

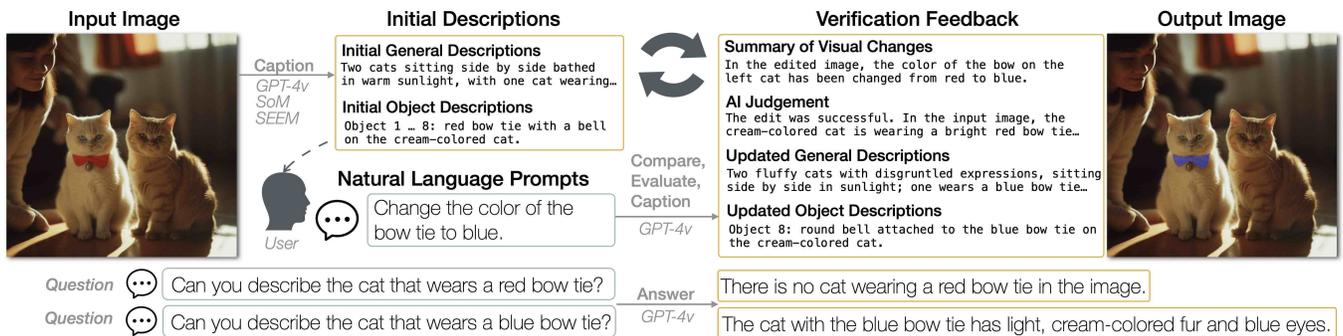
# EditScribe: Non-Visual Image Editing with Natural Language Verification Loops

Ruei-Che Chang  
rueiche@umich.edu  
University of Michigan  
Ann Arbor, MI, USA

Yuxuan Liu  
liurick@umich.edu  
University of Michigan  
Ann Arbor, MI, USA

Lotus Zhang  
hanziz@uw.edu  
University of Washington  
Seattle, Washington, USA

Anhong Guo  
anhong@umich.edu  
University of Michigan  
Ann Arbor, MI, USA



**Figure 1: EditScribe supports non-visual image editing using natural language verification loops. The user first comprehends the image content through initial general and object descriptions, then specifies edit actions using natural language. EditScribe performs the image edit, and provides four types of verification feedback for the user to verify the performed edit, including a summary of visual changes, AI judgement, and updated general and object descriptions. The user can ask follow-up questions to clarify and probe into the edits or verification feedback, before performing another edit.**

## ABSTRACT

Image editing is an iterative process that requires precise visual evaluation and manipulation for the output to match the editing intent. However, current image editing tools do not provide accessible interaction nor sufficient feedback for blind and low vision individuals to achieve this level of control. To address this, we developed *EditScribe*, a prototype system that makes object-level image editing actions accessible using *natural language verification loops* powered by large multimodal models. Using *EditScribe*, the user first comprehends the image content through initial general and object descriptions, then specifies edit actions using open-ended natural language prompts. *EditScribe* performs the image edit, and provides four types of verification feedback for the user to verify the performed edit, including a summary of visual changes, AI judgement, and updated general and object descriptions. The user can ask follow-up questions to clarify and probe into the edits or verification feedback, before performing another edit. In a study with ten blind or low-vision users, we found that *EditScribe* supported participants to perform and verify image edit actions non-visually.

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

ASSETS '24, October 27–30, 2024, St. John's, NL, Canada

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0677-6/24/10...\$15.00

<https://doi.org/10.1145/3663548.3675599>

We observed different prompting strategies from participants, and their perceptions on the various types of verification feedback. Finally, we discuss the implications of leveraging natural language verification loops to make visual authoring non-visually accessible.

## CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI); Accessibility technologies.**

## KEYWORDS

Accessibility, assistive technology, image editing, blind, low vision, generative AI, visual authoring, creativity support tools

## ACM Reference Format:

Ruei-Che Chang, Yuxuan Liu, Lotus Zhang, and Anhong Guo. 2024. *EditScribe: Non-Visual Image Editing with Natural Language Verification Loops*. In *The 26th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '24)*, October 27–30, 2024, St. John's, NL, Canada. ACM, New York, NY, USA, 19 pages. <https://doi.org/10.1145/3663548.3675599>

## 1 INTRODUCTION

Images are crucial visual media used to convey information and serve a variety of critical purposes in our everyday lives (e.g., work, social, entertainment). Blind and low-vision (BLV) individuals often need to create, edit, and share images with sighted peers, for social engagement, visual information access, and many other personal pursuits [18, 40, 63, 65, 85]. Crucial to producing a satisfactory image is reviewing and editing its visual details. For BLV individuals, the necessity for image editing arises in various contexts, such as

enhancing photo quality (e.g., clarity, lighting, and composition) in blind photography [22, 29] and obfuscating private content [19, 20, 30, 84, 86]. However, the BLV community commonly faces access barriers with the image editing process, as existing editing tools do not support comprehensive non-visual understanding of the image content nor accessible visual manipulation control for them to iteratively perform edits and evaluate the visual changes [18, 29, 32, 73, 84].

Recently, HCI and accessibility researchers have increasingly focused on developing accessible tools for the BLV community in visual content creation and editing, including enabling the perception of elemental changes on presentation slide editing platforms [44, 54, 87], making video editing accessible through alternative textual representations for sounds and visuals [38], modifying layouts with tactile supports [47], and allowing users to verify content created by generative AI models with preset verification prompts [37]. Despite these efforts, the domain of image editing remains under-explored, which is visually challenging as it requires users to precisely understand image content, perform desired edit effects, and evaluate changes [20, 84]. Thus in this work, we aim to address the following questions:

*How to enable BLV people to perform image editing non-visually?  
And how to support the evaluation of visual changes after edits?*

To achieve this, we propose *EditScribe*, a prototype system that utilizes large multimodal models (LMMs) to make object-level image editing actions non-visually accessible to BLV people. The core of *EditScribe* is *natural language verification loops*. Using *EditScribe*, the user first comprehends the image content through initial general and object descriptions, then specifies edit actions using open-ended natural language prompts. *EditScribe* performs the image edit, and provides four types of verification feedback for the user to verify the performed edit, including SUMMARY OF VISUAL CHANGES, AI JUDGEMENT, and updated GENERAL and OBJECT DESCRIPTIONS. The user can ask follow-up questions to clarify and probe into the edits or verification feedback before performing another edit. This set of verification feedback was grounded on both visual and textual sources, such as comparing the images before and after an edit to summarize visual changes, using the edited image to provide new and independent image and object descriptions, and combining the above for AI to judge and reason about the success of the edit. We demonstrated the utility of natural language verification loops with five object-level edit actions in *EditScribe*, including blurring an object, removing an object, changing the color of an object, adjusting the brightness of the object, and adding text to the image. We focus on object-level actions as they require precise understanding and manipulation of image details, and are critical to tasks commonly desired by BLV people [20, 22, 73], e.g., image obfuscation for privacy [20, 84] and background removal/blurring to focus on specific objects [22, 36].

We evaluated *EditScribe* with 10 BLV participants to understand *RQ1: How does EditScribe support non-visual image editing? RQ2: How do BLV people prompt EditScribe? RQ3: How do BLV people perceive EditScribe's verification feedback? and RQ4: How do BLV people perceive the final edited images?* We found that BLV participants were able to perform most of the editing tasks using *EditScribe*, and they had different strategies for prompting, such as creating

prompts ranging from detailed to succinct to adapt to discrepant verification feedback, using varying tones or wordings to facilitate the system's understanding, or asking follow-up questions to gauge the reliability and build their trust in *EditScribe*. Also, participants had different preferences for each verification feedback, depending on the contexts of use. Other factors, such as inconsistent or discrepant information, visual experiences and expectations, and tones of verification feedback, also affected their trust and confidence in *EditScribe*. Furthermore, participants felt confident about the edited images by *EditScribe* and were willing to publish based on the context, but preferred a second check on the final edited images using sighted assistance or other AI services. Finally, we discussed how *EditScribe* could be extended to support additional edit actions, provide enhanced verification feedback, and implications for future systems to leverage natural language verification loops for content creation.

## 2 RELATED WORK

Our research is informed by prior work in visual content authoring accessibility, image accessibility, and AI-assisted image editing.

### 2.1 Visual Content Authoring Accessibility

Research around BLV individuals' digital creative activities has recently gained momentum in the field of accessibility. This body of research noted substantial interest from the community in a wide range of digital content creation [18, 40, 64, 65, 85]. Notably, BLV individuals commonly feel the need to engage in visual content authoring, including but not limited to photography, videos, presentation slides, data visualization, websites, and visual-heavy documents [18, 40, 65, 85]. These visual creative activities bring BLV individuals opportunities for information access, socializing, employment, self-expression, entertainment and more [85]. For example, BLV people often share photos with sighted peers or remote agents (e.g., Aira [6], Be My Eyes [7]) for visual interpretation assistance [63], while many also desire access to authoring visual content to better engage with popular social media [22, 61, 85].

However, in authoring visual content, BLV individuals face immense access barriers [18, 22, 29, 32, 52]. Past surveys revealed that digital creative activities that involve visual elements are generally considered more challenging to BLV creators [85]. While basic photography and document formatting tasks have been increasingly supported by technology [11, 12, 17], *editing* visual elements remains particularly challenging [52, 85]—most often, BLV individuals shy away from visual editing tasks or rely on sighted support.

Specific visual editing challenges to BLV individuals were also noted, including limited understanding of visual editing standards, insufficient information about the visuals, difficulties perceiving the effect of edits, and inaccessible editing controls [47, 56, 57, 65, 85]. Still, additional challenges arise across different editing contexts. For example, when editing out private or sensitive information, BLV individuals experience heightened uncertainty about how obscured the result is [20, 84]. In authoring content in collaboration, there are challenges with reviewing others' visual changes [54]. For editing visuals in motion—videos, blind creators also have difficulties skimming through visual information [39].

Given how BLV individuals' needs for visual editing support could vary greatly based on their personal experience and authoring contexts, we consider natural language interaction a potentially more accessible image editing method, as it could allow on-demand visual manipulation and iterative information-seeking. So far, natural language interaction has not been utilized for accessible image editing, and we know little about how to design such interaction. To address this research gap, we developed and evaluated *EditScribe*, a system that utilizes LMMs for non-visual, text-to-image editing make image editing. We further explored how BLV users make use of natural language interaction in different contexts of image editing as well as their design feedback.

## 2.2 BLV Individuals' Access to Images

Key to successful image editing is the access to relevant visual information in the image, which most BLV screen reader users perceive through image descriptions [2, 68, 74]. A high-quality image description should provide a comparable experience to encountering the image visually [68] (e.g., providing the purpose of the image, object or people present, etc. [1, 3, 55, 74]). Also, it is critical to create image descriptions that adapt to individuals' information-seeking goals [1, 3, 76] but not "one size fits all" [68]. For example, for visual elements on social media images, information related to specific persons, location, photo quality and others' responses are all important to describe [50, 73, 76, 88], while for data visualization, granular and objective descriptions (e.g., shape, context) could support users' interpretation freedom [27]. To increase the availability of alternative texts online, AI is considered as a scalable solution [23, 49, 76, 88]. Automated image descriptions and object detection tools have become increasingly available (e.g., [12, 15]). However, as state-of-the-art AI models still produce inaccurate results, these systems need to provide enough information for users to notice potential undesired outcomes, such as by clearly communicating what the system is able to do and how well it could perform the intended task [21, 49, 76, 88].

In addition to the above guidelines, providing image description in the context of visual content *sharing and editing* involves additional considerations. In authoring visual content, BLV individuals tend to desire more information about their images and description accuracy [41, 65, 88]. In particular, spatial layout, appearances of objects, and timely feedback to visual changes critically influence BLV creators' judgement of necessary edits [37, 54, 65, 84]. Image descriptions also need to include information specific to the editing needs in different content authoring contexts (as mentioned in Section 2.1), e.g., focused descriptions on potentially private objects for managing the visual privacy [20, 84], visual changes for collaborative authoring [54].

Image descriptions for content authoring purposes thus tend to involve an abundance of visual details, a key challenge that needs to be designed around [47, 65]. Still, investigation on image descriptions in the context of visual editing is limited, especially around supporting BLV creators' understanding of visual changes on images [84]. In this paper, we draw on prior work and design a set of verification feedback on an image edit. We use it as a probe to gain more insights into how BLV users perceived it to inform potential improvements.

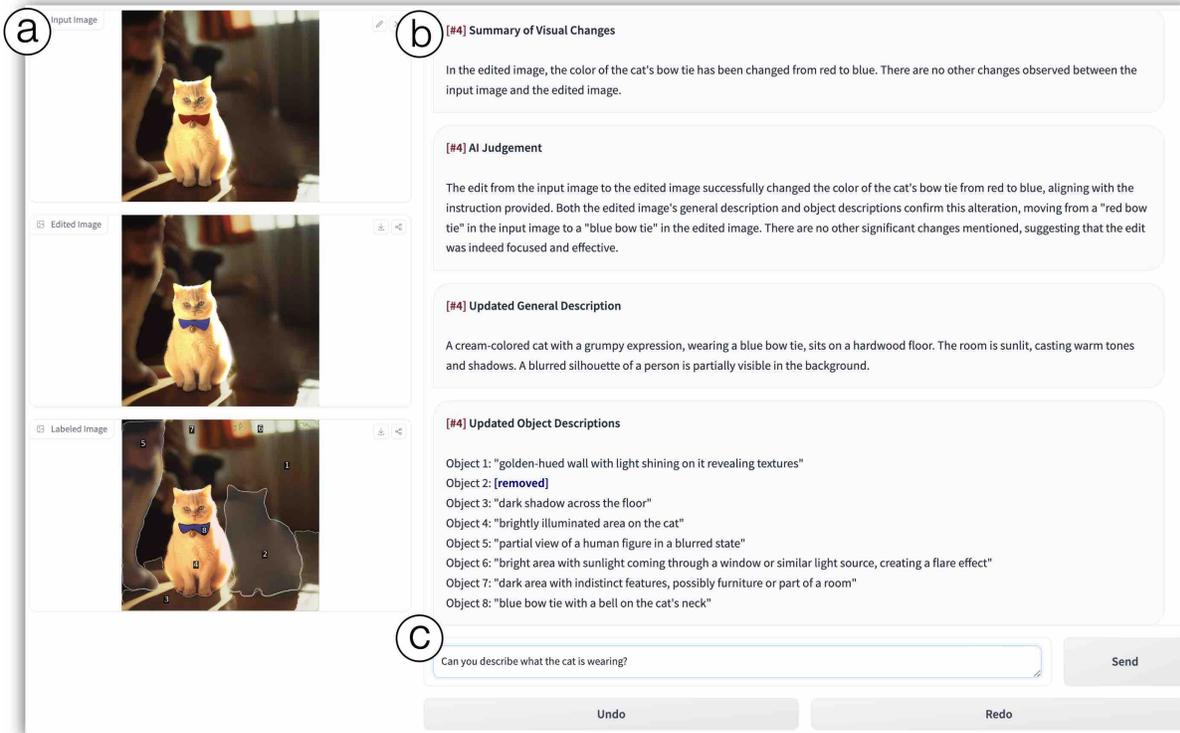
## 2.3 AI-assisted Image Editing

Recent advances on computer vision, large language and multi-modal models [8, 31, 42, 58, 70, 75, 81] have introduced vast opportunity for easing image content authoring. Image editing automation now can perform not only basic photo adjustment, e.g., color adjustment [4, 5], but also advanced feature modifications and creative generation [16, 28, 53, 66]. For example, development in object detection and segmentation algorithms [42, 75] now supports precise image editing, such as foreground extraction, object removal, and inpainting [62, 80]. Generative models [31, 48, 60] opened up possibilities for realistic generation of new images as well as manipulation of image attributes through the latent space [78]. The emergence of text-to-image diffusion models, such as DALL-E 2 [59], Stable Diffusion [60], Prompt-to-Prompt [35], and InstructPix2Pix [24] supports not only text-to-image generation but also text-instructed edits on the image [53, 66, 71].

So far, most AI-based image editing tools target at and have been tested with sighted image editors, primarily for efficiency- and inspiration-aid purposes. Only a limited number of accessibility research has explored how these tools may or may not support BLV individuals, a group whose content creation needs could critically benefit from this type of technology [20, 85]. Huh et al. prototyped a screen reader accessible text-to-image generation tool and evaluated with BLV creators [37]. Their study revealed design insights for accessible text-to-image *generation* (e.g., support for prompt generation and result image understanding). In terms of accessible image *editing* support, research studies have focused primarily on how object detection, image caption, as well as segmentation models may be applied to support the BLV community' visual privacy preservation—specifically, the detection and obfuscation of private and sensitive information in their photos [20, 85]. So far, this research revealed promising benefits these methods could bring to BLV individuals (e.g., independence, sense of control, efficiency), but also challenges (e.g., algorithm inaccuracies, difficulties with envisioning and evaluating obfuscation results, heavy cognitive load). Natural language interaction has the potential to mitigate these challenges by facilitating mutual understanding between the BLV users and the AI models. However, little insight exists around how BLV individuals would leverage this method to elicit feedback from AI models and how AI models could elaborate visual effects in the context of image editing. To understand this loop of communication, we build a prototype system, *EditScribe*, to explore how BLV individuals strategize prompts for the system to understand their editing needs, and how they perceive a set of verification feedback regarding the effects of their edits.

## 3 EDITS CRIBE

*EditScribe* leverages natural language verification loops as its core to support BLV people to understand and edit images non-visually. In this section, we first present how BLV users could use *EditScribe* to edit images in an example scenario, then present the inspiration to design natural language verification loops from prior work, and detail its implementation.



**Figure 2: EditScribe user interface. (a) The images before and after the most recent edit, and an image labeled with masks and indexes for debugging purposes. (b) The accessible chat tagged with different heading levels helps users navigate a history of input prompts and verification feedback. (c) Users can input natural language prompts to perform edits or ask follow-up questions, and undo or redo edits.**

### 3.1 Scenario Walkthrough

Here, we illustrate EditScribe in an example scenario, taking Amanda as the main character, a person who is blind.

Amanda just lost her short-haired British white cat, Elsa, and urgently wants to make a public post asking people to help find it. She navigates her album on the phone and found the latest photo of herself with her cats, Elsa and Rosa. She uploads this photo to EditScribe and obtains a general description and a list of detailed object descriptions (Figure 3a,b) to confirm that the image content matches what she remembered. With these descriptions, she has developed a plan in mind to make a post with this photo.

First, Amanda decides to remove Rosa, her other orange tabby cat (Figure 3c) next to Elsa, to avoid confusion. She prompts “remove the orange cat” and received EditScribe feedback after the edit is completed, including SUMMARY OF VISUAL CHANGES, AI JUDGEMENT, UPDATED GENERAL DESCRIPTIONS, and UPDATED OBJECT DESCRIPTIONS. Amanda confirms Rosa has been removed by cross-checking EditScribe’s feedback, e.g., SUMMARY OF VISUAL CHANGES confirms the edit and AI JUDGEMENT provides detailed reasons, only one cat is described in the UPDATED GENERAL DESCRIPTIONS and the removal of Rosa’s description in UPDATED OBJECT DESCRIPTIONS (Figure 3c). She further confirms with a question “How many cats are in the image?” with answer “One.” Amanda then blurs herself to maintain her privacy by prompting “Blur out the woman in the image”, and also confirms the success of the edit through the consistent feedback from EditScribe (Figure 3d).

To help people better distinguish Elsa, she wants to highlight it and make the photo similar to when Elsa was lost. First, Amanda aims to increase the focus of Elsa and specifies “Make the cat brighter to increase its focus.” The feedback from EditScribe suggests “the intensity of the sunlight on the cat has been increased, giving the cat a more illuminated appearance” (Figure 3e). She wants to ensure the color of Elsa is still maintained and asks “What is the color of the cat?” with the answer “Cream or white”, which aligns with how her friends describe Elsa before. Next, she changes the bow tie’s color from red to blue, which is what Elsa wore when lost “Change the color of the bow tie to blue.” She confirms with EditScribe’s feedback, especially the UPDATED OBJECT DESCRIPTION indicates “blue bow tie with a bell on the cat’s neck” (Figure 3f).

Finally, Amanda wants to add her phone number to the post for others’ reference. She first asks EditScribe “Dividing the image to nine squares, where is a good spot to add text that does not overlap the cat?” with EditScribe answering “Center Right.” She then prompts EditScribe “add text “Please call 12345 if you find Elsa” in the center-right of the image” (Figure 3g). Despite the EditScribe feedback confirming the addition of text, she performs another check “Does the text overlap with the cat?” with answer “No”, and another check “Does the cat with blue bow tie stand out in the image?” with answer “Yes.” Finally, she confidently posts it on her social media and waits for responses.

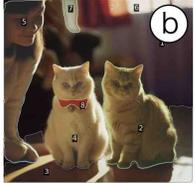
	<p><b>Initial General Descriptions</b></p> <p>Two cats sitting side by side bathed in warm sunlight, with one cat wearing a red bow tie. A woman is seen in the background, smiling down at them. The setting appears to be a cozy indoor space.</p>		<p><b>Initial Object Descriptions</b></p> <p>Object 1: golden-hued wall with light shining on it revealing textures          Object 2: orange tabby cat with light fur, sitting and facing forward          Object 3: dark shadow across the floor          Object 4: pale cream-colored cat with a stern expression, wearing a bell and a bright red bow tie          Object 5: woman with dark hair leaning forward and smiling gently towards the cats          Object 6: bright area with sunlight coming through a window or similar light source, creating a flare effect          Object 7: dark area with indistinct features, possibly furniture or part of a room          Object 8: red bow tie with a bell on the cream-colored cat</p>
<p><b>Edit prompt</b></p> <p>Remove the orange cat.</p> 	<p><b>Summary of Visual Changes</b></p> <p>In the edited image, the orange cat that was initially to the right of the white cat with a red bow tie is no longer present. The area where the orange cat was has been replaced with a blurred background that mimics the surrounding floor and wall colors. The lighting and shadows in the edited image have been adjusted accordingly to account for the absence of the second cat.</p>	<p><b>AI Judgement</b></p> <p>The edit was successful; the orange cat that was present in the input image has been removed in the edited image. The edited image object descriptions no longer list the "orange tabby cat with light fur," which corresponds with the object description #2 from the input image, indicating that the edit accurately followed the instruction given. The general description of the edited image now only mentions a single white cat with a red bow tie, whereas the input image described two cats, confirming the removal of the second cat.</p>	<p><b>Updated General Descriptions</b></p> <p>A white cat with a red bow tie and a bell sits on a wooden floor, illuminated by warm sunlight, with a blurred figure of a smiling woman leaning towards it in the background.</p> <p><b>Updated Object Descriptions</b></p> <p>: : : Object 2: [removed] : : :</p>
<p><b>Edit prompt</b></p> <p>Blur out the woman in the image.</p> 	<p><b>Summary of Visual Changes</b></p> <p>In the edited image, the area where the woman appears has been blurred, reducing the level of detail and making her features indistinct compared to the input image. The rest of the image, including the cat, remains mostly unchanged.</p>	<p><b>AI Judgement</b></p> <p>The edit aligns with the given instruction to blur out the woman in the image. In the edited image, the previously clear figure of a smiling woman described as "#5: 'woman with dark hair leaning forward and smiling gently towards the cats'" in the input image object descriptions, has been altered to "#5: 'partial view of a human figure in a blurred state'" in the edited image object descriptions. The general descriptions also reflect this change, with the edited image describing the figure as "partially revealing a blurred figure of a person," which is consistent with the task of blurring the woman in the image.</p>	<p><b>Updated General Descriptions</b></p> <p>A grumpy-looking light-colored cat with a red bow tie and bell sits on a hardwood floor, with sunlight casting a warm glow and shadows in the background, partially revealing a blurred figure of a person.</p> <p><b>Updated Object Descriptions</b></p> <p>: : : Object 5: partial view of a human figure in a blurred state : : :</p>
<p><b>Edit prompt</b></p> <p>Make the cat brighter to increase its focus.</p> 	<p><b>Summary of Visual Changes</b></p> <p>In the edited image, the brightness and contrast appear to have been increased, which makes the cat stand out more against the background and gives it a more highlighted appearance. The shadows on the cat and surrounding areas are reduced, resulting in less visible detail in the brighter areas. The overall color tone also seems warmer due to the brightness adjustment.</p>	<p><b>AI Judgement</b></p> <p>The edited image shows an increased overall brightness with a notable emphasis on the cat, specifically a "brightly illuminated area on the cat" as mentioned in the object descriptions. This change aligns with the instruction to make the cat brighter to increase its focus. Comparing the general descriptions also supports this assessment, with the edited image being described as having a "warm, glowing atmosphere," suggesting a successful edit in accordance with the provided instruction.</p>	<p><b>Updated General Descriptions</b></p> <p>A cream-colored cat with a grumpy expression, wearing a red bow tie, sits on a sunlit wooden floor; soft focus and lens flare create a warm, glowing atmosphere. There's a blurred figure in the background.</p> <p><b>Updated Object Descriptions</b></p> <p>: : : Object 4: brightly illuminated area on the cat : : :</p>
<p><b>Edit prompt</b></p> <p>Change the color of the bow tie to blue.</p> 	<p><b>Summary of Visual Changes</b></p> <p>In the edited image, the color of the cat's bowtie has been changed from red to blue. The rest of the image appears unchanged when compared to the input image.</p>	<p><b>AI Judgement</b></p> <p>Upon reviewing the input and edited images, it is evident that the color of the bow tie worn by the cat has been changed from red to blue, which is consistent with the editing instruction provided. The overall warm atmosphere and soft focus of the scene remain unchanged between the two images, with the bow tie's color being the only notable difference. The object descriptions also affirm this alteration, with object #8 transitioning from a "red bow tie with a bell" to a "blue bow tie with a bell," indicating that the edit was successful according to the given instruction.</p>	<p><b>Updated General Descriptions</b></p> <p>A cream-colored cat with a grumpy expression wearing a blue bow tie is sitting in a room lit with warm sunlight; the background is softly focused with silhouettes of furniture and partial views of a person.</p> <p><b>Updated Object Descriptions</b></p> <p>: : : Object 8: blue bow tie with a bell on the cat's neck : : :</p>
<p><b>Edit prompt</b></p> <p>Add text "Please call 12345 if you find Elsa" in the center-right of the image.</p> 	<p><b>Summary of Visual Changes</b></p> <p>In the edited image, the text "Please call 12345 if you find Elsa" has been added to the center-right of the scene, overlapping the blurry figure in the background. No other changes are observed between the input image and the edited image.</p>	<p><b>AI Judgement</b></p> <p>The edit appears to be successful as the text "Please call 12345 if you find Elsa" has been added to the center-right of the image, adhering to the instruction provided. The general description of the edited image confirms the presence of the text overlay, which coincides with the main difference from the input image. Object descriptions for both images remained the same aside from the added text, indicating that the only change executed was the inclusion of the specified text in the desired location.</p>	<p><b>Updated General Descriptions</b></p> <p>An orange cat with a blue bow tie sits on a wooden floor, bathed in sunlight, looking directly at the camera. In the background, a blurred figure and room details suggest a domestic setting. Text overlay reads, "Please call 12345 if you find Elsa."</p> <p><b>Updated Object Descriptions</b></p> <p>: : : : : :</p>

Figure 3: Detailed edit prompts and corresponding verification feedback for the Walkthrough and Session 3 in our study. Note that we only show the updated object descriptions for edited objects due to space constraint. The user can access all object descriptions if needed on the EditScribe interface, as demonstrated in Figure 2. In (g), the object descriptions remained the same as before since the user did not specify an object to edit.

### 3.2 Natural Language Verification Loops

When sighted people perform image editing, they utilize visual feedback to verify their actions and make necessary adjustments. However, such visual verification loops are inaccessible to BLV people. Therefore, at its core, EditScribe is enabled by the idea of *natural language verification loops* to support non-visual image editing. Here, we introduce its high-level concept.

LMMs (e.g., GPT-4v [9]) are leveraged in EditScribe as a proxy to communicate between the user and the image. EditScribe interprets the user’s natural language prompt into the action of certain visual edits, and then converts the resulting visual changes back into textual feedback for the user to review and confirm. This loop of cross-modal communication can be complex and requires extensive textual exchanges. First, from applying textual prompts to visual changes, EditScribe interprets the object or region of the user’s interest and desired editing effect by grounding on the user’s prompt and the image content. Second, from interpreting visual changes to textual feedback, EditScribe grounds on cross-modal sources to provide a set of verification feedback, which we describe their rationale:

- (1) **SUMMARY OF VISUAL CHANGES** is designed to resemble and simulate a sighted person editing an image, who can visually discern the main differences between the images before and after edits.
- (2) **AI JUDGEMENT** draws inspiration from a human judge, who, from a second-person perspective, assesses both visual and textual modifications to evaluate and rationalize the success of the edit.
- (3) **UPDATED GENERAL DESCRIPTIONS** is designed to offer an independent perspective on how a sighted person would perceive the new image after an edit [1, 3, 76].
- (4) **UPDATED OBJECT DESCRIPTIONS** is designed as if a sighted person is taking detailed visual inspections of each object to discern the nuances.

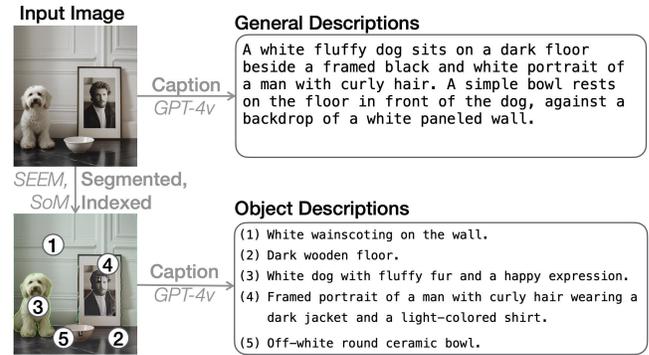
The updated general and object descriptions are also inspired by previous research on presenting image descriptions with varying levels of granularity to minimize cognitive load and enhance efficiency [38, 45, 68, 72]. The provision of four types of feedback, each grounded in different sources, aims to help mitigate the potential AI errors (e.g., misinterpreting user intent or generating hallucinations.) by enabling users to compare the feedback and judge their accuracy. Below, we detail the cross-modal grounding pipeline for understanding user prompts and generating verification feedback.

### 3.3 Cross-modal Grounding Pipeline for Understanding User Prompts and Generating Verification Feedback

To facilitate image content understanding, EditScribe first generates initial general and object descriptions for the user. Then, the user can perform edits with natural language prompts, and get verification feedback after each edit. The user can ask follow-up questions to clarify the verification feedback and verify the edits. They can repeat this process until the edited image satisfies their goal.

**Generating General and Object Descriptions.** Given an input image, EditScribe generates general and object descriptions to

support the initial understanding of image content. Specifically, EditScribe provides general descriptions for the image by prompting GPT-4v [9]. Next, EditScribe generates object descriptions by Set-of-Mark Prompting [79] with input image overlaid with a visible bounding mask (by SEEM [89]) and index on each object:

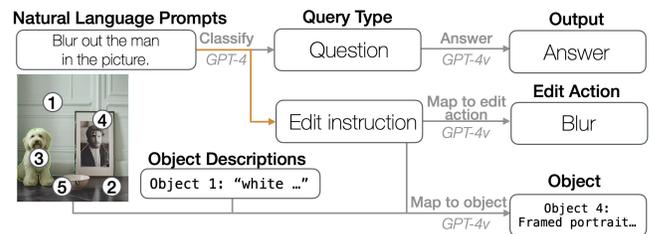


**Figure 4: Pipeline to generate initial general and object descriptions.** EditScribe prompts GPT-4v [9] with input image to generate general descriptions, and uses visual bounding masks (by SEEM [89]) and object indexes to perform Set-of-Mark Prompting to generate object descriptions.

#### Mapping User Prompt to Edit Action or Followup Question.

When receiving a user prompt, EditScribe classifies it as either a *question* or an *edit instruction* that corresponds to one of the EditScribe edit functions by prompting GPT-4. If the prompt is classified as a *question*, EditScribe prompts the user’s prompt to GPT-4v [9] to answer the question. On the other hand, if the prompt is classified as an *edit instruction*, EditScribe extracts the intended edit action and the object of interest from the prompt, for which the user can specify either the object’s name or index for flexible referencing. Having all this information, EditScribe performs the edit to the referenced object (See Section 3.4 for the supported edit actions).

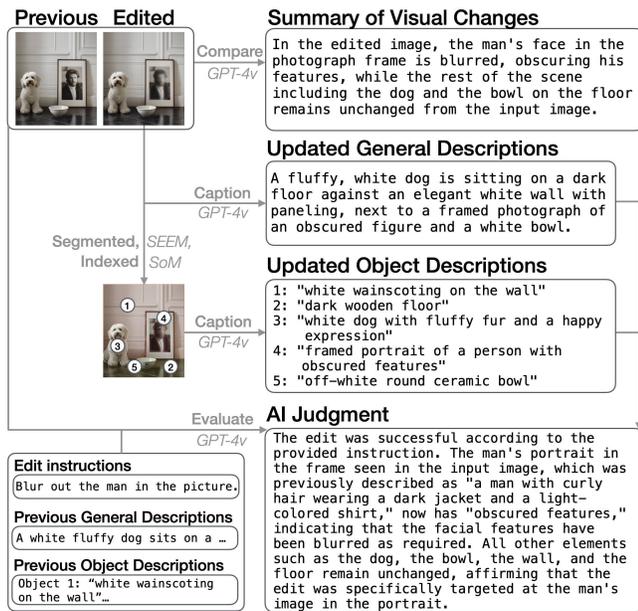
**Verification Feedback.** After an edit is completed, EditScribe generates four types of verification feedback for the user to verify the outcome of the edit, including SUMMARY OF VISUAL CHANGES, AI



**Figure 5: Pipeline to classify user prompts to actionable items.** If the prompt is classified as a *question*, EditScribe prompts the prompt to GPT-4v [9] to answer the question. If the prompt is classified as an *edit instruction*, EditScribe extracts the intended edit action and the object of interest from the prompt.

JUDGEMENT, UPDATED GENERAL DESCRIPTIONS, and UPDATED OBJECT DESCRIPTIONS. These verification information are structured and presented to the user from high to low level as follows:

- (1) **SUMMARY OF VISUAL CHANGES** illustrates an overview of visual changes by comparing the images before and after the edit. It grounds only new and previous images for the prompt.
- (2) **AI JUDGEMENT** explicitly points out if the AI determines the edit is successful or not by grounding and comparing the images before and after the edit, new and previous general and object descriptions, and the edit actions altogether. It also details its reasoning process and identified evidence.
- (3) **UPDATED GENERAL DESCRIPTIONS** are new and independent general descriptions on only the edited image using the same prompt for generating the initial descriptions. It only takes the new image as input for the prompt.
- (4) **UPDATED OBJECT DESCRIPTIONS** are object descriptions on the edited image, using SoM prompting [79]. It takes the new image with labeled masks and indexes for the prompt.



**Figure 6: Pipeline to generate the four types of verification feedback. Summary of Visual changes compares the previous and edited images, while the updated general and object descriptions takes only edited images as input for prompting. AI Judgment takes both previous and edited images as input, as well as the texts, such as user prompts and previous and current general and object descriptions.**

### 3.4 Image Edit Actions

EditScribe supports five editing functions: (i) *blur an object*, (ii) *remove an object*, (iii) *change color of an object*, (iv) *adjust brightness of an object*, and (v) *add text to the image*. We focus on object-level actions as they are critical to tasks commonly desired by BLV people [20, 22, 73], e.g., image obfuscation for privacy [20, 84] and

background removal/blurring to focus on specific objects [22, 36]. These actions require precise understanding and manipulation of image details, and we selected a subset of them that are representative and could inform other tasks. For instance, object removal could inform replacing with new objects, or cropping an image. Blurring and changing color and brightness could inform other actions involving pixel modifications. Inserting texts could inform adding other visual elements such as icons or arrows. We can further augment editing functions in EditScribe by prompting LLMs to classify users' prompts to corresponding edit functions. Below, we exemplify how users could specify prompts for each edit action using natural language, but not limited to these phrasing:

- (1) *Blur an object*: Blurring is a common technique to preserve personal image content privacy [20, 84]. Users can specify prompts relevant to blurring, such as "blur the person out", or "make #2 vague." Based on the prompt, EditScribe will apply the blurry effect on the specified object, using OpenCV's built-in function 'GaussianBlur' [14]. Users can perform this edit action repeatedly on an object until its level of blurriness meets their expectations.
- (2) *Remove an object*: Removing unwanted or personal image content is also common in image editing tasks. With this action, the user can remove an object while preserving the background. Users can specify prompts such as "remove the right person", or "remove the orange cat." EditScribe will then remove the specified object using LaMa [69].
- (3) *Change color of an object*: Changing color is common in basic image editing apps for improving aesthetics (e.g., [4]). It is also a privacy-preserving technique for users to black out or fill an object with a color significantly different from its original appearance [20, 84]. Users can specify prompts, such as "change the cat's collar to blue." EditScribe will then change the color of the specified object to the specified one. This is achieved by modifying the 'Hue' value of the object in the HSV color space (e.g., Hue, Saturation, Value), which represents the color type and is expressed as a degree on the color wheel, ranging from 0 to 360.
- (4) *Adjust the brightness of an object*: Adjusting the brightness is another common editing function. Users can input prompts such as "increase the brightness of the #6 person", or "make the left cat brighter." EditScribe will then adjust the brightness by increasing or decreasing each RGB channel of a pixel with the same increments. Users can gradually adjust an object's brightness until its brightness level meets their expectations.
- (5) *Add text to the image*: Adding text is common for adding an explanation to an image or making any purpose explicit. In our current implementation, users can assign text to one of the nine locations on an image (e.g., center, bottom right) or onto a specific object. Users can specify "add words 'Hello world' on upper third", "place sentence 'Hello world' to center", or "generate text 'Hello' to the top left corner." EditScribe will then place the text to the specified location using OpenCV's built-in function 'addText' [14] with pre-defined colors based on the contrast to the image.

### 3.5 EditScribe Web Interface

EditScribe was implemented using Gradio [10], an open-source Python library for fast prototyping front-end web interfaces. The left side of the interface (Figure 2a) shows the images before and after the most recent edit, and an image labeled with masks and indexes for debugging purposes. On the right side (Figure 2b), there is an accessible chat with verification feedback indexed with different heading levels for BLV users to navigate using a screen reader. For instance, verification feedback of each edit starts with “Verification Output of Edit #4 starts from here”, which is labeled as heading level 1 to help BVI users navigate the edits. On the other hand, each type of feedback title, such as “[#4] Summary of Visual Changes”, is labeled as heading level 2, which allows users to navigate each type of feedback. On the bottom right of the interface (Figure 2c), the user can enter their prompts and questions using natural language, or “undo” or “redo” their edits.

## 4 USER STUDY

We conducted a user study with BLV individuals to understand how natural language verification loops, including user prompts and system verification feedback, may (or may not) support their image editing needs. Specifically, we focus on the following research questions:

- RQ1: How does EditScribe support non-visual image editing?
- RQ2: How do BLV people prompt EditScribe?
- RQ3: How do BLV people perceive EditScribe’s verification feedback?
- RQ4: How do BLV people perceive the final edited images?

### 4.1 Participants

We recruited ten BLV participants (8 male and 2 female) using mailing lists for local accessibility organizations and prior contacts. Participants aged from 26 to 43 (avg. 35.4) and described their visual impairment as total blindness (N=8), having only light perception (N=1) or low vision (N=1). All participants regularly used a screen reader as the primary means to accessing their devices. Some participants had prior experiences in taking photos or editing images, such as having attempted to crop images, adding text or graphics, creating images with Generative AI, or consuming images through tactile displays (Table 1).

### 4.2 Study Procedure and Sessions

The study was conducted remotely over Zoom, and participants were provided with a link to our EditScribe site and used a screen reader to access. The study was approved by our Institutional Review Board (IRB). With participants’ consent, the study was video recorded and took about two hours, and each participant was compensated \$50 for their participation. Participants were asked to optionally provide one of their own images to edit in the study. The study had five sessions, including a tutorial session to familiarize participants with EditScribe, three sessions to edit our provided images, and a session to edit their provided images (Figure 7). We developed several tasks in each session, which we describe next.

**4.2.1 Session 1: Tutorial.** In this session, we aimed to familiarize our participants with EditScribe. Participants were guided through

	<p><b>Session 1: Tutorial</b></p> <p>Tasks:</p> <ol style="list-style-type: none"> <li>1. Change the color of vase to red.</li> </ol> <p>Or, other tasks participants came up with.</p>
	<p><b>Session 2: Performing individual editing actions.</b></p> <p>Tasks:</p> <ol style="list-style-type: none"> <li>1. Change the color of the wall to blue.</li> <li>2. Remove the bowl.</li> <li>3. Blur out the man in the picture.</li> <li>4. Increase the brightness of the dog.</li> <li>5. Add text 'Puppy' to the top right of the image.</li> </ol>
	<p><b>Session 3: Making a flyer to find a missing cat.</b></p> <p>Tasks:</p> <ol style="list-style-type: none"> <li>1. Remove Rosa, the orange cat.</li> <li>2. Blur out Lily, the woman in the image.</li> <li>3. Increase the brightness of Elsa, the white cat.</li> <li>4. Change the color of the red bow tie to blue.</li> <li>5. Add text "Please call 12345 if you find Elsa."</li> </ol>
	<p><b>Session 4: Making a flyer for recruiting a craftsman to make a bathroom shelf.</b></p> <p>Goal: You want to check the overall image content and quality, such as the lighting condition, visual aesthetics, or any private content you don't want to expose.</p> <p>Tasks developed by participants.</p>
	<p><b>Session 5: Editing participants' own images.</b></p> <p>Tasks developed by participants.</p>

**Figure 7: Images and tasks used in the user study sessions. Please refer to Figure 3, 10, and 11 for more details.**

each element on the EditScribe UI and were asked to perform editing tasks of their interests on an example image. Participants were instructed on the EditScribe feedback after an edit, including the four types of outputs, and noted that they could ask follow-up questions to verify the image content as well as undo or redo their edits (Figure 2).

**4.2.2 Session 2: Performing individual edit actions.** In this session, we aimed to understand if participants could successfully perform each edit action of EditScribe. For each edit action, participants were asked to use natural language prompts to edit or verify edits until they confirmed the task was completed. We then asked participants to provide their agreement on the statement “I am confident that the edit is successful” on a scale from 1 to 7, where 1 is strongly disagree and 7 is strongly agree. We randomized the order of the five editing tasks across participants.

**4.2.3 Session 3: Making a flyer to find a missing cat.** In this session, we aimed to explore how participants would use EditScribe in a practical scenario. Participants engaged in the scenario detailed in Section 3.1. Participants were asked to perform the tasks in their preferred order until they confirmed or were satisfied with the results, or if they decided to move on.

**Table 1: Participant demographics information.**

PID	Age	Gender	Vision Level	BLV Onset	Experience with GenAI	Experience in image creation, editing or verification
P1	43	Male	Low vision	Since 30yo	Using daily	Adding basic existing graphics, e.g., arrows, circles, to existing images
P2	42	Male	Total Blindness	Since 9yo	Using daily	None
P3	40	Male	Total blindness	Since 2yo		Cropping or rotating images
P4	40	Female	Total blindness	Birth		None
P5	43	Male	Some light perception	Birth	Using daily.	Resizing or flipping images
P6	29	Male	Total blindness	Birth	Creating and iterating images with Midjourney	Using existing captioning apps or tactile display for verifying content
P7	30	Female	Total blindness	Since 17yo	Using GPT for work	Python or R programming to create or edit graphs for work purposes
P8	34	Male	Total Blindness	Since 18yo		None
P9	26	Male	Total Blindness	Since 16yo	Creating and iterating images with Midjourney and ChatGPT	A photographer. Using the iOS photo app for basic tasks like exposure, applying filters, contrast, and rotating the image
P10	27	Male	Total blindness	Birth		None

**4.2.4 Session 4: Making a flyer for recruiting a craftsman to make a bathroom shelf.** Next, instead of providing individual specific tasks, we provided a high-level goal for participants to develop their editing tasks based on their needs. The image was more complex, with multiple objects and uneven lighting conditions (Figure 7). The instruction was: *You plan to create a flyer and post it on your social media to find someone to custom-make a shelf to fit everything on your sink. You thus first took a photo of the sink in your bathroom. However, it includes many personal items, such as towels, containers, medication bottles, toothbrushes. You want to check the overall image content and quality, such as the lighting condition, visual aesthetics, or any private content you don't want to expose. You should develop three to five edit actions.*

**4.2.5 Session 5: Editing participants' own images.** In this session, we aimed to encourage participants to brainstorm potential use cases and workflows to apply EditScribe on their own images. Five participants provided their images before the study, for which they were familiar with the content and context when the photos were taken. In this open-ended session, participants performed their goals without any restrictions.

### 4.3 Data Analysis

Besides reporting their perceived confidence in each editing task in session 2, participants were asked to think aloud during each session and interviewed afterwards to provide qualitative feedback on their strategies for creating prompts or perceiving EditScribe outputs. All sessions were recorded and transcribed for analysis. The first author independently performed open coding on all transcripts to identify initial codes. The initial coding focused on participants' prompt formation and decision making, feedback toward the system output, and any friction they experienced. Two authors reviewed all coded excerpts and iterated on the set of codes through discussing with the first author. They then conducted affinity diagramming [33] on the initial codes to extract and organize high-level themes.

## 5 FINDINGS

### 5.1 RQ1: How does EditScribe support non-visual image editing?

*Participants were able to successfully complete most of the tasks in the study sessions and found EditScribe promising in supporting everyday scenarios. On the other hand, participants also expressed a need for more edit actions and finer control over image editing.*

**5.1.1 Task performance in the study sessions.** In session 2, all participants were able to complete all the tasks (Figure 7) through natural language verification loops, but with varying confidence about the editing results. In particular, they were generally confident about the results of changing the color of the wall ( $\mu=6.2$ ,  $\sigma=0.8$ ) and adding text to the image ( $\mu=6.4$ ,  $\sigma=0.7$ ), which was generally straightforward and consistent across verification feedback. However, there was occasional confusing verification feedback in the other three tasks, which made participants less confident about the results. First, for the task of blurring out the person in the picture ( $\mu=5.6$ ,  $\sigma=1.4$ ), P8, who scored 2 for this task, was concerned if only the face were blurred but not the entire person *"I would be lower on this, so I gave two. It says the face was blurred, but I expected it to blur out the entire body."* For removing the bowl ( $\mu=5.8$ ,  $\sigma=1.0$ ), P2 was initially not confident in the system and scored 4 for this task, but developed his trust in the system later after asking several followup questions (Section 5.2.3) to verify the changes, stating *"I was uncertain whether the system was reliable so I gave you a neutral rating, but it seemed correct after I tested its reliability, so I could have gave a higher rating later."* Lastly, though participants were able to increase the brightness of the dog ( $\mu=5.3$ ,  $\sigma=1.1$ ), EditScribe occasionally described the white dog as a 'cream-colored' or 'light-colored' dog in UPDATED OBJECT DESCRIPTIONS, even though SUMMARY OF VISUAL CHANGES and AI JUDGEMENT both confirmed the changes of brightness. For instance, P3 specified *"Make the dog brighter color"* and scored five for his confidence after obtaining the updated object descriptions *"Fluffy cream-colored dog sitting"*,

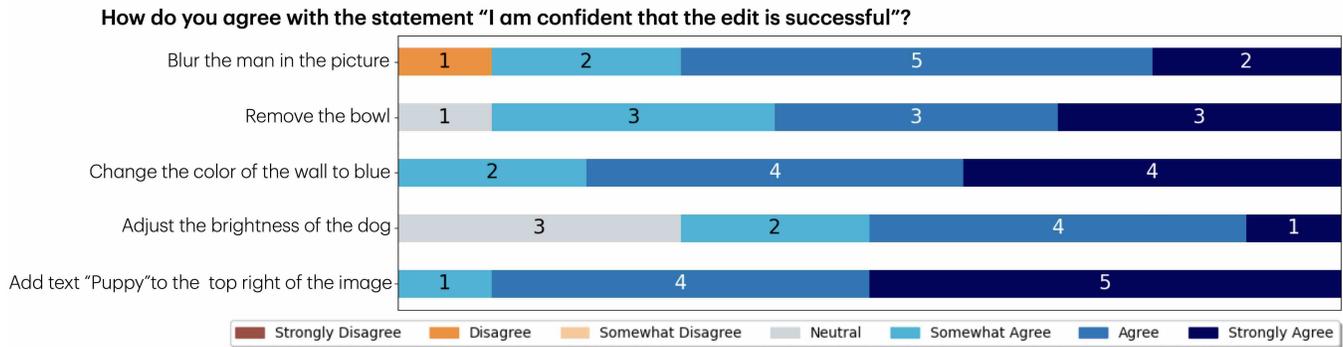


Figure 8: Likert scale responses of participants' perceived confidence for each edit action in study session 2.

which made him confused: *"I said to make it brighter color, which is maybe a little bit different, it's no longer white but cream color. I am not sure if cream is brighter than white."*

In session 3, most participants were able to perform and confirm the completion of the tasks, while a few (P2, P3, P6) encountered issues and had to skip certain tasks. For example, P2 was unable to change the color of the bow tie to blue using his prompts, such as *"change the color of the cat to white cat with blue bow tie"* or *"change the color of the cat's bow tie from red to blue"*. It was because EditScribe misclassified the object of interest as the cat rather than the bow tie and thus changed the color of the cat instead. In another case, P3 encountered discrepant verification feedback or hallucinations after changing the bow tie to blue. For instance, in his two attempts, the bow tie became blue, but with object descriptions changed from *"red bow tie with a bell..."*, to *"blue bow tie with a button or ornament..."*, and *"blue bow tie with polka dots..."*

In session 4, participants developed a number of editing tasks (e.g., increasing the brightness of the sink or towel, blurring out the plastic jars) and were able to complete most of them, with a few failure cases encountered (P4, P8, P10). P4 specified her prompt *"Remove the pill bottle from this image"*, in which the image had multiple bottles, but none of them was recognized as a 'pill bottle' in the object descriptions, which made EditScribe fail to ground on a specific object. In another example, after P8 increased the brightness of the towel, EditScribe hallucinated *"In the edited image, an object resembling a penguin has been added..."* P10 encountered AI JUDGMENT indicating *"the visual difference is indiscernible"* while the UPDATED OBJECT DESCRIPTIONS for the tower changed from 'dark' to 'gray', which confused him.

In session 5, five participants (P1, P3, P4, P6, P8) provided their own images (Figure 9), and each performed two to three edits. Most of the edits were successful, but there were a few instances where participants wanted to edit specific objects they ascertained but were not recognized by EditScribe in sufficient granularity, e.g., the text on P1's T-shirt, P1's partially occluded watch, or the bowl lid by P6. Overall, participants were excited about the opportunities to edit their own images non-visually and came up with more desired edit actions and potential use cases in their daily lives.

5.1.2 *Promise of EditScribe and need for finer controls and more edit actions.* During the study, participants indicated a need for finer control over the edit actions. First, participants expressed a

desire to modify specific parts rather than the entire object. Second, participants desired higher prompt bandwidth, such as editing multiple objects at once or using one detailed prompt to express desired and undesired edit outcomes (Section 5.2.1). Participants also noted advanced functions, such as 'moving object' (P10) or 'applying and describing certain image filters' (P9). Participants also sought higher granularity in adjustments, such as brightness, exact color codes, font family, and sizes. This precise specification could help ensure that the system's output matches users' exact intent, as noted by P9 *"AI is making its own decisions. It seems to complete the tasks but not my intent"* (We will discuss more in Section 6.2).

Also, participants generally appreciated that EditScribe enabled them to edit images and had ideas for supporting everyday scenarios, such as making graphs or flyers for work (P5, P7), adding shapes or arrows to images to highlight content (P1, P2, P7, P8), or posting more images on their social media, as P4 said *"If I get something like this in my hands, I will be taking a lot of pictures and I would totally be one of those people who take pictures of their food every time they make a meal."* P7 also wanted to use EditScribe to edit data graphs: *"It's quite exciting to hear that such a system is being developed. As a data analyst, I program to create my graphs. It would help me to put a red box around the first bar to get people's attention, which I always find difficult and need someone to do it for me manually."* P9 further expressed his eagerness as a photographer and commended that using natural language could benefit more people: *"I am trying to get back into taking photos because there are no accessible ways to do it. It could be helpful to describe the filter I want, and then it generates that filter and description. It could also be useful for even a layperson who doesn't have technical knowledge of editing software, where you can just use natural language to describe what you want and for the image to be returned in that way."*

## 5.2 RQ2: How do BLV people prompt EditScribe?

*Participants found it intuitive to use natural language to specify prompts and ask followup questions, which reduced their learning efforts. They also developed their own prompt strategies throughout the study.*

**P1's image and prompts**

1. Change color of object 1 from green to blue.
2. Blur the face of man.
3. Remove the fence.



**P3's image and prompts**

1. Make the background blue.
2. Add the word "Charlotte" to the bottom center.



**P4's image and prompts**

1. Change door color to sky blue.
2. In the top right part of the image write Doofy!



**P6's image and prompts**

1. Remove the bowl in the image.
2. Blur the cabinet.
3. Change the color of the cabinet to dark



**P8's image and queries**

1. Increase brightness of the shrubs.
2. Remove "cluster of tall, thin palm trees with green fronds".



Figure 9: Participants' own images edited in study session 5, along with their prompts. Faces are blurred for privacy.

5.2.1 *Specifying prompts with varying levels of granularity for better clarity.* We observed that participants specified information at varied levels of granularity in their prompts. Their strategies evolved changeably based on the success of edits or verification feedback, yet the goal remained the same — *to ensure the system understood their prompts.*

Participants used simple prompts, such as “make the dog brighter” used by P9 in session 2, who noted: “*That’s very basic and concise. The idea is not open to interpretation. In theory, I don’t think there should be much confusion about what this instruction means.*” As the study went on, he added more specificity to his prompts, evolving to “make the towel brighter color,” and eventually to “increase the brightness of the cat to make it more visible, without changing the color of the cat.” This demonstrated a progression towards more detailed instructions to better convey their intent and ensure the desired outcome. However, several participants (P6, P8, P9, P10) developed more detailed information about their desired outcomes and specific exclusions in their edits initially. For instance, consider P6’s prompt, who had image generation prompting experiences, in session 3: “Take the object of the cat and increase the brightness of the object so that there is greater contrast between the cat and the rest of the photo.” However, facing failed edits or discrepancies in the verification feedback, P6 attributed those discrepancies to his complex prompts, and decided to simplify them to reduce potential confusion. For instance, when he edited his own image and aimed to remove a lid of the bowl, he changed his prompt from “*Remove the lid in the image so that there is no presence of a bowl or any related accessories,*” to

“*Remove the lid of the bowl,*” and finally to a more succinct “*Remove the lid in the image.*”

We also observed that some participants conveyed high-level goals in their prompts to let the system interpret their intents and perform proper actions. For example, in session 3, participants were asked to obscure the person in the image to maintain privacy and make a white cat noticeable. P7 articulated her prompt: “make the cream-colored cat brighter and stand out more in the image” and explained her rationale: “*I added the second sentence, as my whole intention was to make it more pronounced.*” Additionally, some participants naturally crafted high-level prompts to express their overarching objectives for the scenario, such as “I don’t want to see the lady” (P1), “Increase the focus of the white cat” (P1), “Hide the bowl” (P3). All these high-level prompts were linked correctly to the edit actions and objects.

For referencing the object of interest, participants generally chose to refer to objects by their names as given in the descriptions, such as ‘cat’ or ‘woman.’ However, when errors occurred, they resorted to using detailed descriptors to ensure the system correctly identified the target object, for instance, “Change the color of the white paneled wall to blue” (P6), or “Blur out the man in the framed photo within the image” (P9). Another option, in addition to object names, is to refer to indexes of objects according to the initial object descriptions. Some found referring to indexes of objects helpful in complex scenarios, such as in session 4, where P2 prompted “Blur out object 7” and mentioned “*I had to admit the so many objects, makes it hard to digest. It’s easier to refer to the number and use the*

*natural language to describe what I want to blur out.*” However, some (P2, P9, P10) also expressed concerns about confusing the system, given that the object description may contain multiple objects as descriptors, which the system could incorrectly reference.

**5.2.2 Users’ mental model affecting specified prompts.** We observed that participants’ mental model of how the EditScribe system works influenced the way they formulated their prompts, varying in tone, use of symbols, and specific word choices.

Some participants anthropomorphized EditScribe, treating it as a human-like agent. P3 perceived he was collaborating with human agents, stating, *“two people were working on the image as if one was working on the image and another commentator commenting on the output”*, and *“It’s not obedient, it’s making its own decision.”* P4, similarly, adopted polite tones in their prompts as if she was interacting with someone, exemplified by the use of phrases such as *“Please...”* and *“Can you...”*, and noted, *“I wrote it the way I would speak to a friend editing this picture, like a conversational tone.”* Participants with extensive experience in visual arts and AI technology displayed a higher level of confidence and precision in their language. P1, who used AI technologies in his daily life, showed this trust in EditScribe’s inferential capabilities and prompted *“Change object 1 from green to blue”*. He assumed the AI would have general knowledge of tennis courts and maintain the necessary details, such as the white lines on the tennis court: *“I assumed that AI would probably have taken some inference from other sources of what a blue tennis court is and generalized tennis courts to be blue color.”*

However, uncertainties about EditScribe’s ability to understand nuanced linguistic cues led to concerns over how to format prompts effectively. P3 wondered if quotation marks were necessary to emphasize specific words within his instructions: *“I thought I would have had to put the word puppy in quotation marks to indicate the beginning and end, like a tag.”* Similarly, P6 was concerned that quotation marks included in his prompt might mistakenly appear in the final image and asked follow-up questions to confirm. To distinguish between prompts that were questions and edit instructions, P5, who described himself as an intermediate programmer and tech-savvy, experimented and speculated about the system’s method of recognizing prompts, *“I was trying to figure out how you were determining whether it’s a question or not, like if you’re looking for a question mark, if you’re doing a regular expression, or if you’re having AI determine.”*

**5.2.3 Objectives and strategies of prompting follow-up questions.** Participants often used follow-up questions to verify the results of their edits. For instance, after attempting to remove a bowl in session 2, participants asked *“How many bowls are in the picture?”* (P1), or *“Is there any ceramic material in the picture?”* (P9) to confirm that the removal was completed and no remnants were left behind. This verification step was crucial to ensure the edits met their expectation when discrepancies arose in the outputs, as P9 remarked *“I used summary and AI Judgement to compare to make sure that what I wanted was done. I was more likely to ask follow-up questions when the 2 of them were inconsistent.”* Also, participants were concerned with the aesthetic quality of the images post-edit. They inquired about the overall appearance and integrity of the edits by asking questions such as *“Does the image look doctored?”* (P5), *“Do you see*

*any imperfections in the photo?”* (P5), or *“Describe the brightness of the dog in contrast to the rest of the image.”* (P9).

Asking for spatial information was also prevalent, especially when participants planned to add text to an image or assess the text they added. They sought to understand the layout and space availability, such as *“Are there areas in the image where it’s empty, or less cluttered?”* (P10), or *“What is the size of the word ‘puppy’ compared to the whole of the picture?”* (P2).

Lastly, participants prompted EditScribe of their existing knowledge to gain confidence and gauge the system’s reliability. For instance, after removing a bowl, P2 systematically prompted EditScribe with questions he knew the answers to, such as *“Is there a bowl in the picture?”* (answer: “No”), *“Is there a dog in the picture?”* (answer: “Yes”), *“Is there a man in the picture?”* (answer: “No”). He developed this confidence in EditScribe by confirming known facts, and believed the bowl was removed *“I have doubts about whether the bowl was removed, and then to further try and get more confidence as to whether the AI can answer the question correctly. The first question is to see whether the bowl is removed. The other two questions are just to check whether the AI is reliable.”*

### 5.3 RQ3: How do BLV people perceive EditScribe’s verification feedback?

*Participants found all four types of verification feedback useful depending on the context and information they wanted. They developed different strategies and interpretations for consuming verification feedback.*

**5.3.1 Perception on each verification feedback.** We reported participants’ perceptions of each verification feedback.

**Summary of Visual Changes:** Participants perceived SUMMARY OF VISUAL CHANGES as ‘clear’ (P7) and ‘direct’ (P1) into the actions that have been performed. P4 commented *“Summary is really good at comparing the original and the latest image for the changes. So that’s very helpful in giving me an idea of where I was, and this is where I am.”* However, participants felt that the summary alone might be too superficial and lacked detailed reasoning to fully assess the success of the edits. Therefore, participants combined it with the other feedback.

**AI Judgement:** Based on the above reason, participants tended to combine and compare SUMMARY OF VISUAL CHANGES and AI JUDGEMENT, which they perceived the former as the facts, while the latter as *“internal thinking and reasoning of the system”* (P5). P1 indicated *“AI judgment gives a more detailed description into what you actually assess the thing, even more descriptively which part has been changed.”* However, some (P5, P7) regarded the information of AI JUDGEMENT as verbose and overlapped from other outputs, as P5 suggested *“There might be a point where they become redundant. But at the moment, I appreciate having them all.”*

**Updated General Descriptions:** Most participants found UPDATED GENERAL DESCRIPTIONS useful for gaining a new perspective on how ‘sighted individuals’ may perceive the image, as highlighted by the terms *“new angle”* (P2) and *“new mental image”* (P1). Also, as P9 noted, *“general descriptions determine how much focus that object has when it comes to someone seeing the picture. So it allows the user to construct a better visual idea of what the picture is like.”* Most participants also mentioned that general descriptions offer more

global and comprehensive information, enabling them to picture the image content, as P3 described *“It describes everything in relation to the different objects in relation to each other. It tells me what the image looks like.”* However, general descriptions were generated based solely on the updated image, and the system could generate varying descriptions for the unchanged objects, which confused the user (Section 5.3.2).

**Updated Object Descriptions:** These were perceived as useful for noting detailed edits and serving as a reference for prompts. P1 highlighted this benefit: *“It actually helps me to precisely call out the object so that I can speak this language. I can know how it actually is. I can actually make the amendment.”* However, object descriptions in session 4 had rich information on granular objects in their descriptors, as well as speculative information (e.g., ‘possibly for...’, ‘likely filled with...’, ‘potentially a ...’), which made them confusing to absorb, as noted by P9: *“The quantity of objects is confusing. I wasn’t sure which object I was manipulating, because it seemed like there were more objects than there actually were.”*

**5.3.2 Consistency vs. discrepancy among the verification feedback.** Generally, participants were confident in an edit if all the feedback was consistent. P7 stated *“Summary and AI judgement confirm object 5 has been removed, so it just put object 5 a colon in object description. I also don’t see the bowl in the general description anymore. I think it’s correct as I have multiple avenues of confirmation.”* However, we observed occasional discrepancies in the feedback, such as different types of feedback describing the same object in different ways, prompting participants to seek clarification by asking follow-up questions. For instance, P7, who frequently evaluated and compared SUMMARY OF VISUAL CHANGES and AI JUDGEMENT, thought aloud when tackling a discrepancy: *“The summary and AI judgment don’t really correspond, because summary says the person has been removed, and AI judgment says the woman can no longer be clearly identified. So I think that there’s some discrepancy here. I guess I should check. [prompting ‘Is there a blurred image of human the in background’ (answer ‘yes’)] So it seems like the summary was wrong, and the judgment was correct.”* Discrepancies were also noted between the current and previous verification feedback. For instance, after P3 changed the color of a cat’s bow tie from red to blue, the system feedback added an unexpected detail: polka dots. P3 noted, *“It was like making its own decisions to add polka dots. I don’t know if these were there. Just wasn’t describing it before.”* P9 also mentioned, *“One time it described as teal, one time described as greenish gray. I believe it was describing the same container. So, making sure to use the same kind of descriptors would be quite helpful.”*

**5.3.3 Difficulty picturing editing effects leads to challenges judging edit outputs.** During the study, participants expressed difficulties in picturing certain edited visual effects due to their lack of visual experience, echoed with findings of prior work [84] around the challenge of envisioning complex obfuscation effects (e.g., inpainting an area with surrounding background) by BLV individuals. For instance, P2 was uncertain about the effect of increasing brightness: *“I have to admit I’m not sure what is supposed to happen when I make something brighter.”* Similarly, P6 was confused by the effects of brightening a white dog during session 2, which resulted in the color description changing from white to cream in the new verification feedback. P6 noted *“I think there wasn’t really anything in the*

*description to explain why it chose that particular color or whether increasing the brightness naturally changes it to this color.”* Some participants (P3, P5) also struggled to grasp the intended result of blurring effects on images, as P5 commented *“as a blind person. I’m not entirely sure what Blurred is supposed to do.”*

**5.3.4 Excessive information and tone influence the perception and confidence of the results.** Within the verification feedback, EditScribe sometimes generated details excessive to what users required or presented information using a less confident tone, leading to confusion among participants. For instance, P2 encountered an object description of a man stated as *“Framed portrait of a person with non-descript features due to blurring for privacy.”* P2 found this confusing, as he did not instruct the system to do it for privacy purposes: *“It’s a bit strange. That was totally not in my instruction at all. It provides extra information. Overly explain things.”* Moreover, some participants were also influenced by the negative tones in the verification feedback, which they thought decreased their confidence in the results. For example, P1 was unsure about the success of an edit due to the uncertain tone of the feedback: *“It said the edit appears to be successful, so is this successful or not?”* In contrast, a positive tone in the outputs could enhance confidence, as noted by P3 *“The system seems to be pretty confident, saying the edit is successful. Just based purely on that confident, sounding language. I am willing to give it some credit.”*

## 5.4 RQ4: How do BLV people perceive the final edited images?

*Participants were generally willing to use the final edited image depending on the scenario but expressed the need for further validation from other sources.*

**5.4.1 Determining whether to post on public depending on context.** Participants generally expressed confidence that the system had successfully completed the tasks as instructed, though they were uncertain if it had fully met their intentions. P9, with visual experiences and higher expectations for visual aesthetics, commented *“It was successful in actually doing what I asked it to, but maybe not the intent. I just wanted to know the nature of the text that was added. It’s good that it told me it’s San Serif, but it’s not very useful in a practical way, like the aesthetic quality of the text.”* Consequently, due to concerns about the unseen visual outputs, most participants preferred to seek verification from sighted assistance before posting, as suggested by P1 *“I will probably get somebody just to help me to visually check and confirm before I actually post it. I think there’s a certain level of trust in it. But it’s not 100%.”*

However, some participants noted that they might consider publishing the images in urgent situations or if sighted assistance was unavailable since they perceived the system satisfied the basic requirements. For instance, P7 mentioned the scenario of looking for the cat, where urgency overrides others: *“If I feel that it’s very urgent, and there’s no one else to check over it, then I will go ahead and post because I know that the picture shows my cat, and it shows my number.”* Also, P9, though desired the visual check as mentioned, acknowledged the primary goal of the edit over nuanced visual quality, *“The quality of the edit is secondary. Because what you really need is for people to pay attention to the photo of the cat, to make*

sure that it's recognizable and distinct, and also to have the call to this number. And based on what the AI judgment is telling me, it seems like it performed the task that I asked it to."

**5.4.2 Need for multiple modalities or sources to verify output images.** In addition to seeking sighted assistance, participants employed various strategies to verify the content of images. Many participants were tech-savvy at utilizing AI technologies in their daily routines, including BeMyAI [12], SeeingAI [15], Midjourney [13], and ChatGPT [8]. Therefore, a common method among them was to use different image captioning models or applications to cross-verify the image content, echoing findings from Herskovitz et al. [34] on how BLV people combined and utilized assistive technologies. P6, who usually used an embosser or Braille display to check the content and spatial information of images he created with Midjourney: *"I have an embosser that I can send images to. So I would probably send images to it before posting online. I could recognize something supposedly pretty simplistic, like a ball or a dog, and check the spatial relationship."* For the tasks for adding text, participants also mentioned verifying the text further using optical character recognition (OCR) built into the screen reader or other third-party apps. P6 further pointed out using sonification to make those changes of color or brightness accessible.

**5.4.3 Faith and experiences with AI technologies affect the decision of using final edited images.** Half participants (P2, P4, P5, P6, P9) expressed a lack of confidence not only in EditScribe but also in other AI technologies, stemming from the imperfections they had observed in their experiences. P9, for example, shared his skepticism based on previous experiences with image editing software, particularly with object removal: *"I am less inclined to do it because I do not fully trust that it will do it reliably. But for other functions that make a mistake, it wouldn't be a terrible one."* Similarly, P4 discussed her overall trust in AI technologies, emphasizing the inherent uncertainties faced by BLV users: *"Being totally blind, you could be told the sky is green and the grass is purple. There's no way for me to technically say that's not true. There's an element of you having to put trust into a system and hope it's doing what it's supposed to be doing. It's not really a concern but a lifestyle. It's part of being blind. You have to put a lot of faith in the technology you use and hope it's doing what it's supposed to."* In contrast, we found that participants familiar with AI technologies tended to trust the verification feedback more, and recommended simplifying the information EditScribe provided. P7, a data analyst who regularly programmed and created visualizations for work and used off-the-shelf AI-enabled image captioning apps, suggested: *"I use all these AI every day for my work. We are used to AI, and it worked as it should in general. I think the system [EditScribe] should work this way. Then I think there's no need for so many levels of checks at verification. So most people will actually be fine, with just a short description of the change, and say this to be successful."*

## 6 DISCUSSION AND FUTURE WORK

In this section, we discuss our limitations, lessons learned, and future work to support non-visual image editing.

### 6.1 Enhancing Verification Feedback Loops

While the verification feedback was generally perceived as useful in providing different aspects of edit information, there are many opportunities to make them more informative and personalized to BLV users.

First, although EditScribe's natural language verification loops aim to support BLV image editors' diverse preferences for visual feedback [68, 84] and did prove to bring more flexibility (Section 5.2.3), we noted a need for further customization, such as combining or omitting certain elements for greater conciseness (Section 5.3.1), elaborating specific visual effects (Section 5.3.3), incorporating detailed spatial descriptions (Section 5.2.3). It is also important to consider BLV users' prior visual experiences and experiences with AI technologies (Section 5.4.3), which could provide them with proper levels of feedback *e.g.*, providing professional terms for describing visuals accurately (as for P9 who is a photographer), or providing succinct and direct feedback if users are familiar with and trust AI technologies (as for P7 who used GenAI for daily work).

Second, participants were unclear about the linguistic and editing capabilities of EditScribe. This confusion led to repetitive prompt refinement through trial and error. Drawing on prior research that emphasizes the importance of explaining an automated system's decisions [21, 49, 51, 76, 88], the system could clarify its capabilities of interpreting natural language prompts into specific items (*e.g.*, edit functions, objects), thereby aiding in establishing a user's initial mental model of the system. The system can also improve its effectiveness by building a model of the user, analyzing user prompts, and initiating follow-up questions to elicit user needs and goals. After understanding the user's needs, the system can assign corresponding agents, who may have different domain knowledge, such as interior design, graph design, etc. These agents could provide more efficient and useful feedback tailored to the user's needs and goals. This mutual learning and adaptive approach could allow the system to personalize the visual verification feedback to the user progressively.

Occasional hallucinations and inconsistent descriptors across verification feedback are other significant issues. Potential solutions include incorporating historically-used wordings into prompts for generating consistent feedback, and strategically constraining and guiding the lexical choices of LLMs [43, 83] (*e.g.*, tuning the temperature parameter of GPT). It can also be beneficial to include different VLMs and LLMs to cross-verify the edit and provide verification feedback from different perspectives, as how BLV people suggested and used in their current workflow of verifying visual information (Section 5.4.2 & [34, 37]). Sighted people can also be involved in the image editing process if needed, as prior work has shown different techniques to support collaborative visual [54, 86] or text editing [26, 44]. Furthermore, techniques in different modalities have been proposed and utilized to verify visual content, (*e.g.*, tactile images [27, 45, 47, 87], non-speech sounds [67, 87]). Future work could explore ways to integrate these methods to offer a more consistent and reliable experience for BLV users.

### 6.2 Supporting Richer Edit Actions

In this paper, we demonstrated the utility of natural language verification loops through EditScribe that supports five edit actions. Our

participants expressed the need to have finer controls and more edit actions (Section 5.1), which can be achieved in the future due to EditScribe's flexible structure. For instance, with the power of LLMs, EditScribe can understand more edit actions and allow users to specify a particular value to a certain visual effect (e.g., changing color with color codes) by few-shot prompting (e.g., giving a series of examples as the context of what the model should perform) [25, 82], which could further increase the bandwidth of input prompt and users can perform multiple edits at a time. We can also utilize the models of different granular object segmentation to support finer controls on parts of objects [42, 46, 89], similar to techniques that enable hierarchical image explorations [15, 45, 77]. Other models could enable additional actions, such as recognizing shapes to enable diagram editing, or detecting depth to enable occlusion-aware editing. In contrast to finer edits, another potential integration is to include models that enable global content editing, such as InstructPix2Pix [24] to adjust the overall style of the image or Emu Edit [66] to generate new content consistent with the overall image. Unlike traditional image editing, such generative edits often consider the overall image content and optimize the visual aesthetic, which could create unintended results. Therefore, future work could explore how to describe such global or partial style changes in verification feedback. This should cover not only the content of the images but also any perceived imperfections or artifacts in the edited images (e.g., text cutoffs, visual artifacts).

### 6.3 Limitations of the Study

In our user study, we tried to recruit BLV individuals with diverse backgrounds. Some were familiar with AI technologies, some had extensive visual experiences, and some had image editing experiences. However, our study insights may not represent sufficiently broad perspectives in the BLV community, who have varied experiences and use of technologies. Future studies could explore this broader set of experiences and perspectives through field studies or deployment. Second, we selected specific images across different sessions to showcase various levels of image complexity and practical contexts. However, these chosen images may not accurately represent the types of photos that BLV individuals capture in their daily lives. Although participants in our study session 5 provided and edited their own images, the small sample may not represent the range of characteristics taken by blind people. Future work could further explore how EditScribe would perform with images BLV people personally capture and how they would manage editing and iterating these photos on their own with EditScribe.

## 7 CONCLUSION

We have presented EditScribe, a prototype system demonstrating the concept of natural language verification loops in the context of image editing. EditScribe supports five specific edit actions and allows BLV users to input their editing instructions in natural language. Building on insights from prior work, we developed four types of verification feedback for BLV users to perceive changes made to images. These include a summary of visual changes, AI judgments, and general and object descriptions. Through a user study with ten BLV people, we explored how EditScribe supported non-visual image editing, how participants prompted EditScribe,

and how they perceived verification feedback and the final edited images. Finally, we discussed ways to enhance the verification feedback provided by EditScribe, explored possibilities for expanding the range of edit actions available, and outlined implications for future systems that utilize natural language verification loops.

## ACKNOWLEDGMENTS

We thank our anonymous reviewers and all the participants in our study for their suggestions, as well as Andi Xu for helping facilitate our user studies.

## REFERENCES

- [1] 2015. Specific Guidelines: Art, Photos & Cartoons. <http://diagramcenter.org/specific-guidelines-final-draft.html>
- [2] 2018. How to Write Alt Text and Image Descriptions for the visually impaired. <https://www.perkins.org/resource/how-write-alt-text-and-image-descriptions-visually-impaired/>
- [3] 2018. Web Content Accessibility Guidelines (WCAG) Overview. <https://www.w3.org/WAI/standards-guidelines/wcag/>
- [4] 2022. Auto Color. <https://helpx.adobe.com/ca/premiere-pro/using/auto-color.html>
- [5] 2022. Text to Color Grade. <https://runwayml.com/ai-tools/text-to-color-grade/>
- [6] 2024. Aira. <https://aira.io/>
- [7] 2024. BeMyEyes. <https://www.bemyeyes.com/>
- [8] 2024. ChatGPT. <https://chat.openai.com/>
- [9] 2024. GPT-4 Vision. <https://platform.openai.com/docs/guides/vision>
- [10] 2024. Gradio. <https://www.gradio.app/>
- [11] 2024. How to use Text Analyzer in JAWS to proofread documents. <https://www.perkins.org/resource/how-to-use-text-analyzer-in-jaws-to-proofread-documents/>
- [12] 2024. Introducing Be My AI (formerly Virtual Volunteer) for People who are Blind or Have Low Vision, Powered by OpenAI's GPT-4. <https://www.bemyeyes.com/blog/introducing-be-my-eyes-virtual-volunteer>
- [13] 2024. Midjourney. <https://www.midjourney.com/home>
- [14] 2024. OpenCV. <https://opencv.org/>
- [15] 2024. SeeingAI. <https://www.seeingai.com/>
- [16] 2024. Tap into the power of AI photo editing. <https://www.adobe.com/products/photoshop/ai.html>
- [17] 2024. Use VoiceOver for images and videos on iPhone. <https://support.apple.com/en-ca/guide/iphone/iph37e6b3844/ios>
- [18] Dustin Adams, Lourdes Morales, and Sri Kurniawan. 2013. A qualitative study to support a blind photography mobile application. In *Proceedings of the 6th International Conference on Pervasive Technologies Related to Assistive Environments (Rhodes, Greece) (PETRA '13)*. Association for Computing Machinery, New York, NY, USA, Article 25, 8 pages. <https://doi.org/10.1145/2504335.2504360>
- [19] Tousif Ahmed, Patrick Shaffer, Kay Connelly, David Crandall, and Apu Kapadia. 2016. Addressing Physical Safety, Security, and Privacy for People with Visual Impairments. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. USENIX Association, Denver, CO, 341–354. <https://www.usenix.org/conference/soups2016/technical-sessions/presentation/ahmed>
- [20] Rahaf Alharbi, Robin N. Brewer, and Sarita Schoenebeck. 2022. Understanding Emerging Obfuscation Technologies in Visual Description Services for Blind and Low Vision People. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 469 (nov 2022), 33 pages. <https://doi.org/10.1145/3555570>
- [21] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3290605.3300233>
- [22] Cynthia L. Bennett, Jane E. Martez E. Mott, Edward Cutrell, and Meredith Ringel Morris. 2018. How Teens with Visual Impairments Take, Edit, and Share Photos on Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3173574.3173650>
- [23] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, and Tom Yeh. 2010. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology* (New York, New York, USA) (UIST '10). Association for Computing Machinery, New York, NY, USA, 333–342. <https://doi.org/10.1145/1866029.1866080>
- [24] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. arXiv:2211.09800 [cs.CV]

- [25] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [26] Maitraye Das, Thomas Barlow McHugh, Anne Marie Piper, and Darren Grelge. 2022. Co11ab: Augmenting Accessibility in Synchronous Collaborative Writing for People with Vision Impairments. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 196, 18 pages. <https://doi.org/10.1145/3491102.3501918>
- [27] Danyang Fan, Alexa Fay Siu, Wing-Sum Adrienne Law, Raymond Ruihong Zhen, Sile O'Modhrain, and Sean Follmer. 2022. Slide-Tone and Tilt-Tone: 1-DOF Haptic Techniques for Conveying Shape Characteristics of Graphs to Blind Users. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 477, 19 pages. <https://doi.org/10.1145/3491102.3517790>
- [28] Noor Fatima. 2020. AI in Photography: Scrutinizing Implementation of Super-Resolution Techniques in Photo-Editors. In *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*. 1–6. <https://doi.org/10.1109/IVCNZ51579.2020.9290737>
- [29] Ricardo E. Gonzalez Penuela, Paul Vermette, Zihan Yan, Cheng Zhang, Keith Vertanen, and Shiri Azenkot. 2022. Understanding How People with Visual Impairments Take Selfies: Experiences and Challenges. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) (ASSETS '22). Association for Computing Machinery, New York, NY, USA, Article 63, 4 pages. <https://doi.org/10.1145/3517428.3550372>
- [30] Ricardo E. Gonzalez Penuela, Paul Vermette, Zihan Yan, Cheng Zhang, Keith Vertanen, and Shiri Azenkot. 2022. Understanding How People with Visual Impairments Take Selfies: Experiences and Challenges. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) (ASSETS '22). Association for Computing Machinery, New York, NY, USA, Article 63, 4 pages. <https://doi.org/10.1145/3517428.3550372>
- [31] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. *arXiv:1406.2661 [stat.ML]*
- [32] Susumu Harada, Daisuke Sato, Dustin W. Adams, Sri Kurniawan, Hironobu Takagi, and Chieko Asakawa. 2013. Accessible Photo Album: Enhancing the Photo Sharing Experience for People with Visual Impairment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) (CHI '13). Association for Computing Machinery, New York, NY, USA, 2127–2136. <https://doi.org/10.1145/2470654.2481292>
- [33] R. Hartson and P.S. Pyla. 2012. *The UX Book: Process and Guidelines for Ensuring a Quality User Experience*. Elsevier Science. <https://books.google.ca/books?id=w4I3Y64SWLoC>
- [34] Jaylin Herskovitz, Andi Xu, Rahaf Alharbi, and Anhong Guo. 2023. Hacking, Switching, Combining: Understanding and Supporting DIY Assistive Technology Design by Blind People. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 57, 17 pages. <https://doi.org/10.1145/3544548.3581249>
- [35] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. *arXiv:2208.01626 [cs.CV]*
- [36] Naoki Hirabayashi, Masakazu Iwamura, Zheng Cheng, Kazunori Minatani, and Koichi Kise. 2023. VisPhoto: Photography for People with Visual Impairments via Post-Production of Omnidirectional Camera Imaging. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA) (ASSETS '23). Association for Computing Machinery, New York, NY, USA, Article 6, 17 pages. <https://doi.org/10.1145/3597638.3608422>
- [37] Mina Huh, Yi-Hao Peng, and Amy Pavel. 2023. GenAssist: Making Image Generation Accessible. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) (UIST '23). Association for Computing Machinery, New York, NY, USA, Article 38, 17 pages. <https://doi.org/10.1145/3586183.3606735>
- [38] Mina Huh, Saelyne Yang, Yi-Hao Peng, Xiang 'Anthony' Chen, Young-Ho Kim, and Amy Pavel. 2023. AVscript: Accessible Video Editing with Audio-Visual Scripts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 796, 17 pages. <https://doi.org/10.1145/3544548.3581494>
- [39] Mina Huh, Saelyne Yang, Yi-Hao Peng, Xiang 'Anthony' Chen, Young-Ho Kim, and Amy Pavel. 2023. AVscript: Accessible Video Editing with Audio-Visual Scripts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 796, 17 pages. <https://doi.org/10.1145/3544548.3581494>
- [40] Joonyoung Jun, Woosuk Seo, Jiyeon Park, Subin Park, and Hyungju Jung. 2021. Exploring the Experiences of Streamers with Visual Impairments. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 297 (oct 2021), 23 pages. <https://doi.org/10.1145/3476038>
- [41] Ju Yeon Jung, Tom Steinberger, Junbeom Kim, and Mark S. Ackerman. 2022. "So What? What's That to Do With Me?" Expectations of People With Visual Impairments for Image Descriptions in Their Personal Photo Activities. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (Virtual Event, Australia) (DIS '22). Association for Computing Machinery, New York, NY, USA, 1893–1906. <https://doi.org/10.1145/3532106.3533522>
- [42] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.
- [43] Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. 2024. Style Vectors for Steering Generative Large Language Model. *arXiv preprint arXiv:2402.01618* (2024).
- [44] Cheuk Yin Phipson Lee, Zhuohao Zhang, Jaylin Herskovitz, JooYoung Seo, and Anhong Guo. 2022. CollabAlly: Accessible Collaboration Awareness in Document Editing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 596, 17 pages. <https://doi.org/10.1145/3491102.3517635>
- [45] Jaewook Lee, Jaylin Herskovitz, Yi-Hao Peng, and Anhong Guo. 2022. Image-Explorer: Multi-Layered Touch Exploration to Encourage Skepticism Towards Imperfect AI-Generated Image Captions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 462, 15 pages. <https://doi.org/10.1145/3491102.3501966>
- [46] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. 2023. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767* (2023).
- [47] Jingyi Li, Son Kim, Joshua A. Miele, Maneesh Agrawala, and Sean Follmer. 2019. Editing Spatial Layouts through Tactile Templates for People with Visual Impairments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–11. <https://doi.org/10.1145/3290605.3300436>
- [48] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow Matching for Generative Modeling. *arXiv:2210.02747 [cs.LG]*
- [49] Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding Blind People's Experiences with Computer-Generated Captions of Social Media Images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) (CHI '17). Association for Computing Machinery, New York, NY, USA, 5988–5999. <https://doi.org/10.1145/3025453.3025814>
- [50] Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P. Bigham, and Shaun K. Kane. 2016. "With most of it being pictures now, I rarely use it": Understanding Twitter's Evolving Accessibility to Blind Users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) (CHI '16). Association for Computing Machinery, New York, NY, USA, 5506–5516. <https://doi.org/10.1145/2858036.2858116>
- [51] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D Ragan. 2019. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 97–105.
- [52] Soobin Park. 2020. Supporting Selfie Editing Experiences for People with Visual Impairments. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) (ASSETS '20). Association for Computing Machinery, New York, NY, USA, Article 106, 3 pages. <https://doi.org/10.1145/3373625.3417082>
- [53] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. 2023. Zero-shot Image-to-Image Translation. In *ACM SIGGRAPH 2023 Conference Proceedings* (Los Angeles, CA, USA) (SIGGRAPH '23). Association for Computing Machinery, New York, NY, USA, Article 11, 11 pages. <https://doi.org/10.1145/3588432.3591513>
- [54] Yi-Hao Peng, Jason Wu, Jeffrey Bigham, and Amy Pavel. 2022. Diffscriber: Describing Visual Design Changes to Support Mixed-Ability Collaborative Presentation Authoring. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) (UIST '22). Association for Computing Machinery, New York, NY, USA, Article 35, 13 pages. <https://doi.org/10.1145/3526113.3545637>
- [55] Helen Petrie, Chandra Harrison, and Sundeep Dev. 2005. Describing images on the web: a survey of current practice and prospects for the future. *Proceedings of Human Computer Interaction International (HCI)* 71, 2 (2005).
- [56] Venkatesh Potluri, Tadashi E Grindeland, Jon E. Froehlich, and Jennifer Mankoff. 2021. Examining Visual Semantic Understanding in Blind and Low-Vision Technology Users. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 35, 14 pages. <https://doi.org/10.1145/3411764.3445040>

- [57] Venkatesh Potluri, Maulishree Pandey, Andrew Begel, Michael Barnett, and Scott Reitherman. 2022. CodeWalk: Facilitating Shared Awareness in Mixed-Ability Collaborative Software Development. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) (ASSETS '22). Association for Computing Machinery, New York, NY, USA, Article 20, 16 pages. <https://doi.org/10.1145/3517428.3544812>
- [58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]
- [59] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125 [cs.CV]
- [60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]
- [61] Ethan Z. Rong, Mo Morgana Zhou, Zhicong Lu, and Mingming Fan. 2022. "It Feels Like Being Locked in A Cage": Understanding Blind or Low Vision Streamers' Perceptions of Content Curation Algorithms. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference* (Virtual Event, Australia) (DIS '22). Association for Computing Machinery, New York, NY, USA, 571–585. <https://doi.org/10.1145/3532106.3533514>
- [62] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. 2004. "GrabCut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.* 23, 3 (aug 2004), 309–314. <https://doi.org/10.1145/1015706.1015720>
- [63] Emma Sadjó, Leah Findlater, and Abigale Stangl. 2021. Landscape Analysis of Commercial Visual Assistance Technologies. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) (ASSETS '21). Association for Computing Machinery, New York, NY, USA, Article 76, 4 pages. <https://doi.org/10.1145/3441852.3476521>
- [64] Abir Saha and Anne Marie Piper. 2020. Understanding Audio Production Practices of People with Vision Impairments. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) (ASSETS '20). Association for Computing Machinery, New York, NY, USA, Article 36, 13 pages. <https://doi.org/10.1145/3373625.3416993>
- [65] Anastasia Schaadhardt, Alexis Hiniker, and Jacob O. Wobbrock. 2021. Understanding Blind Screen-Reader Users' Experiences of Digital Artboards. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 270, 19 pages. <https://doi.org/10.1145/3411764.3445242>
- [66] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. 2023. Emu Edit: Precise Image Editing via Recognition and Generation Tasks. arXiv:2311.10089 [cs.CV]
- [67] Alexa Siu, Gene S-H Kim, Sile O'Modhrain, and Sean Follmer. 2022. Supporting Accessible Data Visualization Through Audio Data Narratives. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 476, 19 pages. <https://doi.org/10.1145/3491102.3517678>
- [68] Abigale Stangl, Nitin Verma, Kenneth R. Fleischmann, Meredith Ringel Morris, and Danna Gurari. 2021. Going Beyond One-Size-Fits-All Image Descriptions to Satisfy the Information Wants of People Who Are Blind or Have Low Vision. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) (ASSETS '21). Association for Computing Machinery, New York, NY, USA, Article 16, 15 pages. <https://doi.org/10.1145/3441852.3471233>
- [69] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2021. Resolution-robust Large Mask Inpainting with Fourier Convolutions. arXiv preprint arXiv:2109.07161 (2021).
- [70] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]
- [71] Dani Valevski, Matan Kalman, Eyal Molad, Eyal Segalis, Yossi Matias, and Yaniv Leviathan. 2023. UniTune: Text-Driven Image Editing by Fine Tuning a Diffusion Model on a Single Image. *ACM Trans. Graph.* 42, 4, Article 128 (jul 2023), 10 pages. <https://doi.org/10.1145/3592451>
- [72] Tess Van Daele, Akhil Iyer, Yuning Zhang, Jalyn C Derry, Mina Huh, and Amy Pavel. 2024. Making Short-Form Videos Accessible with Hierarchical Video Summaries. arXiv preprint arXiv:2402.10382 (2024).
- [73] Violeta Voykanska, Shiri Azenkot, Shaomei Wu, and Gilly Leshed. 2016. How Blind People Interact with Visual Content on Social Networking Services. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) (CSCW '16). Association for Computing Machinery, New York, NY, USA, 1584–1595. <https://doi.org/10.1145/2818048.2820013>
- [74] World Wide Web Consortium (W3C). 2022. W3C Image Concepts. <https://www.w3.org/WAI/tutorials/images/>
- [75] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. 2024. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. arXiv:2402.13616 [cs.CV]
- [76] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) (CSCW '17). Association for Computing Machinery, New York, NY, USA, 1180–1192. <https://doi.org/10.1145/2998181.2998364>
- [77] Andi Xu, Minyu Cai, Dier Hou, Rueli-Che Chang, and Anhong Guo. 2024. ImageExplorer Deployment: Understanding Text-Based and Touch-Based Image Exploration in the Wild (W4A '24). Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3677846.3677861>
- [78] Chutian Yang, Xiping He, Qixian Kuang, Ling Huang, and Lingling Tao. 2023. Transformer-based high-fidelity StyleGAN inversion for face image editing. In *Proceedings of the 2023 7th International Conference on Big Data and Internet of Things* (Beijing, China) (BDIOT '23). Association for Computing Machinery, New York, NY, USA, 76–81. <https://doi.org/10.1145/3617695.3617701>
- [79] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V. arXiv preprint arXiv:2310.14441 (2023).
- [80] Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, and Aysegül Dundar. 2023. Inst-Inpaint: Instructing to Remove Objects with Diffusion Models. arXiv:2304.03246 [cs.CV]
- [81] Tao Yu, Runseong Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. 2023. Inpaint Anything: Segment Anything Meets Image Inpainting. arXiv:2304.06790 [cs.CV]
- [82] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*. 841–852.
- [83] Zequn Zeng, Hao Zhang, Ruiying Lu, Dongsheng Wang, Bo Chen, and Zhengjue Wang. 2023. Conzic: Controllable zero-shot image captioning by sampling-based polishing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23465–23476.
- [84] Lotus Zhang, Abigale Stangl, Tanusree Sharma, Yu-Yun Tseng, Inan Xu, Danna Gurari, Yang Wang, and Leah Findlater. 2024. Designing Accessible Obfuscation Support for Blind Individuals' Visual Privacy Management. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 235, 19 pages. <https://doi.org/10.1145/3613904.3642713>
- [85] Lotus Zhang, Simon Sun, and Leah Findlater. 2023. Understanding Digital Content Creation Needs of Blind and Low Vision People. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA) (ASSETS '23). Association for Computing Machinery, New York, NY, USA, Article 8, 15 pages. <https://doi.org/10.1145/3597638.3608387>
- [86] Zhuohao Jerry Zhang, Smirity Kaushik, JooYoung Seo, Haolin Yuan, Sauvik Das, Leah Findlater, Danna Gurari, Abigale Stangl, and Yang Wang. 2023. {ImageAlly}: A {Human-AI} Hybrid Approach to Support Blind People in Detecting and Redacting Private Image Content. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*. 417–436.
- [87] Zhuohao (Jerry) Zhang and Jacob O. Wobbrock. 2023. A11yBoard: Making Digital Artboards Accessible to Blind and Low-Vision Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 55, 17 pages. <https://doi.org/10.1145/3544548.3580655>
- [88] Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. 2017. The Effect of Computer-Generated Descriptions on Photo-Sharing Experiences of People with Visual Impairments. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 121 (dec 2017), 22 pages. <https://doi.org/10.1145/3134756>
- [89] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. 2024. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems* 36 (2024).



	<p><b>Initial General Descriptions</b></p> <p>A dimly lit bathroom countertop with a toothbrush, blue rinse cup, mouthwash, chrome faucet, and various personal care items such as cotton swabs and cleansers; above the sink hangs a folded towel, and a reflection of the countertop items is seen on the mirror behind them.</p>		<p><b>Initial Object Descriptions</b></p> <p>Object 1: dark-colored bathroom wall with a mounted mirror above a backsplash                  Object 2: gray folded towel hanging on a towel bar                  Object 3: clear plastic tray with a ridged surface likely used to hold soap                  Object 4: gray bathroom countertop with a built-in sink                  Object 5: transparent plastic jar with a white lid containing cotton swabs                  Object 6: greenish-gray small circular container with a lid, possibly for holding personal items or accessories                  Object 7: white plastic jar with a label and a white lid, likely filled with cotton balls                  Object 8: teal plastic hand soap dispenser with a pump                  Object 9: blue plastic bottle with a label, potentially a hygiene or cleaning product                  Object 10: small dark bottle, possibly a personal care product like a fragrance or medicinal dropper                  Object 11: red toothbrush with a white handle placed against the backsplash</p>
<p><b>Edit prompt</b></p> <p>Remove the toothbrush.</p> 	<p><b>Summary of Visual Changes</b></p> <p>In the edited image, the toothbrush that was present to the left side of the bathroom countertop in the input image has been removed. Additionally, a slight distortion is noticeable on the surface where the toothbrush was located, but the general background and other objects remain unchanged.</p>	<p><b>AI Evaluation</b></p> <p>The edit appears to have been successful. The toothbrush present in the input image is no longer visible in the edited image. The edited image's object descriptions do not list the toothbrush, confirming its removal in line with the instruction given.</p>	<p><b>Updated General Descriptions</b></p> <p>A dimly lit bathroom countertop with various items including a blue bottle, clear jars with cotton swabs and pads, a metallic faucet, a soap dispenser, and a towel hanging on the left side. The dark blue wall contrasts with the grey countertop.</p> <p><b>Updated Object Descriptions</b></p> <p>: : : Object 11: [removed]. : :</p>
<p><b>Edit prompt</b></p> <p>Increase the brightness of the towel.</p> 	<p><b>Summary of Visual Changes</b></p> <p>In the edited image, the towel on the left side appears to be slightly brighter, with increased visibility of its texture and patterns, indicating an attempt to brighten it as per the instruction. The rest of the scene remains largely unchanged between the input image and the edited image.</p>	<p><b>AI Evaluation</b></p> <p>The edited image shows a light-colored towel instead of the gray folded towel described in the input image, indicating that the towel has indeed been brightened in accordance with the instruction. This is further supported by the change in the object description from a "gray folded towel hanging on a towel bar" (#2) in the input image to a "light-colored towel hanging on a rack" (#2) in the edited image. Therefore, the edit was successful as it aligns with the provided instruction to brighten the towel.</p>	<p><b>Updated General Descriptions</b></p> <p>A dimly lit bathroom countertop with various items including bottles, jars, a soap dispenser, and a faucet above a sink. On the left, there is a hanging towel and to the right, a mirror reflecting the dark room.</p> <p><b>Updated Object Descriptions</b></p> <p>: : Object 2: light-colored towel hanging on a rack. : : :</p>
<p><b>Edit prompt</b></p> <p>Blur out the object 5: transparent plastic jar with a white lid containing cotton swabs</p> 	<p><b>Summary of Visual Changes</b></p> <p>In the edited image, the transparent plastic jar with a white lid, which presumably contains cotton swabs, has been blurred compared to its clear appearance in the input image. This modification obscures the details and contents of the jar, aligning with the intention of the instruction. No other significant changes are apparent between the input image and the edited image.</p>	<p><b>AI Evaluation</b></p> <p>The edit appears to be successful based on the provided instruction. The object described in the input image as a "transparent plastic jar with a white lid containing cotton swabs" (#5) has been altered in the edited image to a "translucent bottle possibly made of frosted glass with a narrow neck and a metal cap" (#5), indicating that the object has been blurred as instructed. This change can also be visually discerned when comparing the two images directly: the originally clear plastic jar in the input image is now obscured in the edited image.</p>	<p><b>Updated General Descriptions</b></p> <p>A dimly lit bathroom counter with personal care items scattered around a sink, underneath a hanging towel and a closed overhead cabinet.</p> <p><b>Updated Object Descriptions</b></p> <p>: : Object 5: a translucent bottle possibly made of frosted glass with a narrow neck and a metal cap. : : :</p>
<p><b>Edit prompt</b></p> <p>Change the color of the blue bottle to red.</p> 	<p><b>Summary of Visual Changes</b></p> <p>In the edited image, the color of the previously blue bottle has been changed to red. The other elements in the image remain unaltered compared to the input image.</p>	<p><b>AI Evaluation</b></p> <p>The edit of changing the color of the blue bottle to red in the second image appears to be successful. The general descriptions of both images indicate the presence of a bottle with varying color descriptions—blue in the input image and red in the edited image—which aligns with the instruction given. Furthermore, the object descriptions confirm this change, with the object initially described as a "blue plastic bottle" (#9 in the input image) now being described as a "red cylindrical container with a red cap" (#9 in the edited image), fulfilling the editing task as instructed.</p>	<p><b>Updated General Descriptions</b></p> <p>A dimly lit bathroom vanity with a variety of items including a red bottle, a white container, a silver faucet, medicine bottles, and a dispenser. A towel hangs to the left, and a cabinet is faintly visible above.</p> <p><b>Updated Object Descriptions</b></p> <p>: : Object 9: red cylindrical container with a red cap, likely a personal care product. : : :</p>
<p><b>Edit prompt</b></p> <p>Add the text "Please call 12345 if you can help" to the center top</p> 	<p><b>Summary of Visual Changes</b></p> <p>The edited image has text added to the upper center area that reads "Please call 12345 if you can help". This text is white with a slight shadow effect to enhance visibility against the dark background. No other changes are observed between the input image and the edited image.</p>	<p><b>AI Evaluation</b></p> <p>The edit was successful as per the instruction given. Comparing the visual content of both images, the edited image includes the addition of the text "Please call 12345 if you can help" positioned at the center top, which is in accordance with the instruction provided. The general and specific object descriptions between the input and edited images remain consistent, indicating that no other alterations were made besides the requested text addition.</p>	<p><b>Updated General Descriptions</b></p> <p>A dimly lit bathroom counter with various items including a red bottle, two white bottles, a silver tap, a soap dispenser, and a small container with a tamper-evident lid. A hanging gray towel and a dark mirror are in the background. Text above reads, "Please call 12345 if you can help."</p> <p><b>Updated Object Descriptions</b></p> <p>: : : :</p>

Figure 11: Detailed edit prompts and corresponding verification feedback for the Walkthrough and Session 4 in our study.