

Vipera: Blending Visual and LLM-Driven Guidance for Systematic Auditing of Text-to-Image Generative AI

Yanwei Huang
yanwei.huang@connect.ust.hk
HKUST
Hong Kong S.A.R, China

Wesley Hanwen Deng
hanwend@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Sijia Xiao
xiaosijia@cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Motahhare Eslami
meslami@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Jason I. Hong
jasonh@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Arpit Narechania
arpit@ust.hk
HKUST
Hong Kong S.A.R, China

Adam Perer
adamperer@cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Abstract

Despite their increasing capabilities, text-to-image generative AI systems are known to produce biased, offensive, and otherwise problematic outputs. While recent advancements have supported testing and auditing of generative AI, existing auditing methods still face challenges in supporting effectively explore the vast space of AI-generated outputs in a structured way. To address this gap, we conducted formative studies with five AI auditors and synthesized five design goals for supporting systematic AI audits. Based on these insights, we developed Vipera, an interactive auditing interface that employs multiple visual cues including a scene graph to facilitate image sensemaking and inspire auditors to explore and hierarchically organize the auditing criteria. Additionally, Vipera leverages LLM-powered suggestions to facilitate exploration of unexplored auditing directions. Through a controlled experiment with 24 participants experienced in AI auditing, we demonstrate Vipera’s effectiveness in helping auditors navigate large AI output spaces and organize their analyses while engaging with diverse criteria.¹

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI; Interactive systems and tools.**

Keywords

AI Auditing, Generative Text-To-Image Models, Visual Analytics

1 Introduction

Generative text-to-image (T2I) models are gaining popularity for their ability to enhance the efficiency and expressiveness of creative activities. However, the associated risks are significant; these models can produce images that may perpetuate biases, cause offense, or disseminate misleading information [46, 54]. To surface and address these issues, researchers and practitioners have turned

to *AI auditing*—a process of repeatedly testing an algorithm with inputs and observing the corresponding outputs—in order to better understand its behavior and potential external impacts. [7, 45, 60].

Recent research in HCI and responsible AI (RAI) has explored tools and processes to support diverse auditors in AI audits [11, 13, 15, 37, 44, 64]. However, auditing generative T2I models in a systematic and scaled way remains challenging. To start with, manual evaluation of these systems can be time-consuming, often limiting assessments to a small number of prompts or outputs, which can hardly uncover the comprehensive issues of the system. Developers often face challenges in exploring potentially productive auditing directions and gathering scaled data for each identified issue [17, 37]. Reviewing and aggregating audit reports also remains a demanding task [11, 15, 48].

To tackle these challenges, automated approaches have also been proposed, often involving AI-supported output labeling and applied to textual models [4, 57, 62]. However, two primary obstacles persist when it comes to auditing T2I models. First, images encompass a broad range of semantics, resulting in a vast auditing space that is difficult to characterize with a concise list of criteria as often used for evaluating texts [68]. For example, an image may feature multiple characters, each of whom can be evaluated based on the clothing, and the clothing can be further assessed by its cultural nature and relationship to the character’s profession. To systematically explore this space, it is crucial to keep auditors aware of the *unknown unknowns* - criteria that are yet to be explored [17, 34]. Additionally, our observations from a formative study with five auditors indicate that many auditors often rely on intuition or personal experience during the auditing process without a formal approach. There is a lack of structured methodologies for navigating and exploring this extensive auditing space.

In this paper, we propose Vipera², an interactive system for streamlining and enhancing the systematicness of large-scale T2I model auditing. Vipera facilitates structured multi-faceted analysis through the visual cue of *scene graph* [32], where the auditing

¹The work is in submission.

²Vipera stands for “Visual Intelligence-Promoted End User Auditing”

criteria are organized hierarchically, coordinated with visual statistics, and associated with scenic semantics of images. Additionally, Vipera incorporates LLM-powered auditing suggestions to uncover new avenues for analysis and facilitate exploration within the auditing space. In particular, it provides audit analysis support by highlighting differences between images to suggest new criteria, and prompt suggestions to inspire exploration of new topics, with optional user-customizable keywords for targeted guidance. To support auditors in documenting and reporting their findings, Vipera allows users to bookmark visual evidence immediately upon getting insights, incorporates a dedicated note view for structured note construction, and supports LLM-powered note completion for better efficiency.

Through a controlled user study with 20 general auditors and 4 expert auditors, we demonstrate Vipera’s effectiveness in helping auditors navigate the large auditing space and organize their analyses while engaging with diverse criteria. In particular, we discovered that the blended guidance prompted participants to explore more prompts, images, and auditing criteria compared with conditions without guidance (see Section 6.1.2). We also show that visual and LLM-driven guidance are highly complementary, and blending them can reduce user workload, improve performance, and encourage a more systematic auditing process (see Section 6.2.3). Furthermore, we have revealed several interesting patterns in their auditing process that serve as the basis for future personalized auditing tools (see Section 6.2.4). For example, we observed breadth-oriented and depth-oriented patterns in both the creation of prompts and criteria, suggesting a need for personalized auditing guidance.

Ultimately, our work makes the following contributions:

- Vipera, a system that blends two distinct guidance modalities for the auditing of T2I models: *visual* guidance from an interactive, statistics-augmented scene graph, and *LLM-powered* guidance of prompts and auditing criteria.
- An empirical study on 24 participants with various background in auditing that evaluates the system’s usability as well as the distinct and combined effects of these guidance modalities, demonstrating that their integration leads to more systematic and thorough auditing with reduced cognitive load compared to three ablated versions of the system.
- A set of design implications for better blending visual and LLM-driven guidance to support systematic T2I auditing, as well as for incorporating tools like Vipera into real-world organizational settings and beyond.

2 Related Work

2.1 Auditing generative AI (at scale)

Generative AI (GenAI) systems have sparked increasing societal concerns due to potential harmful behaviors, such as social biases and violence [20]. This has led to increased focus on detecting and mitigating these issues through benchmarks like SafetyBench [80] and AgentHarm [3]. However, the limited diversity of inputs and contexts in these benchmarks hampers real-world performance assessments [34].

AI audits have gained prominence as a method for uncovering biased, discriminatory, or otherwise harmful behaviors in algorithmic systems [5, 7, 10, 27, 45, 47, 53, 60, 65, 69]. More recent research have demonstrated the value of conducting AI audits to ensure more safe and responsible GenAI, often even involving general end users in the auditing process, as they can uncover overlooked cases and provide insights [12, 17, 37, 64]. For example, Mack et al. engaged 25 people with disabilities to review and reflect on images generated by T2I systems they uncovered several nuanced societal stereotypes that extended beyond existing taxonomies, such as “perpetuating broader narratives in society around disabled people as primarily using wheelchairs, being sad and lonely, incapable, and inactive” [42]. Similarly, Shelby et al. conducted workshops with 15 cross-cultural artists to synthesize folk theories related to the use of T2I models, their potential harms, and harm reduction strategies perceived by artists [63]. Finally, Deng et al. created WeAudit system to engage 45 users in individually and collectively auditing GenAI, with scaffolds to support users in providing actionable insights to AI developers [17].

Prior research has primarily engaged people in auditing *limited sets of GenAI outputs*, while acknowledging the needs for better toolings for scalable, systemic auditing [17, 63]. However, research in HCI show that crowdsourced auditing remains challenging and costly due to data scale and often narrow evaluation criteria [11, 15]. To this end, recent studies have begun to focus on scalable auditing of generative AI, primarily addressing two key challenges. First, evaluating large numbers of outputs within limited timeframes: here, human-in-the-loop auditing combines large language models for automatic evaluation with user-facing summaries [58, 62, 82], and annotations and visualizations have been introduced to support interpretation [14, 25]. Second, early auditing tools often relied on rigid quantitative metrics, but recent efforts allow users to define their own criteria [35].

However, aforementioned work on improving systematic auditing has mainly focused on auditing text-to-text models. Extending them to text-to-image models introduces new challenges: Images contain rich semantic information (e.g., objects and styles) that can lead to a diverse range of auditing criteria, yet there is a notable lack of tools to assist auditors in interpreting auditing results for iterative assessments in the image context [52]. To fill this gap, **our work introduces Vipera, which integrates visual cues and AI-driven auditing suggestions, to support structured and systematic exploration of auditing criteria in the image domain.** In next section, we expand on the unique challenges of auditing image context.

2.2 Visual analytics for sensemaking image collection

Sensemaking of large image collections plays a key role in auditing T2I models at scale, and visual analytics approaches have proven effective in this context [1]. One common method involves clustering images based on pixel or vector embeddings for visualization, which supports various analytical tasks such as search and exploration, comparison, and visual summarization [50, 61, 79]. To incorporate image semantics alongside visual features, recent works identify

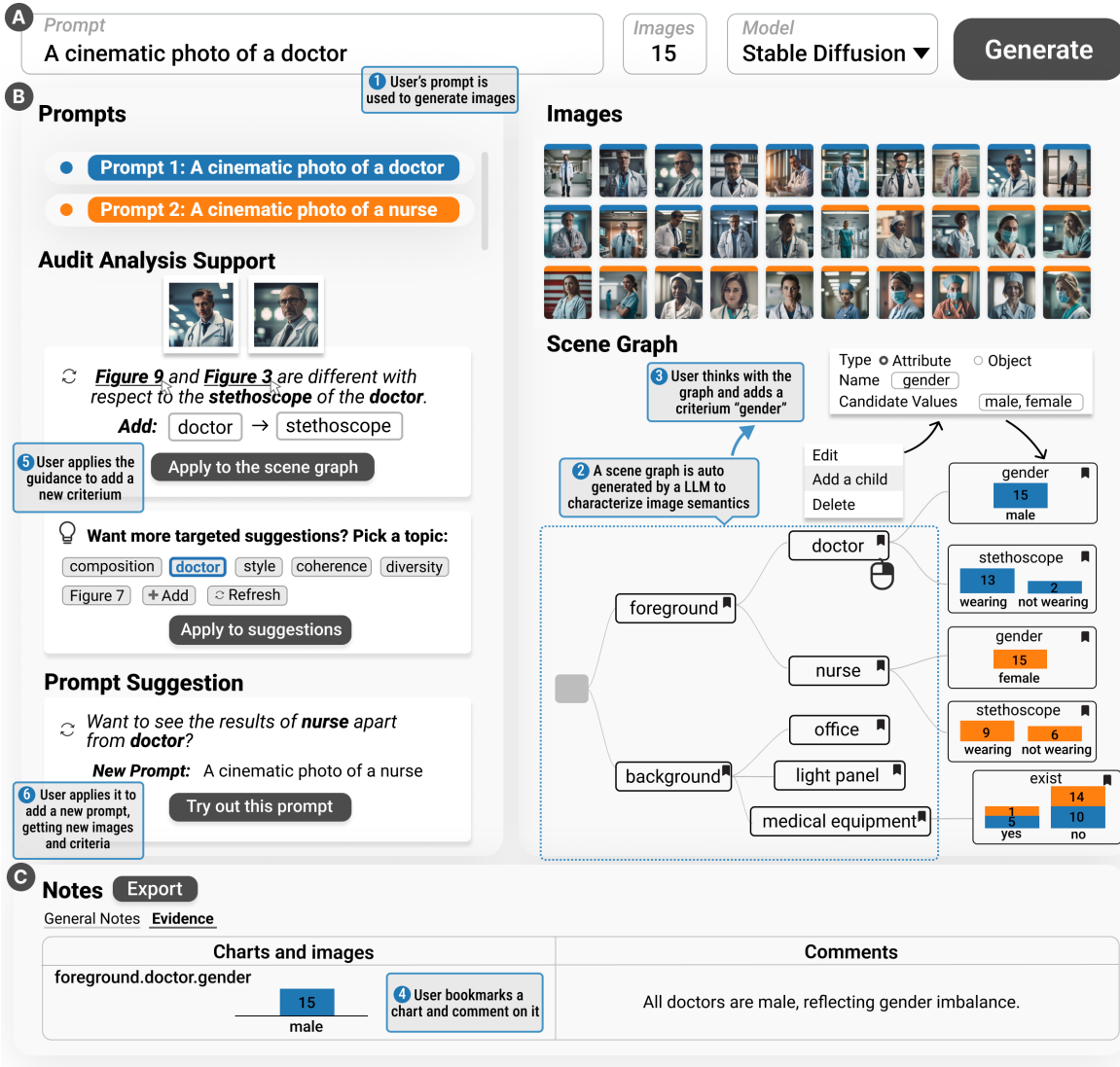


Figure 1: The Vipera interface. (A) The input box for creating prompts and specifying the number of images. (B) The analysis view for interactive auditing. Left: The prompts and AI-powered suggestions of auditing criteria and prompts. Right: The generated images and an interactive scene graph summarizing the image semantics to inspire users and guide the auditing process. (C) The note view for composing an auditing report.

semantic objects within images and extract concise textual representations—such as keywords, descriptions, or captions. These text analysis results facilitate tasks like semantic categorization and pattern mining [39, 75, 77]. However, characterizing the relationships between objects using plain text can be challenging. To address this, *scene graphs* have emerged as a more expressive representation, where nodes denote semantic objects and edges illustrate their relationships. This representation enhances detailed understanding of image collections and improves analysis performance [23, 32].

With the advent of generative T2I models, a recently emerged task is to visualize the relationship of between prompts and corresponding images. For instance, DreamSheets organizes prompts

and images in a spreadsheet layout to facilitate comparison [2]. Similarly, PromptTHis introduces an *image variant graph* to illustrate the semantic transitions of image clusters as users create new prompts [26]. However, these approaches often struggle to support sensemaking in large-scale image sets, where variations among outputs for a single prompt can lead to overly condensed visual clusters.

Drawing from prior work and situating it in the context of AI auditing, Vipera first embeds bar charts within a tree-based scene graph, revealing distributional patterns for images related to individual prompts and allowing for comparisons across prompts. Besides, Vipera leverages the scene graph as a visual aid to

foster systematic and structured AI auditing while providing inspiration to auditors. Additionally, **Vipera incorporates various view coordinations to facilitate navigation between auditing results and raw images for enhanced sensemaking.** Our work further contributes by evaluating these visualizations in the context of systematic AI auditing, offering critical insights for developing more effective auditing tools for text-to-image systems.

3 Design Study & Goals

3.1 Formative study

To inform Vipera’s design, we conducted a formative study with two main objectives: (1) understanding the common practices and challenges auditors face when evaluating generative text-to-image (T2I) models, and (2) testing our hypothesis that a scene graph as a visual aid could enhance insights and structure in the auditing process.

3.1.1 Study setup.

Participants. The study involved five participants (P1-P5) recruited from our collaborating institutions, including three Ph.D. students, one Postdoctoral researcher, and one software engineer, all of whom had prior experience in auditing generative AI or algorithms.

Prototypes. Following parallel prototyping [21], we designed two prototypes for the formative study, referred to as ViperaBase and Baseline in the following discussions. As illustrated in Figure 2, ViperaBase features a scene graph view, i.e., the zoomable node-link diagram, alongside a prompt input box and image view, whereas the Baseline includes all views except the scene graph view. In this view, blue nodes represent objects, and green nodes indicate their attributes, with labeled edges illustrating relationships. The size of each node reflects the number of images containing that object or attribute. Hovering over an attribute node displays a bar chart of evaluation results.

The technical workflow for both prototypes began when a user submitted a prompt to generate a set of images using the Stable Diffusion v3-medium model. For the ViperaBase prototype, we then utilized LLaVA v1.5 to process each generated image and extract its semantics into an individual scene graph. These individual graphs were subsequently aggregated to construct the final, comprehensive scene graph view. This aggregation process involved merging identical nodes (e.g., all “doctor” nodes) and summing their frequencies to inform the node size encoding.

Protocol. We first conducted semi-structured interviews to explore participants’ current auditing practices. Participants were then tasked with auditing the Stable Diffusion model using both prototypes in order (Baseline first, followed by ViperaBase). Each auditing session lasted five minutes, during which participants selected a prompt from the following prompts: a) *A cinematic photo of a doctor*, b) *A family having a picnic in the park*, c) *A peaceful nature scene with diverse wildlife*, and d) *An award-winning chef preparing a gourmet meal*. We designed these prompts to cover various topics (e.g., humanities, nature) and potential issues (e.g., bias on gender, profession, and race) while preventing users from the cold-start issue. The prompt would be used to generate 30 images, and the participants were asked to explore these images for as many insights

as possible and note their insights and evidence in brief phrases or sentences. After interacting with each prototype, participants shared their reasoning processes. Finally, they provided feedback on their perceptions of both prototypes and requirements for Vipera. All study sessions were conducted online, and the average duration of each session was approximately 50 minutes.

3.1.2 Findings. Our primary observations and insights from the study are summarized as follows.

Typical scales and motivations for T2I auditing. All participants reported that the scale of auditing varies based on the specific task. While the most common average scale for auditing was reviewing around 10 images at once, some participants indicated that they had audited hundreds of images simultaneously in certain cases. Motivations for large-scale auditing differed among participants: auditors who were also developers often engaged in systematic testing. For example, P1 reported a pattern of category-based image generation and auditing: *“The occupations I used were teacher, dishwasher, janitor, and then there were higher-class occupations like CEO, lawyer, and doctor. I asked the model to generate a series of images for each occupation, and then I used a visual assistant AI to gather demographic data about the generated images. This allowed me to collect information on what the model perceives a janitor and a CEO to look like. I ended up generating around 30 images for each of the 10 occupations, resulting in a total of about 300 images.”* In contrast, non-developer auditors preferred smaller scales typical of their daily work but would analyze a larger image corpus if the initial images have issues or did not meet their requirements. As P3 noted, *“I often focus on the image quality first... My goals would be generating [ideal] images rather than testing the model.”*

Auditing practices. Generally, the most common method for auditing was described as “just looking at the images”. As P4 noted, *“Because usually just like simply looking at them, you can tell [the issues].”* However, participants acknowledged the time-consuming nature of examining images at scale. Specifically, P1 and P4 utilized generative AI tools with visual understanding capabilities, such as GPT-4 and LLaVA [40], to gain observations or insights. P1 described his workflow: *“I use the visual assistant AI called LLaVA to examine images a lot... If you just create a pipeline where you create a bunch of images and then send all these images to the visual assistant AI, you can extract things like demographics on the images or just any random thing.”* Nonetheless, these tools presented challenges, including inaccuracies or lackluster results (as noted by P4) and visual anomalies—like distorted human hands—that could only be identified by human reviewers (observed by P1). P4 specifically mentioned: *“So it’s not 100% accurate. That’s the biggest challenge... it only works about like 90% of the time.”*

Auditing criteria. The study shows that ViperaBase helped users identify unnoticed abnormalities compared to Baseline. Participants explained that this was because the additional scene graph serves as a “check-list”, enabling them to examine the images in a disciplined manner, compared to merely relying on their experience and intuition in their usual workflow. P2 elaborated on the benefits: *“I think it’s good at just like telling you what is in the image. There’s a lot of things that you don’t really notice when you’re just staring at the image, staring at 20 images total. But when you get an AI to analyze every single image and compile all the data, you can really get a sense*



Figure 2: The ViperaBase prototype used in the formative study, showing the generated images (right) and a scene graph (left) for the user prompt (top). The scene graph is a node-link diagram where nodes indicate objects (or their attributes) within the images and edges indicate the semantic relationships. Bar charts will be shown when hovering on attribute nodes.

of what the model is printing out.” Moreover, all of them agreed that the node size encoding was helpful in uncovering minority objects, while comparing the sizes between adjacent nodes was also likely to yield interesting insights. As P3 noted: “So the fact that like the size of the circles corresponding to the amount of that item inside the image gives you a really good idea of the content of the images... So I think just like the most helpful thing is just the size of the circle.” However, three participants mentioned that the dense scene graph was too time-consuming to go through, and they hoped to prune or customize the graph, such as cutting out nodes with very few data points (P1, P5) or adding evaluation criteria related to the keywords in their prompts (P3).

3.2 Design goals

Based on the findings and feedback from the design study, we have derived the following design goals (DGs):

DG1: Leverage intuitive and interactive visual aids for structured, customizable auditing and effective result sense-making. According to the formative study, the visual aids, including the scene graph and the bar charts, might bring cognitive load to users, despite their effectiveness in inspiring audit directions and helping users understand the images. According to the user feedback, we infer two crucial reasons that account for this. First, the graphical layout and the size of the scene graph might overwhelm users, making them confused about where to focus on

during the analysis. The other was the lack of interactiveness in the visual aids: the users were forced to reactively check the suggested criteria instead of proactively creating them, limiting their creativity and activeness. Moreover, it is hard for users to associate the statistics back with the images due to the lack of coordinated views. Therefore, Vipera should grant users full agency through diverse interactions while allowing them to navigate between visual cues and data at any time.

DG2: Enable intuitive comparison of how the distribution of large-scale image outputs changes across prompts. Comparing different images has been demonstrated as a rich source of inspiration for auditing in prior literature [17, 64]. Likewise, we hypothesize that comparing the distributions and prompts may similarly provide abundant inspiration for users. To support this, Vipera should facilitate easy comparison of prompts and image distributions while highlighting the differences through proper visual assistance.

DG3: Promote divergent thinking by highlighting image details and potential prompts that auditors might have overlooked. Our formative study revealed that auditors often rely on intuition and personal experience, which can lead to confirmation bias and a failure to explore the model’s “unknown unknowns”[41]. This tendency limits the scope of an audit to familiar issues. To break this pattern, Vipera should proactively surface unexpected or novel avenues for investigation. This involves drawing attention

to subtle image differences that human reviewers might miss and suggesting new prompts that diverge from the user’s current line of inquiry, thereby encouraging a more comprehensive and less predictable auditing process.

DG4: Integrating the LLM-powered guidance with visual guidance. While LLM-powered suggestions are effective for inspiration, they can feel abstract or disconnected from the data if presented in isolation. Conversely, visual aids provide concrete evidence but may lack context or actionable next steps. To create a seamless and trustworthy workflow, Vipera should tightly couple these two modalities. This means that LLM-driven guidance (e.g., a suggested criterion) should be directly linked to the visual guidance (e.g., the scene graph), while the visual guidance characterizes the user’s current state of auditing and provides contexts for the generation of targeted and visually explainable guidance. This multimodal loop aims to ground the AI’s reasoning, making its suggestions more transparent and justifiable, while also making the visual data more actionable for the user.

DG5: Assist auditors in documenting, synthesizing, communicating, and acting upon their audit findings. The ultimate goal of an audit is to produce a coherent, evidence-backed report that can inform stakeholders and drive action. The process of documenting insights while simultaneously conducting an exploratory analysis is cognitively demanding and often disjointed. To address this, Vipera should integrate documentation directly into the analytical interface. This involves providing tools for auditors to seamlessly capture their thoughts, bookmark key visual evidence (such as charts or specific images), and structure their findings as they emerge, ultimately streamlining the transition from exploration to a communicable, actionable report.

4 The Vipera System

This section presents the design and implementation details of Vipera. As shown in Figure 1, three major components are included: a input view for creating prompts and generating images (A), an analysis view (B) for customizing the auditing process, seeking guidance, and analyzing the results, and a note view (C) for report generation.

4.1 Interface Design

4.1.1 Input view. The input view incorporates a input box for writing prompts, a numerical input box for specifying the number of images to be generated, and a model selector for choosing the model to be audited. The user can revisit this view at any time during the auditing to create new prompts or images.

4.1.2 Analysis view. The analysis view consists of four subviews: prompts, images, scene graph, and audit suggestions.

Prompts. The prompts view displays all prompts created by the user, each encoded by a different color. By default, the results of all prompts will be displayed and analyzed.

Images. The image view shows the thumbnails of the generated images, with the border color indicating the corresponding prompt. Users can click on the images to see the full-sized version, where a bookmark button is provided for saving images of interest for later report generation. In addition, they can hover on the image to see the labels for the auditing criteria defined in the scene graph, and

meanwhile the associated nodes in the graph will be highlighted in the scene graph (Figure 3). When they notice inaccurate labels, they are also allowed to edit the labels by right-clicking on the image, selecting the edit option in the consequent menu to enter a modal, and manually editing the labels in it.

Scene graph. The scene graph serves as both a visual summary of the semantic contents within the images and a visual cue for users to organize their auditing (DG1). Objects in the scene graph are represented as nodes and organized in a tree layout based on their semantic relationship. The tree layout is used to replace the graphical one in ViperaBase to for better intuitiveness (DG1), helping users organize their thoughts and navigate in the auditing space.

To use the scene graph for auditing, users can add nodes to it by right-clicking on existing nodes and selecting the “add” option in the context menu to add a node. A user can specify the following information when adding a node:

- **Node type:** Vipera supports two types of nodes: *object* nodes and *attribute* nodes. Object nodes represent physical objects or concepts in the images, while attribute nodes are leaf nodes that represent specific properties or characteristics of their parent object node. For example, non-leaf nodes in Figure 1 such as *foreground* and *doctor* are object nodes that describe the image semantics, while leaf nodes like *gender* are attribute nodes that can used to evaluate their parent nodes.
- **Node name:** The name of the node to be added.
- **Scope:** The scope of each node defines the prompts or images to which the current evaluation criteria (for attribute nodes) or the criteria of descendant attribute nodes (for object nodes) apply. This will help prevent evaluating all images against all criteria, which could be overwhelming and lead to intent misalignment (e.g., evaluating the nurse images with the criteria “doctor’s gender”). Users can choose to apply the criteria to all prompts, specific prompts, all images, or specific images. By default, all images and prompts will be selected.
- **Scope type:** Since Vipera supports iterative creation of prompts and auditing criteria, the scope of nodes in the scene graph can be dynamic to accommodate the prompts and images to come. Given this, Vipera supports two scope types: *fixed* and *auto-extended*. If an auto-extended scope is specified, the criteria will be applied to newly generated images, which is not the case for a fixed scope.
- **Candidate values (optional):** To prevent the LLM from generating overly diversified labels for the images that hardly leads to meaningful distinctions, Vipera allows user to specify a list of predefined candidate label values for classification (e.g., *male* and *female* for the attribute node *gender*). If provided, the LLM will be constrained to choose from this list when labeling images.

When an attribute node is added, the node will be used as a criterium for a labeler LLM to label the images. The result labels will be visualized in stacked bar charts with colors indicating the corresponding prompts to facilitate comparison (DG2). Users can

hover on the bars to navigate back to the matched images highlighted in the image view (Figure 3). Additionally, they can edit the information of a node or choose to relabel the images based on an attribute node via the context menu. A relabeling will also be triggered when the user modifies the candidate values or the scope. In this case, two options are provided: relabeling all images or only relabeling images that are affected (i.e., contain a label that is not within the candidate values or not in the scope). The user can change the option in the relabeling modal.

Suggestions. Informed by DG3, Vipera provides two classes of auditing guidance, *audit analysis support* and *prompt suggestion*, to inspire users of new auditing criteria and prompts respectively.

- **Audit analysis support.** This module highlights two images with notable differences and suggests auditing additional auditing criteria as nodes. To mitigate intent misalignment caused by random selection, Vipera enables users to choose from a diverse set of LLM-generated keywords or create their own custom keywords for more targeted suggestions. These keywords are injected into the prompt, and suggestions are presented to the user only when the LLM’s self-reported confidence score exceeds a predefined threshold after several iterations, indicating a reliable identification of topic-relevant differences between the images.
- **Prompt suggestion.** This module promotes divergence thinking by suggesting insightful prompts, replacing words or phrases in existing prompts. Upon users deciding to try out the prompt, Vipera will duplicate the existing auditing criteria while generating new images, allowing for an effective comparison of results between different prompts. For instance, when the user applies the suggestion replacing the word “doctor” with “nurse” in the prompt “a cinematic photo of a doctor”, Vipera will automatically add an object node “nurse” to the sibling of “doctor” while duplicating the descendent nodes of “doctor” in this new branch. Such a transformation makes sure that the labeling results of newly generated images for nurses will be displayed in the correct branch.

4.1.3 Note view. To facilitate the authoring of auditing reports (DG5), a note view is incorporated, which consists of two tabs: a tab of *General Notes* that includes an input box for noting down the overall insights, and a tab of *Evidence* that shows the bookmarked charts and images. Detailed user comments can be appended to specific charts or images as fine-grained insights. Moreover, Vipera features LLM-powered auto-completion when users write in this view to further enhance the auditing efficiency. The prompts, bookmarked items, and existing notes will be injected into a LLM to generate the completion, which can be easily applied upon users tap on the *tab* key on the keyboard.

4.2 User workflow and technical pipeline

Vipera has been developed and deployed as a web-based application. Figure 4 shows the workflow of Vipera³. After generating images from a user’s prompt, Vipera initiates the scene graph construction

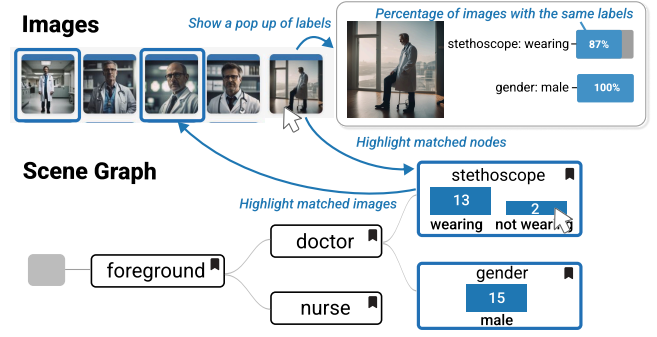


Figure 3: View coordinations between the image view and the scene graph. When the user hovers on an image, a pop-up will appear showing its labels, and the relevant attribute nodes in the scene graph will be highlighted. When the user hovers on a bar in an embedded bar chart in the attribute node, the corresponding images will be highlighted in the image view.

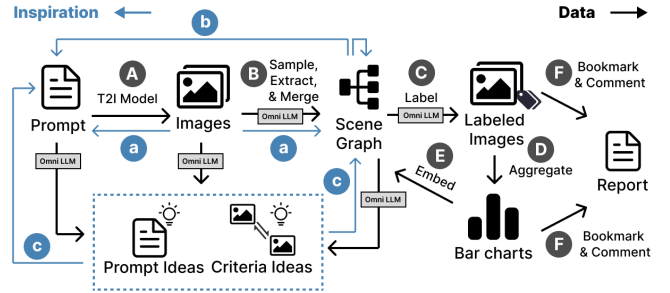


Figure 4: Vipera’s technical pipeline. Black edges indicate the data flow: (A) The prompt from the user is fed to T2I model to generate images. (B) A scene graph is generated from the images (see Figure 5 for details). (C) Upon an attribute node is added to the scene graph, the images will be labeled using the specified criteria. (D) The labels will be aggregated and visualized as a stacked bar chart. (E) The charts will be embedded into the attribute nodes of the scene graph for rendering. (F) The images and charts will be added to the auditing report upon being bookmarked by the user, with the user’s comments attached to them. Blue edges indicate the inspiration flow for iterative auditing: Prompts and auditing criteria can be inspired by inspecting the images (a), thinking with the scene graph (b), and applying the LLM-driven guidance (c).

process, which is detailed in Figure 5. It starts by randomly sampling four images (or all, if fewer than four are available) to serve as a representative basis for the initial scene graph (Figure 5A). While this small sample may not capture the entire corpus, it provides a manageable starting point for the user. For each sampled image, Vipera prompts an omni-modal LLM (Gemini 2.5 Flash in our implementation) to extract a tree-based scene graph that captures the image’s semantics (Figure 5B). The LLM is instructed to identify key physical objects and notable features and organize them into a

³All system prompts used in the pipeline are provided in the appendix.

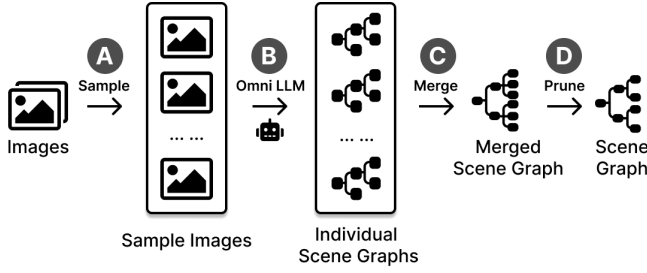


Figure 5: The detailed technical pipeline for generating the initial scene graph. (A) A random subset is sampled from the images. (B) For each image, an individual scene graph is generated through an omni-modal LLM. (C) The result scene graphs are merged into one. (D) The result is pruned by keeping a maximum of five leaf nodes.

tree structure. To provide a consistent high-level organization, we constrain the first level of the tree to “foreground” and “background” nodes in our study, which is an intuitive and widely applicable distinction. The resulting scene graphs from the four images are then merged, with their leaf nodes combined (Figure 5C). To avoid overwhelming the user, a maximum of five leaf nodes are randomly selected for the final aggregated scene graph (Figure 5D, **DG1**).

With the scene graph serving as a visual summary, users can add object or attribute nodes, where attribute nodes will serve as auditing criteria. When an attribute node is added, Vipera begins to evaluate and label the images using this criterium (Figure 4C). First, it updates the scene graph with the new node and ensures that all images within the new node’s scope are also included in the scope of its ancestor nodes up to the root. Then, it extracts a *partial graph schema*, which is a subset of the scene graph containing the new node and its path to the root. This schema is used to prompt an omni-modal LLM (GPT-5-mini in our implementation) to generate a label for each image within the specified scope. If the user has defined candidate values for the attribute, the LLM is constrained to choose from that list. The resulting labels are then aggregated, and their distribution is visualized as a stacked bar chart (Figure 4D) embedded into the added node using the D3.js library⁴ (Figure 4E). Furthermore, users can bookmark the images and the charts within scene graph nodes at any time during auditing and comment about them in the note view for an auditing report (Figure 4F).

Additionally, users can benefit from diverse sources of inspirations for iterative exploration of prompts and auditing criteria in Vipera, illustrated as blue arrows in Figure 4. Specifically, they may inspect the images carefully to get insights and new inspirations for prompts and criteria (Figure 4a). Such inspirations may also come from the scene graph, where the user either thinks with its structure about potential nodes to be added or examine the embedded charts for insights (Figure 4b). Finally, they may adopt the LLM-powered suggestions at any time for new auditing ideas (Figure 4c). Such an inspiration loop is built upon the artifacts from the data loop and serves as the engine for the user workflow, promoting iterative, mixed-initiative, and comprehensive auditing.

⁴<https://d3js.org/>

4.3 Usage Scenario

Now we illustrate Vipera’s workflow using the story of Bob, an auditor tasked with evaluating the Stable Diffusion model for potential biases.

To begin his audit, Bob enters the prompt “A cinematic photo of a doctor” into the input view and generates 15 images (Figure 1, step 1). Vipera populates the analysis view with the prompt and the images, while showing a corresponding scene graph that characterizes the semantics of images (Figure 1, step 2). The graph organizes the content into *foreground* (e.g., *doctor*, *nurse*) and *background* (e.g., *office*, *medical equipment*) nodes.

With the scene graph as an overview of the images’ contents, Bob organized his thoughts and decided to start from the *doctor* node. He right-clicked on this node and chose to add a child in the context menu (Figure 1, step 3). The new node was an attribute node named *gender* with candidate values *male* and *female*. Vipera automatically labeled the images based on the gender of the doctor and visualized the results in the bar chart embedded in the new node. Realizing all doctors were male from this chart, Bob bookmarked it as evidence and documented his insights in the notes view (Figure 1, step 4).

Afterwards, Bob felt tired to come up with new auditing ideas and decided to refer to the AI suggestions. He first examined the Audit Analysis Support, which highlighted two images with a key difference and suggested adding a *stethoscope* attribute to the *doctor* node. Bob found the suggestion insightful and applied the suggestion (Figure 1, step 5). The resulting chart showed that some doctors were not wearing a stethoscope. To verify this, he hovered over the corresponding bar, which highlighted the specific images in the image view (Figure 3), allowing for rapid, coordinated sensemaking.

Next, Bob noticed the **Prompt Suggestion** recommending he replace “doctor” with “nurse” to explore a related profession. He accepted the prompt, and Vipera generated a new set of images (Figure 1, step 6). Crucially, the system also added a new *nurse* branch to the scene graph and automatically duplicated the existing criteria (*gender* and *stethoscope*) under it. With Vipera’s color-coded bar charts, Bob compared the results of these two prompts on the scene graph easily and acquired new insights in occupational stereotypes. From this point, Bob can continue his audit through multiple iterations of structured exploration and comparison.

5 User Study

We have conducted a user study to evaluate the usability and effectiveness of Vipera. Our main goal was to understand how Vipera’s core features—the Scene Graph and AI-powered suggestions—influence the process and outcomes of different AI auditing tasks. We were also interested in the usage patterns revealed from the user’s auditing process.

5.1 Study Design and Methodology

Participants. We recruited 24 participants (P1-P24, 14 males and 10 females). Among them, 20 were general student auditors recruited from our institution (4 undergraduates, 7 Master’s, and 9 Ph.D students) who reported an average familiarity with AI auditing ($M=3.04/5$, $SD=1.16$) and use T2I models with an average frequency ($M=2.92/5$, $SD=0.929$). The remaining 4 participants were expert auditors recruited from both academic and industry who

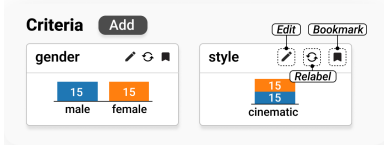


Figure 6: The additional Criteria View added to Systems A and C. Users are allowed to add auditing criteria, and the results will be visualized as bar charts in the respective cards. Each criterium card includes three buttons for editing the criterium, relabeling the images, and bookmarking the criterium, respectively.

had rich experience and knowledge in AI auditing. The ages of the participants ranged from 20-29, with an average of 23.9.

Baseline. There were four systems (A-D) used in the study, including three baseline systems adapted from Vipera (A-C) and the original Vipera system (D).

- *System A:* The system was adapted from Vipera by removing both the AI auditing support and the scene graph. To equip the system with minimum auditing capabilities, we added an additional *criteria* view to this system allowing criteria creation and result visualization, as depicted in Figure 6. User can add auditing criteria, set candidate values and scope for labeling as in Vipera.
- *System B:* The system was adapted from Vipera by merely removing the AI auditing support.
- *System C:* The system was adapted from Vipera by removing the scene graph and replacing it with the same criteria view as in System A.

Tasks. Participants were asked to audit the Stable Diffusion XL model. They were assigned one of four initial prompts designed to elicit a range of potential biases and quality issues: (a) *A couple on their wedding day*, (b) *A family having a picnic in the park*, (c) *Worldwide athletes in the Olympic Games*, or (d) *An award-winning chef preparing a gourmet meal*. These prompts are able to cover diverse topics and image features, such as individuals (a, d), and crowd (b, c), country (c), occupation bias (a, c), cultural issues and potential bias (c, d). Participants were instructed to explore the model’s behavior freely from this starting point.

Procedure. Each session lasted approximately 90 minutes. After providing informed consent, participants completed a demographic questionnaire. They were then equally divided into two groups (Group I and Group II), with the distribution of general and expert auditors kept consistent across both groups. The study employed a mixed design where participants in Group I used systems A, B, and D, while those in Group II used systems A, C, and D. Before using each system, they were provided a tutorial on the system, followed by a warm-up session of up to 5 minutes to familiar themselves with the interface. Subsequently, they were given 15 minutes to perform an auditing task, with the goal of acquiring as many insights as possible. To balance between controlled comparison and free exploration, all participants began with their assigned prompt but were free to write subsequent prompts on the same topic. Each starting prompt was assigned to an equal number of participants. They were also instructed to note down their insights

in phrases or short sentences as soon as they get them to eventually form an auditing report after using each system. We employed the think-aloud protocol [38] to capture their reasoning process. After completing the tasks with all three systems, participants filled out a questionnaire that included NASA-TLX scales for each system and several 7-point Likert scales to rate the usefulness of system components. The study concluded with a semi-structured interview to gather qualitative feedback. Participants received a \$10 gift card as the compensation for their time. All sessions were both screen- and audio-recorded. The study proposal has been approved by the IRB committee of our institution.

Data collected and analysis. We collected multiple forms of data for each participant: three auditing reports (one for each system used), pre- and post-study questionnaires, system interaction logs, and full session recordings. Our analysis approach was twofold. Quantitative data from the questionnaire responses and system interaction logs were analyzed using statistical methods to assess performance and usability. Qualitative data from the auditing reports and transcripts from the think-aloud protocol and semi-structured interviews, were analyzed using thematic analysis [9] to identify user strategies, reasoning patterns, and key feedback themes.

6 Results

In this section, we share the results from our controlled experiments. In particular, analysis of the NASA-TLX questionnaire responses revealed that participants reported lower workload and improved performance when using Vipera—especially with the AI auditing support—compared to the baseline. Auditing logs show that Vipera helped participants create more auditing criteria through interaction with AI support and visual guidance. Our interviews with practitioners also revealed nuanced trade-offs between scaffolding systematic auditing while managing cognitive demand, personalizing recommendations, and leveraging resources tailored to individual auditors.

6.1 Mixed-method Analysis of Questionnaires and Audit Logs

6.1.1 Questionnaires. The detailed distribution of participants’ rating on NASA-TLX scales is illustrated in Figure 7. Overall, participants reported less workload in all NASA-TLX dimensions and better performance on average when using Vipera (D) compared with the baseline (A). The difference was significant in the Mental Demand ($t=-2.055$, $p=0.0257$), Physical Demand ($t=-2.061$, $p=0.0254$), and Performance ($t=2.685$, $p=0.0066$) dimensions. These results show that Vipera is generally effective in reducing users’ perceived workload and improve the auditing performance.

For participants in Group I, we observed a general pattern of reduced Physical Demand and Temporal Demand and increased Performance across the sequence from A to D as shown in the Group I panel. Specifically, Performance improved significantly from A to B ($t = 3.023$, $p = 0.0058$), and from B to D the reported increase in Performance was also significant ($t = 1.876$, $p = 0.0437$). Note that the latter p-value is modest and multiple comparisons were performed; if corrections for multiple testing are applied the significance of some comparisons may change. It is worth noting that the Mental Demand in A and B has the same median with the average of B

slightly higher, and meanwhile the median of both of them are higher than D. Overall, these results suggest that the scene graph alone may increase mental demand, while subsequent addition of the AI auditing support (D) reduces workload and improves performance, but order and multiple-comparison concerns warrant cautious interpretation.

Meanwhile, for participants in Group II, System C generally shows a lower workload than System A across multiple dimensions, including Mental Demand, Physical Demand, Temporal Demand, and Effort. Figure 7 also indicates that Performance and Frustration across A, C, and D are broadly comparable with overlapping distributions. Statistical comparisons show that participants reported significantly lower Mental Demand ($t = -5.063$, $p = 0.0002$), Physical Demand ($t = -1.959$, $p = 0.038$), Temporal Demand ($t = -2.077$, $p = 0.031$), and Effort ($t = -2.803$, $p = 0.0086$) for System C compared with System A. They also reported significantly higher Temporal Demand for System D compared with C ($t = 2.569$, $p = 0.0130$). These results indicate that the AI auditing support can effectively reduce users' mental workload and effort when added alone, while the scene graph may introduce additional temporal demand when added alongside the AI auditing support. Most participants attributed this extra temporal demand to the significant amount of time spent on exploring promising auditing criteria when the scene graph is present. A few participants also mentioned the fact that it is naturally difficult to find useful criteria and get insights in systems used at a later time when they have already explored the model extensively in the previous systems.

In addition, participants generally perceived all components within Vipera as helpful, with an average rating of 3.50 ($SD=0.885$) and 3.67 ($SD=0.868$) on the scene graph and AI auditing support, respectively. They also appreciated the automatic labeling of images ($M=3.375$, $SD=0.970$) and note-taking features ($M=3.71$, $SD=1.083$).

6.1.2 Auditing logs. We measured the number of prompts, generated images, used criteria, and bookmarked charts or images used by each user in each system, as illustrated in Figure 8. Note that we use the number of bookmarks as a proxy to evaluate the number of insights despite potential misrepresentation, given that it is hard to define an atomic insight for calculation. The results show that the introduction of guidance had a notable effect on user behavior. Specifically, users created more auditing criteria in all guided systems (B, C, and D) compared to the baseline (System A), with the highest number of criteria being used in the full Vipera system (System D). A similar trend was observed for prompts and images, though the increase was primarily driven by the systems featuring AI auditing support (Systems C and D). In contrast, the addition of the scene graph alone (System B) did not lead to a noticeable increase in the number of prompts or images. Conversely, we observed a consistent and significant decrease in the number of bookmarked charts across the guided systems. This effect was most pronounced in Systems C and D, suggesting that the presence of AI support may have shifted users' focus from documenting existing views to actively exploring new criteria and prompts. This interpretation aligns with participants' feedback in the interview, which indicated that more effort was spent on exploration when using Vipera's advanced features.

One interesting discovery is that the introduction of AI auditing support (System A vs. C; System B vs. D) consistently leads to an increase in the number of prompts, images, and criteria, as well as a consistent decrease in bookmarked charts or images. This shows that auditing support powered by AI is beneficial and inspirational for most participants despite the extra effort in exploration. By contrast, the effect brought by the introduction of the scene graph (System A vs. B; System C vs. D) was more diverse and random, varying from person to person.

We also evaluated the auditing pattern and efficiency of participants. Overall, we found that in Systems C and D, respectively, 87.2% and 78.4% of the criteria were authored by AI, while only 61.8% and 50.8% of the prompts were generated by AI. This reveals that participants desired more agency and control in writing prompts than criteria. We also observed a high consistency in users' prompts: the average pairwise BERT [19] cosine similarity between all user-generated prompts within each session was 0.904 (on a 0–1 scale), showing that most participants intentionally utilized similar prompts to facilitate comparison. Meanwhile, we observed that each prompt led to 1.2, 1.2, 0.9, and 0.8 bookmarks on average in Systems A–D, respectively, showing a decrease in auditing efficiency despite the broadened auditing scope.

Notice that we also calculated the average number of prompts, images, criteria, and bookmarks for each starting prompt, but we didn't find any major and consistent impact of the starting prompt on these metrics.

6.1.3 Auditing reports. We also conducted thematic analysis on the auditing insights authored by the participants. The results are shown in Table 1. We find that almost all participants paid significant attention to evaluating the *quality* of images, assessing them from pixel-level (e.g., resolution), component level (e.g., human anatomy), to a macro level (e.g., theme and style). We summarize them into three subcategories: *Exquisiteness*, *Authenticity*, and *Style*. Many of them also moved beyond single-image evaluation, focusing on multi-image features like distribution and diversity.

Another common focus is *Prompt Interpretation*, where participants investigated if the generated images aligned with the prompt, common sense, or their understanding. This reveals the gap between the knowledge of existing T2I models and both the objective world and the specific auditor.

Furthermore, a few auditors mentioned their reasoning for the LLM's behavior in their insights. With the help of visualizations in Vipera, like stacked bar charts, participants were enabled to compare images and their distributions from different prompts easily. Some of them summarized the LLM's tendency in image creation or even explained it, proposing assumptions about the LLM's training sets.

It's worth noting that although we clarified in the system tutorial that the task was to audit the T2I model specifically, a few participants still misunderstood the goal and commented on other system components, such as the scene graph and the AI labeler. Relatedly, while many participants noticed the errors of the AI labeler, only about half of them tried to correct the errors or asked for relabeling with modified expressions, with many participants ending up trusting the wrong results. This pattern extends prior work on aligning user intents with LLM-based judges [49, 62], showing that even

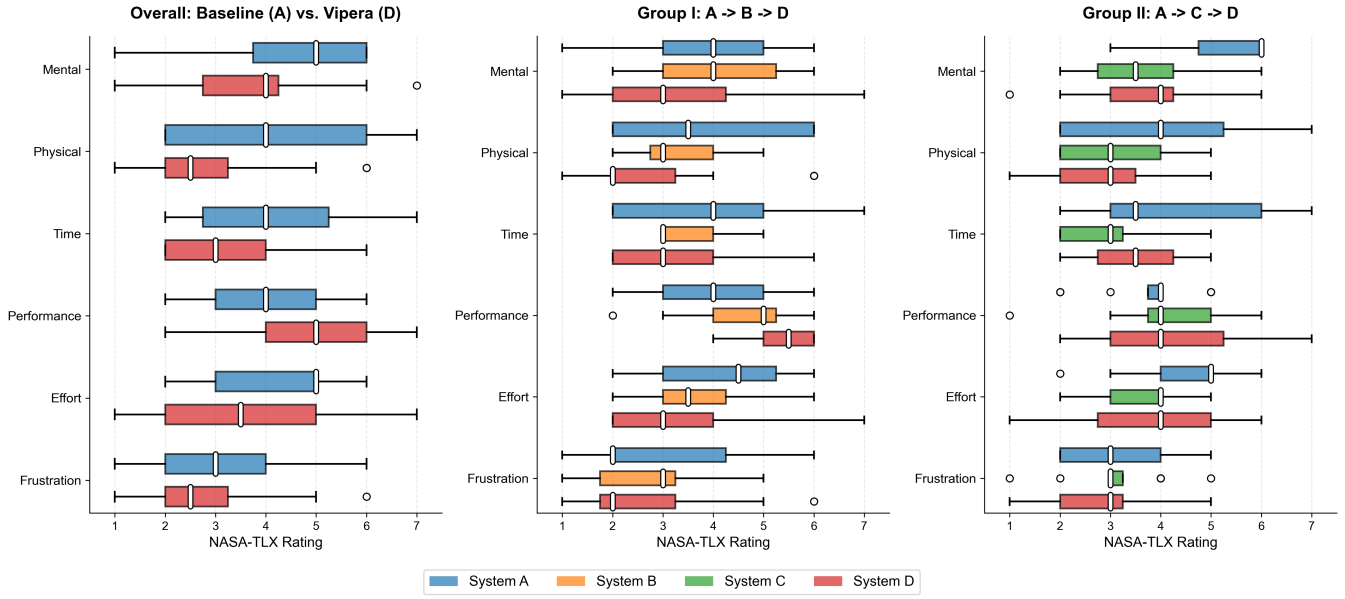


Figure 7: Participants' ratings on the NASA-TLX scale. The first subplot shows a rating comparison ratings of all participants on the Baseline (A) and Vipera (D), while the other two subplots shows detailed rating comparison for the three systems used by participants in each group. The white marks on the boxes indicate the medians.

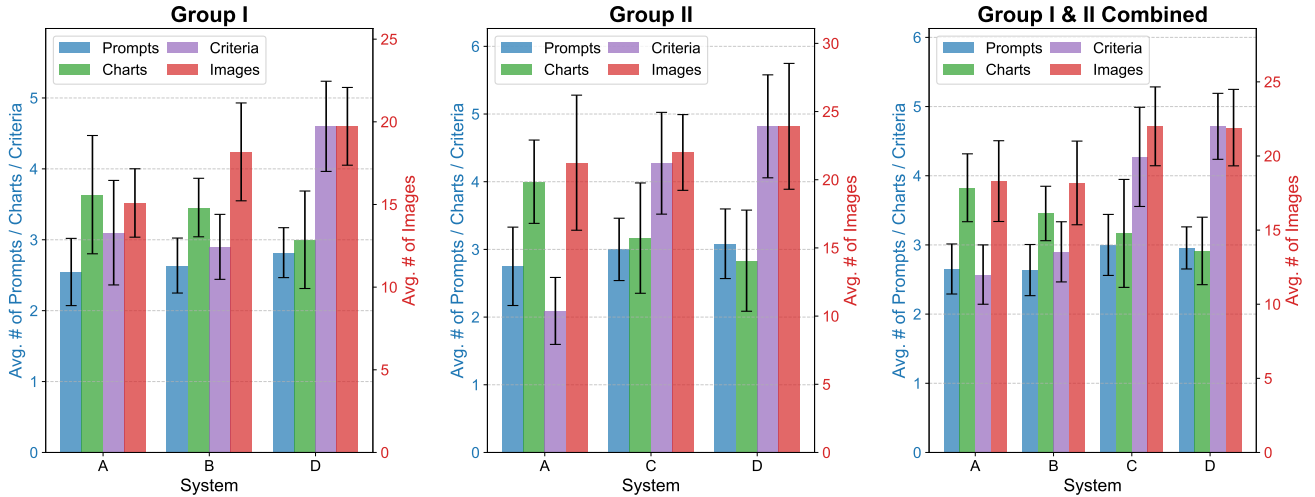


Figure 8: The number of prompts, images, used criteria, and bookmarked charts or images in each system used by the participants. Error bars indicate standard error of the mean.

when users are prompted to evaluate a specific component, failures in automated judges can go unchallenged, so such tools should be treated and deployed with caution.

6.2 Qualitative Findings from Interviews

Our main insights from the semi-structured interviews are summarized below.

6.2.1 The scene graph encourages systematic auditing while bringing additional cognitive demand and pressure. We received divided opinions on the effect of the scene graph. Participants in favor of this visual guidance suggested that it provided them with a general overview of the images, served as a good source of inspiration, and encouraged systematic thinking and attention

Table 1: Thematic analysis results of the participants’ auditing insights.

Category	Subcategory	Example quotes (participant ID)
A. Quality	A1. Exquisiteness	“The details are slightly lacking” (P08); “The picture looks beautiful” (P16); “Mosaics appearing” (P19)
	A2. Authenticity	“The person looks like a dummy” (P02); “The dishes are floating” (P12); “People’s faces look distorted” (P05)
	A3. Style	“Prefer to see more realistic images but the majority are artistic” (P21)
	A4. Diversity	“The characters are highly homogeneous” (P19); “Racial bias” (P02); “AI may discriminate unwealthy people” (P13)
B. Prompt Interpretation	B1. Alignment with the prompt	“The content of the picture aligns with the ‘worldwide’ theme” (P15); “No ‘athletes’ in the image” (P03)
	B2. Alignment with common sense	“There are six fingers in one hand” (P08); “People’s mood is wrong; they should be happy on the wedding day” (P05)
	B3. Alignment with user understanding	“Even if I added the word ‘beautiful’ to the prompt, the images were still ugly” (P02); “It misunderstood me” (P03)
C. Reasoning for LLM Behavior	C1. Prompt comparison	“The pictures are better when more details are included in the prompt” (P06)
	C2. Express the tendency of the LLM	“The chance of errors increases when more elements exist” (P16); “LLM didn’t understand subjective keywords well” (P02)
	C3. Explain for the LLM behavior	“I probably want a more diverse set of colors. Could be attributed to most of the training images having this color” (P21)
D. Critic on the system		“AI gives me more options... But the recommended prompt may have little novelty” (P14); “The classification of images is problematic” (P06)

to detail, ultimately making them feel more productive and confident. For instance, P1 noted, “*I was used to focusing on the macro scenario in the image and treating it as a whole, but the scene graph made me realize that images could be decomposed into individual elements.*” P21 similarly mentioned that the scene graph gave him “*a systematic understanding of the images’ compositions.*” Moreover, we observed that some participants acquired inspirations from interacting with the scene graph. P3 mentioned that apart from the tree structure and the node contents, the results of open labeling (i.e., without predefined candidate values) could also be inspiring, reminding him of “corner cases”. However, others found the scene graph’s complexity to be a barrier, citing a steep learning curve and information overload. This structured approach also had a psychological cost for some. P06 and P09, for instance, felt pressured by the graph, stating it limited the “free-form authoring” he preferred and undermined his sense of agency. Additionally, several participants expressed neutral views. P21 and P08 argued that the usefulness of the scene graph was most prominent when the prompts or images were abstract (i.e., with unclear scenes or in an artistic style), which were easier to assess using the scene graph compared with direct inspection. P21 further indicated that the scene graph’s usefulness was content-dependent, asserting that nodes related to human characters or biases were more beneficial than those focusing on minor details.

6.2.2 AI-powered auditing suggestions, though inspiring and effort-saving, needed to be personalized and upskilled.

While most participants agreed on the inspiration and saved effort from AI, they suggested various opportunities for improvement. Some participants (P02, P09, P11, P12, P22) believed that the performance of AI auditing support might decrease over time, with the LLM’s creativity “converged” at some point, providing recommendations that were repetitive, not interesting, or did not make sense. P1 and P21 noted that the LLM tended to modify the original prompt too much when suggesting new prompts, causing significant changes to the images and hence hindering comparison. Others expressed their desire for more personalized suggestions that take the system state and interaction history into account. In addition, we noticed that while Vipera allowed users to customize LLM suggestions through selected keywords or topics, few participants viewed this feature as useful. P07 and P20 described the process of manually aligning the LLM as demanding and reactive. By contrast, they preferred a more proactive, LLM-driven approach that anticipates their preference and adapts automatically.

6.2.3 Effective auditing relies on drawing on diverse sources of insight and integrating them to perform a comprehensive assessment. Based on the interviews and our inspection of the task process, we have observed six major sources of insights that motivate user interactions during the task: direct image inspection, the scene graph, the AI auditing support, comparing results from different prompts, personal experience, and serendipity. Almost all of them were commonly integrated by all participants with no significant differences between general and expert auditors. Contrary to our assumption that expert auditors may prefer advanced

auditing techniques, P22, an expert auditor, highly relied on direct inspection. He shared that *“examining the images is super effective...If I were asked to start by using the second (C) or third system (D) without actually trying the first (A), I would probably just rely on AI guidance without noticing the insights from checking the images, since I am a lazy person.”*

6.2.4 The auditing approaches varied among participants, with both breadth-oriented and depth-oriented patterns in prompts and criteria. Participants showed various patterns during their auditing process and expressed distinct rationales in the interview. Most participants followed a pattern where they attempted a diverse range of criteria on only a few prompts. For instance, 9 participants used fewer than 2 prompts on average among all three systems, with four of them spending most time exploring distinct criteria. P3 explained that he began the task with several rough auditing criteria in mind, so he tended to exhaust those criteria on existing images before moving on to new ones. P18 shared that she found the images from the initial prompt to be of extremely low quality, leading her to spend significant time evaluating them thoroughly without time for exploring other prompts. By contrast, some participants followed another approach, iteratively refining prompts based on new insights in a manner similar to prompt engineering. P5 and P6 explained that, rather than assessing images from multiple perspectives, they tried to obtain satisfactory images after finding the initial outputs unsatisfactory and therefore focused on how model behavior changed across successive prompts. Although some participants ultimately obtained images they were happy with, P6 did not and became frustrated, noting down his constraints in the auditing report. We also noticed one participant, P1, who tried to balance between breadth and depth when exploring prompts and criteria. He noted, *“whenever I felt that I dived too much in a specific direction, I told myself to think about other ways”*.

7 Limitations

Our study has several limitations. First, we have identified several threats to the validity of our study. For instance, our controlled experiment, conducted in a lab-like setting with a fixed duration, may not fully capture the open-ended and long-term nature of real-world auditing tasks. Meanwhile, our participant pool consisted primarily of student auditors, whose workflows may differ from those of professional red teamers or compliance officers. The generalizability of our findings could be strengthened by future longitudinal deployments with a more diverse range of experts and domain-specific prompts. Furthermore, we used in our analysis bookmarks as a proxy for insights, a metric that future work could refine. The fixed order of system presentation could also introduce potential learning or fatigue effects, as the difficulty of finding new insights naturally increases over the course of a session. Finally, while Vipera can be extended to audit multiple T2I models in parallel, doing so likely requires further work to adapt guidance and reconcile differences across models.

8 Discussion

Based on our results, in this section we discuss opportunities to better support the design, development, and evaluation of AI auditing tools that enable more systematic auditing. In particular, we

highlight how visual guidance can scaffold auditors before, during, and after their work by making complex results more interpretable and actionable. We also examine how AI-driven inspirations and guidance can be designed to better adapt to users’ goals, drawing on our empirical findings. In addition, we shed light on best practices for integrating different forms of guidance to support human–AI collaboration in auditing. Finally, we discuss the potential to embed tools like Vipera into on-the-ground responsible AI workflows in industry settings and beyond. Together, these opportunities outline design directions for building future auditing tools capable of scaling to the complexity and diversity of generative AI systems, while enhancing and complementing auditors’ capacity of sensemaking AI outputs at scale.

8.1 Opportunities from involving visual guidance in algorithm auditing

Prior works on AI auditing have explored various types of guidance ranging from community, algorithmic, to expert [17, 64]. Our findings reveal the numerous opportunities brought by providing *visual* guidance. Specifically, we show that users could benefit from visual guidance throughout the auditing process: before auditing, it can help users organize their thoughts and provide inspiration (see Section 6.2.1). During auditing, it can serve as an interactive environment for seamless auditing. After auditing, it can organize the results for systematic analysis. Furthermore, we demonstrate the effectiveness of augmenting auditing with the scene graph in improving the confidence and production of users.

Drawing from prior works in data visualization [76], future auditing interfaces could explore different visual designs that adapt with input modality, the visual literacy of auditors, and the themes or topics of auditing. For instance, visualizations could take the form of timelines to capture changes across iterative audits, given similar practices in visualizing iterations of data or code [22, 30, 70]. Auditing interface could also leverage keyword or topic visualizations such as word cloud [29], cluster maps, or document cards [67] to reveal patterns across diverse user reports. Finally, there are opportunities where existing theories and techniques for visualization recommendation (e.g., [55]) or proactive visual analytics support (e.g., [73, 81]) can be integrated, augmenting the auditing process with guided visual data analysis.

8.2 Leveraging AI for inspirations and guidance in AI auditing

While providing algorithmic guidance has been shown helpful for auditors’ working process [64],

Building on insights from how participants perceived and interacted with AI-generated auditing guidance (see Section 6.2.2), our study extends existing suggestions (e.g., [17, 37, 64]) for designing AI-powered auditing guidance and suggests several advanced lessons.

First, as we shown that AI-generated auditing guidance should be personalized and built upon the inference of user’s goals or preferences. Future design could consider taking advantage of users’ interactions, learning from users’ feedback, and providing guidance that matches users’ focus, as demonstrated in prior literature [28, 66].

Second, a number of participants suggested that AI-generated guidance should incorporate mechanisms to avoid repetitiveness and maintain novelty and relevance over prolonged use. In future interaction design, users should be informed when guidance on auditing criteria (or prompts) is going to converge so that they can try to generate new prompts (or criteria) purposefully. Alternatively, higher sampling temperature or nucleus sampling could be used periodically to surface diverse suggestions.

Finally, as we found that AI-generated guidance still lacks high-level rationales, effective auditing strategies should be injected into the training process of LLMs beyond relevant knowledge. They can benefit from learning expert practices, such as slightly varying the prompt with controlled variates at each time. This necessitates studies on effective auditing techniques as well as the development of tools capable of eliciting such strategies.

8.3 Integrating different types of guidance for effective human-AI collaboration

Our findings indicate that combining visual and LLM-driven guidance is more effective than either modality alone for reducing user workload and improving auditing performance (Section 6.2.3). These modalities play complementary roles and mitigate each other’s weaknesses: visual guidance conveys concrete, situational information (e.g., UI state, highlighted anomalies, and semantic relationships among auditing criteria) that helps both users and models ground their reasoning, while LLM-driven guidance synthesizes context, generates high-level hypotheses, and proposes next-step strategies. Specifically, the LLM can substantially reduce the additional cognitive load introduced by visual guidance by summarizing and prioritizing visual information; conversely, visual guidance helps the LLM and users by communicating system state and on-screen evidence in an immediately perceivable form. Together, they streamline iterative sense-making: visual cues attract attention and reduce search cost, while textual LLM explanations contextualize those cues, translate model inferences into actionable audit steps, and justify recommendations in familiar language.

Practically, this multimodal loop improves transparency and trust by making the AI’s reasoning traceable (users can link suggestions to visual cues), supports progressive disclosure of complexity (e.g., intricate criteria can be hierarchically represented and revealed through the scene graph), and accommodates diverse user preferences and expertise levels. To fully realize these benefits, system designs should tightly couple the modalities, such as enabling cross-modal interactions like clicking a visual highlight to obtain an LLM justification. Future work could draw from recent systems [24, 81] to design such interactions.

8.4 Incorporating intelligent auditing system into real world

Through designing, developing, and evaluating Vipera, our work provides initial empirical evidence and design implications for creating better auditing tools that can systematically surface and support sensemaking of potentially problematic AI-generated image outputs (see Section 6.1 and 6.2). However, as noted in our Limitations section (see 7), future work is much needed to evaluate Vipera in more

ecologically valid environments with real-world auditors in industry settings and beyond, aligning with many calls from prior work on studying responsible AI practices in real-world contexts [6, 33]. To better situate Vipera in current developer workflows in light of the calls from recent RAI tool development in HCI [16, 72], future deployments of Vipera could integrate into Jupyter notebooks or other in-house developer tools, rather than existing as a stand-alone system, thereby reducing adoption friction.

In addition, because AI auditing and other RAI practices should be collaborative efforts among cross-functional teams [18, 31, 51, 71], future researchers could extend Vipera and similar auditing interfaces to support communication across different roles. For instance, inspired by prior work in HCI and RAI [8, 78], the “Note view” feature (see 4.1.3) could be more explicitly designed to help auditors or model evaluators communicate auditing results to business analysts or leadership without technical backgrounds.

Finally, a large body of HCI and RAI research has shown that practitioners often encounter pushback from leadership when advocating for more responsible technologies [36, 43, 56]. Moreover, despite the growing number of RAI tools, organizational studies of industry practices highlight how the profit-driven and fast-paced nature of industry work frequently discourages meaningful RAI engagement [59, 74]. While our results demonstrate greater time efficiency for auditors (see Section 6.1), future work should explicitly examine how tools like Vipera would be socially situated within organizational contexts, and what roles computational tools for systematic GenAI auditing can—and should—play within broader RAI efforts in industry.

8.5 Future work

In addition, we have summarized several promising directions for future research. First, participants expressed a desire for more personalized and proactive guidance. Future versions could incorporate a user model that learns an auditor’s focus to anticipate their needs and surface more relevant suggestions. Second, our observation that users often did not correct errors made by the AI labeler points to a need for improving human-AI collaboration. Future interfaces could integrate confidence scores, label explanations, and more intuitive correction mechanisms to foster greater user engagement and trust. Finally, the system could be enhanced by expanding the granularity of auditing guidance and cross-modal interactions, such as linking anomalies in charts to the pixel-level areas within the raw image data to create a more fine-grained sensemaking loop.

9 Conclusion

We introduce Vipera, an innovative system designed to enhance the systematic auditing of these models by providing structured multifaceted analysis and LLM-powered suggestions. Vipera further leverages multiple visual aids, including scene graphs and stacked bar charts, to facilitate intuitive sensemaking of auditing results and assist auditors in streamlining and organizing their auditing. Our user study confirms Vipera’s usability and effectiveness, demonstrating its potential to help auditors navigate the complex auditing landscape while uncovering new insights. We believe Vipera will contribute significantly to the advancement of end-user auditing

and foster more responsible human-AI collaboration in creative applications.

Acknowledgments

The research was supported by National Science Foundation (NSF) program on Fairness in AI in collaboration with Amazon under Award No. IIS-2040942, as well as an award from Notre Dame–IBM Technology Ethics Lab. We would like to thank all the anonymous reviewers for their constructive comments.

References

- [1] Shehzad Afzal, Sohaib Ghani, Mohamad Mazen Hittawe, Sheikh Faisal Rashid, Omar M Knio, Markus Hadwiger, and Ibrahim Hoteit. 2023. Visualization and visual analytics approaches for image and video datasets: A survey. *ACM Transactions on Interactive Intelligent Systems* 13, 1 (2023), 1–41.
- [2] Shm Garanganoo Almeda, JD Zamfirescu-Pereira, Kyu Won Kim, Pradeep Mani Rathnam, and Bjoern Hartmann. 2024. Prompting for Discovery: Flexible Sense-Making for AI Art-Making with DreamSheets. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
- [3] Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, et al. 2024. AgentHarm: A benchmark for measuring harmfulness of llm agents. *arXiv preprint arXiv:2410.09024* (2024).
- [4] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L. Glassman. 2024. ChainForge: A Visual Toolkit for Prompt Engineering and LLM Hypothesis Testing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3613904.3642016
- [5] Joshua Asplund, Motahhare Eslami, Hari Sundaram, Christian Sandvig, and Karrie Karahalios. 2020. Auditing race and gender discrimination in online housing markets. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 24–35.
- [6] Agathe Balayn, Mireia Yurrita, Jie Yang, and Ujwal Gadiraju. 2023. “Fairness Toolkits, A Checkbox Culture?” On the Factors that Fragment Developer Practices in Handling Algorithmic Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 482–495.
- [7] Abeba Birhane, Ryan Steed, Victor Ojewale, Briana Vecchione, and Inioluwa Deborah Raji. 2024. AI auditing: The broken bus on the road to AI accountability. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 612–643.
- [8] Edyta Bogucka, Marios Constantinides, Sanja Šćepanović, and Daniele Quercia. 2024. Co-designing an AI impact assessment report template with AI practitioners and AI compliance experts. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. 168–180.
- [9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
- [10] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.
- [11] Ángel Alexander Cabrera, Abraham J Druck, Jason I Hong, and Adam Perer. 2021. Discovering and validating ai errors with crowdsourced failure reports. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–22.
- [12] Ángel Alexander Cabrera, Abraham J. Druck, Jason I. Hong, and Adam Perer. 2021. Discovering and Validating AI Errors With Crowdsourced Failure Reports. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021). doi:10.1145/3479569
- [13] Wang Claire, Wesley Hanwen Deng, Jason Hong, Ken Holstein, and Motahhare Eslami. 2024. Designing a Crowdsourcing Pipeline to Verify Reports from User AI Audits. *Work in Progress of the AAAI Conference on Human Computation and Crowdsourcing* (2024).
- [14] Dazhen Deng, Chuhan Zhang, Huawei Zheng, Yuwen Pu, Shouling Ji, and Yingcai Wu. 2025. AdversFlow: Visual Red Teaming for Large Language Models with Multi-Level Adversarial Flow. *IEEE Transactions on Visualization & Computer Graphics* 31, 01 (Jan. 2025), 492–502. doi:10.1109/TVCG.2024.3456150
- [15] Wesley Hanwen Deng, Boyuan Guo, Alicia Devrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. 2023. Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.
- [16] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring how machine learning practitioners (try to) use fairness toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 473–484.
- [17] Wesley Hanwen Deng, Claire Wang, Howard Ziyu Han, Jason I Hong, Kenneth Holstein, and Motahhare Eslami. 2025. WeAudit: Scaffolding User Auditors and AI Practitioners in Auditing Generative AI. *arXiv preprint arXiv:2501.01397* (2025).
- [18] Wesley Hanwen Deng, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio. 2023. Investigating practices and opportunities for cross-functional collaboration around AI fairness in industry practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 705–716.
- [19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.
- [20] Nathalie Diberardino, Clair Baleshta, and Luke Stark. 2024. Algorithmic Harms and Algorithmic Wrongs. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 1725–1732. doi:10.1145/3630106.3659001
- [21] Steven P Dow, Alana Glassco, Jonathan Kass, Melissa Schwarz, Daniel L Schwartz, and Scott R Klemmer. 2010. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Transactions on Computer-Human Interaction (TOCHI)* 17, 4 (2010), 1–24.
- [22] Klaus Eckelt, Kiran Gadhave, Alexander Lex, and Marc Streit. 2025. Loops: Leveraging Provenance and Visualization to Support Exploratory Data Analysis in Notebooks. *IEEE Transactions on Visualization and Computer Graphics* 31, 1 (2025), 1213–1223. doi:10.1109/TVCG.2024.3456186
- [23] Fei Fang, Miao Yi, Hui Feng, Shenghong Hu, and Chunxia Xiao. 2017. Narrative collage of image collections by scene graph recombination. *IEEE transactions on visualization and computer graphics* 24, 9 (2017), 2559–2572.
- [24] Lin Gao, Jing Lu, Zekai Shao, Ziyue Lin, Shengbin Yue, Chiokit Leong, Yi Sun, Rory James Zauner, Zhongyu Wei, and Siming Chen. 2025. Fine-Tuned Large Language Model for Visualization System: A Study on Self-Regulated Learning in Education. *IEEE Transactions on Visualization and Computer Graphics* 31, 1 (2025), 514–524. doi:10.1109/TVCG.2024.3456145
- [25] Katy Ilonka Gero, Chelse Swoopes, Ziwei Gu, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. Supporting Sensemaking of Large Language Model Outputs at Scale. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3613904.3642139
- [26] Yuhang Guo, Hanning Shao, Can Liu, Kai Xu, and Xiaoru Yuan. 2024. PromptTHis: Visualizing the Process and Influence of Prompt Editing during Text-to-Image Creation. *IEEE Transactions on Visualization and Computer Graphics* (2024).
- [27] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. 2014. Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of the 2014 Conference on Internet Measurement Conference*. 305–318.
- [28] Jeffrey Heer. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1844–1850.
- [29] Florian Heimerl, Steffen Lohmann, Simon Lange, and Thomas Ertl. 2014. Word cloud explorer: Text analytics based on word clouds. In *2014 47th Hawaii international conference on system sciences*. IEEE, 1833–1842.
- [30] Fred Hohman, Kanit Wongsuphasawat, Mary Beth Kery, and Kayur Patel. 2020. Understanding and Visualizing Data Iteration in Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI ’20). Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376177
- [31] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.
- [32] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li-Fei-Fei. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3668–3678.
- [33] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting interpretability: understanding data scientists’ use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.
- [34] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021. Dynabench: Rethinking benchmarking in NLP. *arXiv preprint arXiv:2104.14337* (2021).
- [35] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024. EvalLM: Interactive Evaluation of Large Language Model Prompts on User-Defined Criteria. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3613904.3642216
- [36] Apoorva Nalini Pradeep Kumar, Justus Bogner, Markus Funke, and Patricia Lago. 2024. Balancing Progress and Responsibility: A Synthesis of Sustainability Trade-Offs of AI-Based Systems. In *2024 IEEE 21st International Conference on Software*

- Architecture Companion (ICSA-C)*. IEEE, 207–214.
- [37] Michelle S. Lam, Mitchell L. Gordon, Danaë Metaxa, Jeffrey T. Hancock, James A. Landay, and Michael S. Bernstein. 2022. End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2 (Nov. 2022). doi:10.1145/3555625
- [38] Clayton Lewis and Robert Mack. 1982. Learning to use a text processing system: Evidence from “thinking aloud” protocols. In *Proceedings of the 1982 Conference on Human Factors in Computing Systems* (Gaithersburg, Maryland, USA) (*CHI '82*). Association for Computing Machinery, New York, NY, USA, 387–392. doi:10.1145/800049.801817
- [39] Yiran Li, Junpeng Wang, Prince Aboagye, Chin-Chia Michael Yeh, Yan Zheng, Liang Wang, Wei Zhang, and Kwan-Liu Ma. 2024. Visual Analytics for Efficient Image Exploration and User-Guided Image Captioning. *IEEE Transactions on Visualization and Computer Graphics* (2024).
- [40] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744 [cs.CV] <https://arxiv.org/abs/2310.03744>
- [41] David C. Logan. 2009. Known knowns, known unknowns, unknown unknowns and the propagation of scientific enquiry. *Journal of Experimental Botany* 60, 3 (03 2009), 712–714. doi:10.1093/jxb/erp043 arXiv:<https://academic.oup.com/jxb/article-pdf/60/3/712/1237855/erp043.pdf>
- [42] Kelly Avery Mack, Rida Qadri, Remi Denton, Shaun K Kane, and Cynthia L Bennett. 2024. “They only care to show us the wheelchair”: disability representation in text-to-image AI models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–23.
- [43] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.
- [44] Matheus Kunzler Maldaner, Wesley Hanwen Deng, Jason Hong, Ken Holstein, and Motahhare Eslami. 2024. MIRAGE: Multi-model Interface for Reviewing and Auditing Generative Text-to-Image AI. *Demo of the AAAI Conference on Human Computation and Crowdsourcing* (2024).
- [45] Danaë Metaxa, Joon Sung Park, Ronald E Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, Christian Sandvig, et al. 2021. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human-Computer Interaction* 14, 4 (2021), 272–344.
- [46] Ranjita Naik and Besmira Nushi. 2023. Social Biases through the Text-to-Image Generation Lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, New York, NY, USA, 786–808. doi:10.1145/3600211.3604711
- [47] Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism*. NYU Press.
- [48] Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. 2024. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. *arXiv preprint arXiv:2402.17861* (2024).
- [49] Qian Pan, Zahra Ashktorab, Michael Desmond, Martin Santillan Cooper, James Johnson, Rahul Nair, Elizabeth Daly, and Werner Geyer. 2024. Human-Centered Design Recommendations for LLM-as-a-judge. *arXiv preprint arXiv:2407.03479* (2024).
- [50] Xingjia Pan, Fan Tang, Weiming Dong, Chongyang Ma, Yiping Meng, Feiyue Huang, Tong-Yee Lee, and Changsheng Xu. 2019. Content-based visual summarization for image collections. *IEEE transactions on visualization and computer graphics* 27, 4 (2019), 2298–2312.
- [51] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency*. 39–48.
- [52] Rune Pettersson. 1993. *Visual information*. Educational Technology.
- [53] Marcelo OR Prates, Pedro H Avelar, and Luis C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications* 32, 10 (2020), 6363–6381.
- [54] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Mengdi Wang, and Prateek Mittal. 2023. Visual adversarial examples jailbreak large language models. *arXiv preprint arXiv:2306.13213* (2023).
- [55] Xin Qian, Ryan A. Rossi, Fan Du, Sungchul Kim, Eunye Koh, Sana Malik, Tak Yeon Lee, and Joel Chan. 2021. Learning to Recommend Visualizations from Data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Virtual Event, Singapore) (*KDD '21*). Association for Computing Machinery, New York, NY, USA, 1359–1369. doi:10.1145/3447548.3467224
- [56] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.
- [57] Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. 2023. Supporting Human-AI Collaboration in Auditing LLMs with LLMs. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, New York, NY, USA, 913–926. doi:10.1145/3600211.3604712
- [58] Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. 2023. Supporting Human-AI Collaboration in Auditing LLMs with LLMs. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, New York, NY, USA, 913–926. doi:10.1145/3600211.3604712
- [59] Mark Ryan, Eleni Christodoulou, Josephina Antoniou, and Kalypto Iordanou. 2024. An AI ethics ‘David and Goliath’: value conflicts between large tech companies and their employees. *AI & SOCIETY* 39, 2 (2024), 557–572.
- [60] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry* (2014).
- [61] Johanna Schmidt, M Eduard Gröller, and Stefan Bruckner. 2013. VAICo: Visual analysis for image comparison. *IEEE transactions on visualization and computer graphics* 19, 12 (2013), 2090–2099.
- [62] Shreya Shankar, J.D. Zamfirescu-Pereira, Bjoern Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3654777.3676450
- [63] Renee Shelby, Shalaleh Rismani, and Negar Rostamzadeh. 2024. Generative AI in Creative Practice: ML-Artist Folk Theories of T2I Use, Harm, and Harm-Reduction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
- [64] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021). doi:10.1145/3479577
- [65] Hong Shen, Leijie Wang, Wesley H Deng, Ciell Brusse, Ronald Velgersdijk, and Haiyi Zhu. 2022. The model card authoring toolkit: Toward community-centered, deliberation-driven AI design. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 440–451.
- [66] Leixian Shen, Haotian Li, Yifang Wang, Xing Xie, and Huamin Qu. 2025. Prompting generative AI with interaction-augmented instructions. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–9.
- [67] Hendrik Strobelt, Daniela Oelke, Christian Rohrdantz, Andreas Stoffel, Daniel A Keim, and Oliver Deussen. 2009. Document cards: A top trumps visualization for documents. *IEEE transactions on visualization and computer graphics* 15, 6 (2009), 1145–1152.
- [68] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3613904.3642400
- [69] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Queue* 11, 3 (2013), 10–29.
- [70] April Yi Wang, Will Epperson, Robert A DeLine, and Steven M Drucker. 2022. Diff in the loop: Supporting data comparison in exploratory data analysis. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–10.
- [71] Qiaosi Wang, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox. 2023. Designing responsible ai: Adaptations of ux practice to meet responsible ai challenges. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [72] Zijie J. Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael Madaio. 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–40.
- [73] Luoxuan Weng, Xingbo Wang, Junyu Lu, Yingchaojie Feng, Yihan Liu, Haozhe Feng, Danqing Huang, and Wei Chen. 2025. InsightLens: Augmenting LLM-Powered Data Analysis with Interactive Insight Management and Navigation. *IEEE Transactions on Visualization and Computer Graphics* (2025).
- [74] David Gray Widder, Laura Dabbish, James D Herbsleb, and Nikolas Martelaro. 2024. Power and Play: Investigating “License to Critique” in Teams’ AI Ethics Discussions. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–23.
- [75] Xiao Xie, Xiwen Cai, Junpei Zhou, Nan Cao, and Yingcai Wu. 2018. A semantic-based method for visualizing large image collections. *IEEE transactions on visualization and computer graphics* 25, 7 (2018), 2362–2377.
- [76] Fernando Yanez, Cristina Conati, Alvitia Ottley, and Carolina Nobre. 2025. The State of the Art in User-Adaptive Visualizations. In *Computer Graphics Forum*, Vol. 44. Wiley Online Library, e12571.
- [77] Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti Hearst. 2003. Faceted metadata for image search and browsing. In *Proceedings of the CHI conference on Human factors in computing systems*. 401–408.

- [78] Nur Yildirim, Mahima Pushkarna, Nitesh Goyal, Martin Wattenberg, and Fernanda Viégas. 2023. Investigating how practitioners use human-ai guidelines: A case study on the people+ ai guidebook. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [79] Jan Zahálka, Marcel Worring, and Jarke J Van Wijk. 2020. II-20: Intelligent and pragmatic analytic categorization of image collections. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 422–431.
- [80] Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023. Safetybench: Evaluating the safety of large language models with multiple choice questions. *arXiv preprint arXiv:2309.07045* (2023).
- [81] Yuheng Zhao, Xueli Shu, Liwen Fan, Lin Gao, Yu Zhang, and Siming Chen. 2025. ProactiveVA: Proactive Visual Analytics with LLM-Based UI Agent. *arXiv preprint arXiv:2507.18165* (2025).
- [82] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2024. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA.