivotal role as foundational platforms that support a wide range of everyday activities. The scale of this ecosystem is immense: recent reports estimate more than 1.1 billion active websites, with an additional 252,000 new sites launched daily [2, 3].

Developing web applications, also referred to as full-stack development, involves two stages: front-end development and back-end development. The former is concerned with the graphical user interface (GUI), such as layouts, content, and interactive elements [36]. The latter focuses on the server-side logic that supports the application, including APIs, data storage, and request handling [5, 6].

Full-stack development is complicated and time-consuming. It requires proficiency in both client-side and server-side technologies and constant adaptation to evolving tools and frameworks [45]. For novices, the breadth of skills needed, e.g., HTML, CSS, JavaScript, databases, and server-side languages, creates a steep learning curve that hinders turning ideas into applications [45]. Even for experienced developers, the sheer quantity of frameworks, programming languages, and deployment models leads to cognitive overload [11, 29, 39, 41]; industry data show that learning a new development environment can take as long as a year [25]. These challenges necessitate an automated solution framed as **Req-to-App: automatically translating user requirements (e.g., text descriptions, design sketches) into a functional web application**.

Despite its importance, practical solutions for building full-stack web applications from user requirements remain underexplored.

The closest line of work focuses on the simpler “design-to-code” task, which only produces front-end webpages rather than full-fledged applications. Table 1 compares front-end and full-stack development and highlights related research. Within this scope, prior studies have focused on converting GUI designs into code for Android apps [4, 11, 38, 41, 59], or on generating synthetic datasets and simple designs [9] to train deep learning models (e.g., CNN, LSTM). More recently, Multimodal Large Language Models (MLLMs) have shown promise in generating webpage code from screenshots [23, 44, 46, 47, 50, 60]. However, full-stack development introduces additional challenges beyond layout rendering, including implementing backend logic, managing servers and databases, handling APIs and resources, and ensuring seamless integration across multiple components. Existing design-to-code methods thus fall short in addressing these requirements, leaving the challenges of full-stack automation largely unexplored.

Apart from research works, commercial tools such as Bolt.new¹ and Lovable.dev² are able to generate complete websites from user input. Yet, their performance remains limited: a recent empirical study [34] reports that applications generated by these frameworks with state-of-the-art LLMs fail to implement required functionalities or even fail to compile in over 70% of cases.

¹<https://bolt.new>

²<https://lovable.dev>

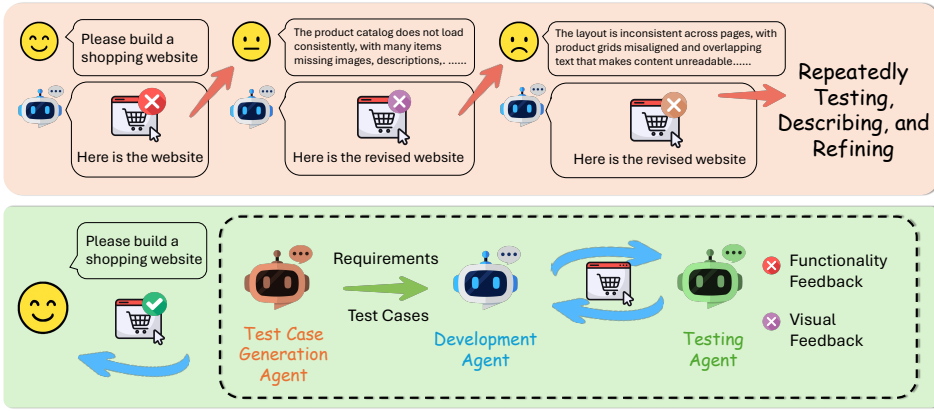


Fig. 1. Comparison between our proposed TDD framework (lower) and current industry tools (upper).

As a result, users must manually devise test cases, check functionality, describe errors or visual mismatches through the chat interface, and repeatedly request refinements until the application meets expectations—a process that is slow, labor-intensive, and frustrating [8].

Test-driven development (TDD) is a software engineering practice where developers iteratively write a test for a specific feature, implement code to satisfy that test, and then refine the code to improve quality [35]. Inspired by TDD, we believe that an agent-based system can collaboratively emulate such a workflow to handle Req-to-App without further user intervention. Specifically, we expect the development agents to: **1) generate test cases from user requirements to validate correctness and functionality, 2) produce full-stack code and test it against the generated tests, and 3) iteratively refine the code based on test results until all cases pass and the application meets the desired standards.** Figure 1 shows a comparison between our proposed pipeline and current industry tools.

Unfortunately, creating a TDD framework in the context of web development presents several unique challenges. **First**, test case generation methods in traditional TDD are not directly applicable to the Req-to-App scenario because (1) user requirements are usually high-level and under-specified (e.g., “Please create a shopping website”), making it difficult to derive comprehensive, executable test cases (e.g., user registration, shopping cart utilities); (2) web testing focuses more on non-crash defects such as interaction and logical errors, which often evade conventional TDD techniques [33]. This requires test cases to capture user interaction flows (e.g., entering a username and password, clicking login) rather than simple input–output pairs. **Second**, web application test cases must simultaneously account for user interaction flows and UI layout checks, which traditional script-based or unit-test TDD strategies cannot adequately capture [57]. Addressing these requirements calls for end-to-end agents capable of performing fine-grained element recognition, identifying usability issues that are technically functional but practically unusable (e.g., low contrast, overlapping elements), and remaining robust to unexpected scenarios such as *login walls*—barriers that require authentication before granting access to subsequent functionality. **Finally**, beyond testing, the framework must deliver rich feedback covering both functional correctness and visual presentation to effectively guide refinement—an aspect that has also been unexplored in existing work. Table 2 provides a comprehensive comparison between TDD on traditional code tasks and TDD in Req-to-App scenario.

In this work, we introduce **TDDev, the first TDD-enabled LLM-agent framework for full-stack web development**, which automatically generates high-quality web applications from natural language descriptions or image designs without requiring further user intervention. Specifically, we first design a test case generation agent that leverages requirement engineering and soap-opera test techniques to derive detailed requirements and test cases from high-level user requirements. Then, we implement a development agent that integrates file handling, planning, and memory capabilities for full-stack development. Finally, we create a testing agent capable of simulating user interactions based on provided test cases. Together, these agents are orchestrated in a TDD workflow that generates high-quality web applications from user requirements in an end-to-end manner.

To evaluate the effectiveness of TDDev, we propose a dataset **Req-to-App-MM**, which is an augmented version of WebGen-Bench [34], a benchmark designed to assess LLM agents' ability to generate multi-file website codebases from text requirements. We extended WebGen-Bench to a multimodal setting—where users can provide both design images and text for higher-fidelity requirements. We evaluated TDDev on the extended benchmark, and the results show that TDDev achieves a 14.4% improvement on overall accuracy compared to state-of-the-art baselines, demonstrating its effectiveness in producing reliable, high-quality web applications without requiring manual intervention.

In summary, our contributions are as follows:

- We introduce TDDev, the first TDD-enabled LLM-agent framework for full-stack web development, addressing the unique challenges of integrating TDD into web application generation.
- We design and implement an orchestration of three specialized agents: (1) a test case generation agent that transforms under-specified user requirements into executable test cases, (2) a development agent with planning, memory, and file-handling capabilities for multi-file full-stack projects, and (3) a testing agent that simulates realistic user interactions and provides feedback on both functionality and visual presentation.
- We conduct extensive experiments across diverse web development scenarios, showing that TDDev substantially improves generation quality, achieving a 14.4% higher overall accuracy, and significantly decreases human development effort in the user study.
- We release the implementation of TDDev, to foster future research on TDD-driven web application generation and support practical development tasks.

2 Background

2.1 Task Formulation: Req-to-App

This task takes a piece of text description and optionally a visual webpage design as input, aiming to generate code that satisfies the description. Let T_0, I_0 represent the text description and image (optional) input by the user. Given T_0, I_0 , an agent generates a web application $App = Agent(I_0, T_0)$. The functionality of App should closely match T_0 , and the GUI of App should resemble I_0 .

2.2 Related Work

2.2.1 UI Code Generation. UI code generation produces front-end code from screenshots or design images. Early approaches typically rely on CNNs and Computer Vision (CV) techniques for automated GUI prototyping [7, 9, 11, 13, 15, 38, 42, 56]. The advent of MLLMs has enabled more advanced approaches, such as Design2Code [44]. To address element omission, distortion, and misarrangement, DCGen [47] introduced a divide-and-conquer strategy, while LayoutCoder [50] and UICopilot [23] adopted layout-aware techniques. DeclarUI [59] further incorporated page transition graphs for mobile UI generation. ScreenCoder [26] implement a modular agent system

and reinforcement learning (RL) framework for front-end development. Interaction2Code and DesignBench [53, 54] added interaction-aware generation and repair, and MRWeb [46] explored resource-aware generation. EfficientUICoder [55] proposed a token compression framework for efficient UI code generation. Despite these advances, existing methods remain limited to front-end webpages without functionality. Recently, an empirical study, WebGenBench [34], benchmarked industry tools on full-stack development, revealing that applications generated with state-of-the-art LLMs fail to implement required functionalities—or even compile—in over 60-70% of cases, underscoring the need for more effective approaches to improve the performance of LLM-based agents in automatic full-stack web development.

2.2.2 LLM Agents. The rise of Large Language Model (LLM) agents [18, 51] has opened new opportunities for automation in software engineering. These agents have demonstrated effectiveness in code generation [16], program repair [10], and program testing [19]. Building on LLM agents, commercial tools such as Bolt.new³ and Lovable.dev⁴ are able to generate complete websites from user input. However, their performance is limited and requires extensive user intervention, which is frustrating [8, 34]. Figure 1 shows a comparison between our proposed pipeline and current industry tools.

2.2.3 GUI Testing. Automated GUI testing has been studied through multiple approaches. Record-and-replay methods are straightforward to implement but tend to be fragile and require frequent maintenance as applications evolve [57]. Random-based tools such as Monkey [1] reduce manual effort but often achieve limited functional coverage. Model-based testing [21, 37] introduces more structure by deriving cases from formal models, though its effectiveness is restricted by limited model accuracy and continuous model updates, and it generally overlooks GUI semantics. Learning-based methods [28, 30, 43], often built on reinforcement learning, can discover testing policies but require substantial training data and adapt less effectively to rapidly changing applications with a lack of deeper semantic understanding [32]. Recently, MLLM-based methods [32, 33] have started to leverage visual semantics and functional structures, offering a promising direction for GUI testing. Their planning, however, remains largely exploratory, making it difficult to achieve systematic coverage of targeted functionalities (e.g., evaluating the usability of a shopping cart). Moreover, most prior work has centered on the Android platform, whereas our focus is on general web applications, where testing must address both overall UI layout and fine-grained component behavior within the broader Req-to-App task, rather than the more tailored, touch-oriented layouts of mobile interfaces.

2.2.4 Test Driven Development. Improving code generation with tests has become a popular direction in traditional code tasks. Wang et al. [48] use test execution feedback during training to detect errors, while Chen et al.'s Codex [12] highlights the limits of single-sample generation and motivates iterative testing. Compiler feedback has been exploited to ensure compilable code [49]. AutoCodeRover [58] applies test cases for spectrum-based fault localization. Mathews et al. [35] empirically demonstrated that applying TDD principles to LLM-based code generation is beneficial using human-written tests. Liu et. al [31] designed an automated test-driven checker development approach with LLM. However, all of these works focus on traditional code tasks, and most of them on program repair. Due to the fundamental difference between web development and traditional code generation, as shown in Table 2, none of these methods applies to the full-stack development scenario.

³<https://bolt.new>

⁴<https://lovable.dev>

Table 2. Comparison between traditional TDD and web application TDD.

Aspect	Traditional Code TDD	Web Application TDD
Scope	Single function, class, or algorithm in isolation.	Integration of front-end and back-end.
Test Cases	Clear unit tests with well-defined input-output pairs.	Derived from vague user requirements and expanded into functional and visual validations.
Test Methods	Automated unit tests or compiler checks.	End-to-end user simulation and UI visual evaluation.
Artifacts	Single source file or self-contained program.	Multiple interdependent files (HTML, CSS, JavaScript, server code, database, configuration).
Feedback	Binary results (pass/fail) or compilation errors.	Rich feedback on both functional correctness and visual presentation.
Example	Sorting an array, computing shortest path, or matrix multiplication.	Shopping website with product catalog, cart, login integration, and responsive UI.
Related Work	[12, 31, 35, 48, 49, 58]	Ours

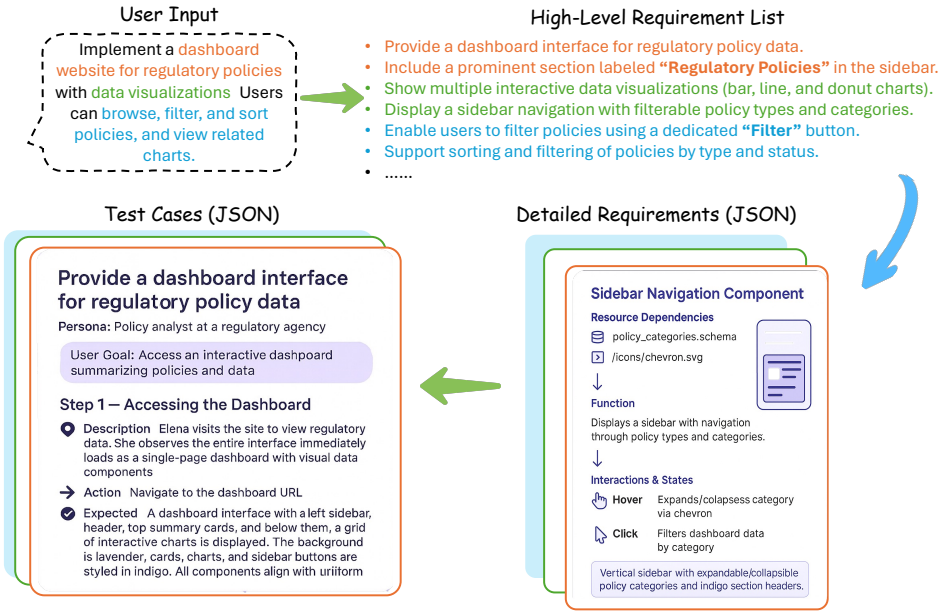


Fig. 2. The workflow of the test generation agent.

3 Methodology

This section introduces the mechanism of TDDev, a framework that automatically generates high-quality web applications from natural language descriptions or image designs. TDDev begins with a test case generation agent, which applies requirement engineering and soap-opera testing techniques to derive detailed requirements and test cases from high-level user inputs. It then employs a development agent equipped with file handling, planning, and memory capabilities to perform full-stack development. Finally, a testing agent inspects the generated application by simulating user interactions according to the test cases. These agents are orchestrated within a TDD workflow that iteratively refines the application based on testing feedback until the output meets the desired standards.

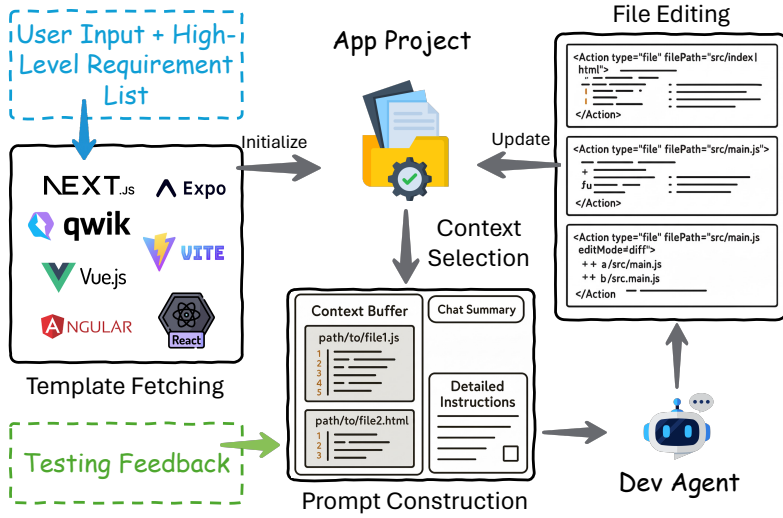


Fig. 3. Workflow of the development agent.

3.1 Test Generation Agent

This stage aims to refine the input into a set of actionable and concrete requirements, each paired with a corresponding test case. Figure 2 illustrates the workflow. The agent takes as input a high-level web development instruction, optionally accompanied by a screenshot of the expected final web application. It then progressively transforms it into concrete specifications ready for testing.

The workflow is orchestrated across three specialized agents. The **requirement decomposition agent** first parses the input and produces a structured list of discrete requirements. Each requirement is expressed as a concise statement, typically describing a functionality, layout constraint, or design element. Beyond explicit instructions, the agent also infers implicit requirements necessary to make the application complete and feasible.

Next, the **requirement elaboration agent** enriches each high-level item with detailed specifications, covering functionalities, static UI design, and dynamic interactions. At this stage, relevant data sources are also identified, which may include predefined datasets, databases, or external APIs. Depending on the data type, the agent either provides direct content or generates database setup instructions with defined schemas. The refined outputs are represented in JSON format, ensuring determinism and executability in subsequent implementation.

Finally, the **test case generation agent** produces one test case for each requirement, ensuring complete coverage and alignment with user expectations. Inspired by the concept of *soap opera testing* [27], the agent begins by imagining a user persona with specific goals and then generates step-by-step instructions describing user actions and expected outcomes. These structured cases not only capture realistic usage scenarios but also reduce ambiguity and minimize testing errors.

Overall, this process simulates industrial development workflows by systematically bridging high-level user requirements and concrete testable specifications, thereby ensuring strong alignment between the final application and the original intent [14].

3.2 Development Agent

Forked from Bolt.diy, we implement a development agent equipped with file management capabilities to support end-to-end application development. The agent takes as input both the initial

Table 3. 13 different starter templates from GitHub covering various frameworks and use cases.

Template	Description
Expo App	Cross-platform mobile app development
Basic Astro	Static website generation
NextJS Shadcn	Full-stack Next.js with shadcn/ui components
Vite Shadcn	Vite with shadcn/ui components
Qwik TypeScript	Resumable applications
Remix TypeScript	Full-stack web applications
Slidev	Developer presentations using Markdown
SvelteKit	Fast web applications
Vanilla Vite	Minimal JavaScript projects
Vite React	React with TypeScript
Vite TypeScript	Type-safe development
Vue.js	Vue applications
Angular	Angular applications with TypeScript
SolidJS	Lightweight reactive applications

natural language requirement from the user and the structured high-level requirement list produced by the test generation agent. It then initializes a project from a suitable template, iteratively determines which files should be edited, and applies modifications until the application satisfies the requirements. Figure 3 illustrates the workflow of development agent.

Template Fetching. To bootstrap development, the agent leverages 13 open-sourced starter projects from GitHub, covering a wide range of frameworks and use cases (Table 3). During initialization, the agent prompts an LLM to classify the application type based on the input requirements and select the most appropriate template. The chosen template is then cloned into the working directory, serving as the seed project for subsequent development.

Context Selection. Once initialized, the agent adapts the seed project to the user requirements or feedback. To do so, it constructs a focused context prompt for code generation and file editing. Context construction proceeds in three steps: 1) File filtering: heuristics specific to each template exclude irrelevant files (e.g., `node_modules`, build artifacts, hidden directories). 2) Context buffer extraction: the system collects the current state, including (i) all available file paths, (ii) already loaded files, and (iii) the chat summary with the user’s latest request. 3) Relevance selection: the agent invokes a dedicated LLM instance to analyze the context buffer and determine which files are relevant. The agent is instructed to respond in a structured XML format, explicitly marking included and excluded files. An illustrative example of the selection prompt is shown below:

Prompt For Selecting Relevant Context

You are a software engineer working on a project. You have access to the following files:
 [Available Files]
 You have the following code loaded in the context buffer that you can refer to:
 [Context Buffer]
 Select only the files relevant to the current task. Respond in the following XML format:
 [Response Format]

Prompt Construction. The system parses the LLM response to extract included File and excluded File directives, updating the context buffer by adding newly relevant files and removing those no longer needed for the current development task. The updated context buffer is formatted into a

structured prompt. Each included file is wrapped in a file tag with the full file content, creating a comprehensive view of the codebase that the LLM can reference when making modifications. The prompt for development is shown below:

Prompt for Development

You are a software engineer working on a project. Below is the artifact containing the context loaded into context buffer for you to have knowledge of and might need changes to fulfill current user request.

<Context Buffer>

<file filePath="path/to/file1.js"> [File Content With Line Number] </file>

<file filePath="path/to/file2.html"> [File Content With Line Number] </file>

</Context Buffer>

Here is the chat history till now:

[Chat Summary]

Please make necessary updates to the project based on the instructions:

[Detailed Instructions & Response Format]

File Editing. After the development prompt is constructed, the LLM generates a structured response containing one or more tags that specify file operations required to fulfill the user request. Each action corresponds to a file creation, modification, or update operation within the project workspace.

For most cases, file modifications follow a full-content replacement strategy. The LLM outputs the complete updated file content within a <Action type="file" filePath="path/to/file"> tag. This approach guarantees consistency, eliminates ambiguity about the final file state, and avoids potential merge conflicts, ensuring that the edited file can be safely written into the workspace without relying on prior versions. In addition, the system optionally supports a diff-based editing mode for scenarios where only small, localized changes are required. In this mode, the LLM encloses changes within file modification tags using unified diff syntax, specifying exactly which lines to add, modify, or remove. Unlike the default full-content replacement, diff-based updates are explicitly invoked and optimized for efficiency when the modified portions are small relative to the file size.

Before applying edits, the system cleans formatting artifacts—such as Markdown code fences and escaped HTML entities—for workspace compatibility. It also enforces file-locking by listing protected files in the system prompt, ensuring the LLM skips any modifications that would compromise project integrity.

3.3 Testing Agent

The testing agent takes as input the entire project directory produced by the development agent and the test cases generated by the test case generation agent. It executes these test cases on the application and provides structured feedback to guide further refinement.

Deployment Verification. The testing agent first launches the application using PM2⁵, a widely used process manager for Node.js applications, and hosts it on a local browser. During the launch phase, it performs an initial verification by capturing a screenshot to check for errors such as blank screens or crash messages. If a launch failure is detected, all subsequent tests are aborted, and the agent immediately returns feedback containing logs and diagnostic information. When an expected screenshot is provided, the captured and reference screenshots are compared to identify visual discrepancies, which are then included in the feedback.

⁵<https://pm2.keymetrics.io/>

User Simulation. To simulate the interaction flows specified in the soap opera test cases (Section 3.1), we integrate Browser-Use [40], an LLM-powered autonomous browser agent with 69.7k GitHub stars. Browser-Use enables natural language control of a browser, supporting navigation, form filling, data extraction, and multi-step workflows (e.g., “Book a flight from Hong Kong to New York on the United Airlines website”). For efficiency, test cases are executed in parallel: multiple application instances are launched, and each instance is tested independently by a Browser-Use agent. The agent strictly follows the soap opera testing paradigm, enacting the defined user persona, executing step-by-step actions, verifying outcomes against expectations, and logging any deviations.

Feedback Construction. After executing the test cases, the testing agent processes the Browser-Use logs to produce comprehensive feedback. Each testing report details the failed steps, executed actions, expected versus actual outcomes, error categories, supplementary technical information, and actionable recommendations to guide correction and prevent recurrence.

Error Handling. In addition, the testing agent is equipped with autonomous error-handling capabilities. It incorporates a bounded retry mechanism to handle transient failures, which prevents infinite execution loops. We observe that the agent exhibits sufficient environmental awareness and problem-solving capabilities. For instance, upon encountering a login modal with “account does not exist” error, it can autonomously attempt to register and authenticate before proceeding with the predefined test steps.

4 Experiment

4.1 Research Questions

We evaluate the effectiveness of TDDev on automating full-stack web development through answering the following research questions (RQs):

- **RQ1:** How effective is TDDev in automating full-stack web development?
- **RQ2:** Does the multi-step design of the test case generation agent improve effectiveness?
- **RQ3:** Can feedback from the testing agent enhance the quality of generated applications?
- **RQ4:** In what ways does TDDev provide advantages for developers over existing open-source tools?

4.2 Baselines

We evaluate two widely used open-sourced and proprietary industry-level code-agent frameworks as baselines: Bolt.diy⁶ and Cursor⁷. Bolt.diy, the open-source version of Bolt.new, is a browser-based framework for generating and previewing web applications, with 17.7k stars on GitHub. It provides a user interface with a Linux-like WebContainer and prompts the model to select frontend and backend frameworks (e.g., Vite, React, Remix) before importing and extending a template. Cursor, a proprietary AI-assisted integrated development environment (IDE) built on Visual Studio Code, provides features such as code generation and an agent mode for end-to-end task execution. It has become one of the most widely adopted proprietary coding assistants, reportedly reaching 360k active users by 2024 [17]. In our experiments, we employ Cursor’s agent mode, manually supplying the WebGen-Bench requirement as the initial prompt using a subset of 10 WebGen-Bench test data. Whenever the agent raises follow-up questions, we either select the default option or allow the agent to decide autonomously, continuing this process until no further queries are made.

⁶<https://bolt.new/>

⁷<https://cursor.com/>

Table 4. Comparison of development and testing costs across different frameworks per round. One round refers to a full cycle of code generation and feedback.

Method	Develop (Token/Time)	Test (Token/Time)	Cost (Claude-4-Sonnet)
Cursor	Unknown / ~4 min	Manual	20 USD/month
Bolt.diy	~10K / ~4 min	Manual	0.18 USD/round
TDDDev	~10K / ~4 min	~10K / ~4 min	0.36 USD/round

4.3 Backbone Models

We evaluate the frameworks on two state-of-the-art (SOTA) general-purpose proprietary: GPT-4.1 and Claude-4-Sonnet. We also discuss the performance of our framework on open-sourced models such as Qwen-2.5-VL and DeepSeek-V3.1 in Section 6.

4.4 Experiment Setup

All experiments are run on an iMac with 10 Core Intel Core i9 processor, 32GB RAM. All LLM models are accessed through the official API services. Temperatures are set to 0 for all models. Max tokens are set to the maximum allowable value for each model.

4.5 Cost & Efficiency

As shown in Table 4, all methods require roughly four minutes to develop, but differ in testing and cost structures. Bolt.diy and Cursor rely on manual testing, with low per-round cost (0.18 USD) or flat subscription (20 USD/month) but requiring constant user attention and manual intervention. In contrast, TDDDev incurs a higher per-round token cost (0.36 USD) but fully automates testing and refinement, eliminating manual effort and allowing developers to focus on other tasks, making it more efficient despite slightly higher computational cost.

4.6 Evaluation

4.6.1 Dataset Construction. We use the test set of WebGen-Bench [34], a benchmark designed to assess LLM agents' ability to generate multi-file website codebases from scratch, as our core evaluation data. However, WebGen-Bench is text-only, with user requirements expressed purely in natural language. To extend it to a multimodal setting—where users can provide both design images and text for higher-fidelity requirements—we augment the benchmark by generating images for each text description using Gemini-2.5-Flash-Image [20], a multimodal LLM capable of producing realistic designs from prompts. These images serve as additional visual requirements for the agents. We name this dataset Req-to-App-MM. During evaluation, we further assess visual consistency between the generated websites and the corresponding design images. Figure 4 shows an instance of the data, and the statistics of the dataset are shown in Table 5. To perform the evaluation, we randomly select 10 websites while preserving the category distribution of the test set.

4.6.2 Functionality Metrics. We adopt a UI agent-based evaluation for functionality correctness. We utilize BrowserUse, one of the SOTA web navigation UI agents⁸, to execute test operations and verify outcomes. Each test case's operation and expected result are encoded into a standardized prompt, which directs the agent to simulate interactions, analyze trajectories and screenshots, and return *YES*, *NO*, or *PARTIAL* for requirement fulfillment. When the interaction limit is reached without completion, a decision prompt induces the agent to make a final judgment. We employ

⁸<https://browser-use.com/posts/sota-technical-report>

Fig. 4. An example data instance. Only 1 out of 7 test cases are shown.

Input	Category	Test Case (1 out of 7 shown)	Design Image
Please implement a dashboard website for displaying regulatory policies. The website should have data visualization capabilities to clearly display complex regulatory policies. Users should be able to browse and analyze different policies, view related data and charts, and be able to filter and sort as needed. Assign lavender to the background and indigo to the UI elements.	primary_category: Content Presentation subcategories: Data Visualization, Static Page Generation	task#1: "Verify that the website uses 'lavender' as the background color and 'indigo' as the component color, as specified in the design requirements." expected_result#1: "The website has a background color of 'lavender' and components (such as buttons, cards, headers, etc.) are styled with the color 'indigo', accurately reflecting the design instruction." task#2: (omitted for brevity)	

Table 5. The number of website-generation instructions in each technical category in Req-to-App-MM. Each main category contains multiple subcategories. A sample may belong to one main category and multiple subcategories.

Main Categories	Num.	Sub Category	Num.
Content Presentation	28	Static Page Generation	20
		Dynamic Content Rendering	18
		Data Visualization	36
		Media Display	6
User Interaction	49	Form Systems	40
		Authentication	18
		Real-time Features	20
		E-commerce	22
		AI Integration	19
Data Management	24	CRUD Operations	29
		API Integration	20
		Big Data	12
		File Handling	5
Total	101		

Claude-4-Sonnet as the agent’s engine. All the instructions for the evaluation agent are adopted from [34].

4.6.3 Visual Metrics. Apart from functionality correctness, we also evaluate the visual quality of generated websites. We use a set of metrics covering rendering success, content relevance, layout harmony, design modernness, and similarity with the design image. These metrics are formulated into a prompt and assessed by Claude-4-Sonnet, which assigns a score from 1 to 5 (higher is better) to quantify the overall aesthetics and visual consistency of each website. The procedure and prompts for visual quality evaluation are adopted from [34]. The prompt for deciding visual similarity is shown below:

Visual Similarity Evaluation Prompt

You are an expert web designer and evaluator. Your task is to assess how well a generated website matches a given design image. Consider the following aspects: 1. **Layout**: Does the overall structure of the website (e.g., positioning of sections, grids, spacing) align with the design image? 2. **Components**: Are the required UI elements (e.g., buttons, forms, images, navigation bars) present and placed correctly? 3. **Color Scheme**: Are the background, text, and component colors consistent with the design image? 4. **Size and Proportion**: Do elements (e.g., images, text blocks, buttons) have similar relative sizes and proportions compared to the design image? 5. **Visual Fidelity**: Does the overall look and feel of the website closely resemble the design image?

Instruction: Compare the website rendering against the design image. Provide an overall similarity score from 1 to 5 (1 = very poor match, 5 = almost identical).

5 Result & Analysis

5.1 RQ1: Effectiveness of TDDDev

5.1.1 Overall Effectiveness. This section evaluates the overall performance of TDDDev compared with the baselines. The result is presented in Table 6. Accuracy is computed using the formula $\text{Accuracy} = \frac{N_{\text{Yes}} + 0.5 \times N_{\text{Partial}}}{N_{\text{Total}}} \times 100\%$, where N_{Yes} and N_{Partial} denote the number of test cases assessed as YES and PARTIAL, respectively, and N_{Total} is the total number of test cases.

We can see that TDDDev consistently outperforms both Bolt.diy and Cursor across most metrics. With GPT-4.1, TDDDev achieves the highest accuracy (78.2%), a relative improvement of +30% over Cursor (60.2%) and more than triple the performance of Bolt.diy (25.6%). Similarly, with Claude-4-Sonnet, TDDDev attains 70.2%, surpassing Cursor (63.8%) and substantially outperforming Bolt.diy (44.5%). Importantly, TDDDev has the lowest No rate (9.5% with GPT-4.1 and 15.0% with Claude-4-Sonnet), showing that it is less likely to fail functional requirements outright.

Reliability is another advantage: TDDDev shows no failures to start in either configuration, while both baselines exhibit frequent launch failures (e.g., 80.0% for Bolt.diy with GPT-4.1 and 40.0% for Cursor with GPT-4.1). This highlights TDDDev's robustness in generating runnable applications without manual intervention.

On user-facing criteria, TDDDev also delivers competitive results. Its appearance scores are consistently high (4.0 with GPT-4.1), and with Claude-4-Sonnet it achieves the best visual similarity (4.3), indicating faithful reproduction of UI designs. Although the similarity score is lower with GPT-4.1 (2.3), this suggests that the choice of model can affect the trade-off between functional accuracy and visual fidelity.

5.1.2 Detailed Result. Besides overall results, we calculate the accuracy for each category of instructions and test cases in Table 7. The result shows that TDDDev consistently achieves the strongest or near-strongest performance across most categories. For instruction categories, TDDDev with GPT-4.1 excels in handling user interaction (91.7%) and data management (78.6%), outperforming both Bolt.diy and Cursor. Claude-4-Sonnet achieves the best performance in content presentation (71.4%), while maintaining competitive scores in other categories. For test case categories, TDDDev also demonstrates superiority. With GPT-4.1, it achieves the highest accuracy in functionality (66.7%) and data display (87.5%), while Claude-4-Sonnet reaches perfect performance in design validation (100%).

Table 6. Evaluation of three code-agent frameworks using different proprietary and open-source models. Accuracy is computed using a weighted score, where YES samples are weighted by 1 and PARTIAL samples are weighted by 0.5; the total score is then divided by the number of test cases. The highest accuracy and appearance scores are marked in **bold**. Appearance score and Vis. Similarity score are between 1-5, other metrics are in percentage.

Test Name	Yes Rate	Partial Rate	No Rate	Fail to Start	Accuracy	Appearance	Vis. Similarity
TDDev							
GPT-4.1	65.9	24.6	9.5	0.0	78.2	4.0	2.3
Claude-4-Sonnet	55.0	30.0	15.0	0.0	70.2	3.7	4.3
Bolt.diy							
GPT-4.1	23.1	7.7	69.2	80.0	25.6	4.0	3.0
Claude-4-Sonnet	26.9	34.6	38.5	20.0	44.5	3.4	2.6
Cursor							
GPT-4.1	44.7	26.3	28.9	40.0	60.2	3.3	3.0
Claude-4-Sonnet	50.0	25.0	25.0	20.0	63.8	3.8	3.3

Table 7. Category-wise evaluation results. The first three columns represent categories of website-generation instructions, while the last three represent categories of test cases. All metrics are in percentage. N.A. implies all generated applications failed to start in the category.

Test Name	Instruction Categories			Test Case Categories		
	Content	User Interact.	Data Manage.	Functionality	Data Display	Design Validation
TDDev						
GPT-4.1	64.3	91.7	78.6	66.7	87.5	83.3
Claude-4-Sonnet	71.4	75.0	64.3	61.1	68.8	100.0
Bolt.diy						
GPT-4.1	42.9	8.3	N.A.	8.3	40.0	50.0
Claude-4-Sonnet	47.0	36.9	57.1	28.3	55.0	61.1
Cursor						
GPT-4.1	61.9	48.6	78.6	55.0	59.1	58.3
Claude-4-Sonnet	65.5	60.6	71.4	55.0	67.4	66.7

Table 8. Comparison of TDDev with straightforward test generation versus proposed test generation. Accuracy is reported as percentages. The best performance in each category is marked in **bold**.

Setting	Strtfwd. Test Gen + TDDev	Multi-Step Test Gen + TDDev
Functionality	33.3	61.1
Data Display	75.0	68.8
Design Validation	100.0	100.0
Overall Acc.	59.1	70.2
Vis. Similarity (1-5)	4.3	4.3

Answer to RQ1: TDDev is effective in automating full-stack web development. It achieves higher accuracy, stronger reliability, and more consistent visual quality than baseline frameworks across both overall and category-wise evaluations and across two different SOTA LLM backbones.

5.1.3 RQ2: Ablation Study on Test Case Generation Agent. In TDDev, we design a multi-step test case generation agent that consists of three stages: high-level requirement extraction, detailed requirement completion, and soap opera test generation. This design aims to progressively decompose user instructions into testable requirements while ensuring coverage of both front-end and back-end aspects. To assess the necessity of such a design, we compared TDDev with the multi-step agent against a variant using a straightforward test case generation agent. In the latter, the model directly produces test cases from the initial user input using the following prompt:

Straightforward Test Case Generation Prompt

You are an expert AI Product Manager. Your task is to analyze the following user-provided instruction for webpage development and break it down into a detailed list of requirements and test cases.
Your analysis must: 1) Capture every detail from the user's instructions without omission, no matter how minor. 2) Go beyond front-end representation by inferring and specifying necessary backend components (e.g., databases, APIs, integrations) to ensure the feature is fully functional and realistic. 3) Include details on database setup, API calls, and integration requirements wherever applicable. 4) Prefer real integrations over placeholders; when exact details are unavailable, clearly specify a placeholder rather than inventing information.
The final output should be a JSON array of this format: [JSON Format Specification]

The comparison in Table 8 reveals several insights. First, the multi-step agent yields a substantially higher functionality accuracy (61.1% vs. 33.3%), showing that progressive refinement better captures system-level requirements and functional dependencies. While the straightforward approach achieves slightly higher accuracy in data display (75.0% vs. 68.8%), this gain is offset by its weakness in functionality, suggesting that it focuses more on surface-level outputs rather than end-to-end correctness. Both approaches achieve perfect performance in design validation (100.0%), reflecting that visual styling requirements are easier to capture regardless of test case generation method.

Our further inspection shows that the performance gain in functionality is two-fold: (1) the high-level requirement list guides the development agent toward a deeper understanding of the task and more comprehensive planning, thereby improving code generation quality; and (2) the detailed test cases derived from refined requirements enable the testing agent to provide more targeted and fine-grained feedback, which helps the development agent refine implementations more effectively. These results demonstrate that decomposing test case generation into multiple steps is more effective than a single-shot approach, as it leads to stronger functional reliability and balanced coverage across categories while maintaining high design fidelity.

Answer to RQ2: The multi-step design of the test case generation agent is effective. It improves functionality accuracy and overall reliability by guiding the development agent with structured requirements and enabling the testing agent to deliver more targeted feedback.

5.2 RQ3: Ablation Study on Testing Agent Feedback

To study the effectiveness of the testing agent feedback, we compare the performance of TDDev without feedback and with different rounds of iterative feedback. In the no-feedback setting, the testing agent is disabled and only the initial output of the development agent is evaluated. In the feedback settings, we vary the number of feedback iterations by adjusting $max_iter = 1, 2, 3$ to examine how additional refinement cycles impact performance.

The results in Table 9 indicate several key trends. First, introducing 1-2 rounds of feedback does not immediately improve functional correctness or overall test accuracy—in fact, accuracy drops

Table 9. Effectiveness of the testing agent in TDDev. We compare TDDev without feedback and with different rounds of test feedback ($max_iter = 1, 2, 3$). TDD Pass Rate is the pass rate of generated test cases in the TDD workflow. Test Acc. is the overall accuracy of WebGen-Bench test cases.

Evaluation Setting	No Feedback	1 Round	2 Rounds	3 Rounds
Functionality	44.4	33.3	33.3	61.1
Data Display	56.3	56.3	43.8	68.8
Design Validation	100.0	100.0	100.0	100.0
TDD Pass Rate	19.6	16.3	30.4	33.8
Test Acc.	58.3	25.3	25.2	70.2
Vis. Similarity (1-5)	2.3	3.7	3.3	4.3

to 25.3%, suggesting that 1-2 rounds may be insufficient for the model to meaningfully refine the code (e.g., the agent may introduce new bugs when refining the code). However, as the number of rounds increases, improvements emerge. With three rounds of feedback, TDDev achieves the best performance across nearly all metrics: functionality rises to 61.1%, data display to 68.8%, and test accuracy to 70.2%, substantially outperforming the no-feedback baseline (58.3%).

Visual quality also benefits from iterative refinement. The visual similarity score increases from 2.3 (no feedback) to 4.3 with three rounds, indicating that feedback not only strengthens correctness but also helps align the generated application more closely with design specifications. Importantly, design validation remains perfect (100.0%) across all settings, showing that feedback mainly influences deeper functional and data-related aspects rather than surface-level styling.

Answer to RQ3: Testing agent feedback is effective in enhancing the quality of generated applications. Multiple rounds of feedback enable the agent to iteratively fix bugs and improve functionality, accuracy, and visual fidelity.

5.3 RQ4: Advantages for Developers Over Existing Open-Source Tools

We assess the practicality of TDDev in accelerating the development workflow. Specifically, we implement TDDev into a user-friendly demo tool and recruit three developers (two research staff who have previously developed at least two web applications, and one front-end developer from a startup company) to participate in our study, following the methodology of Chen et al. [11]. Each participant is tasked with developing a web application from a requirement in WebGen-Bench. Participants are asked first to develop the application using TDDev and then to develop the same application again with an open-source industry framework, Bolt.diy. In both cases, they refine the application until it reaches a satisfactory state.

We record the manual intervention time (inputting prompts, testing, and revising code) for each participant, as shown in Table 10, and capture their development processes through screen recordings.

The results highlight a key difference between the two methods. With Bolt.diy, participants spent an average of 4.7 minutes on manual interventions across a total of 15.2 minutes, requiring three rounds of interaction and 74 additional words of prompting. Although 4.7 minutes of manual effort may appear modest, this time is spread across the entire 15.2-minute session. For example, a participant first inputs the initial prompt, then after 3 minutes of agent execution must test and provide feedback, and again after another refinement cycle, more feedback is required. As a result, the agent demands the user's continuous attention, preventing them from focusing on other tasks.

Table 10. Comparison of manual intervention time.

Method	Manual/Total Time (min)	Intervention Frequency	Additional Prompt Len. (word)
Bolt.diy	4.7/15.2	3.0	74.0
TDDev	0.0/18.7	0.0	0.0

By contrast, with TDDev, participants required no manual interventions. Once the initial requirement was provided, the system autonomously handled development, testing, and refinement. Although the total development time was slightly longer (18.7 minutes), users had the full duration free to focus on other work without interruption.

This finding underscores a central advantage of TDDev: it shifts the workload from continuous prompt engineering and manual testing to autonomous, feedback-driven refinement. Developers are freed from micromanaging iterations and can concentrate on higher-level design decisions or parallel tasks, leading to a more efficient and less disruptive workflow.

Answer to RQ4: TDDev provides advantages over existing open-source tools by eliminating the need for manual interventions. Existing open-source tools require scattered user input and constant attention, while TDDev allows developers to remain fully disengaged during generation, reducing workload and enabling them to focus on other tasks while the system completes development autonomously.

6 Discussion

6.1 Accuracy of UI Agent Testing

Our evaluation relies on UI agent testing; to evaluate the accuracy of the UI agent testing process, we randomly sampled 5 generated web applications and manually verified the testing results produced by the agent, covering 28 test cases in total. We focus on Claude-4-Sonnet, given its high accuracy, and compare two UI agents: BrowserUse (used in our work) and WebVoyager [24] (used in WebGen-Bench). Manual testing serves as ground truth and requires precision; thus, two annotators independently labeled each case, with disagreements resolved by a third annotator. The Alignment Rate is defined as $\text{Alignment Rate} = \frac{N_{\text{Manual=Agent}}}{N_{\text{Total}}} \times 100\%$, where $N_{\text{Manual=Agent}}$ denotes the number of test cases where the agent-generated result matches the manual annotation. Table 11 reports performance measured by manual inspection and agent testing, along with alignment rates.

From these results, we observe: **(1) the Yes/No rates and Accuracy reported by BrowserUse closely track manual testing**, with only minor fluctuations of 0.8–6.2%; **(2) the overall alignment rate is satisfactory**, with 23 out of 28 cases perfectly aligned, yielding an 82.8% alignment rate.

To further interpret the relatively high Partial rate and the fluctuations in Yes/No/Accuracy, we examined the misaligned cases. We found that all agent judgment errors originate from *partially correct* cases, which often involve subjective strictness (e.g., whether an error should be raised when invalid input is detected). The agent tended to map human-annotated Yes or No cases into Partial, and occasionally labeled human-perceived Partial cases as Yes or No. **When restricting to strictly Yes/No cases, the alignment rate improves to 100% (23 out of 23).**

Table 11. Performance (%) reported by different testing entities and alignment rate (%) between UI agent testing results and manual testing results. Numbers in brackets indicate the difference from manual testing results.

Testing Method	Yes Rate	Partial Rate	No Rate	Acc	Align.	Yes/No Align.
WebVoyager	42.9 (-23.8)	25.0 (+11.7)	32.1 (+12.1)	55.4 (-18.0)	46.4	66.7
BrowserUse (TDDev)	62.1 (-4.6)	24.1 (+10.8)	13.8 (-6.2)	74.1 (+0.8)	82.8	100.0
Manual	66.7	13.3	20.0	73.3	–	–

6.2 Failure Analysis and Future Works

6.2.1 Scaling Feedback Rounds. During the experiment in RQ3, we observed that increasing the number of feedback rounds beyond five introduces greater variance in overall accuracy. On the one hand, additional rounds give the agent more opportunities to refine code and resolve existing bugs. On the other hand, as the codebase becomes increasingly complex after successive edits, it becomes more susceptible to the introduction of new errors. Consequently, excessive feedback cycles may not consistently yield performance gains.

This variation can be mitigated by straightforward selection strategies, such as executing multiple rounds (e.g., ten rounds) and selecting the best-performing output. In future work, we plan to explore more efficient mechanisms, such as branching and rollback strategies, that can scale feedback rounds while maintaining stability.

6.2.2 Open-Source LLMs. Beyond the widely adopted proprietary models Claude-4-Sonnet and GPT-4.1, we also conducted preliminary experiments with two popular open-source models: DeepSeek-V3.1, known for strong coding ability, and Qwen-2.5-VL-72B, known for strong visual capacity. Although TDDev is adaptable to these models, we identified several challenges:

- **Output Formatting:** TDDev requires strict output formats. The Test Generation Agent must produce JSON outputs, while the Development Agent depends on XML-structured outputs for parsing context and executing edits. Both DeepSeek-V3.1 and Qwen-2.5-VL-72B frequently failed these formatting constraints, leading to parsing errors or truncated outputs, which in turn caused workflow failures.
- **Coding Ability:** Despite Qwen-2.5-VL-72B’s strong visual reasoning, its limited coding ability resulted in a high number of fail-to-start cases.
- **Visual Ability:** DeepSeek-V3.1, while strong in code generation, lacks visual reasoning capacity and thus cannot independently power the visual-driven testing agent. When paired with Claude-4-Sonnet to provide testing feedback, performance remained limited due to formatting inconsistencies.

These observations suggest that current open-source LLMs face difficulties in agent workflows that demand structured output, code generation, and multimodal reasoning. As part of future work, we aim to identify or train models with stronger agentic capacities, particularly in enforcing output formats and balancing coding with visual reasoning.

7 Threat to Validity

We have identified the following threats to the validity of our work:

Effectiveness of the Agentic System Design. TDDev is built upon an intricate agentic workflow, and its performance may be constrained by the effectiveness of individual components. To mitigate this concern, we conducted an extensive ablation study, which validates the contributions of the test

case generation module, the testing module, and the overall TDD pipeline. These results provide evidence that the system design is both robust and effective.

Reliability of UI Agent Testing. Our evaluation relies on UI agent testing, which requires high levels of reliability and accuracy. To ensure trustworthiness, we supplemented automated testing with manual inspection. The results show that the performance reported by the UI agent closely aligns with human judgments, with an overall alignment rate that we consider satisfactory.

8 Conclusion

In summary, this work presents TDDDev, the first TDD-enabled LLM-agent framework capable of generating end-to-end full-stack web applications from natural language or visual specifications. By integrating automated requirement extraction, test case generation, iterative code refinement, and user interaction simulation, TDDDev overcomes the limitations of existing approaches that focus solely on front-end generation. Our experiments demonstrate that TDDDev not only ensures both functional correctness and visual fidelity but also significantly reduces failure rates compared to state-of-the-art baselines. These results highlight the potential of TDD-driven LLM agents to advance the automation of full-stack development, paving the way for more reliable, efficient, and scalable web application generation.

Data Availability

The code for TDDDev is available at <https://github.com/yxwan123/TDDDev> for replication and future research.

References

- [1] 2023. UI/Application Exerciser Monkey. <https://developer.android.com/studio/test/other-testing-tools/monkey>. Android Studio documentation, last updated 2023-04-12.
- [2] 2024. 17+ Surprising WordPress Statistics You Should Not Miss [2024]. *WPDeveloper* (2024). <https://wpdeveloper.com/wordpress-statistics-2024> Accessed: 2024-05-30.
- [3] 2024. How Many Websites Are There in 2024? (13 Latest Statistics). *TechJury* (2024). <https://techjury.net/blog/how-many-websites-are-there/> Accessed: 2024-05-30.
- [4] A. A. Abdelhamid, S. R. Alotaibi, and A. Mousa. 2020. Deep learning-based prototyping of android GUI from hand-drawn mockups. *IET Software* 14, 7 (2020), 816–824.
- [5] Amazon Web Services. 2025. *Front End vs Back End – Difference Between Application Layers*. <https://aws.amazon.com/compare/the-difference-between-frontend-and-backend/>
- [6] Amazon Web Services. 2025. *What is Full Stack Development?* <https://aws.amazon.com/what-is/full-stack-development/>
- [7] Batuhan Aşıroğlu, Büşta Rümeysa Mete, Eyyüp Yıldız, Yağız Nalçakan, Alper Sezen, Mustafa Dağtekin, and Tolga Ensari. 2019. Automatic HTML code generation from mock-up images using machine learning techniques. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*. Ieee, 1–4.
- [8] Joel Becker, Nate Rush, Beth Barnes, and David Rein. 2025. Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. *arXiv:2507.09089* [cs.SE] <https://arxiv.org/abs/2507.09089>
- [9] Tony Beltramelli. 2018. pix2code: Generating code from a graphical user interface screenshot. In *Proceedings of the ACM SIGCHI symposium on engineering interactive computing systems*. 1–6.
- [10] İsmel Bouzenia, Premkumar Devanbu, and Michael Pradel. 2024. Repairagent: An autonomous, llm-based agent for program repair. *arXiv preprint arXiv:2403.17134* (2024).
- [11] C. Chen, T. Su, G. Meng, Z. Xing, and Y. Liu. 2018. From UI design image to GUI skeleton: a neural machine translator to bootstrap mobile GUI implementation. In *Proceedings of the 40th International Conference on Software Engineering*. 665–676.
- [12] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).
- [13] W.-Y. Chen, P. Podstreleny, W.-H. Cheng, Y.-Y. Chen, and K.-L. Hua. 2022. Code generation from a graphical user interface via attention-based encoder–decoder model. *Multimedia Systems* 28, 1 (2022), 121–130.

- [14] Betty H. C. Cheng and Joanne M. Atlee. 2007. Research Directions in Requirements Engineering. *Future of Software Engineering (FOSE '07)* (2007), 285–303. <https://api.semanticscholar.org/CorpusID:14304156>
- [15] A. A. J. Cizotto, R. C. T. de Souza, V. C. Mariani, and L. dos Santos Coelho. 2023. Web pages from mockup design based on convolutional neural network and class activation mapping. *Multimedia Tools and Applications* (2023), 1–27.
- [16] Yihong Dong, Xue Jiang, Jiaru Qian, Tian Wang, Kechi Zhang, Zhi Jin, and Ge Li. 2025. A Survey on Code Generation with LLM-based Agents. *arXiv preprint arXiv:2508.00083* (2025).
- [17] Takafumi Endo. 2024. How Cursor Became the Fastest Growing SaaS by Empowering the Rise of the Vibe Coder. <https://medium.com/@takafumi.endo/how-cursor-became-the-fastest-growing-saas-by-empowering-the-rise-of-the-vibe-coder-48ca266e429a>. Accessed: 2025-08-29.
- [18] Tianqing Fang, Zhisong Zhang, Xiaoyang Wang, Rui Wang, Can Qin, Yuxuan Wan, Jun-Yu Ma, Ce Zhang, Jiaqi Chen, Xiyun Li, et al. 2025. Cognitive Kernel-Pro: A Framework for Deep Research Agents and Agent Foundation Models Training. *arXiv preprint arXiv:2508.00414* (2025).
- [19] Robert Feldt, Sungmin Kang, Juyeon Yoon, and Shin Yoo. 2023. Towards autonomous testing agents via conversational large language models. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 1688–1693.
- [20] Google AI. 2024. API Overview. <https://ai.google.dev/gemini-api/docs/api-overview> Accessed: 2024-06-06.
- [21] Tianxiao Gu, Chengnian Sun, Xiaoxing Ma, Chun Cao, Chang Xu, Yuan Yao, Qirun Zhang, Jian Lu, and Zhendong Su. 2019. Practical GUI Testing of Android Applications Via Model Abstraction and Refinement. *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)* (2019), 269–280. <https://api.semanticscholar.org/CorpusID:89608086>
- [22] Yi Gui, Zhen Li, Yao Wan, Yemin Shi, Hongyu Zhang, Yi Su, Shaoling Dong, Xing Zhou, and Wenbin Jiang. 2024. VISION2UI: A Real-World Dataset with Layout for Code Generation from UI Designs. *ArXiv abs/2404.06369* (2024). <https://api.semanticscholar.org/CorpusID:269010048>
- [23] Yi Gui, Yao Wan, Zhen Li, Zhongyi Zhang, Dongping Chen, Hongyu Zhang, Yi Su, Bohua Chen, Xing Zhou, Wenbin Jiang, and Xiangliang Zhang. 2025. UICopilot: Automating UI Synthesis via Hierarchical Code Generation from Webpage Designs. *Proceedings of the ACM on Web Conference 2025* (2025). <https://api.semanticscholar.org/CorpusID:277998658>
- [24] Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. 2024. WebVoyager: Building an End-to-End Web Agent with Large Multimodal Models. In *Annual Meeting of the Association for Computational Linguistics*. <https://api.semanticscholar.org/CorpusID:267211622>
- [25] IT Revolution. 2021. *Why the Full Stack Engineer Is Problematic*. <https://itrevolution.com/articles/why-the-full-stack-engineer-is-problematic/>
- [26] 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 Multimodal Agents. *arXiv preprint arXiv:2507.22827* (2025).
- [27] Cem Kaner. 2013. An Introduction to Scenario Testing. <https://api.semanticscholar.org/CorpusID:59641340>
- [28] Yuanhong Lan, Yifei Lu, Zhong Li, Minxue Pan, Wenhua Yang, Tian Zhang, and Xuandong Li. 2024. Deeply Reinforcing Android GUI Testing with Deep Reinforcement Learning. *2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE)* (2024), 854–866. <https://api.semanticscholar.org/CorpusID:267523834>
- [29] Valéria Lelli, Arnaud Blouin, and Benoît Baudry. 2015. Classifying and Qualifying GUI Defects. *2015 IEEE 8th International Conference on Software Testing, Verification and Validation (ICST)* (2015), 1–10. <https://api.semanticscholar.org/CorpusID:2288032>
- [30] Yuanchun Li, Ziyue Yang, Yao Guo, and Xiangqun Chen. 2019. Humanoid: A Deep Learning-Based Approach to Automated Black-box Android App Testing. *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)* (2019), 1070–1073. <https://api.semanticscholar.org/CorpusID:210693353>
- [31] Jun Liu, Yuanyuan Xie, Jiwei Yan, Jinhao Huang, Jun Yan, and Jian Zhang. 2024. Write Your Own CodeChecker: An Automated Test-Driven Checker Development Approach with LLMs. *arXiv preprint arXiv:2411.06796* (2024).
- [32] Zhe Liu, Chunyang Chen, Junjie Wang, Mengzhuo Chen, Boyu Wu, Xing Che, Dandan Wang, and Qing Wang. 2023. Make LLM a Testing Expert: Bringing Human-Like Interaction to Mobile GUI Testing via Functionality-Aware Decisions. *2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE)* (2023), 1222–1234. <https://api.semanticscholar.org/CorpusID:264439493>
- [33] Zhe Liu, Cheng Li, Chunyang Chen, Junjie Wang, Boyu Wu, Yawen Wang, Jun Hu, and Qing Wang. 2024. Vision-driven Automated Mobile GUI Testing via Multimodal Large Language Model. *ArXiv abs/2407.03037* (2024). <https://api.semanticscholar.org/CorpusID:270923733>
- [34] Zimu Lu, Yunqiao Yang, Houxing Ren, Haotian Hou, Han Xiao, Ke Wang, Weikang Shi, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2025. WebGen-Bench: Evaluating LLMs on Generating Interactive and Functional Websites from Scratch. *arXiv preprint arXiv:2505.03733* (2025).

- [35] Noble Saji Mathews and Meiyappan Nagappan. 2024. Test-Driven Development and LLM-based Code Generation. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering* (Sacramento, CA, USA) (ASE '24). Association for Computing Machinery, New York, NY, USA, 1583–1594. doi:10.1145/3691620.3695527
- [36] Sara A. Metwalli. 2025. *What Is Front-End Development? (vs. Back-End, Tools)*. <https://builtin.com/software-engineering-perspectives/front-end-development>
- [37] Jose Lorenzo San Miguel and Shingo Takada. 2016. GUI and usage model-based test case generation for Android applications with change analysis. *Proceedings of the 1st International Workshop on Mobile Development* (2016). <https://api.semanticscholar.org/CorpusID:5574875>
- [38] K. Moran, C. Bernal-Cárdenas, M. Curcio, R. Bonett, and D. Poshyvanyk. 2018. Machine learning-based prototyping of graphical user interfaces for mobile apps. *IEEE Transactions on Software Engineering* 46, 2 (2018), 196–221.
- [39] Kevin Moran, Boyang Li, Carlos Bernal-Cárdenas, Dan Jelf, and Denys Poshyvanyk. 2018. Automated Reporting of GUI Design Violations for Mobile Apps. *2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)* (2018), 165–175. <https://api.semanticscholar.org/CorpusID:3634687>
- [40] Magnus Müller and Gregor Žunič. 2024. *Browser Use: Enable AI to control your browser*. <https://github.com/browser-use/browser-use>
- [41] T. A. Nguyen and C. Csallner. 2015. Reverse engineering mobile application user interfaces with remaui (t). In *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. 248–259.
- [42] Tuan Anh Nguyen and Christoph Csallner. 2015. Reverse engineering mobile application user interfaces with remaui (t). In *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 248–259.
- [43] Minxue Pan, An Huang, Guoxin Wang, Tian Zhang, and Xuandong Li. 2020. Reinforcement learning based curiosity-driven testing of Android applications. *Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis* (2020). <https://api.semanticscholar.org/CorpusID:220497623>
- [44] Chenglei Si, Yanzhe Zhang, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. 2024. Design2Code: How Far Are We From Automating Front-End Engineering? *ArXiv abs/2403.03163* (2024). <https://api.semanticscholar.org/CorpusID:268248801>
- [45] Simplilearn. 2025. *Is Full Stack Development Hard? Insights & Tips!* <https://www.simplilearn.com/is-full-stack-development-hard-article>
- [46] Yuxuan Wan, Yi Dong, Jingyu Xiao, Yintong Huo, Wenxuan Wang, and Michael R. Lyu. 2024. MRWeb: An Exploration of Generating Multi-Page Resource-Aware Web Code from UI Designs. *ArXiv abs/2412.15310* (2024). <https://api.semanticscholar.org/CorpusID:274965541>
- [47] Yuxuan Wan, Chaozheng Wang, Yi Dong, Wenxuan Wang, Shuqing Li, Yintong Huo, and Michael Lyu. 2025. Divide-and-Conquer: Generating UI Code from Screenshots. *Proc. ACM Softw. Eng.* 2, FSE, Article FSE094 (June 2025), 24 pages. doi:10.1145/3729364
- [48] Xin Wang, Xiao Liu, Pingyi Zhou, Qixia Liu, Jin Liu, Hao Wu, and Xiaohui Cui. 2022. Test-Driven Multi-Task Learning with Functionally Equivalent Code Transformation for Neural Code Generation. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*. 1–6.
- [49] Xin Wang, Yasheng Wang, Yao Wan, Fei Mi, Yitong Li, Pingyi Zhou, Jin Liu, Hao Wu, Xin Jiang, and Qun Liu. 2022. Compilable neural code generation with compiler feedback. *arXiv preprint arXiv:2203.05132* (2022).
- [50] Fan Wu, Cuiyun Gao, Shuqing Li, Xinjie Wen, and Qing Liao. 2025. MLLM-Based UI2Code Automation Guided by UI Layout Information. *ArXiv abs/2506.10376* (2025). <https://api.semanticscholar.org/CorpusID:279319153>
- [51] 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 (2025), 801–824.
- [52] Jingyu Xiao, Yuxuan Wan, Yintong Huo, Zhiyao Xu, and Michael R Lyu. 2024. Interaction2Code: How Far Are We From Automatic Interactive Webpage Generation? *arXiv preprint arXiv:2411.03292* (2024).
- [53] Jingyu Xiao, Yuxuan Wan, Yintong Huo, Zhiyao Xu, and Michael R. Lyu. 2024. Interaction2Code: How Far Are We From Automatic Interactive Webpage Generation? *ArXiv abs/2411.03292* (2024). <https://api.semanticscholar.org/CorpusID:273821629>
- [54] Jingyu Xiao, Ming Wang, Man Ho Lam, Yuxuan Wan, Junliang Liu, Yintong Huo, and Michael R. Lyu. 2025. DesignBench: A Comprehensive Benchmark for MLLM-based Front-end Code Generation. *ArXiv abs/2506.06251* (2025). <https://api.semanticscholar.org/CorpusID:279244894>
- [55] Jingyu Xiao, Zhongyi Zhang, Yuxuan Wan, Yintong Huo, Yang Liu, and Michael R Lyu. 2025. EfficientUICoder: Efficient MLLM-based UI Code Generation via Input and Output Token Compression. *arXiv preprint arXiv:2509.12159* (2025).
- [56] Y. Xu, L. Bo, X. Sun, B. Li, J. Jiang, and W. Zhou. 2021. image2emmet: Automatic code generation from web user interface image. *Journal of Software: Evolution and Process* 33, 8 (2021), e2369.
- [57] Shengcheng Yu, Chunrong Fang, Ziyuan Tuo, Quanjun Zhang, Chunyang Chen, Zhenyu Chen, and Zhendong Su. 2023. Vision-Based Mobile App GUI Testing: A Survey. *ArXiv abs/2310.13518* (2023). <https://api.semanticscholar.org/CorpusID:264406197>

- [58] Yuntong Zhang, Haifeng Ruan, Zhiyu Fan, and Abhik Roychoudhury. 2024. AutoCodeRover: Autonomous Program Improvement. *arXiv preprint arXiv:2404.05427* (2024).
- [59] Ting Zhou, Yanjie Zhao, Xinyi Hou, Xiaoyu Sun, Kai Chen, and Haoyu Wang. 2024. Bridging Design and Development with Automated Declarative UI Code Generation. *arXiv preprint arXiv:2409.11667* (2024).
- [60] Ting Zhou, Yanjie Zhao, Xinyi Hou, Xiaoyu Sun, Kai Chen, and Haoyu Wang. 2024. Bridging Design and Development with Automated Declarative UI Code Generation. <https://api.semanticscholar.org/CorpusID:272708114>