Multimodal DeepResearcher: Generating Text-Chart Interleaved Reports From Scratch with Agentic Framework

Zhaorui Yang[1*], Bo Pan[1*], Han Wang[1*], Yiyao Wang[1], Xingyu Liu[1], Luoxuan Weng[1]
Yingchaojie Feng[2], Haozhe Feng[3], Minfeng Zhu[4†], Bo Zhang[4†], Wei Chen[1†]
[1]State Key Lab of CAD&CG, Zhejiang University
[2]National University of Singapore
[3]Tencent TEG
[4]Zhejiang University

Abstract

Visualizations play a crucial part in effective communication of concepts and information. Recent advances in reasoning and retrieval augmented generation have enabled Large Language Models (LLMs) to perform deep research and generate comprehensive reports. Despite its progress, existing deep research frameworks primarily focus on generating text-only content, leaving the automated generation of interleaved texts and visualizations underexplored. This novel task poses key challenges in designing informative visualizations and effectively integrating them with text reports. To address these challenges, we propose Formal Description of Visualization (FDV), a structured textual representation of charts that enables LLMs to learn from and generate diverse, high-quality visualizations. Building on this representation, we introduce Multimodal DeepResearcher, an agentic framework that decomposes the task into four stages: (1) researching, (2) exemplar report textualization, (3) planning and (4) multimodal report generation. For the evaluation of the generated reports, we develop MultimodalReportBench which contains 100 diverse topics as inputs, and a set of dedicated metrics for report and chart evaluation. Extensive experiments across models and evaluation methods demonstrate the effectiveness of Multimodal DeepResearcher. Notably, utilizing the same Claude 3.7 Sonnet model, Multimodal DeepResearcher achieves an 82% overall win rate over the baseline method.

Extended version — https://arxiv.org/pdf/2506.02454

Introduction

Large language models (LLMs) have demonstrated broad capabilities in solving diverse tasks such as question answering, coding and math (Bai et al. 2022; Guo et al. 2025; Huang et al. 2025). Augmented with searching and reasoning capabilities (Xie et al. 2023; Nakano et al. 2022; Li et al. 2025a), LLMs can perform deep research and effectively leverage up-to-date external information beyond static parameters (Li et al. 2025a). Recently, this paradigm has garnered significant attention with its remarkable efficacy in generating grounded, comprehensive reports from scratch (Shao et al. 2024; Huot et al. 2025). However, existing deep research frameworks from both academia (Jin et al. 2025; Zheng et al. 2025b) and industry (Google 2024; xAI 2025) predominantly focus on generating textual content, neglecting the display beyond text modality. The text-heavy nature of these reports impedes effective communication of concepts and information (Ku et al. 2025; Zheng et al. 2025a), which limits their readability and practical utility.

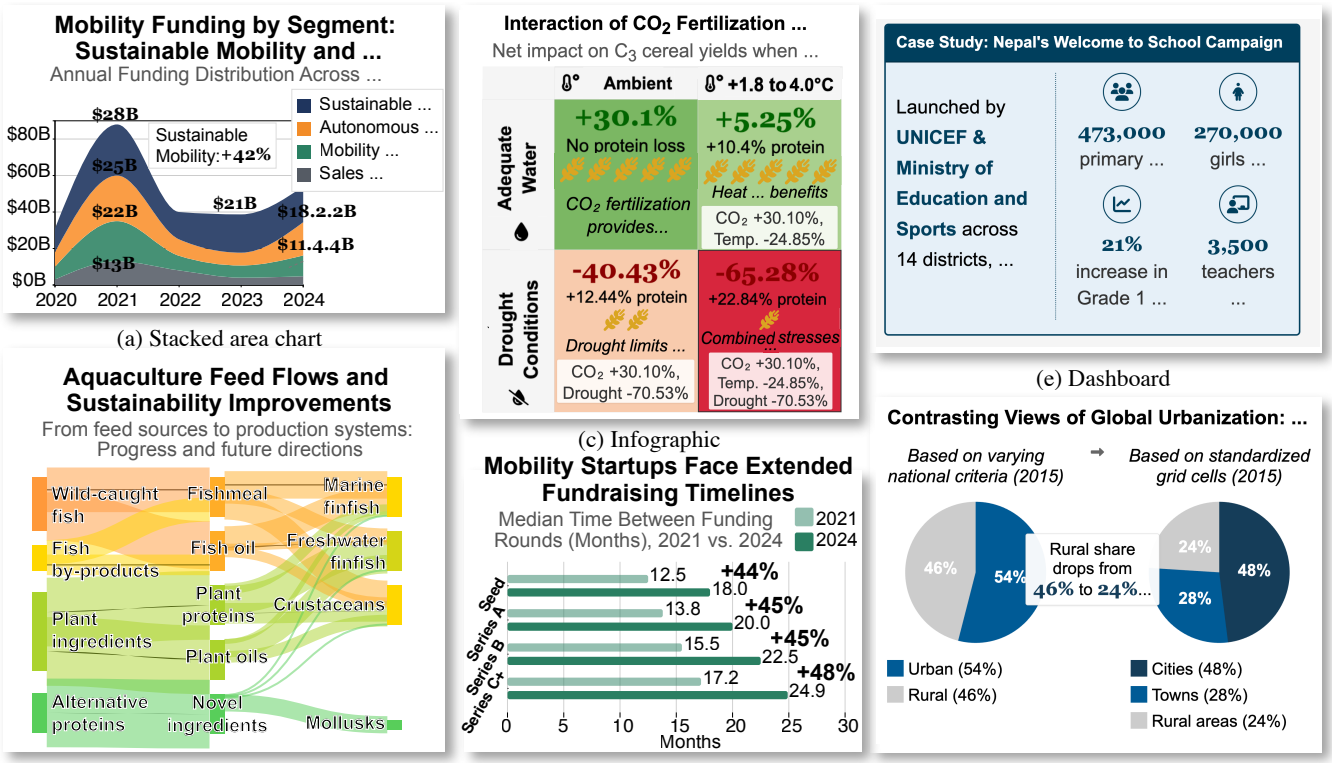
In real-world scenarios, visualization serves as a crucial part of reports and presentations, offering remarkable capabilities for conveying data insights (Otten, Cheng, and Drewnowski 2015), facilitating the identification of implicit patterns (Yang et al. 2024), and enhancing audience engagement (Barrick, Davis, and Winkler 2018; Zheng et al. 2025a). Human experts typically craft meticulously designed visualizations with consistent styles to effectively communicate ideas and insights. They then integrate these visualizations within appropriate textual context (He et al. 2025b) to create coherent text-chart interleaved reports.

However, the end-to-end generation of multimodal reports remains challenging. Although LLMs are capable of generating individual charts through coding (Yang et al. 2024; Seo et al. 2025; Han et al. 2023), effectively representing and integrating these visualizations with textual content still poses a challenge. While in-context learning appears to be a promising approach for guiding such generation, there lacks an appropriate representation to integrate text-chart interleaved content within the context of LLMs.

To address this challenge, we introduce the Formal Description of Visualization (FDV), a structured representation method inspired by the grammar of graphics (Wilkinson 1999), a classical visualization theory. FDV comprehensively captures visualization designs through four perspectives (i.e., overall layout, plotting scale, data, and marks). This representation provides universal and high-fidelity descriptions that enables in-context learning of multimodal reports from human experts, and can be generated to produce diverse and high-quality charts.

Building upon FDV, we introduce Multimodal DeepResearcher, an agentic framework that generates text-chart interleaved reports from scratch. The framework operates through four stages: (1) researching, which gathers com-

[*]These authors contributed equally.
[†]Corresponding Authors.



(a) Stacked area chart

(b) Sankey diagram

(c) Infographic

(d) Horizontal bar chart

(e) Dashboard

(f) Pie chart

Figure 1: Examples of **visualization charts** generated by Multimodal DeepResearcher. **Accompanying texts are omitted** for brevity. As shown in the figure, it can produce diverse, high-quality charts beyond basic charts (e.g., line or bar chart).

prehensive information through searching and reasoning; (2) exemplar report textualization, which textualizes multimodal reports from human experts using our proposed Formal Description of Visualization (FDV) for in-context learning; (3) planning, which establishes a content outline and visualization style guide; and (4) multimodal report generation, which produces the final interleaved report through drafting, coding and iterative chart refinement. Some examples of the generated charts are presented in Figure 1.

We evaluate Multimodal DeepResearcher with MultimodalReportBench, which comprises 100 topics used as inputs. Our experiments include both proprietary and open-source models with automatic and human evaluation. The evaluation encompasses both report-level and chart-level assessments, each employing five dedicated metrics. As a baseline, we adapted DataNarrative (Islam et al. 2024), a relevant framework that generates simple placeholders for charts from tabular inputs, to perform our task. Both automatic and human evaluations consistently demonstrate Multimodal DeepResearcher's superior performance compared to the baseline. Notably, when using Claude 3.7 Sonnet as the generator, Multimodal DeepResearcher achieves an impressive 82% overall win rate.

Our contributions can be summarized as follows:

- We introduce a novel task that generates a text-chart interleaved multimodal report from scratch and a corresponding dataset and evaluation metrics.
- We propose Formal Description of Visualization (FDV), a structured textual representation of visualizations that enables the in-context learning and generation of multimodal reports.
- We introduce Multimodal DeepResearcher, an end-to-end agentic framework that generates high-quality multimodal reports, which largely outperforms the baseline method.

Related Work

Deep Research Recently, the combination of retrieval techniques (Li et al. 2025b; Zhao et al. 2024) and reasoning (Guo et al. 2025) has enabled LLMs to transcend their parametric constraints by leveraging external knowledge. Pioneering works have designed specialized prompts and workflows for complex research tasks, as exemplified by OpenResearcher (Zheng et al. 2024) and Search-o1 (Li et al. 2025a). Subsequent research explored reinforcement learning for end-to-end reasoning and information retrieval (Jin et al. 2025; Zheng et al. 2025b). However, these studies primarily focus on generating and evaluating text-only results, whereas this work advances the field by generating text-chart interleaved reports that significantly enhance information comprehension and communication with visualizations.

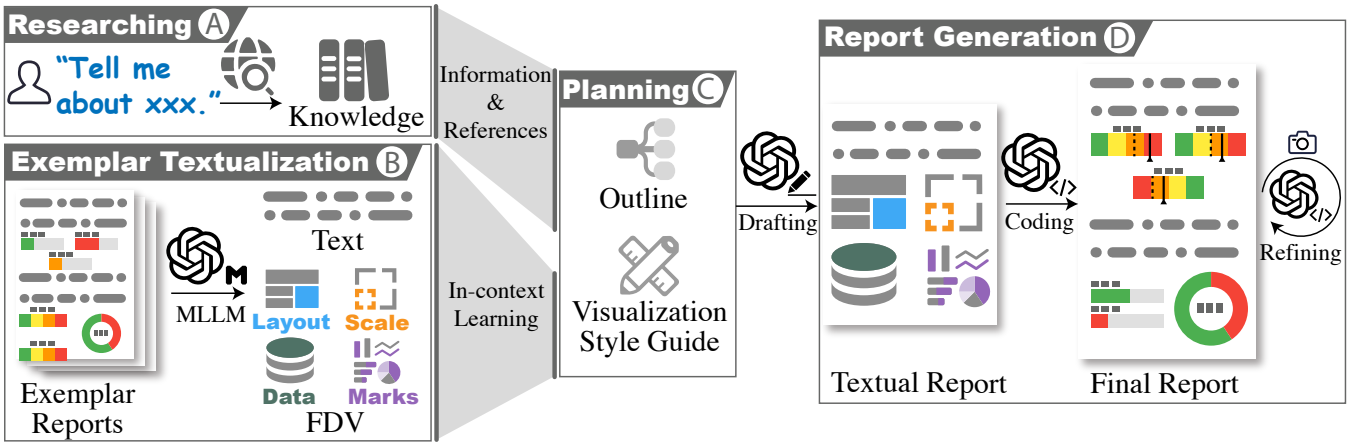


Figure 2: The framework of the Multimodal DeepResearcher. It decomposes the task of multimodal report generation into four stages: Ⓐ Iterative researching about given topic; Ⓑ Exemplar textualization of human experts using proposed Formal Description of Visualization (FDV); Ⓒ Planning; Ⓓ Report Generation, which generates the final report with crafting, coding and iterative refinement.

LLM for Data Visualizations Current work has focused on enhancing individual chart quality through various approaches, including multi-stage pipelines (Dibia 2023), iterative debugging with visual feedback (Yang et al. 2024), chain-of-thought prompted query reformulation (Seo et al. 2025), and models fine-tuned with domain-specific data for chart generation (Han et al. 2023; Tian et al. 2024). Another line of work has explored how to articulate generation intent, such as multimodal prompting with sketches and direct manipulations (Wen et al. 2025), multilingual natural language interfaces (Maddigan and Susnjak 2023), and conversational context management (Hong and Crisan 2023). Corresponding evaluation methodologies have also been proposed (Li et al. 2024a; Chen et al. 2025). However, previous work has predominantly focuses on generating individual charts with limited data. To the best of our knowledge, we are the first to explore generating and evaluating text-chart interleaved reports with multiple visualizations, based on in-the-wild and heterogeneous information.

LLM for agentic generation LLMs have been widely applied to various generation tasks due to their ability to process complex textual information (Ku et al. 2024; Nijkamp et al. 2023b,a; Jimenez et al. 2024; Yang et al. 2025b). For challenging tasks that require multiple steps, researchers have designed LLM agents that decompose problems into reasoning, planning, and execution stages (Luo et al. 2025). These agents have demonstrated remarkable success across scientific research (Lu et al. 2024; Si, Yang, and Hashimoto 2024; Li et al. 2024b; Bogin et al. 2024), video generation (He et al. 2025a), and computer system interaction (Xie et al. 2024; Deng et al. 2023; Zhang et al. 2023). This paradigm extends effectively to the visualization domain as well. TheoremExplainAgent (Ku et al. 2025) uses agents to generate educational videos, and PPTAgent (Zheng et al. 2025a) automatically creates slides for presentation with integrated text and visuals. Most relevant to our work, DataNarrative (Islam et al. 2024) explores generating simple specifications for data-driven visualizations and evaluating these specifications as proxies for actual charts. However, this approach remains limited to simple chart types such as bar chart and line chart, which restricts its practical utility.

Method

We formulate the task of multimodal report generation as follows: given a topic t and a set of multimodal exemplar reports R containing interleaved texts and charts, the system is expected to generate a multimodal report as in R based on t. To solve this task, we introduce Multimodal DeepResearcher, an agentic framework which decomposes it into four steps: (1) researching through iterative web search and reasoning, (2) exemplar report textualization, which textualizes multimodal exemplar reports from human experts using proposed Formal Description of Visualization (FDV), (3) planning, and (4) Multimodal report generation. We present an overview of Multimodal DeepResearcher in Figure 2.

Researching

To leverage online information beyond parametric knowledge, Multimodal DeepResearcher conducts iterative research on a given topic t, generating a comprehensive set of learnings L. These learnings encompass both information acquired through web sources and their corresponding references. The process involves iterative execution of two primary operations: (1) web search and (2) subsequent reasoning based on search results. Initially, the agent prompts the LLM to generate relevant keywords $K = k_1, \cdots, k_{n_K}$ based on the given topic t. The agent then conducts web searches using these keywords and retrieves webpages $P = p_1, \cdots, p_{n_P}$. Subsequently, the agent analyzes these webpages, synthesizes the information into learnings L, and formulates a research question q for the next iteration. Based on this research question and the original topic, the research agent performs the next research cycle. After n_R rounds

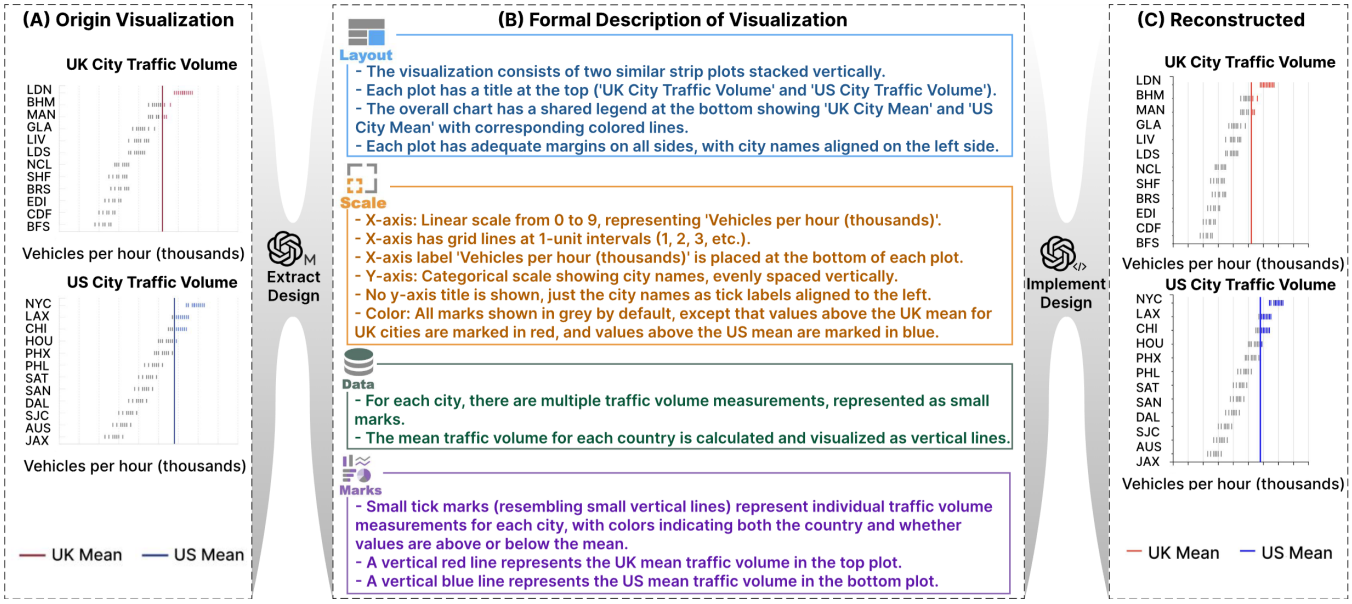


Figure 3: The illustration Formal Description of Visualization (FDV) for the exemplar textualization process. (A) Original traffic volume visualizations for UK and US cities; (B) The Formal Description of Visualization (FDV) that systematically captures the visualization's layout, scale, data, and marks using a structured format; and (C) The reconstructed visualization based on the formal description. This process textualizes high-quality text-chart interleaved reports by transforming visual elements into structured textual representations that preserve the visualization's essential characteristics.

of iteration, a final list of learnings and references are produced. More details are provided in the extended version.

Exemplar Textualization

Human experts typically produce reports with both texts and visualizations to enhance communication and audience engagement (Zheng et al. 2025a; Yang et al. 2024). To generate high-quality multimodal content comparable to expert-created reports, we employ in-context learning with exemplar reports crafted by human experts. This approach necessitates an effective methodology for converting multimodal exemplar reports R into textual exemplar reports $\tilde{R}$.

To address this challenge, we propose Formal Description of Visualization (FDV), a structured description method for visualization charts inspired by the grammar of graphics (GoG) theory (Wilkinson 1999), which theoretically provides universal and high-fidelity descriptions for any visualization designs. As shown in Figure 3 (B), FDV characterizes each visualization chart from four perspectives: (1) Overall layout, detailing the constituent subplots and their spatial arrangements; (2) Plotting scale, describing the scaling logic behind each "data to visual channel (e.g., position, color)" mapping and their annotations; (3) Data, describing both the numeric data and text elements used to generate the visualization. (4) Marks, describing the design specifications of each visual element. The reverse process of textualization can be achieved via coding, which reconstructs the visualization from FDV, as shown in Figure 3 (C).

In the exemplar textualization process, Multimodal DeepResearcher first extracts all visualization charts from the report, then prompts a multimodal large language model to extract the FDV representations of each chart. The FDV representations are then used to replace the charts. The algorithm for the process is presented in Algorithm 2. Further details of FDV and prompts are provided in the extended version.

Algorithm 1: Textualization of multimodal reports

Inputs: Multimodal exemplar reports R.
Requires: Multimodal large language model M_v, replace function *replace*.
Outputs: Textualized exemplar reports $\tilde{R}$.
Initialize $\tilde{R} = \emptyset$
for r **in** R **do**
  Init. $\tilde{r} = r$
  for each image i in r **do**
    // Extract FDV from image
    $\text{FDV}_i = M_v(i)$
    // Replace image with extracted FDV
    $\tilde{r} = \tilde{r}.replace(i, \text{FDV}_i)$
  end for
  $\tilde{R} = \tilde{R} \cup \{\tilde{r}\}$
end for
Return: $\tilde{R}$

Planning

After iterative researching about the topic t, Multimodal DeepResearcher creates a plan before generating the final report. Specifically, it constructs an outline O of the report to generate based on the learnings L, topic t and textual exemplar report $\tilde{R}$. The outline comprises a hierarchical structure

Algorithm 2: Algorithm for refining charts

Inputs: chart c represented as code.
Requires: Browser tool T, LLM M_t, Multimodal LLM M_v.
Outputs: Refined chart $\tilde{c}$.
Hypars: Number of max retry times N_{max}.
Initialize satisfied = False, $c_0 = c$, $C = \{c\}$.
for $i = 1$ **to** N_{max} **do**
  // Get console message and image
  msg, $i = T(c)$
  // Critic M_v evaluates the chart
  satisfied, feedback $= M_v(i)$
  if satisfied == True **then**
    break
  end if
  // actor M_t refines previous chart
  $c_i = M_t(c_{i-1}, \text{msg}, \text{feedback})$
  $C = C \cup \{c_i\}$
end for
$\tilde{c} = c_0$
if $|C| > 1$ **then**
  // Selects from the last two charts
  $\tilde{c} = M_v(C[-1], C[-2])$
end if
Return: $\tilde{c}$

of sections, each with a descriptive title and a brief summary. To learn the style of visualizations in exemplar reports $\tilde{R}$ and maintain a consistent style of charts, Multimodal DeepResearcher also prompts the LLM to generate a visualization style guide G. The visualization style guide provides guidelines that control the overall style of visualizations in the report (e.g., color palette, font hierarchy). More details of this process can be found in the extended version.

Final Report Generation

The final stage of Multimodal DeepResearcher is to generate the multimodal report with interleaved textual content and visualizations. The report is generated with outputs of previous stages, i.e., learnings L, exemplar textual reports $\tilde{R}$, outline O and visualization style guide G.

Multimodal DeepResearcher first prompts the LLM to generate a textual report with Formal Description of Visualization (FDV) as a placeholder for the underlying visualization chart to be generated. The format of this textual report is expected to be the same as those in textual exemplar reports used for in-context learning. Then, Multimodal DeepResearcher extracts all occurrences of FDVs, and prompts the LLM to implement the design via coding. Since visualizations represented by FDV have extensive flexibility, which may exceed the expressive capabilities of typical declarative visualization libraries (Heer and Bostock 2010) (e.g., matplotlib), we directed the LLMs to utilize D3.js, the most widely used imperative visualization programming to implement the target visualization designs.

To further improve the quality of visualizations generated, we include an actor-critic mechanism to revise and refine the code for generating the charts motivated by recent advancements of agents (Yang et al. 2024). In this scenario, the actor is the LLM M_t that generates code for chart, and feedback comes from both console and a critic model.

Console feedback is collected using chrome developer tool provided as Python package. It first tries to load each visualization, collecting all console message with errors or warnings during loading. After all elements are loaded, it takes a screenshot to obtain the visualization chart rendered.

After getting the screenshot of each visualization chart, Multimodal DeepResearcher employs a multimodal LLM (MLLM) M_v to serve as a critic, which provides visual feedback. The MLLM takes the chart rendered as input, examines its visual quality, and delivers corresponding feedback. It further determines whether the current chart needs improvement. If improvement is needed, the actor refines its code based on the feedback and console message. This iterative refinement continues until the critic is satisfied, or a predefined upper limit of retry times is reached, which we set as 3 to avoid infinite refinement cycles. When the refinement process finishes, the critic selects the final chart from the last two candidates during refinement.

The refine process is detailed in Algorithm 2. The prompts and a comprehensive *full report* generated by Multimodal DeepResearcher is presented in the extended version.

Experiments

Data Selection

To systematically evaluate the multimodal report generated by Multimodal DeepResearcher, we constructed MultimodalReportBench, a benchmark comprising 100 real-world topics curated from public websites that feature multimodal reports crafted by human experts, i.e., Pew Research (Pew 2025), Our World in Data (OWID 2025) and Open Knowledge Foundation (OKF 2024). Pew Research informs the public about issues, attitudes and trends shaping the world through research report. Our World in Data presents empirical data and research on global development challenges through web publications. The Open Knowledge Foundation is dedicated to promoting open data and content across all domains, ensuring information accessibility. These sources contain exemplary multimodal reports, making their topics appropriate for our task.

The topics are then used as inputs for multimodal report generation. To ensure that our dataset applies to the real-world scenario, we meticulously curated topics spanning 10 categories, such as travel, energy and education. The distribution of topic categories is provided in the extended version. We also collected 6 multimodal reports with no overlapping in topics to serve as exemplar reports for in-context learning.

Baseline Selection

Our task requires generating a multimodal report from scratch, which is infeasible with direct prompting or existing deep research frameworks. Most existing visualization generation works either focus on single-chart generation (Dibia 2023; Yang et al. 2024; Tian et al. 2024) or requires human

interactions (Fu, Bromley, and Setlur 2025; Li, Wang, and Qu 2024; Shao et al. 2025), which deviates from our setting of automated generation. Most similar to our work, DataNarrative (Islam et al. 2024) generates simple data-driven visualization specifications based on data tables as input, and evaluates the textual specification as a proxy of chart. We incorporate our researching module and adapt its framework accordingly to establish our baseline. For fair comparison, we utilize the learnings generated with our researching stage and plans instead of tables as the input. It then goes through generate-verify-refine process, consistent with the original framework. Since the original framework lacks mechanisms for transforming design specifications into actual charts, we extract all design specifications and generate corresponding visualizations using the same pipeline as Multimodal DeepResearcher does.

Framework Implementation

Multimodal DeepResearcher is an agentic framework with multiple stages. In this section, we describe the implementation details of each stage. In the researching stage, we perform web search and conduct reasoning with GPT-4o-mini. GPT-4o-mini is also utilized for planning. Claude 3.7 Sonnet is utilized as the MLLM for the textualization of exemplar reports. The generation of the final multimodal report requires both a large language model to craft textual report, and a multimodal large language model to provide visual feedback for the chart. Our experiments encompasses two model configurations: (1) State-of-the-art proprietary models, with Claude 3.7 Sonnet serving as both the LLM and multimodal LLM. (2) Open-source models, specifically Qwen3-235B-A22B (Yang et al. 2025a) and Qwen2.5-VL-72B-Instruct (Bai et al. 2025). To ensure fair comparison, all the settings are consistent in both Multimodal DeepResearcher and the DataNarrative baseline where applicable.

Automatic Report Evaluation

Given the multimodal nature of the outputs in our task, evaluation necessitates assessment of both texts and visualizations. To accomplish this, we conducted both report-level and chart-level evaluation to comprehensively assess the quality of all reports. For automatic report evaluation, we task the evaluator (i.e., GPT-4.1) with pairwise comparison of reports, generated from the save topic with both methods. Since report generation constitutes an open-ended, subjective task, reference-based metrics typically fail to align with human-perceived standards (Liu et al. 2023). Therefore, we established a comprehensive criteria incorporating both texts and visualizations in reports, which primarily consists of five metrics:

Informativeness and Depth. Evaluates whether the report delivers comprehensive, substantive and thorough information through both texts and accompany visualizations.

Coherence and Organization. Evaluates whether the report is well-organized, and whether the visualizations connect meaningfully to the text.

Verifiability. Evaluates whether the information of the reports can be verified with citations. Apart from textual links to references, we also prompt the evaluator to check the annotation present in visualizations that may contain source information.

Visualization Quality. Evaluates the quality of visualization charts in the report, including visual clarity and textual labels and annotations.

Visualization Consistency. Evaluates whether the visualizations in the report maintain a consistent overall style. The style contains the color palettes, typography and information hierarchy in visualizations.

During evaluation, we provide the evaluator with the topic, learnings which contain both knowledge acquired through web search, references, and both reports. Specifically, we employ rubric scoring on a 1-5 scale with detailed guidelines. The scores are then compared to determine which method is better or they tie. To mitigate potential positional bias, we randomize the order of reports. The complete prompts for evaluation are provided at extended version.

	Ours vs DataNarrative		
Evaluation Metrics	Ours Win	Ours Lose	Tie
w. *Claude 3.7 Sonnet*			
Informativeness and Depth	**75%**	25%	0%
Coherence and Organization	**76%**	21%	3%
Verifiability	**86%**	5%	9%
Visualization Quality	**80%**	16%	4%
Visualization Consistency	**78%**	17%	5%
Overall	**82%**	16%	2%
w. *Qwen3-235B-A22B & Qwen2.5-VL-72B-Instruct*			
Informativeness and Depth	**50%**	50%	0%
Coherence and Organization	41%	**51%**	8%
Verifiability	**66%**	21%	13%
Visualization Quality	**48%**	46%	6%
Visualization Consistency	**52%**	42%	6%
Overall	**55%**	40%	5%

Table 1: Automatic evaluation results of the multimodal report: Multimodal DeepResearcher (Ours) vs. DataNarrative.

Results. As illustrated in Table 1, Multimodal DeepResearcher consistently outperforms DataNarrative across both proprietary and open-source model configurations. With Claude 3.7 Sonnet, it achieves an overall win rate of 82%. Specifically, Multimodal DeepResearcher outperforms with a high win rate in Verifiability (86%), Visualization Quality (80%) and Visualization consistency (78%). A similar pattern is observed with open-source models, where Multimodal DeepResearcher achieves a win rate of 55%. Notably, the performance advantage is more pronounced with Claude 3.7 Sonnet than with open-source models. This gap arises as Multimodal DeepResearcher requires multifaced capabilities, including planning, writing, coding, and refinement. Therefore, Multimodal DeepResearcher benefits more from a stronger model, whereas DataNarrative's simpler architecture limits its capacity to leverage model improvements. The results demonstrate the efficacy of its in generating multimodal reports. We also presented the raw scores obtained and results with other evaluators in the extended version.

Evaluation Metrics	Ours Win	Ours Lose	Tie
Informativeness and Depth	**100%**	0%	0%
Coherence and Organization	**95%**	0%	5%
Verifiability	**100%**	0%	0%
Visualization Quality	**75%**	20%	5%
Visualization Consistency	**90%**	0%	10%
Overall	**100%**	0%	0%

Table 2: Human evaluation of the generated reports: Multimodal DeepResearcher (Ours) vs. DataNarrative.

Evaluation Metrics	Ours	DataNarrative
w. *Claude 3.7 Sonnet*		
Readability	**8.97**	8.52
Layout	**9.23**	8.48
Aesthetics	**9.12**	8.38
Data Faithfulness	**9.83**	9.59
Goal Compliance	**9.75**	9.24
w. *Qwen3-235B-A22B & Qwen2.5-VL-72B-Instruct*		
Readability	**7.05**	6.85
Layout	**6.70**	6.40
Aesthetics	**7.22**	6.74
Data Faithfulness	7.93	**7.99**
Goal Compliance	**7.17**	6.94

Table 3: Evaluation of chart quality. The evaluator assigns a score between 1 to 10 for each metric, and the results are average across all reports.

Human Evaluation

For human evaluation, we utilized the same set of metrics as in automatic report evaluation. We selected a random subset of 20 topics for evaluation. Specifically, 5 annotators performed pairwise comparison of reports generated by both Multimodal DeepResearcher and DataNarrative with Claude 3.7 Sonnet. As with automatic evaluation, we randomized the order to avoid potential positional bias. Results are presented in Table 2. Surprisingly, Multimodal DeepResearcher achieves an overall win rate of 100%. Specifically, two annotators preferred all 20 reports generated by Multimodal DeepResearcher, one annotator preferred 19 out of 20, another annotator preferred 18, and the last annotator preferred 15. Comparing with the results given by GPT-4.1, the agreement between them is 80%. These results further validate the effectiveness of Multimodal DeepResearcher.

Chart Evaluation

To provide a more fine-grained evaluation of our framework, we further conducted assessments of individual charts to examine their quality and fidelity. Following established practices in data-driven visualization (Dibia 2023; Chen et al. 2025), which provided explainable evaluations for charts, we curated five metrics: (1) *Readability*, (2) *Layout*, (3) *Aesthetics*, (4) *Data Faithfulness* and (5) *Goal compliance*. For each chart, we employed the evaluator to score based on the chart along with its original design specification. We then average the scores of all charts within each report. As demonstrated in Table 3, Multimodal DeepResearcher consistently outperforms DataNarrative, with particularly notable improvements in layout and aesthetics.

Ablated Components	Lose	Win	Tie
- w/o Exemplar Learning	70%	20%	10%
- w/o Planning	85%	15%	0%
- w/o Refinement of charts	80%	20%	0%

Table 4: Results of ablation studies across three different setups. We report the lose, win and tie rates for each setup against the complete Multimodal DeepResearcher. Claude 3.7 Sonnet serves as both the LLM and MLLM here.

Ablation Studies

To assess the efficacy of individual components of Multimodal DeepResearcher, we conducted ablation experiments on a random subset of 20 topics. Specifically, we compared 3 variants against Multimodal DeepResearcher: (1) w/o in-context learning from exemplar reports (2) w/o planning (3) w/o iterative refinement of charts. To ensure fair comparison, all other settings and hyperparameters remained consistent across variants. As shown in Table 4, removing any component results in significant performance degradation. Specifically, eliminating exemplar learning from human reports yields a 70% lose rate, direct generation without planning leads to 85%, and removing chart refinement process loses in 80% cases. We further employed alternative evaluators to examine the effect of exemplar learning. The overall win rate for Multimodal DeepResearcher is 70% when using GPT-5 as the evaluator, and 60% with Gemini-2.5-Pro. These findings demonstrate the contribution of each component in Multimodal DeepResearcher.

Analysis

Visualization Analysis

In this section, we analyze the characteristics of visualizations generated with Multimodal DeepResearcher and the baseline. While the average number of charts per report between our framework (9.3) and DataNarrative (9.4) is comparable, the visualizations generated by Multimodal DeepResearcher are notably more diverse. As illustrated in Figure 4, although both methods prioritized basic chart types such as line chart and bar chart, Multimodal DeepResearcher demonstrates superior capability in generating sophisticated and complex visualizations.

For instance, across the 100 selected topics, Multimodal DeepResearcher produces 15 flowcharts and 18 dashboards, while DataNarrative generates merely 2 flowcharts and 1 dashboard. Another example involves the "Others" category, which encompasses hard-to-categorize visualizations such as infographics and mind maps. Our framework generates 280 such charts, substantially exceeding the 96 produced

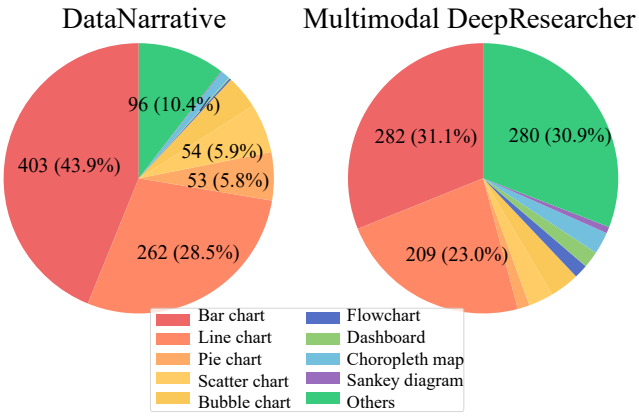


Figure 4: Distribution of visualization charts generated with DataNarrative and Multimodal DeepResearcher (Ours). The first column in the legend (denoted by red and yellow colors) represents conventional chart types.

by DataNarrative. This disparity demonstrates that our approach can accommodate to diverse real-world scenarios. We provide a collection of examples for each type generated by our framework in Figure 1 and extended version.

Error Analysis

Despite the remarkable efficacy of Multimodal DeepResearcher, the integration of visualizations poses new challenges. In this section, we categorize the identified common errors into the following two categories.

Overlapping Element overlap represents the most prevalent error in the charts, primarily due to the inherent complexity of determining precise spatial positioning for all chart components without real-time visual feedback during the coding process. This error can be generally attributed to two factors: (1) excessive information in FDV that complicates proper arrangement within limited space. (2) suboptimal placement of legends, labels and annotations. Examples of both scenarios are provided in the extended version.

Hallucination Hallucination is a fundamental challenge for LLMs (Shao et al. 2024), which also extends to the generation of visualizations (Islam et al. 2024). Despite explicit instructions to avoid creating fake data, models occasionally hallucinate when data is insufficient of unavailable. Figure 11 in the extended version exemplifies this issue through a choropleth map chart. In this case, the model erroneously marked regions with inadequate data using red color, to denote the decline of a certain metric.

Efficiency Analysis

Another challenge for Multimodal DeepResearcher lies in balancing utility and efficiency. The system requires iterative refinement of multiple charts within reports. As demonstrated in our ablation study, this process significantly enhances the overall quality of generated reports. However, it also introduces computational overhead. In our experiments, we refine each chart for at most 3 iterations. After filtering out instances affected by network issues, the average generation time for a single report is 767.20 seconds, compared to DataNarrative's 372.94 seconds. Further analysis reveals that the refinement process accounts for the majority of execution time, requiring interaction with headless browsers, evaluation by multimodal large language models, and code regeneration. We plan to explore more precise critique mechanisms in future work.

Conclusion

In this work, we investigate the challenge of generating multimodal reports from scratch. We introduce the Formal Description of Visualization, a structured representation of charts that enables in-context learning from human-created exemplar reports. Based on this, we propose Multimodal DeepResearcher, an end-to-end framework for the generation of multimodal reports. While extensive experiments using both automatic and human evaluation confirm the efficacy of our framework, several challenges remain, including improving visualization quality, reducing hallucination, and balancing utility with efficiency.

Acknowledgments

This work is supported by National Key R&D Program of China under Grant No. 2024YFB4505500 & 2024YFB4505503, National Natural Science Foundation of China (No. 62132017, No. 62421003 and No. 62402434), and Zhejiang Provincial Natural Science Foundation of China (No. LD24F020011 and No. LQ24F020006). The work is partially conducted during Zhaorui Yang's internship at the Machine Learning Platform Department, Tencent TEG. We thank Tencent Cloud BI for their support in the commercial implementation of the data reporting feature.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv:2204.05862.

Barrick, A.; Davis, D.; and Winkler, D. 2018. Image Versus Text in PowerPoint Lectures: Who Does It Benefit? *Journal of Baccalaureate Social Work*, 23(1): 91–109.

Bogin, B.; Yang, K.; Gupta, S.; Richardson, K.; Bransom, E.; Clark, P.; Sabharwal, A.; and Khot, T. 2024. SUPER: Evaluating Agents on Setting Up and Executing Tasks from Research Repositories. arXiv:2409.07440.

Chen, N.; Zhang, Y.; Xu, J.; Ren, K.; and Yang, Y. 2025. VisEval: A Benchmark for Data Visualization in the Era of Large Language Models. *IEEE Transactions on Visualization and Computer Graphics*, 31(1): 1301–1311.

Deng, X.; Gu, Y.; Zheng, B.; Chen, S.; Stevens, S.; Wang, B.; Sun, H.; and Su, Y. 2023. Mind2Web: Towards a Gener-

alist Agent for the Web. In *Advances in Neural Information Processing Systems*, volume 36, 28091–28114.

Dibia, V. 2023. LIDA: A Tool for Automatic Generation of Grammar-Agnostic Visualizations and Infographics using Large Language Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 113–126.

Fu, Y.; Bromley, D.; and Setlur, V. 2025. DataWeaver: Authoring Data-Driven Narratives through the Integrated Composition of Visualization and Text. In *Computer Graphics Forum*, e70098. Wiley Online Library.

Google. 2024. Gemini Deep Research. https://blog.google/products/gemini/google-gemini-deep-research/. Accessed: 2025-05-15.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Han, Y.; Zhang, C.; Chen, X.; Yang, X.; Wang, Z.; Yu, G.; Fu, B.; and Zhang, H. 2023. ChartLlama: A Multimodal LLM for Chart Understanding and Generation. arXiv:2311.16483.

He, L.; Song, Y.; Huang, H.; Liu, P.; Tang, Y.; Aliaga, D.; and Zhou, X. 2025a. Kubrick: Multimodal Agent Collaborations for Synthetic Video Generation. arXiv:2408.10453.

He, Y.; Cao, S.; Shi, Y.; Chen, Q.; Xu, K.; and Cao, N. 2025b. Leveraging Foundation Models for Crafting Narrative Visualization: A Survey. arXiv:2401.14010.

Heer, J.; and Bostock, M. 2010. Declarative Language Design for Interactive Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 16(6): 1149–1156.

Hong, M.-H.; and Crisan, A. 2023. Conversational AI Threads for Visualizing Multidimensional Datasets. arXiv:2311.05590.

Huang, S.; Cheng, T.; Liu, J. K.; Xu, W.; Hao, J.; Song, L.; Xu, Y.; Yang, J.; Liu, J.; Zhang, C.; et al. 2025. OpenCoder: The Open Cookbook for Top-Tier Code Large Language Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 33167–33193.

Huot, F.; Amplayo, R. K.; Palomaki, J.; Jakobovits, A. S.; Clark, E.; and Lapata, M. 2025. Agents' Room: Narrative Generation through Multi-step Collaboration. In *International Conference on Learning Representations*.

Islam, M. S.; Laskar, M. T. R.; Parvez, M. R.; Hoque, E.; and Joty, S. 2024. DataNarrative: Automated Data-Driven Storytelling with Visualizations and Texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 19253–19286.

Jimenez, C. E.; Yang, J.; Wettig, A.; Yao, S.; Pei, K.; Press, O.; and Narasimhan, K. R. 2024. SWE-bench: Can Language Models Resolve Real-world Github Issues? In *International Conference on Learning Representations*.

Jin, B.; Zeng, H.; Yue, Z.; Yoon, J.; Arik, S.; Wang, D.; Zamani, H.; and Han, J. 2025. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. arXiv:2503.09516.

Ku, M.; Chong, T.; Leung, J.; Shah, K.; Yu, A.; and Chen, W. 2025. TheoremExplainAgent: Towards Video-based Multimodal Explanations for LLM Theorem Understanding. arXiv:2502.19400.

Ku, M.; Jiang, D.; Wei, C.; Yue, X.; and Chen, W. 2024. VIEScore: Towards Explainable Metrics for Conditional Image Synthesis Evaluation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12268–12290.

Li, G.; Wang, X.; Aodeng, G.; Zheng, S.; Zhang, Y.; Ou, C.; Wang, S.; and Liu, C. H. 2024a. Visualization Generation with Large Language Models: An Evaluation. arXiv:2401.11255.

Li, H.; Wang, Y.; and Qu, H. 2024. Where are we so far? understanding data storytelling tools from the perspective of human-ai collaboration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–19.

Li, R.; Patel, T.; Wang, Q.; and Du, X. 2024b. MLR-Copilot: Autonomous Machine Learning Research based on Large Language Models Agents. arXiv:2408.14033.

Li, X.; Dong, G.; Jin, J.; Zhang, Y.; Zhou, Y.; Zhu, Y.; Zhang, P.; and Dou, Z. 2025a. Search-o1: Agentic Search-Enhanced Large Reasoning Models. arXiv:2501.05366.

Li, X.; Jin, J.; Zhou, Y.; Zhang, Y.; Zhang, P.; Zhu, Y.; and Dou, Z. 2025b. From matching to generation: A survey on generative information retrieval. *ACM Transactions on Information Systems*, 43(3): 1–62.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2511–2522.

Lu, C.; Lu, C.; Lange, R. T.; Foerster, J.; Clune, J.; and Ha, D. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. arXiv:2408.06292.

Luo, J.; Zhang, W.; Yuan, Y.; Zhao, Y.; Yang, J.; Gu, Y.; Wu, B.; Chen, B.; Qiao, Z.; Long, Q.; et al. 2025. Large Language Model Agent: A Survey on Methodology, Applications and Challenges. arXiv:2503.21460.

Maddigan, P.; and Susnjak, T. 2023. Chat2VIS: Fine-Tuning Data Visualisations using Multilingual Natural Language Text and Pre-Trained Large Language Models. arXiv:2303.14292.

Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; et al. 2022. WebGPT: Browser-assisted question-answering with human feedback. arXiv:2112.09332.

Nijkamp, E.; Hayashi, H.; Xiong, C.; Savarese, S.; and Zhou, Y. 2023a. CodeGen2: Lessons for Training LLMs on Programming and Natural Languages. In *International Conference on Learning Representations*.

Nijkamp, E.; Pang, B.; Hayashi, H.; Tu, L.; Wang, H.; Zhou, Y.; Savarese, S.; and Xiong, C. 2023b. CodeGen: An Open Large Language Model for Code with Multi-Turn Program Synthesis. In *International Conference on Learning Representations*.

OKF. 2024. Open Knowledge Foundation. https://ourworldindata.org/. Accessed: 2025-05-15.

Otten, J. J.; Cheng, K.; and Drewnowski, A. 2015. Infographics and public policy: using data visualization to convey complex information. *Health Affairs*, 34(11): 1901–1907.

OWID. 2025. Our World In Data. https://ourworldindata.org/. Accessed: 2025-05-15.

Pew. 2025. Pew Research Center. https://www.pewresearch.org/. Accessed: 2025-05-15.

Seo, W.; Lee, S.; Kang, D.; Yuan, Z.; and Lee, S. 2025. VisPath: Automated Visualization Code Synthesis via Multi-Path Reasoning and Feedback-Driven Optimization. arXiv:2502.11140.

Shao, Y.; Jiang, Y.; Kanell, T.; Xu, P.; Khattab, O.; and Lam, M. 2024. Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6252–6278.

Shao, Z.; Shen, L.; Li, H.; Shan, Y.; Qu, H.; Wang, Y.; and Chen, S. 2025. Narrative player: Reviving data narratives with visuals. *IEEE Transactions on Visualization and Computer Graphics*.

Si, C.; Yang, D.; and Hashimoto, T. 2024. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. arXiv:2409.04109.

Tian, Y.; Cui, W.; Deng, D.; Yi, X.; Yang, Y.; Zhang, H.; and Wu, Y. 2024. Chartgpt: Leveraging llms to generate charts from abstract natural language. *IEEE Transactions on Visualization and Computer Graphics*.

Wen, Z.; Weng, L.; Tang, Y.; Zhang, R.; Liu, Y.; Pan, B.; Zhu, M.; and Chen, W. 2025. Exploring Multimodal Prompt for Visualization Authoring with Large Language Models. arXiv:2504.13700.

Wilkinson, L. 1999. *The grammar of graphics*. Berlin, Heidelberg: Springer-Verlag. ISBN 0387987746.

xAI. 2025. Grok 3. https://x.ai/news/grok-3. Accessed: 2025-05-15.

Xie, T.; Zhang, D.; Chen, J.; Li, X.; Zhao, S.; Cao, R.; Hua, T. J.; Cheng, Z.; Shin, D.; Lei, F.; et al. 2024. OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 52040–52094.

Xie, T.; Zhou, F.; Cheng, Z.; Shi, P.; Weng, L.; Liu, Y.; Hua, T. J.; Zhao, J.; Liu, Q.; Liu, C.; et al. 2023. OpenAgents: An Open Platform for Language Agents in the Wild. arXiv:2310.10634.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025a. Qwen3 Technical Report. arXiv:2505.09388.

Yang, J.; Jimenez, C. E.; Zhang, A. L.; Lieret, K.; Yang, J.; Wu, X.; Press, O.; Muennighoff, N.; Synnaeve, G.; Narasimhan, K. R.; Yang, D.; Wang, S. I.; and Press, O. 2025b. SWE-bench Multimodal: Do AI Systems Generalize to Visual Software Domains? In *International Conference on Learning Representations*.

Yang, Z.; Zhou, Z.; Wang, S.; Cong, X.; Han, X.; Yan, Y.; Liu, Z.; Tan, Z.; Liu, P.; Yu, D.; et al. 2024. MatPlotAgent: Method and Evaluation for LLM-Based Agentic Scientific Data Visualization. In *Findings of the Association for Computational Linguistics: ACL 2024*, 11789–11804.

Zhang, C.; Yang, Z.; Liu, J.; Han, Y.; Chen, X.; Huang, Z.; Fu, B.; and Yu, G. 2023. AppAgent: Multimodal Agents as Smartphone Users. arXiv:2312.13771.

Zhao, P.; Zhang, H.; Yu, Q.; Wang, Z.; Geng, Y.; Fu, F.; Yang, L.; Zhang, W.; Jiang, J.; and Cui, B. 2024. Retrieval-Augmented Generation for AI-Generated Content: A Survey. arXiv:2402.19473.

Zheng, H.; Guan, X.; Kong, H.; Zheng, J.; Zhou, W.; Lin, H.; Lu, Y.; He, B.; Han, X.; and Sun, L. 2025a. PPTAgent: Generating and Evaluating Presentations Beyond Text-to-Slides. arXiv:2501.03936.

Zheng, Y.; Fu, D.; Hu, X.; Cai, X.; Ye, L.; Lu, P.; and Liu, P. 2025b. DeepResearcher: Scaling Deep Research via Reinforcement Learning in Real-world Environments. arXiv:2504.03160.

Zheng, Y.; Sun, S.; Qiu, L.; Ru, D.; Jiayang, C.; Li, X.; Lin, J.; Wang, B.; Luo, Y.; Pan, R.; et al. 2024. OpenResearcher: Unleashing AI for Accelerated Scientific Research. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 209–218.