From Provenance to Aberrations: Image Creator and Screen Reader User Perspectives on Alt Text for AI-Generated Images

Maitraye Das
Northeastern University
Boston, Massachusetts, USA
ma.das@northeastern.edu

Alexander J. Fiannaca
Google
Seattle, Washington, USA
afiannaca@google.com

Meredith Ringel Morris
Google DeepMind
Seattle, Washington, USA
merrie@google.com

Shaun Kane
Google Research
Boulder, Colorado, USA
shaunkane@google.com

Cynthia L. Bennett
Google Research
New York, New York, USA
clbennett@google.com

ABSTRACT

AI-generated images are proliferating as a new visual medium. However, state-of-the-art image generation models do not output alternative (alt) text with their images, rendering them largely inaccessible to screen reader users (SRUs). Moreover, less is known about what information would be most desirable to SRUs in this new medium. To address this, we invited AI image creators and SRUs to evaluate alt text prepared from various sources and write their own alt text for AI images. Our mixed-methods analysis makes three contributions. First, we highlight creators' perspectives on alt text, as creators are well-positioned to write descriptions of their images. Second, we illustrate SRUs' alt text needs particular to the emerging medium of AI images. Finally, we discuss the promises and pitfalls of utilizing text prompts written as input for AI models in alt text generation, and areas where broader digital accessibility guidelines could expand to account for AI images.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in accessibility**.

KEYWORDS

AI art, Text-to-Image, Accessibility, Alt text, Blind, Screen reader users

ACM Reference Format:
Maitraye Das, Alexander J. Fiannaca, Meredith Ringel Morris, Shaun Kane, and Cynthia L. Bennett. 2024. From Provenance to Aberrations: Image Creator and Screen Reader User Perspectives on Alt Text for AI-Generated Images. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3613904.3642325

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642325

1 INTRODUCTION

In 2021, the DALL-E [2, 72] text-to-image (T2I) generator from OpenAI was released to great fanfare. People's ability to create artistically compelling images from a natural language prompt [68] opened the door to new forms of creative expression, including by people who may not traditionally have created visual art. For example, a creator might prompt a T2I system with the text "A robot family portrait taken in the 1950s," and then receive back several AI-generated images which depict the prompt in different ways, such as a robot family posing in a black-and-white film-like image—a father robot holding a cigar, mother robot wearing an apron, and two robot children. Image generation technology has advanced rapidly in the past few years, with systems like DALL-E 3 [12], Midjourney [6], Stable Diffusion [8], Adobe Firefly [10], and Imagen [75] providing a variety of options for creators. In August 2023, the Everypixel Journal blog estimated that over 15 billion images had been generated within a year [83], signaling extreme growth in the realm of AI image creation.

However, we are unaware of AI-generated images that are inherently accessible to screen reader users (SRUs), although many SRUs are interested in visual media [54, 73, 80, 88]. Indeed, nonvisual access to images is an ongoing accessibility challenge. Despite best practices guiding web accessibility standards for decades, the annual WebAIM survey of the most-visited one million web pages [14] found 30% of images had missing or uninformative alternative text descriptions i.e., alt text, upon which screen reader users rely.[1] T2I systems are no exception to this, as they do not generate alt text for the output images either.

To address image accessibility, researchers have published alt text quality assessments and user needs in different contexts [49, 51, 81, 87]. Researchers have also proposed (and more products now provide) descriptions from different sources (e.g., image captioning AI models [1, 3, 16]). Others have innovated various ways of presenting alt text such as through touch-based exploration [53, 65] and visual question answering (VQA) [13]). While this prior work has led to advancement in image accessibility, AI-generated images contain unique features and are being used in new ways which have

[1] Sometimes "alt text" is used to denote a concise description added into the "alt" HTML attribute of an image uploaded to the web, while "image description" refers to longer, detailed information about an image [17, 48]. We however use these two terms interchangeably irrespective of description length, following prior work that did not make such distinctions and used "alt text", "image descriptions", or "image captions" interchangeably [32, 64, 88].

received little attention in alt text research and may necessitate specific guidance (e.g., unconventional combinations of materials, styles [30] and uncanny content [67]). Additionally, the ease with which AI images can be generated are raising concerns about whether consumers can easily understand image provenance (e.g., if someone captured a photo in real life or created it with AI) [25, 60] to avoid potential misinformation. Although SRUs are at increased risk of consuming misinformation [58, 78, 79], accessible image provenance is still understudied.

Given their emerging state, alt text of AI images are a timely and compelling area of accessibility research. While not focusing on AI images specifically, previous research has leveraged alt text composition and evaluations to uncover how various stakeholders conceive of the quality of these descriptions and how that translates to the content and structure they expect [41, 51, 64, 81, 87]. As such, we recruited both screen reader users (SRUs) who are the primary consumers, and sighted T2I creators to write and evaluate four different versions of alt text of AI images. These are: (1) the text prompts used by creators to generate the images from T2I models; (2) the alt text authored by creators who knew that target readers could not see the images; (3) the alt text written by accessibility experts who regularly read and write alt texts; and (4) the alt text generated by a vision-to-language (V2L) captioning model [31]), which was state-of-the-art at the time of our data collection. Among these four versions, prompts and creator-authored alt text were produced and submitted by the sighted creators before the study along with corresponding AI images. Although alt texts written by experts and generated by V2L models are already being used for image accessibility more broadly, the first two versions we evaluated (prompts and creator-authored alt text) are novel but understudied sources of alt text. Text prompts have the potential to inform alt text at scale. For example, previous research showed that tools could "crawl" the internet and surface text found alongside the same images posted elsewhere (e.g., captions, metadata) to screen readers as alt text [45]. Indeed, in August 2023, Google Workspace automatically populated alt text of AI-generated images with the text prompt the user input to generate the image [18]. Additionally, while content creators seldom write image descriptions [38], T2I interfaces might be augmented to support them to write descriptions as they generate images, as other research on alt text authoring in productivity tools has explored [57]. Taken together, these four versions served as a foundation to explore alt text for T2I images and discuss greater accessibility concerns emerging with this new medium—with both T2I creators and SRUs i.e., alt text readers. Specifically, our analysis was guided by the following research questions:

- ***RQ1:*** *What are the characteristics of alt texts for AI images prepared from different sources (creator-written, expert-written, the T2I prompt, and a V2L model)?*
- ***RQ2:*** *How do creators and screen reader users evaluate alt texts from different sources?*
- ***RQ3:*** *Can text prompts be a good source of alt text for AI images?*
- ***RQ4:*** *What should be described in alt texts for AI images? How, if at all, does this differ from alt text considerations for traditional images?*

Our study makes three contributions to accessibility and human-centered AI research:

- We contribute a set of 64 AI images with their alt texts prepared from four different sources that are useful for comparing across sources and stakeholder groups (e.g., experts, T2I creators, and SRUs).
- We share results from a quantitative and qualitative evaluation of four alt text versions for a subset of the images (32), along with 'ideal' alt text versions written by the image creator and two SRUs.
- We synthesize considerations for making AI images non-visually accessible and discuss future accessibility research directions as generative AI proliferates, such as regarding accessible image provenance and T2I interfaces.

2 RELATED WORK

Below we provide background on AI-generated images and overview image accessibility research motivating our study.

2.1 Background: AI-Generated Images

Generative AI refers to deep learning models capable of generating digital content such as text, image, audio, and video [29]. Early image-focused machine learning efforts centered around image understanding (e.g., object detection, image captioning), though the development of Generative Adversarial Networks (GANs) [36, 42, 43, 50] demonstrated that deep learning models could be trained to produce realistic images. Later, text conditioning [59] allowed users to generate images based on natural language descriptions. This insight, together with advances in large language models (LLMs) (e.g., BERT [34], GPT [69], T5 [70]) and contrastive language-vision models (e.g., CLIP [68]), led to the development of a new generation of text-to-image (T2I) models. These models include DALL·E, DALL·E2, and DALL·E3 [12, 71, 72] from OpenAI, Imagen [76] and Parti [89] from Google Research, Stable Diffusion [8, 74], and Midjourney [6]. Given an input natural language prompt [68], these models can generate a vast range of outputs—from highly realistic to entirely surreal images.

Images generated by these T2I models have led to a range of concerns around the provenance and authenticity of AI-generated content. For instance, the popular term "deep fakes" [86] refers to AI images or videos of people (often celebrities or politicians [22]) which are meant to be passed off as authentic. These concerns have led to research efforts around the detection of deep fake images [25, 35, 60, 90]. Recently, DeepMind has introduced a watermarking technique so that provenance of images generated by the Imagen model could be validated [44].

In parallel, HCI scholars have started exploring how T2I models might support creativity and artistic practices. Researchers investigated T2I model use in collaborative design tasks like preparing slide decks [52]— a timely context of inquiry given T2I's integration into mainstream productivity tools [4, 18]. Others found that creators treated T2I models as an artistic medium and spent significant time "engineering" text prompts [30, 37, 56]. Notably, Huh et al. [47] introduced GenAssist, which focused on making *AI image generation* more accessible to blind and low vision creators by giving comparative descriptions of multiple images outputted from T2I

models so that creators could select images for their content needs or re-engineer the prompt. Our study complements this work by uncovering alt text needs of screen reader users while *consuming AI images* that may be shared or posted online by others.

2.2 Alt Text Research in HCI

Alt text best practices have existed for decades, coinciding with the broader Web Content Accessibility Guidelines (WCAG) [20]; however it is underutilized on the web [14, 38]. Researchers and developers have attempted to scale alt text with different approaches including captioning AI models [3, 16, 88], gathering alt text from different sources [41], crowdsourcing [77], reverse image searches [45], visual question-answering [13], and supporting content creators in authoring alt texts [9, 57].

Importantly, the salience of alt texts to their contexts of use impacts how SRUs rank them [51]. Hence, researchers have distilled what types of information SRUs want in alt texts [64] in what contexts [80, 81]. Others developed targeted guidance for specific use cases such as online shopping [82], scientific papers [49, 87], social media profile pictures [63] and other emerging media like memes [40] and GIFs [39]. Researchers also expanded alt text guidance to describe people's identity (e.g., gender, race) in photographs [24] and fictional representations [48]. Collectively, these studies highlight a tricky balance in alt text generation—enabling users to skip uninteresting details but dive deeper into interesting ones. To this end, Morris et al. [62] proposed a "rich representations" structure for dynamic and multisensory nonvisual experiences of images based on their content and user need. For example, layered descriptions could offer on-demand increasing detail, allow image exploration by touch, and augment text descriptions with non-speech audio and haptics [62]. Recently, researchers have built systems applying similar rich representations and touch-based exploration [53, 65].

A subset of image description guidelines pertains to visual art [15, 54, 66], which overlaps with one of the many emerging applications of AI-generated images called AI or prompt artistry [30]. Further, HCI researchers have developed systems to increase nonvisual access to art [23, 28, 73] for example, by pairing verbal descriptions with musical scores and nature sounds [55, 73]. Guidance for describing art differs from website alt text recommendations , such as by including more detail and describing the art methodically (e.g., from top left to bottom right) [15]. Our paper shares synergy with this work on visual arts accessibility, and expands it to the novel medium of AI images, which are likely to be consumed in many additional contexts.

3 METHODS

We conducted an alt text evaluation study with AI image creators and screen reader users (SRUs) between March – May 2023. Drawing on prior work [51, 57, 64, 81, 87], our mixed methods procedures included numeric scoring, semi-structured interviewing, and an alt text writing task. Our quantitative evaluation revealed patterns in the length and parts of speech of the various alt texts, while the alt text writing exercise and qualitative feedback from participants complemented, enriched and nuanced the quantitative data.

3.1 Participants

We recruited 32 adult, U.S.-based participants who provided informed consent to attend our remote study. First, we recruited 16 creators (denoted as C#) from a network of T2I model users at our institution. Creators submitted five AI images that they had generated (elaborated in Section 3.3). A subset of these images were used for the alt text evaluation tasks (Section 3.5).

Regarding their overall experience with AI image generation, these creators had produced ten to hundreds of images during the month before the study using various T2I models including DALL.E2 [5], Midjourney [6], Imagen [75], and Parti [7], with their usage ranging from 1–12 months (median 8.5, mean 7.25, *sd* 3.66). All creators were sighted. Seven were unfamiliar with alt text, six knew the term but did not have experience writing it, while three had written alt text infrequently for work-related tasks.

Next, we recruited 16 SRUs (denoted as S#) from a pool that had consented to receive information about studies from our institution. All SRUs read alt text regularly. Nine did not know what AI images were, six had read about them in the media or talked about them with friends, and only one had generated images with AI. Two SRUs were prospective AI image creators but reported the T2I interfaces they tried were inaccessible.

Table 5 in the Appendix shows participants' demographic information. Participants received a gift card prorated according to their time spent: 120 and 90 minutes for creators and SRUs, respectively.

3.2 Stimuli: Four Alt Text Versions

As stimuli for the evaluation tasks (details in Section 3.5), we prepared four different versions of alt text for the AI images submitted by the creators.

(1) *Prompt:* We were interested in exploring how text prompts that are inputted into T2I models might inform alt text for their respective image outputs. Hence, in addition to collecting AI images from the creators, we obtained the respective text prompts they had entered to generate those images (details in Section 3.3). These text prompts were used as one version of alt text.
(2) *Creator-Original:* Creators wrote and submitted their own version of alt text for each AI image they submitted to the study (detailed in Section 3.3). Given their contextual knowledge about their intent, creators are well-positioned to write alt text about their own images. However, they often lack knowledge and experience writing alt text [87]. Thus, this version of alt text represented creators' position of familiarity with the image but less familiarity with alt text authoring guidelines. We understood this decision as a validity tradeoff as this then positioned creators to rank their own alt text. However, alt text research lacks perspectives from creators, so we accepted this tradeoff in order to learn more about their alt text writing process.
(3) *Expert:* One alt text variation was written collaboratively by the first and the last authors who had extensive experience both reading and writing alt texts. Following best practices [11, 19], in this alt text version, the experts summarized the most important information first followed by details of the image. They did not read the text prompts beforehand to

avoid bias and gaining knowledge about the creators' intent. Hence, the experts erred on the side of being more comprehensive following Gleason et al. [38]'s rubric of "great" alt text where "almost everything is described." Finally, experts included cultural and media references in alt text when they were depicted in images, in line with prior research on describing such content in memes [40]. In contrast with the Creator-Original alt text, this version represented alt text expertise and unfamiliarity with the image.

(4) *V2L:* Finally, since automated alt texts are proliferating into mainstream products [1], we generated an alt text version with a vision-to-language (V2L) captioning model [31], which was state-of-the-art at the time of our study (March 2023).

Below we describe the procedures completed by all participants—16 creators and 16 SRUs. The creators and SRUs completed somewhat different procedures, so we describe them separately in Sections 3.3 and 3.4. Finally, in Section 3.5, we detail the alt text evaluation tasks that all participants completed during remote sessions conducted on Google Meet. Each session was recorded with the participant's consent and attended by the first and the last authors—one who led the activities and question-answering and another who assisted by taking notes and asking followup questions.

3.3 Creators' Study Procedure

Creators participated in a two-phase study, as described below.

3.3.1 Asynchronous Pre-Work. A few days before joining the alt text evaluation sessions (Section 3.3.2), creators completed one hour of pre-work which consisted of the following activities. They submitted five AI images that they had created, which they consented to be published in this research. Creators were allowed to submit any images that reflected the breadth of their AI image generation practice, and we did not put any restrictions about curating images for particular contexts. We acknowledge that this is a limitation of our study, given the importance of context in image alt text [51, 81]. We accepted this tradeoff to have the opportunity of studying alt text of AI images generated and selected by creators themselves, rather than researchers. However, to instill some consistency in our image set, we required that the submitted images be generated with only one text prompt (e.g., not using negative prompts [37]) and not edited post-production.

Along with the images, creators submitted the respective text prompt, the name of the T2I model used, and an alt text for each of these five images. The alt texts were written by the creators themselves based on brief instructions we summarized from online guidelines [11, 19]. Our instructions clarified that alt text benefits screen reader users who may not be able to see the images. We advised creators to include a high level summary description in the first sentence and that important details should be communicated in subsequent sentences. Finally, our instructions clarified that we were interested in understanding how alt text could convey what the creators themselves thought was important about the image.

Upon receiving the pre-work materials from the creators, we prepared and compiled four alt text versions for each image (described in Section 3.2).[2] We selected four out of the five images submitted by each creator for their evaluation session, removing images that were repetitive (e.g., we removed one if a creator submitted multiple images of nature scenes) or did not comply with our instructions (e.g., a creator submitted a screenshot of a T2I interface).

3.3.2 Alt Text Evaluation Session. We conducted a 60-minute evaluation session with each creator, where they first answered questions about their experience using T2I models followed by completing a series of alt text evaluation tasks for four different AI images (elaborated in Section 3.5). Overall, the 16 creators evaluated alt text variations for 64 (=4×16) images. We showed the images and alt texts to creators using slides and screen sharing.

We concluded the evaluation sessions by asking creators to offer advice to others on writing alt text, and to identify two images that were the easiest and hardest for them to describe. We aggregated these two images from each of the 16 creators to curate a dataset of 32 images for the SRUs' evaluation sessions. Each of the 16 SRUs evaluated alt text variations for four images selected from these 32 images such that each image is evaluated by two SRUs. Thus, by reducing the final dataset, we were able to compare feedback on each image from the creator and two SRUs.

3.4 Screen Reader Users' Study Procedure

Each SRU attended a 90-minute session, where they first answered questions about positive and negative experiences with reading alt text and prior knowledge of AI images. Since AI images were new to most of the SRUs, we read a plain language description of T2I models and text prompts. SRUs then read alt text for a sample AI image. We used the same sample image for all SRUs, which depicted a couch made of potatoes, and demonstrated that AI images may be unrealistic. They then completed the alt text evaluation task elaborated in Section 3.5.

To create a consistent experience irrespective of SRUs' visual disabilities, we did not show the images visually to them during evaluation. Additionally, to mitigate nonvisual access barriers to screen share, we shared a doc of alt text versions directly with SRUs. All participants only had access to the alt text versions for one image at a time; the researcher advanced slides for creators, and removed and replaced alt text versions in the doc when a SRU transitioned to evaluate a new image.

3.5 Evaluation Tasks and Questions

During evaluation sessions, each participant (creator or SRU) evaluated four alt text versions for four different AI images. The four alt text versions for each image were presented at once and labeled 1–4 to hide their sources. The order of the alt text versions were counterbalanced across images to reduce potential order effects. We instructed participants to evaluate the alt text as if it were available on a public website. We deliberately chose this broad context to gather general insights on alt texts for AI images, given the novelty of this medium and the evolving nature of their use cases.

After reading all four alt text versions for an image, participants answered four questions.

(1) What are your first impressions of each alt text?
(2) Compare these four different versions of alt text. What does each do well or not do well to convey the most important aspects of the image to someone who cannot see it?

[2]Images and alt texts are added in supplementary materials, with creators' permission.

(3) Rank each version of alt text from best to worst.
(4) Write your 'ideal' version of alt text for this image.

To write the ideal version, participants could copy from or edit any of the four alt texts and/or write completely new text. We allocated only a few minutes on this task, since our goal was not to produce another alt text commensurate with the other versions but to gather feedback on what information participants considered most important to include in alt text [64, 87].

SRUs completed these evaluation tasks with two notable differences. First, after a SRU finished reading the alt text versions of an image, we asked if they had any questions about the accuracy of alt texts. This access facilitated more parity between SRUs and creators, who could compare alt texts with the image visually, and to prevent SRUs over-trusting alt texts [58]. However, we only allowed SRUs to ask specific questions, not request a comprehensive description. For example, S30 asked if a house was present in I24 since it was listed in the creator and expert alt texts but not in the prompt or V2L alt texts. We confirmed that the image did indeed depict a house. The second difference regarded the ideal alt text writing activity. SRUs had a variety of experience using Google Docs, so we offered them the option to dictate their ideal alt text while one researcher typed and read it back for verification.

We concluded evaluation sessions with debrief questions about participants' most and least favorite alt texts they encountered during the study, how if at all alt text for AI images should be different from alt text for other images, and suggestions for making AI images more accessible. Finally, we revealed the sources of the alt text versions, which elicited additional feedback, particularly on the suitability of prompts to inform alt text generation.

3.6 Data Analysis

We adopted a mixed-methods analysis approach, and thus describe how we analyzed each data type. Our dataset for the quantitative analysis consisted of 32 images that were evaluated by the creator and two SRUs. We analyzed alt text content (e.g., character count and parts of speech) and rankings using Python libraries. These libraries included NLTK [26] for extracting parts of speech tags, NumPy [46] and Pandas [85] for data cleaning and preprocessing, and SciPy [84] for running statistical tests. We used non-parametric tests since our normality tests and histograms revealed that the data was not normally distributed. The exact tests performed (i.e., Kruskal-Wallis, Friedman test, Wilcoxon Signed Rank test, Spearman correlation test) depended on the data type (e.g., ordinal or continuous) and the particular question being answered through the tests, as reported in Sections 4.1 and 4.2. As with relevant prior research [64, 87], the quantitative analysis revealed patterns in participant preferences across the alt text versions, which would have been difficult to derive from the qualitative feedback alone.

In parallel, we qualitatively analyzed session transcripts following a thematic analysis process [27]. Two researchers closely read and coded a non-overlapping subset of transcripts and wrote analytic memos. They first developed codes deductively drawing on research questions and then inductively coded for other factors that participants reported as important in alt text for AI images. We refined the codes through regular group discussions, and developed the broader topics presented in Sections 4.3 and 4.4. Our qualitative analysis also involved closely reviewing alt texts of all 64 images that were evaluated by at least one creator. Like before, two researchers independently examined images and alt text content to categorize the text prompts; they inductively coded for phrases related to topics that came up during the analysis such as types of text prompts, information about provenance, aberration, medium, and style; and then resolved any disagreements through discussion.

4 FINDINGS

Below we organize our findings around the four research questions that guided our exploration: the (1) characteristics and (2) rankings of the alt text versions, (3) the suitability of text prompts to inform alt text generation, and (4) considerations for alt text of AI images.

4.1 RQ1: What are the characteristics of alt texts for AI images prepared from different sources (creator-written, expert-written, the T2I prompt, and a V2L model)?

We analyzed characteristics of alt texts generated for each image: the four 'original' versions prepared before the evaluation sessions (prompt, V2L, expert, and creator-original) and the three 'ideal' versions generated during the evaluation sessions (one creator-ideal and two SRU-ideal alt texts). Figure 1 shows alt text versions for two images.

4.1.1 Alt Text Characteristics. As expected, the content of the alt text differed considerably across their sources. Each alt text version had some distinguishing characteristics: Prompts often included keywords or technical terms, the V2L model generated one-sentence terse descriptions, and the alt texts authored by experts and creators were more detailed.

As noted in prior work [51, 87], quality and preference for alt texts is influenced by the length and amount of detail in those descriptions. Therefore, we calculated the character count (including spaces) of alt text versions. Additionally, to understand how descriptive terms were used in each alt text version [64], we analyzed the percentage of nouns, verbs, adjectives, adverbs, and other words (e.g., preposition, conjunction, pronouns and other parts of speech tags extracted using the NLTK library [26]); see Table 1.

Kruskal-Wallis tests revealed significant differences between alt text versions in terms of character count and percentage of parts of speech. Post hoc analysis with Bonferroni correction revealed that the expert alt text was significantly longer than other versions ($p < 0.001$ for all pairs) except creator-ideal alt text, while the prompt and V2L alt text were significantly shorter than other versions ($p = 0.0007$ for prompt and creator-original, $p < 0.001$ for other pairs, no significant difference between prompt and V2L). Moreover, prompts had a significantly higher percentage of nouns than other versions ($p < 0.001$ for all pairs), potentially because prompts are meant to specify the contents of the desired image, which often contained jargon in order to make the specification more clear to the particular model they were using. Additionally, the significantly higher proportion of adjectives in expert alt text than in other versions (except creator-ideal) may be owed to them having more visual details ($p = 0.003$ for prompt and expert-written, $p < 0.001$ for other pairs).

(a) Image I25.

Prompt: Detailed painting of a strange creature with a body of a hamster and head of a finch.

V2L: A painting of a rodent with a bird beak.

Expert: A drawing of a parrot-otter hybrid animal standing on its hind legs viewed at an angle from the left. The animal has an orange-ish brown otter-like body with a yellow belly and whiskers, and parrot-like short curved gray beak and brown toes on its hind legs. Its short front legs resemble that of an otter but have a parrot-like green color, and it has a gray fish-like short tail. The background is dark gray, graduating to white in the middle. "Model_name" is written near the bottom right corner.

Creator-original: (C11) Painting of a standing hamster-like creature with sharp beak. The background is light gray, creature has mostly brown hair with yellow abdomen. The metallic gray beak resembles one of eagle. Paws are green, legs resemble talons, there is also a tail that looks like second part of a fish body.

Creator-ideal: (C11) A painting of a parrot-gopher hybrid animal standing on its hind legs viewed at an angle from the left. The animal has an orange-ish brown gopher-like body with a yellow belly and whiskers, and parrot-like short curved gray beak and brown toes on its hind legs. Its short front legs resemble that of a gopher but have a parrot-like green color, and it has a gray fish-like short tail. The background is dark gray, graduating to white in the middle. "Model_name" is written near the bottom right corner.

SRU-ideal: (S21) Detailed painting of a strange creature with a body of a hamster and head of a finch.

SRU-ideal: (S31) A drawing of a parrot-otter hybrid animal standing on its hind legs viewed at an angle from the left. The animal has an orange-ish brown otter-like body with a yellow belly and whiskers, and parrot-like short curved gray beak and brown toes on its hind legs. Its short front legs resemble that of an otter but have a parrot-like green color, and it has a gray fish-like short tail. The background is light gray. Image made by "Model_name."

(b) Image I6.

Prompt: cute pygmy hippo. book of kells, lavishly decorated illuminated manuscript. high quality scan, ink on vellum. The decoration combines traditional Christian iconography with the ornate swirling motifs typical of Insular art. Figures of humans, animals and mythical beasts, together with Celtic knots and interlacing patterns in vibrant colors, enliven the manuscript's pages. The pigments for the illustrations included red and yellow ochre, green copper pigment (sometimes called verdigris), indigo, and possibly lapis lazuli.

V2L: A painting of a hippo in a fancy frame.

Expert: Centered is a brown hippopotamus in this primarily green old book-style painting. It has two golden ornamental patches on its body and is standing on an olive green surface. Surrounding the hippo are a couple of oval rings with intricate brown and green chain-like motifs in between. Outside this, another layer of green, brown, and yellow motifs, including some tiny avocados, fill up the rectangular painting. "Model_name" is written near the bottom right corner.

Creator-original: (C6) Illustration of a pygmy hippo framed with ornamental celtic knotwork, in the style of the book of kells, ink on vellum.

Creator-ideal: (C6) Centered is a brown pygmy hippopotamus in this primarily green painting in the style of the 13th century Illuminated manuscript the Book of Kells. The hippo has two golden ornamental patches on its body and is standing on an olive green surface. Surrounding the hippo is a decorative oval frame with intricate brown and green celtic knotwork chain-like motifs in between. Outside this, another layer of green, brown, and yellow motifs fill up the rectangular painting. "Model_name" is written near the bottom right corner.

SRU-ideal: (S19) Illustration of a pygmy hippo framed with ornamental celtic knotwork, in the style of the book of kells, ink on vellum. Generated by AI model "Model_name."

SRU-ideal: (S30) Centered is a brown hippopotamus in this primarily green old book-style painting. It has two golden ornamental patches on its body and is standing on an olive green surface. Surrounding the hippo are a couple of oval rings with intricate brown and green chain-like motifs in between. Outside this, another layer of green, brown, and yellow motifs, including some tiny avocados, fill up the rectangular painting. The decoration combines traditional Christian iconography with the ornate swirling motifs typical in celtic artwork. "Model_name" is written near the bottom right corner.

Figure 1: Images I25 and I6 with different alt text versions.

Table 1: Characteristics of alt text versions. Each cell shows mean value across all images and standard deviation in parentheses.

Alt text version	Character count	% Nouns	% Verbs	% Adjectives	% Adverbs	% Other
Prompt	98.7 (98.2)	48.2 (15.5)	7.2 (7.5)	11.7 (10.7)	0.4 (1.4)	32.5 (15.1)
V2L	46.7 (6.7)	36.1 (7.4)	7.2 (6.8)	7.5 (8.2)	0.4 (2.5)	48.7 (7.7)
Expert	402.8 (156.4)	32.1 (3.7)	12.5 (3.1)	18.1 (4.1)	1.2 (1.6)	36.1 (3.3)
Creator-Original	215.8 (157.8)	34.5 (8.4)	11.1 (5.6)	11.7 (5.4)	1.5 (2.4)	41.2 (6.7)
Creator-Ideal	356.0 (199.3)	33.0 (5.0)	11.8 (4.2)	16.5 (5.4)	1.0 (1.4)	37.8 (4.8)
SRU-Ideal	270.8 (143.8)	36.3 (6.2)	11.5 (4.4)	13.8 (6.3)	1.1 (1.6)	37.4 (5.6)

Table 2: Number of 'ideal' alt texts with the closest edit distance to different 'original' alt text versions.

Ideal version written by	# Closest to Expert	# Closest to Creator-Original	# Closest to V2L	# Closest to Prompt
Creators	17	9	4	2
SRUs	27	20	8	9

Table 3: Average rankings for alt text versions. Lower numbers = higher rank. Standard deviation is in parentheses.

Alt text version	Rank (creator)	Rank (SRU)	Rank (all)
Prompt	3.18 (0.98)	3.03 (1.02)	3.08 (0.98)
V2L	3.22 (0.97)	2.97 (0.92)	3.05 (0.94)
Expert	1.59 (0.67)	1.91 (1.03)	1.80 (0.94)
Creator-Original	2.0 (0.92)	2.09 (1.03)	2.06 (0.99)

4.1.2 Creation of 'Ideal' Alt Text. As shown in Table 1, on average, creators' ideal alt text was significantly longer than the original alt text they provided ($p = 0.003$). That is, creators tended to expand their descriptions after reviewing other alt texts. Comparing parts of speech between creators' original and ideal alt texts, we find the biggest change occurred with an increase from 11.7% to 16.5% adjectives ($p = 0.0007$). This potentially happened because creators gained perspective on visual details that they had missed earlier. C8 noted, *"[Expert alt text] pointed out things that are visible in the picture that I wasn't aware of when I was writing my alt text."* Thus, for creators, exposure to other alt texts materialized what could be described in the image [87].

Additionally, SRU-ideal alt texts were significantly longer than prompt ($p < 0.001$) and V2L ($p < 0.001$) alt texts, but shorter than expert alt texts ($p < 0.001$). The percentage of adjectives in SRU-ideal alt texts (13.8%) were significantly lower than that in expert-written ($p = 0.0007$) but higher than in V2L alt texts ($p < 0.001$). These numbers suggest that SRUs preferred slightly more descriptive language than what the V2L model generated, but not as much as experts included.

When creating the 'ideal' alt texts, in nearly every case, participants adapted one of the existing alt texts. To identify this source alt text, we compared the 'ideal' versions with the four 'original' versions and reported the alt text with the lowest Levenshtein string edit distance (see Table 2). We observed that both creators and SRUs most often based their 'ideal' alt texts on the expert versions, followed by the creator-original versions. On average, the difference between the source and 'ideal' alt text (i.e., the number of characters changed from the source alt text) was 101.7 (SD=74.9) for creators and 90.5 (73.7) for SRUs.

4.2 RQ2: How do creators and SRUs evaluate alt texts from different sources?

Each of the four 'original' alt texts (prompt, V2L, creator-original, and expert) received three rankings: one from the creator and two from SRUs. The 'ideal' alt texts were not ranked, as they were produced after the ranking activity.

4.2.1 Preferred Alt Text Versions. Participants ranked the four 'original' alt text versions of each image from best (1) to worst (4). Table 3 shows the average ranking for each alt text version. To measure variation in preferred alt text, we calculated the average rank given by participants and compared them using a Friedman test (which is appropriate for ordinal, not normally distributed data), separately for the creator and SRU groups. Post hoc analysis was performed using a Wilcoxon Signed-Rank test with Bonferroni correction. We observed a significant difference between ranks for the four alt text versions according to both creators' and SRUs' evaluation ($\chi^2(3) = 24.02, p < 0.01$). In both groups, there were significant differences between creator-original and prompt ($p < 0.01$), creator-original and V2L ($p < 0.01$), expert and prompt ($p < 0.01$), and expert and V2L ($p < 0.01$) alt texts. In summary, both creators and SRUs provided similar average rankings, preferring the expert alt text and creator-original to the prompt or V2L alt texts.

We referred to the qualitative feedback to understand motivations for these rankings and edits. While the two longer versions (creator-original and expert) tended to be ranked higher and to influence the 'ideal' versions, participants had different perspectives on how much was too much verbosity, even though they often preferred more description than the prompts and V2L versions. S22 and S30's responses to I24 (Figure 2) of a house in a nature scene exemplify how each of the longer versions (creator-original and expert) served their needs in different ways, similar to prior research [81]. S22 praised the creator alt text saying, *"It has a pretty good idea of what you're seeing. It describes the house and the sky and the mountains in enough detail that I get the idea."* But S30 considered the creator alt text to be a *"basic overview"* and preferred the expert-authored one because, *"It gives you all of the things that are visible. It describes the house, and it gives more of a description of the surroundings. It puts you more in the picture."*

Moreover, we found that the few prompts and V2L alt texts that were highly ranked tended to contain descriptive or action-oriented

Figure 2: Image I24. Prompt: Fairy-tale like mountain scenery. **V2L:** A lush green hillside with mountains in the background. **Creator-original:** A mountain scenery with red house in the middle. the sky is blue with snow-covered mountain range in the background with green pasture in the foreground. **Expert:** A red hut with a conical roof and a white tower-like structure atop is centered in a green meadow, with a few small green trees in the foreground. The slanted gray roof of another hut is visible at a distance along with snow-covered mountain ranges even further away. The background shows a layer of thick green forest in front of green foothills against a bright blue sky with white clouds. "Model_name" is written near the bottom right corner.

Table 4: Correlation between compositions of alt text and ranking by creators and SRUs

Rater	Char count	% Nouns	% Verbs	% Adjectives	% Adverbs	% Other
Creators	-0.59**	0.32**	-0.25*	-0.33**	-0.35**	0.23*
SRUs	-0.37**	0.19**	-0.29**	-0.08	-0.15*	0.06

*$p < 0.05$, **$p < 0.01$

phrases. For instance, S27 ranked I3's prompt (Figure 3a) the highest, because *"it's poetic and it's a really neat, exciting image description. It has a lot of really powerful image words like 'gigantic fiery wings flapping through the cosmos' and 'eyes burning like the sun.'"*

Overall, we found consistencies in alt text composition and rankings across participants. However, SRUs had strong and differing preferences about verbosity once alt texts exceeded a few sentences: Shorter versions tended to rank higher when their word economy was dominated with description and action-oriented phrases.

4.2.2 Compositions of Preferred Alt Text. While the prior results highlight participant preferences for alt text from different sources, we were also interested to know whether certain attributes of the alt text were associated with a better rank. To that end, we examined the correlation between text composition and overall ranking via Spearman tests which is appropriate for ordinal, not normally-distributed data (see Table 4).

Notably, description length is moderately correlated with rank, indicating that longer text is ranked better. Generally, higher proportions of verbs, adjectives, and adverbs were correlated with better rankings, while higher proportions of nouns and other words were correlated with worse rankings. We posit that the presence of nouns was not necessarily problematic, but that the presence of descriptive words qualifying these nouns was preferred.

4.3 RQ3: Can text prompts be a good source of alt text for AI images?

Given the emergence of T2I models becoming available for everyday use [4, 18], we questioned how, if at all, text prompts could inform or even serve as alt text, since the ability to repurpose a prompt as an alt text could potentially help with the challenge of scaling alt texts [14, 45]. First, and unsurprisingly, participants agreed that prompts differed in purpose from alt text, which meant that prompts often lacked information needed to develop a complete visual imagery of the generated content. S19 explained: *"Describing what I want the model to generate an image of, to me, is not the same thing as the image… So I don't see the actual prompt being used as the alt text."* However, creators reflected on how they built on prompts while writing alt texts for the pre-work activities. C6 shared, *"I was mainly staring at my image and staring at my prompt, and I mostly changed some wording so they sound like a description that a human would read rather than my prompt, and corrected details."* When asked what advice they had for other T2I creators, C6 added, *"You may prompt the AI model with the style of some artist, but when you're writing alt text, you can't leave it 'in the style of.' You'll have to reference what it actually means for the particular image."*

While prompts were not always a suitable alternative for alt text, they often included useful content about the visual properties of the resulting image, as the creators elaborated. Our analysis of prompts resulted in four categories of unique content compared to the other versions of alt texts, which we detail below.

4.3.1 Irrelevant Content. By 'irrelevant' we highlight those prompts that did not contain any phrase that could inform the composition of alt texts for the images. We found three such prompts that all evaluators considered to be *"absolutely useless"* (S32) as alt texts. These prompts were often experimental, such as abstract adages like 'The meaning of life' (I26, Figure 3b) or C2 submitting his last name (I34, Figure 3c) *"to see what AI will draw."* Although irrelevant prompts made up a small sample of those submitted for our study, we note this category since users are continuously experimenting with novel T2I interfaces, shaping new ways of prompt engineering [30, 37, 56], and thus increasing the likelihood of prompts being irrelevant to the image outputs.

4.3.2 Generic Overview. Another category of prompts were *"generic overviews"* (S21, S25). Unlike irrelevant alt texts, these prompts (18/64) specified *"a very high-level description"* (C3) of the generated images, but they lacked even minimal detail, such as I24's prompt, "Fairy-tale like mountain scenery." While these generic prompts were considered to be *"more valuable than not having any alt text"* (S25), participants critiqued that these prompts *"did not actually articulate what's in the image"* (C13). S30 referred to I24's prompt (Figure 2), saying: *"It gives you an idea that it's supposed to*

(a) Image I3. Prompt: A space phoenix with gigantic fiery wings flapping through the cosmos and eyes burning like the sun.

(b) Image I26. Prompt: The meaning of life

(c) Image I34. Prompt: Creator's last name

(d) Image I18. Prompt: Photo of a white fender Stratocaster :: explosion of thick fire smoke paint ink :: psychedelic style :: white background::2 –v 4 –upbeta.

(e) Image I4. Prompt: A renaissance oil painting of happy monkey, head tilted, winking, enjoying an Aperol Spritz.

Figure 3: Five images with prompts.

be idyllic in some way, but it doesn't really give you a lot of context as to how... Given that the [creator- and expert-written alt texts] very specifically note a red house in this image and that it's centered, you know you're missing things in [prompt]." This limitation of prompts as alt texts was due to the nature of T2I models and how these models generate images in an unpredictable manner, often veering off from the prompts. This means that AI images commonly contain content that is not specified in the original prompts but important to include in alt texts.

4.3.3 Undepicted Phrases. Not only do the generated images contain unspecified content, sometimes prompt specifications are not rendered. 10/64 prompts in our dataset included such undepicted phrases. C15 explained, *"The image generation left out some items, so the prompt wouldn't be as accurate... There's additional text that isn't showing up in the actual picture."* While sighted individuals could visually determine which phrases of a prompt had not been rendered by the T2I model, SRUs expressed more confusion upon encountering these undepicted phrases. For example, regarding I18 (Figure 3d) of an ink explosion where the phrase 'white fender stratocaster' mentioned in the prompt was not rendered in the image, S17 commented, *"It doesn't seem like it's the same picture because [other alt texts] don't mention anything about a guitar but the [prompt] says the Stratocaster (a type of guitar)... Clearly those (prompts) are no help."* Thus, while prompts could convey *"what the artist originally would have wanted you to see"* (C15), they fell short in reporting exactly what the T2I model portrayed.

4.3.4 Jargon. A unique characteristic of prompts was that they were often laden with technical and artistic jargon, appearing in half of the prompts in our study (32/64). We considered a term jargon if it referred to a particular style or image specification. These spanned styles (e.g., Van Gogh style, arcane style) to quality (e.g., high res, hyperdetailed), aesthetic (e.g., cybernetic, psychedelic), camera type (e.g., DSLR), and shot specifications (e.g., dynamic pose). Generally, SRUs felt that jargon was distracting and required niche knowledge to understand. Excessive jargon made some prompts entirely unsuitable as alt text (e.g., I18's prompt, Figure 3d). S20 articulated the compounding impact of too much jargon and ingesting alt text through audio on comprehension, *"I would say [prompt] is probably the worst description just because it uses so many words that listening to it with the screen reader, it gets jumbled and confusing."*

Nevertheless, most prompts (29/32) still contained a few jargon terms that participants cited as useful. For instance, S28's ideal alt text for I4 (Figure 3e) closely resembled the prompt which contained some useful jargon ("renaissance oil painting"). We noticed that

SRUs positively received the relatively familiar phrases of image mediums and artistic styles compared to other unusual jargon terms (e.g., "–v4 –upbeta").

Overall, we found mixed reactions to prompts from our participants. If prompts were not irrelevant, they risked being generic overviews, or having inaccurate, jargon-laden details, some of which were not even depicted in the resultant images. Despite these limitations, prompts often reflected the rendered image at a high level and gave context clues which were useful to creators in writing their alt text, and may be useful for other alt text authors when choosing descriptive language.

4.4 RQ4: What should be described in alt texts for AI images? How, if at all, does this differ from alt text considerations for traditional images?

We identified four factors that shaped what information SRUs desired to know and creators wanted to convey in alt texts for AI images: (1) image provenance, (2) T2I aberrations, (3) visual uncertainty and creator intent, and (4) image medium, style, and ambience. Tables 6 and 7 in the Appendix present a comparative summary and examples of how alt text from different sources included information related to these factors.

4.4.1 Provenance. One of the most critical aspects of describing AI images was relaying provenance, i.e., origin of the images, echoing recent responsible AI discussions [25, 60]. Images generated by some models (e.g., DALL-E2) contained provenance information in the form of a visual watermark, although some other models (e.g., Midjourney) did not. Within our dataset, expert alt texts always included descriptions of the watermark if visible on the image. Out of 27 images with a visible watermark, creators included it in 5 'original' versions and in 16 'ideal' versions. SRUs included watermark information in even more, 23/27 images with a watermark.

Although a few SRUs did not find descriptions of watermarks *"necessary for the picture"* (S17), saying that they *"probably wouldn't even pay attention to it"* (S21), most participants were *"very much in favor of those AI-generation call-outs"* (S19) as a matter of transparency. Some creators thought that conveying provenance information is critical to highlight the human (and AI) contribution behind the generated images. C9 explained, *"When I edited this image, I changed that signature [watermark] to include my name… So it's like a co-creation between machine and human. So I want information about who makes what… who the artist was, which algorithm was used… because that's important context."* SRUs shared that knowing watermark information would be especially important while incorporating AI images into their own content or encountering these images in collaborative scenarios (e.g., on social media) so that they could develop a common ground with others. S25 explained, *"If I was going to use this in my own content, I'd want to know that watermark is there… because I could easily post an image and have no clue about a watermark that someone's talking about."* Even those who did not express a strong preference for watermark information in alt text still advocated for it to be described to promote equitable access for blind and low vision people. Others motivated their preference for provenance information in alt texts to contemporary examples of viral AI images propagating misinformation and causing *"damaging experiences"* (C1).

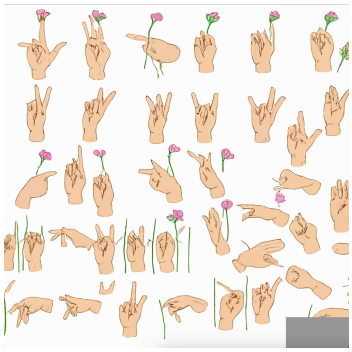
> S22: *"I think they [AI images] have a lot of potential for danger if they're not labeled correctly. I saw an image of Obama and Angela Merkel dancing on the Beach together and it was made by Midjourney and it looked very very realistic. It can lead to fake information out there that can have a really bad impact on our society."*

However, visual watermarks are not always present in AI images, either because the T2I model did not generate one or it was edited out. Even in those cases, certain styles may provide subtle hints to sighted people if an image is AI-generated. For instance, images may exhibit unnaturally smooth and uniform texture, resembling glossy plastic-like skin. Hence, S30 commented that provenance information—irrespective of whether visually obvious due to watermark or deducible from subtle stylistic cues—should be equitably available to all users. They said, *"Whatever's in the visual image should be in the alt text. If it has a watermark that should be in the alt text. If it's very visually clear that it's an AI-generated image, then maybe that can also be included in the alt text."*

Further, SRUs emphasized that provenance declarations should be in plain language so that those who do not have extensive knowledge of T2I models can also easily understand this information. This issue became evident when several SRUs asked questions about the meaning of watermarks or misinterpreted them. For example, regarding the expert alt text of I7 (Figure 4e) featuring a painter robot, S17 remarked, *"I don't know what that (model_name) is… I was thinking, is that the name of the robot (image subject)?... It's kind of confusing … At least let us know that that's just the watermark."* Indeed, some SRUs tried out different provenance phrases while writing their ideal versions. S31 said, *"I certainly believe that there should be some sort of disclosure… It should be even more explicit than just even 'watermark in corner'... I would have typed at the end something like 'Image made by [model_name]."*

In sum, provenance was a critical aspect of AI images; several creators and SRUs added it into their 'ideal' versions. Declaring provenance in an accessible manner could raise awareness for content creators and image consumers, as some AI images may propagate misinformation. Importantly, several SRUs were unfamiliar with AI images and specific model names. Their requests for plain language provenance that does not just mimic what might be visible in the image (e.g., describing the text or logos in a watermark, or visual features associated with AI image generation) is in tension with alt text best practices which caution against interpretive descriptions [15]. Further, since watermarks may be removed from the images, accessible provenance may exceed established alt text standards.

4.4.2 Aberrations and Uncanny Content. T2I models frequently produce "aberrant creations" [67], i.e., images with surrealistic, distorted depictions that subvert familiar presentation of bodies and objects. These AI-generated images are often considered to have 'grotesque' qualities, convey unsettling and nightmarish aesthetics, and cross the "uncanny valley" [61] whereby they closely mimic human (or other familiar objects') features yet fail to match them fully. We observed that uncanny content can be generated both intentionally (e.g., someone prompting to create a cow-frog hybrid



(a) Image I27. Prompt: Sign language flowers.

(b) Image I55. Prompt: 15 fingers on a hand.

(c) Image I13. Prompt: Birthday party for a mouse.

(d) Image I14. Prompt: Four turtles squeezed in a yellow cab with the background of New York City in the style of illustration.

(e) Image I7: Prompt: Thomas Kinkade painting of an anime cyborg painting a picture of itself on a canvas while looking at a mirror behind the canvas.

Figure 4: Five AI images, each with a visible watermark (blurred for anonymity). Images 4a-4d show T2I model aberrations.

animal) or unintentionally due to model flaws (e.g., misshapen human hands and fingers when those are not defined in the prompt). We use the term 'aberration' to refer to unintentional instances of uncanny content, such as fingers in a human hand fused together (I55, Figure 4b) or illegible text (I13, Figure 4c). Experts described aberrations in 12/64 images among which 5 alt texts explicitly called out that those could be possible "aberrations of the AI model." Creators described aberrations in 10/64 alt texts (both 'original' and 'ideal' versions), although they labeled the respective descriptions as "aberrations" in only 4 ideal versions, a decision likely informed by observing the expert-written versions.

Diving deeper into participants' reactions, we found that some SRUs appreciated knowing the presence of *"weird visual things"* (S32) in AI images. Context of use was cited as an important factor governing whether participants perceived aberration descriptions would be useful [51]. C9 provided an example of when describing aberrations would be particularly important, *"There's times when you're showing an image because they are interesting or funny… And so if that image is being shared because of these funny hallucinations, I would hope that alt text reflects that."* Additionally, C1 considered aberration descriptions as a form of provenance, saying, *"Why is there an extra finger there? It might tell you that this image is not real and that it's an AI-generated image."* This example calls back to S30's comment about alt text reflecting subtle aspects of the image other than watermarks that may clue consumers that it was AI-generated (Section 4.4.1). Furthermore, aberration information was particularly important to those SRUs who wanted to use T2I models for their own content creation. They needed to evaluate how outputs and artifacts would make a difference in whether they would like to use a generated image or not. S23 explained, *"I would want to know when I'm generating and when I'm choosing an image. Because it's my brand, and I want to be careful with what I post."*

However, since aberrations were often unrealistic, some SRUs struggled to develop mental images from their respective descriptions. S23 commented about one such call out in I13's (Figure 4c) creator-original alt text, *"The pedestal has a cake on it but it melts into a squashed present — I'm having trouble understanding what that looks like or what that means."* Similarly, creators also found it *"hard to articulate"* (C13) images with uncanny content. C3 reflected on their experience of writing alt text for a made-up movie poster (I17, Figure 6a):

> *"These are things that you might not see normally. And because you can come up with anything with T2I, it may require more description about relative positions, placements, colors, to give you a feel of what this actually looks like as opposed to an image of a well-known movie poster — it would be very easy to describe because the viewer may have a sense already of what this is like through their background."*

Aligning with C3's reflection, S21 shared a preference for extended descriptions of novel, unrealistic content. Regarding I6 (Figure 1a) featuring a hybrid animal, they said, *"If I had no clue what the two animals were listed in it… or if it's something that's completely made up, then I would want more detail… without really being able to compare it to anything."* As these examples illustrate, uncanny content in AI images needs more detailed descriptions.

Some SRUs appreciated aberration descriptions but cautioned about subjective judgments. For instance, although S25 liked the description of uncanny features in I27 (Figure 4a), he considered calling them out as aberrations to be an overreach. He explained, *"The hand shapes are unusual and they don't have clearly recognizable palms or fingers—I think that's plenty of information to explain that to a user. But then adding the information about an AI model, to me, that's just opinion."* This aligns with the fact that while SRUs described aberrations in 3 'ideal' alt texts (out of 4 SRU-evaluated images that had aberrations reported in at least one 'original' alt text), they did not call them out as aberrations.

Moreover, some participants considered aberration descriptions *"useless verbosity"* (S19) because aberrations might make up a relatively small part of the image, and not contribute to its overall meaning. For instance, the expert-written alt text for I14 (Figure 4d) described "The car has a black and white checkerboard design in the front where the license plate and logo should have been." C14 critiqued this description: *"There's too much emphasis on the white checkerboard design whereas I just think that's a glitch in the generation model, so I wouldn't have described it that much."* Thus, expert alt texts brought more attention to oddities in images which, to these participants, drew the focus away from the key points.

Overall, participants were variably interested in aberration descriptions depending on contexts. In cases when they wanted this information, they expressed that descriptions needed to be detailed in case the reader did not have mental models of the uncanny and unrealistic content.

4.4.3 Visual Uncertainty and Creator Intent. The blending of real and imagined content in AI images increases visual uncertainty, and their descriptions may vary depending on individual interpretation. For example, certain artistic styles can make it difficult to identify objects or creatures that bear visual similarity in real-life, particularly when the describer lacks context about the image. I9 (Figure 5a) showed a black-and-white logo of a llama that was interpreted as an alpaca by experts. Other diverging descriptions reflected V2L model limitations, e.g., in I1 (Figure 5b), a drawing of a Japanese village was described as a Chinese village, reflecting known cultural biases in object recognition [33].

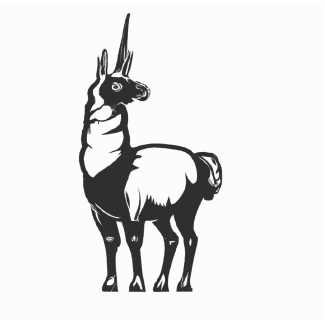
Creators' descriptions of such visual uncertainty reflected the information they had, including their artistic intent and background knowledge. While writing alt text for I13 (Figure 4c), C13 noted, *"It doesn't quite make the mark of being a present because of the aberrations of the AI model. So the [experts] chose to say 'object,' I'm going with 'present'."* Similarly, regarding the animal in I9 (Figure 5a), C9 called the experts' interpretation *"not accurate… Maybe it really is an alpaca but I asked for a llama. So I'm thinking that this is supposed to be a llama, not an alpaca."* Especially for images generated with a specific goal in mind, correctly capturing the creator's intent in alt texts was crucial. C9 added, *"The backstory for this creation is that my friend really likes llamas and unicorns and has a stuffed animal that is a llama unicorn. I was trying to create an image to mimic her stuffed animal and put it on a gift and send it to her. That's why, it was key to the context of how I'm utilizing this."* Some SRUs resonated with this sentiment. S23 said: *"I would lean more toward the intent of what the author wanted to do, because that's what's the important part for me."* Thus, creator intent became an important factor in resolving visual uncertainty and a tie-breaker in choosing one from varied interpretations of AI images.

However, SRUs often *"preferred accuracy over interpretations"* (S23) even if that required using generic terms to describe what's in the image. Regarding I29 (Figure 5c), an object was described as a 'flamethrower', a 'bat', and a 'weapon' in different alt text versions. While authoring their ideal version, S23 picked the term 'weapon,' explaining that *"I could be fine with 'flamethrower.' But I just like [to say] that it's a weapon if it's not explicitly sure that it's a flamethrower."* S23 also chose the more generic term 'drawing' while authoring ideal alt text for I5 (Figure 5d) where the prompt and the expert alt text describe the image to be a 'pencil sketch' and a 'charcoal drawing,' respectively. Similarly, regarding I21 (Figure 5e) where the V2L inferred 'a woman's face' and the creator described it as 'closeup of a face,' S27 commented, *"I would just go for the safe mode and put ['closeup of a face'] because that's the most descriptive that I think I could get and still not mix everything up."*

Yet, sometimes generic or low-level object or scene descriptions made to avoid interpretations led to further confusion among SRUs, especially for complex images requiring dense descriptions. I16's (Figure 5f) expert-authored alt text lacked a straightforward description of the surrounding, which made it difficult for S18 to understand the setting of this image. In contrast, the word 'gym' in the first sentence of the creator-written alt text helped them imagine this setting more easily. S18 explained, *"A cat in shorts is lying on a bench in a gym… that gives me a visual of the gym already. This sentence was very clear and precise of where the cat is at. But then in the [expert] alt text, the last sentence [mentioned] workout equipment. Could [it] be a gym or [not]?"*

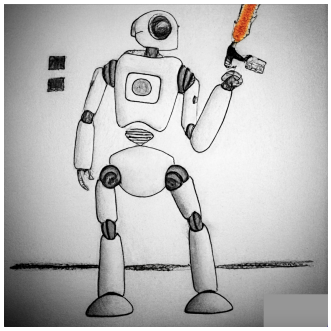
Qualifying language comforted some participants with its promise of conveying visual uncertainty without over-interpreting. Regarding I27 (Figure 4a), S25 said, *"Maybe something that says 'appears to be' can help because… to me, it implies that, 'this kind of looks like sign language but I don't know if they're actually saying something in sign language.'"* Similarly, referring to some figures on I50 (Figure 5g) resembling humans on a beach, C10 commented, *"It looks very uncanny because I'm expecting a person but a person does not have this form, so maybe [saying] 'people-like silhouette' is better."*

Overall, unclear objects or images with content that required domain knowledge to describe led to different interpretations in alt text versions. While some creators leveraged their background knowledge and artistic intent to write descriptions (which were easier to read in some cases), SRUs wanted agency to interpret themselves when they became aware of these different interpretations. However, low-level descriptions could be verbose and confusing, preventing some SRUs from creating a gestalt perception of the image. Hence, some participants suggested qualifying language as an acceptable medium.



(a) Image I9. Prompt: Draw a professional black and white graphic logo of a llama standing with a unicorn horn.

(b) Image I1. Prompt: Arcane style, streets of old Kyoto with a castle in the background in Feudal Japan.



(c) Image I29. Prompt: Humanoid robot with a flamethrower hyperdetailed charcoal drawing.

(d) Image I5. Prompt: Pencil sketch of a headless horseman by Rembrandt.

(e) Image I21. Prompt: Beautiful work of art entitled duality.

(f) Image I16. Prompt: A furry bodybuilder cat doing a bench press. high res dslr photo.

(g) Image I50. Prompt: Scenic sunset on the beach with people doing water activities.

Figure 5: Seven images with prompts, each showing some visual uncertainty that led to varied interpretations.

4.4.4 Image Medium, Style, and Ambience. Web accessibility guidelines recommend against calling out image styles in alt texts [21], because the assumption is the descriptions pertain to photos, making the callout redundant, although some guidelines on describing visual art recommends it [11, 19]. T2I models can generate a variety of image mediums (e.g., photo, painting, drawing, illustration) and styles (e.g., hyperrealistic, pencil sketch, portrait, abstract painting) and blend multiple styles in unusual manners (e.g., unrealistic content portrayed as a realistic photograph) that belie straightforward assumptions around online image properties. Moreover, AI images are used more widely than artistic contexts (e.g., in marketing materials or slides), raising questions of whether this information would be considered important enough to include in those situations.

We found that several participants appreciated when alt texts mentioned the medium and style of AI images. S19 commented about the expert alt text for I6 (Figure 1b), *"I like that it conveys to me that it's not an image of a living hippo. It's not like a still photo but it's an illustration."* Similarly, S25 commented, *"I do appreciate that styling information. Is this a black-and-white image that I might think is an older image? Or is this animated, more playful?"* Interestingly, even some who were initially against medium and style descriptors in alt texts also reflected on the importance of this information upon learning more about the variety of T2I outputs. S21 noted about I25 (Figure 1a), *"I didn't know I had the option to make it look like a photo or drawing… I was just thinking like normal photo, I don't need that [medium/style descriptor] there, because it's repetitive, but if I'm specifically choosing to make it look like a picture or photo or drawing, then that's good to know that it does look that way."*

A closely related factor was the overall 'feel' and ambience of an image. Some creators reflected on their own alt text authoring practice and realized how they leaned towards conveying the *"emotional quality"* (C7) and *"mood or the atmosphere that the image is projecting"* (C5), at times more than they focused on describing intricate details. C3 echoed this sentiment regarding I17 (Figure 6a): *"I noticed that I put a lot of those [words] in there to invoke some feeling about what I feel when I look at these images... I guess I was trying to give the person who would be experiencing this through the alt text some other ways to experience. So that's why I mentioned the image was kind of like a scary movie."* Critiquing expert alt text for I44 (Figure 6b), C7 stated,

> *"It got sort of the details but is kind of missing the point of the image… When I look at images, it's a kind of a depth-first process where you take in the dominant features, first the vibe, the warmth, the colors, maybe the style, and then you start to drill down into the arrangement of things, and then maybe you finally end up with things like words or figures, or what's happening in scenes… And the impressive majesty of this image,*

(a) Image I17. Prompt: A 1970s poster for a scifi robot movie called Table Hunter. **Creator-ideal: (C3)** A movie poster-style image of a blocky, clunky humanoid looking robot with mostly blue and some red and yellow body parts and a dark gray computer screen on the chest area is standing near a yellow table while pointing at it with its long, pointy fingers. One yellow chair is tucked in to the table. "Table hunter" is written in a large yellow shaky font near the top of the image's black background, and off white borders the image's left and right sides. "Model_name" is written near the bottom right corner. The image gives off the feeling that it is a scary movie.

(b) Image I44. Prompt: Sunrise Rugged Landscape Concept Art. **Creator-ideal: (C7)** A warm slightly impressionist painting of a sunrise over a verdant valley framed by large rocky cliffs. The sun is breaking slightly from a gap between the cliffs in the top center of the image. The sky is overcast with layers of clouds graduating from deep blue and purple near the top to orange and yellow near the horizon as if the sun is shining through the valley. A river spirals through a dark boulder field in the foreground. "Model_name" is written near the bottom right corner.

(c) Image I19. Prompt: Sad person in front of a tombstone. **Creator-ideal: (C5)** A figure in a dark hoodie sits in front of a tombstone facing away from the marker, their head in their hand and their face not visible beneath the hood. Their posture suggests sadness or grieving.

Figure 6: Three images with prompts where creator-ideal alt texts contain phrases related to mood and ambiance.

> *it's kind of like Yosemite Valley. You don't start with the color of the leaves or something on a particular tree, you really start at that whole picture."*

SRUs' preferences, however, contradicted creators' inclinations. Several SRUs critiqued alt texts that used subjective phrases (e.g., 'beautiful', 'happy', 'sad', 'fancy') and instead preferred that alt text *"maintain some degree of neutrality"* (S19) in describing image content. S25 critiqued I19's alt text (Figure 6c): *"We don't know that the person's actually sad. That automatic assumption is making me not like this specific alt text."* Similarly, although C3 preferred the term "scary" in the alt text to convey the ambience of I17 (Figure 6a), S31 did not want that interpretative description in their 'ideal' alt text. Instead, they only included descriptions of content that led to the "scary" ambience, i.e., "yellow shaky letters." They explained, *"The idea that it's a scary movie, I don't want that in the description. Because it's unclear where that comes from… I think that the yellow shaky letters are what was sort of giving that mysterious scary vibe. And so that's why I included it in mine (ideal version)."*

In sum, descriptions of medium, style, and ambience were interesting to SRUs, some of whom discovered the variety of images they could create with T2I through the evaluation process. Recent work also highlights the importance of medium, style, and emotions in alt text for SRUs who are interested in creating AI images [47]. However, these descriptions could cross into subjective interpretations, which interfered with SRUs preferences to draw their own conclusions about the overall feel of an image.

5 DISCUSSION

We have thus far shared a characteristics comparison and quality evaluation of alt texts for AI images, developed from different sources—the text prompts, image creators, accessibility experts, and a V2L model. We additionally surfaced unique considerations for alt text of AI images from creators and SRUs, including the suitability of text prompts to inform alt text.

We found that both creators and SRUs ranked the longer creator- and expert-authored alt text higher on average, and drew on those while writing their ideal versions. However, some curt prompts and V2L alt texts were also ranked highly, particularly when they contained descriptive language like the human-authored alt texts and useful jargon denoting image medium and styles. Notably, creators' ideal alt texts were longer than their 'original' versions, suggesting that the expert examples they encountered during evaluation sessions influenced them to add more detail, as examples have influenced the quality of alt text in other research [57].

We also uncovered differences in alt text content and perspectives between creators and SRUs. First, SRUs' ideal versions were shorter than creator-ideal or expert alt texts. These insights suggest that desired details to one participant may seem verbose to another, in line with research that has long called for diverse and rich representations of alt text according to context and user preference [51, 62, 81]. Creators' artistic intent unsurprisingly influenced their descriptions in some cases, adding helpful context for describing visually uncertain objects. However, using creator intent to decide how to interpret visual uncertainty contrasted with SRUs' preferences. Finally, some creators quickly described ambience with

subjective terms whereas SRUs wanted objective descriptions of what was visible so they could make their own judgment.

We additionally drew out unique considerations important to creators and SRUs as the emerging domain of T2I proliferates into mainstream visual media. We learned that text prompts should not be repurposed as alt text, but they often contained specifications that enriched descriptions more than descriptions from sources unfamiliar with the image (e.g., V2L model and experts). Notably, provenance (i.e., how these images were generated) [25, 60], was emphasized by almost all participants as extremely important to convey in an accessible manner. Reasons varied from the right to know, to mitigating misinformation [22, 86], and its importance for blind and low vision content creators to assess the suitability of AI images for their needs. Relatedly, participants pointed out opportunities for alt text to describe aberrations which might be unfamiliar to SRUs who struggled to develop mental models of these unrealistic content. Finally, we highlight features that describe the overall image including medium, style, and ambience as potentially more salient to include in alt text for AI images, which can render in a greater variety than the online photos that SRUs had encountered most frequently.

Below we discuss how these results may inform future research and practice regarding accessible AI images. First, we outline the need for accessible provenance, among the most important and yet unaddressed findings from our study. We follow with opportunities to support alt text production in T2I generation pipelines, and recommendations for updated alt text guidelines to reflect the unique content in AI images.

5.1 Accessible Provenance

Provenance, i.e., information on an image's origin, was among the most discussed topics during our study. While several creators did not include watermark information in their alt texts, upon reading it in expert-authored versions, they often added it to their 'ideal' versions. Some creators however did not recognize the utility of provenance in alt text, rightfully believing that alt text should describe the subjects and actions of the images, rather than metadata including watermarks that happened to be represented visually in the image. However, participants overwhelmingly agreed that provenance should be available to everyone for reasons from interest in what prompts were generating the images for creativity purposes, to combating misinformation.

First, there is an opportunity for content creators and alt text authors to include provenance information in manually-written alt text. Alt text guidelines, which do recommend that text in images be included [21], should clarify that text can include watermarks. We suggest placing this information at the end of alt text with a qualifier that it is a watermark, to reduce confusion around whether the label is related to provenance or the image content, and to be easily skippable when consumers do not need it, such as image galleries for a particular T2I model where their provenance is obvious. In certain cases where a watermark is not visible but the alt text author perceives image features associated with T2I outputs, some SRUs wanted to know these associations; however, others found such information subjective. Alt text could incorporate qualifying language (such as "possibly") before a plain language phrase such as "AI-generated image" at the end of the description. We note that since the alt text itself might be AI-generated, it is important to ensure the phrase clarifies that the provenance declaration refers to the image.

However, simply updating alt text guidelines and relying on content creators is insufficient. There remain open questions about how to communicate provenance information when watermarks are unfamiliar to SRUs who may not have mental models about the practice of watermarking, or who may lack knowledge about specific model names or logos. Further, some T2I models do not add watermarks, and they can be easily edited out. To this end, we recommend that accessibility, provenance, and T2I research collaborate, as new provenance detection tools are released [44]. For example, it isn't clear how provenance detection tools work with screen readers, or whether they will communicate provenance in a manner that does not assume the consumer has knowledge of T2I model names. Participants recommended phrases such as "AI-generated image" or "Image generated by [model_name] AI model" as plain language phrases that did not expect consumers to be familiar with T2I models. Further, as captioning models and LLMs become widely used to generate alt text, there is opportunity for AI developers to work with accessibility researchers to explore how such models could detect and communicate provenance, and differentiate watermarks from other meaningful text in the image. There are additional opportunities for provenance detection tools to become embedded in apps and browsers so all users may surface information about an image's provenance if they wish. This would resolve concerns SRUs had when provenance was not visible in the image and therefore technically outside alt text's purview, and when describers are unfamiliar with AI images who may not realize that visual features are aberrations that could indicate their provenance.

While AI image provenance is of peak importance now as people experiment with T2I, we suspect that its necessity will change over time, and become more or less useful in different contexts. Still, our research has highlighted opportunities for accessibility and consistency in how image credit and provenance are communicated irrespective of what role AI had. For example, news media and photography have an established practice of naming the photographer in the image caption, which is typically perceivable to screen readers. Further, one creator expressed interest in editing the watermark to include their name and the tools they used to refine the image from the T2I output version. Attributing multiple sources for image and content creation could also help to destigmatize stereotypes that automation (e.g., using a T2I model to generate images) is a result of lack of effort, when it increases agency particularly in accessibility use cases [47]. In addition to providing ways for image creators to disclose what tools they used, alt text reading tools might automatically detect the usage of filters or image editing techniques, and offer opt-in information to screen readers.

Finally, while we have focused on AI image provenance, we note there is a broader opportunity and we argue, crucial need, to discover and design accessible provenance and explainability for AI-generated content across mediums. For example, LLM-generated answers to users' chatbot queries could be confused with search results as they are presented in the same text medium. While research on blind and low vision people's strategies for determining

provenance is extremely limited, we know they are at risk of not detecting misinformation [78, 79], and have overtrusted AI-generated image captions [58]. It is reasonable to expect that screen reader users employ popular information inspection tactics such as fact checking reputable sources. However, we have already seen that misinformation warnings on social media may be inaccessible [78] and in the present study, we learned that a popular provenance technique, watermarking, was not perceiveable to screen readers, and some watermark descriptions were incomprehensible to SRUs who were unfamiliar with T2I models. In some cases like AI-generated image captions, SRUs may not have access to sources considered as "ground truth," the images corresponding with the captions in this case. This research could involve many threads only some of which we suggest including understanding existing techniques screen reader users are employing as they use generative AI or sift through search results, the differential impacts on screen reader users when they consume misinformation as compared to people who are visually processing it, and designing provenance and explainability to be both screen reader accessible and comprehensible to users unfamiliar with specific AI models.

5.2 Alt Text for AI-Generated Images

Alt text has long been a digital accessibility priority and research topic (e.g., [20, 21, 32, 38, 41, 65, 80, 81, 88]). Our paper contributes to this research in two ways: by offering comparisons of alt text generated from different sources, and by informing alt text for AI images, a new media that has not received much accessibility attention.

First, by comparing different versions of alt text, we distilled several implications for content creators, alt text authors, and AI developers on AI image content and structure. Participants highlighted types of content they found relevant to AI image descriptions, which may be incorporated into alt text guidelines or model fine tuning. For example, while some guidelines consider declaring the image medium repetitive to information available to a screen reader (e.g., announcing that an element is a graphic), similar to visual art [11, 19], participants wanted to know the medium of AI images since they represented such a variety. SRUs could not assume that images resembled photographs they commonly encountered on social media or news articles. Next, as with other visual media such as memes and GIFs [39, 40], unrealistic content necessitated more detailed descriptions. However, aligning with research on preferred identity descriptions [24], we suspect that as AI images become more common, which types of content need more detailed descriptions will evolve, presenting yet another reason why best practices must be updated regularly.

Next, echoing other research conducted with sighted alt text authors [64, 87], we learned that creators wrote high ranking alt text with minimal instruction. However, additional support is necessary, particularly as visual media evolves, including T2I models. Further, low rates of alt text remain [14], and sighted people still struggle to know what to include, diverging from SRU preferences in some cases [51, 63]. Additionally, creators referred to the four original versions as they composed their ideal alt text, as was found in previous research which used AI-generated captions as examples for alt text writers [57]).

T2I creation tools may suggest alt text in several ways, for example, by ingesting prompts as context clues, asking creators to contextualize prompts and edit alt text suggestions, and summary tools that are tailored to address potential differences in what sighted and blind consumers might think is important to include in alt text (e.g., pointing out subjective language and suggesting qualifying language). Summarization features may be useful to provide information ordering recommendations or efficiency edits.

By comparing alt text from different sources, we explored how text prompts input to create AI images might inform alt text production. While prompts were rife with technical jargon, sometimes worded awkwardly, and did not accurately or comprehensively reflect the output image, some prompts received high rankings from SRUs for their efficient, vivid descriptions. Additionally, they served as a resource for creators composing their initial versions; one reported that they went back and forth between their prompt and our instructions. Moreover, context impacts how even high-quality descriptions are ranked by SRUs [51], yet context authored by the image creator is rarely available. Text prompts may serve as context both for creators and outside sources to write alt text. Several design choices could impact how prompts are used. For example, creators could categorize their prompt which would then set how influential it should be in alt text production, similar to temperature controls users can adjust to control model output. Instead of directly influencing the image output, it would communicate to models or alt text authors how literally the prompt should be taken. For example, experimenters entering names or abstract contents might not suggest that prompts should influence alt text. Similar to the prompt verification component of GenAssist [47], our prompt content categories of irrelevant, generic overview, undepicted phrases, and jargon could scaffold VQA tasks, automatically comparing images to prompts to detect what terms from the prompt should be carried over to the alt text, what terms should be corrected, and which terms, such as jargon, might be substituted with a more common word to describe the style or ambience.

Our insights further allude to the importance of studying different contexts in which alt text readers may encounter AI images, as these images become more prevalent over time [51, 81]. For instance, some SRUs wanted more information about aberrations if the images were shared as examples of *"funny hallucinations"* or for quality checking purposes if they were creating the images themselves. Additionally, SRUs who were content creators were keen to receive provenance information, though this would be most important if they were selecting from pre-generated images than if they were using T2I models themselves [47]. Future research is necessary to understand the extent to which contextualized preferences from prior work [51, 81] apply to AI images. We posit there will be some overlap, but that differences might arise in higher stakes situations. This might be explored by investigating a range of contexts based on how much the image source and accuracy matters (e.g., users might want to ensure the images accurately depict products or that they are from a reputable sources for images used in the context of marketing and education).

Finally, as with other studies [41, 45], we found that showing alt text from multiple sources could be useful, particularly when SRUs are concerned about accuracy. In our study, SRUs used different versions to develop their own mental pictures, and differences

among them encouraged skepticism [47, 53]. These inaccuracies not only occurred in V2L captions, but also in prompt specifications not rendering in output images and differing interpretations between creator and expert-written alt texts. This raises an opportunity for research on risks of over-trusting information due to access barriers which has thus far focused on AI-generated captions [58], to expand to ensuring all information can be inspected accessibly. In the near term, new image exploration features could assist SRUs to consume alt text from multiple sources to their advantage. Summarization could be used to create more efficient descriptions, and could surface similarities and differences among descriptions, as done in recent work [47]. However, consuming multiple sources of alt text, even if summarized, still places extra burden on alt text readers when sighted people can quickly detect differences in text translations of visual content. Future research is necessary to innovate ways to improve alt text accuracy and lowering the burden to accessibly inspect alt texts.

5.3 Limitations

We acknowledge that sighted creators evaluating their own prompts and alt texts might have biased our results. However, in doing so, creators could elaborate on how their original intent behind creating the image influenced their prompts and alt texts in comparison to other versions of alt texts. We chose this tradeoff since creator perspectives are under-represented in alt text research. Another limitation of our study is that we analyzed alt text needs for AI images more broadly given the novelty of this medium, but did not systematically capture SRUs' alt text needs in particular contexts [51, 81], which is an important avenue of future work. Moreover, the V2L alt text in our study, although generated by a state-of-the-art model at the time of our data collection, was limited to short one-sentence description only. We encourage future researchers to explore more recent sources of alt text e.g., LLM-powered long-form descriptions that have become publicly available on platforms like Be My AI [13], Google Bard, ChatGPT Plus.

6 CONCLUSION

AI images are proliferating online. Yet they risk perpetuating a long trend of visual media remaining inaccessible to blind people, as current state-of-the-art models do not generate alt text or encourage creators to do so. Hence, we invited creators and screen reader users (SRUs) to evaluate and write their own alt text of AI images. We found that creators wrote high-ranking alt text with limited instructions, but their alt texts still misaligned with SRUs' preferences in some aspects, such as creators' tendency to use subjective language to describe ambience. We also learned that while text prompts may be useful for alt text generation, their ability to inform quality alt text composition varies greatly depending on how experimental and jargon-laden they are. In turn, we recommended new alt text guidelines and interface designs to encourage alt text authorship, as well as new research questions to ensure image provenance and novel visual elements, such as aberrations and new combinations of materials and styles are accessible. We are optimistic that this research will promote blind people's inclusion in novel visual media, including AI images, as not only consumers but as agentive creators.

REFERENCES

[1] 2018. *Powered by AI: Automatic alt text to help the blind 'see' Facebook.* Retrieved September 9, 2023 from https://tech.facebook.com/artificial-intelligence/2018/06/using-artificial-intelligence-to-help-blind-people-see-facebook/

[2] 2021. *DALL·E: Creating images from text.* Retrieved September 6, 2023 from https://openai.com/research/dall-e

[3] 2021. *Use VoiceOver Recognition on your iPhone or iPad.* Retrieved September 9, 2023 from https://support.apple.com/en-us/HT211899

[4] 2022. *AI-generated imagery is the new clip art as Microsoft adds DALL-E to its Office suite.* Retrieved December 9, 2023 from https://www.theverge.com/2022/10/12/23400270/ai-generated-art-dall-e-microsoft-designer-app-office-365-suite

[5] 2022. *DALL-E 2.* Retrieved December 6, 2023 from https://openai.com/dall-e-2

[6] 2022. *Midjourney.* https://www.midjourney.com/home/

[7] 2022. *Parti text-to-image model.* https://sites.research.google/parti/

[8] 2022. *Stable Diffusion Launch Announcement.* Retrieved September 6, 2023 from https://stability.ai/blog/stable-diffusion-announcement

[9] 2022. *We're making images on Twitter more accessible. Here's how.* Retrieved September 9, 2023 from https://blog.twitter.com/en_us/topics/product/2022/making-images-twitter-more-accessible

[10] 2023. *Adobe Unveils Firefly, a Family of new Creative Generative AI.* Retrieved September 9, 2023 from https://news.adobe.com/news/news-details/2023/Adobe-Unveils-Firefly-a-Family-of-new-Creative-Generative-AI/default.aspx

[11] 2023. *Alt text for accessibility. Level Access.* Retrieved December 6, 2023 from https://www.levelaccess.com/blog/alt-text-for-accessibility/

[12] 2023. *DALL-E 3.* Retrieved December 6, 2023 from https://openai.com/dall-e-3

[13] 2023. *Introducing: Be My AI.* Retrieved September 9, 2023 from https://www.bemyeyes.com/blog/introducing-be-my-ai

[14] 2023. *The WebAIM Million.* Retrieved September 6, 2023 from https://webaim.org/projects/million/

[15] n.d. *Collection Image Descriptions: National Gallery of Art.* Retrieved December 9, 2023 from https://www.nga.gov/visit/accessibility/collection-image-descriptions.html

[16] n.d. *Get image descriptions on Chrome.* Retrieved September 9, 2023 from https://support.google.com/chrome/answer/9311597

[17] n.d. *How to Write Alt Text and Image Descriptions for the visually impaired.* Retrieved December 6, 2023 from https://www.perkins.org/resource/how-write-alt-text-and-image-descriptions-visually-impaired/

[18] n.d. *Now Available: Duet AI for Google Workspace.* Retrieved December 9, 2023 from https://workspace.google.com/blog/product-announcements/duet-ai-in-workspace-now-available

[19] n.d. *Web Accessibility Initiative: Tips and Tricks in Images Tutorial.* Retrieved December 6, 2023 from https://www.w3.org/WAI/tutorials/images/tips/

[20] n.d. *Web Content Accessibility Guidelines (WCAG).* Retrieved September 6, 2023 from https://www.w3.org/WAI/standards-guidelines/wcag/

[21] n.d. *WebAIM: Alternative Text.* Retrieved September 11, 2023 from https://webaim.org/techniques/alttext/

[22] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. 2019. Protecting World Leaders Against Deep Fakes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.*

[23] Dragan Ahmetovic, Nahyun Kwon, Uran Oh, Cristian Bernareggi, and Sergio Mascetti. 2021. Touch Screen Exploration of Visual Artwork for Blind People. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21).* ACM, 2781–2791. https://doi.org/10.1145/3442381.3449871

[24] Cynthia L. Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P. Bigham, Anhong Guo, and Alexandra To. 2021. "It's Complicated": Negotiating Accessibility and (Mis)Representation in Image Descriptions of Race, Gender, and Disability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21).* ACM, Article 375, 19 pages. https://doi.org/10.1145/3411764.3445498

[25] Aparna Bharati, Daniel Moreira, Patrick J. Flynn, Anderson de Rezende Rocha, Kevin W. Bowyer, and Walter J. Scheirer. 2021. Transformation-Aware Embeddings for Image Provenance. *IEEE Transactions on Information Forensics and Security* 16 (2021), 2493–2507. https://doi.org/10.1109/TIFS.2021.3050061

[26] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[27] Virginia Braun and Victoria Clarke. 2021. *Thematic Analysis: A Practical Guide.* Sage Publications, London.

[28] Matthew Butler, Erica J Tandori, Vince Dziekan, Kirsten Ellis, Jenna Hall, Leona M Holloway, Ruth G Nagassa, and Kim Marriott. 2023. A Gallery In My Hand: A Multi-Exhibition Investigation of Accessible and Inclusive Gallery Experiences for Blind and Low Vision Visitors. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (<conf-loc>, <city>New York</city>, <state>NY</state>, <country>USA</country>, </conf-loc>) *(ASSETS '23).* ACM, Article 9, 15 pages. https://doi.org/10.1145/3597638.3608391

[29] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun. 2023. A Comprehensive Survey of AI-Generated Content (AIGC): A History

of Generative AI from GAN to ChatGPT. *arXiv* (2023). https://arxiv.org/abs/2303.04226

[30] Minsuk Chang, Stefania Druga, Alexander J. Fiannaca, Pedro Vergani, Chinmay Kulkarni, Carrie J Cai, and Michael Terry. 2023. The Prompt Artists. In *Proceedings of the 15th Conference on Creativity and Cognition* (Virtual Event, USA) *(C&C '23)*. ACM, 75–87. https://doi.org/10.1145/3591196.3593515

[31] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Alexander Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. PaLI: A Jointly-Scaled Multilingual Language-Image Model. In *International Conference on Learning Representations*. https://arxiv.org/abs/2209.06794

[32] Sanjana Shivani Chintalapati, Jonathan Bragg, and Lucy Lu Wang. 2022. A Dataset of Alt Texts from HCI Publications: Analyses and Uses Towards Producing More Descriptive Alt Texts of Data Visualizations in Scientific Papers. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. ACM, Article 30, 12 pages. https://doi.org/10.1145/3517428.3544796

[33] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does Object Recognition Work for Everyone?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[34] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). ACL, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[35] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The DeepFake Detection Challenge (DFDC) Dataset. *arXiv* (2020). https://arxiv.org/abs/2006.07397

[36] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. 2017. CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms. *arXiv* (2017). https://arxiv.org/abs/1706.07068

[37] Yingchaojie Feng, Xingbo Wang, Kam Kwai Wong, Sijia Wang, Yuhong Lu, Minfeng Zhu, Baicheng Wang, and Wei Chen. 2023. PromptMagician: Interactive Prompt Engineering for Text-to-Image Creation. *IEEE Transactions on Visualization and Computer Graphics* (2023). https://arxiv.org/abs/2307.09036

[38] Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M. Kitani, and Jeffrey P. Bigham. 2019. "It's Almost like They're Trying to Hide It": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. ACM, 549–559. https://doi.org/10.1145/3308558.3313605

[39] Cole Gleason, Amy Pavel, Himalini Gururaj, Kris Kitani, and Jeffrey Bigham. 2020. Making GIFs Accessible. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. ACM, Article 24, 10 pages. https://doi.org/10.1145/3373625.3417027

[40] Cole Gleason, Amy Pavel, Xingyu Liu, Patrick Carrington, Lydia B. Chilton, and Jeffrey P. Bigham. 2019. Making Memes Accessible. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, PA, USA) *(ASSETS '19)*. ACM, 367–376. https://doi.org/10.1145/3308561.3353792

[41] Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M. Kitani, and Jeffrey P. Bigham. 2020. Twitter A11y: A Browser Extension to Make Twitter Images Accessible. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. ACM, 1–12. https://doi.org/10.1145/3313831.3376728

[42] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[43] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative Adversarial Networks. *Commun. ACM* 63, 11 (oct 2020), 139–144. https://doi.org/10.1145/3422622

[44] Sven Gowal and Pushmeet Kohli. 2023. *Identifying AI-generated images with SynthID*. Retrieved September 6, 2023 from https://www.deepmind.com/blog/identifying-ai-generated-images-with-synthid

[45] Darren Guinness, Edward Cutrell, and Meredith Ringel Morris. 2018. Caption Crawler: Enabling Reusable Alternative Text Descriptions Using Reverse Image Search. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, 1–11. https://doi.org/10.1145/3173574.3174092

[46] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature* 585, 7825 (Sept. 2020), 357–362. https://doi.org/10.1038/s41586-020-2649-2

[47] Mina Huh, Yi-Hao Peng, and Amy Pavel. 2023. GenAssist: Making Image Generation Accessible. In *Proceedings of the ACM Symposium on User Interface Software and Technology* (San Francisco, California, USA) *(UIST '23)*. ACM. https://arxiv.org/abs/2307.07589

[48] Emory James Edwards, Kyle Lewis Polster, Isabel Tuason, Emily Blank, Michael Gilbert, and Stacy Branham. 2021. "That's in the Eye of the Beholder": Layers of Interpretation in Image Descriptions for Fictional Representations of People with Disabilities. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. ACM, Article 19, 14 pages. https://doi.org/10.1145/3441852.3471222

[49] Crescentia Jung, Shubham Mehta, Atharva Kulkarni, Yuhang Zhao, and Yea-Seul Kim. 2022. Communicating Visualizations without Visuals: Investigation of Visualization Alternative Text for People with Visual Impairments. *IEEE Transactions on Visualization and Computer Graphics* 28, 01 (jan 2022), 1095–1105. https://doi.org/10.1109/TVCG.2021.3114846

[50] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv* (2018). https://arxiv.org/abs/1710.10196

[51] Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022. Context Matters for Image Descriptions for Accessibility: Challenges for Referenceless Evaluation Metrics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. ACL, Abu Dhabi, United Arab Emirates, 4685–4697. https://doi.org/10.18653/v1/2022.emnlp-main.309

[52] Chinmay Kulkarni, Stefania Druga, Minsuk Chang, Alex Fiannaca, Carrie Cai, and Michael Terry. 2023. A Word is Worth a Thousand Pictures: Prompts as AI Design Material. *arXiv* (2023). https://arxiv.org/abs/2303.12647

[53] Jaewook Lee, Yi-Hao Peng, Jaylin Herskovitz, and Anhong Guo. 2021. Image Explorer: Multi-Layered Touch Exploration to Make Images Accessible. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. ACM, Article 69, 4 pages. https://doi.org/10.1145/3441852.3476548

[54] Franklin Mingzhe Li, Lotus Zhang, Maryam Bandukda, Abigale Stangl, Kristen Shinohara, Leah Findlater, and Patrick Carrington. 2023. Understanding Visual Arts Experiences of Blind People. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. ACM, Article 60, 21 pages. https://doi.org/10.1145/3544548.3580941

[55] Huaigu Li, Jon Bellona, Leslie Smith, Amy Bower, and Jessica Roberts. 2023. "Let the Volcano Erupt!": Designing Sonification to Make Oceanography Accessible for Blind and Low Vision Students in Museum Environment. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (<conf-loc>, <city>New York</city>, <state>NY</state>, <country>USA</country>, </conf-loc>) *(ASSETS '23)*. ACM, Article 79, 6 pages. https://doi.org/10.1145/3597638.3614482

[56] Vivian Liu and Lydia B Chilton. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. ACM, Article 384, 23 pages. https://doi.org/10.1145/3491102.3501825

[57] Kelly Mack, Edward Cutrell, Bongshin Lee, and Meredith Ringel Morris. 2021. Designing Tools for High-Quality Alt Text Authoring. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. ACM, Article 23, 14 pages. https://doi.org/10.1145/3441852.3471207

[58] 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)*. ACM, 5988–5999. https://doi.org/10.1145/3025453.3025814

[59] Elman Mansimov, Emilio Parisotto, Jimmy Ba, and Ruslan Salakhutdinov. 2016. Generating Images from Captions with Attention. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1511.02793

[60] Daniel Moreira, Aparna Bharati, Joel Brogan, Allan Pinto, Michael Parowski, Kevin W. Bowyer, Patrick J. Flynn, Anderson Rocha, and Walter J. Scheirer. 2018. Image Provenance Analysis at Scale. *IEEE Transactions on Image Processing* 27, 12 (2018), 6109–6123. https://doi.org/10.1109/TIP.2018.2865674

[61] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. 2012. The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100. https://doi.org/10.1109/MRA.2012.2192811

[62] Meredith Ringel Morris, Jazette Johnson, Cynthia L. Bennett, and Edward Cutrell. 2018. Rich Representations of Visual Content for Screen Reader Users. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, 1–11. https://doi.org/10.1145/3173574.3173633

[63] Martez E Mott, John Tang, and Edward Cutrell. 2023. Accessibility of Profile Pictures: Alt Text and Beyond to Express Identity Online. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. ACM, Article 49, 13 pages. https://doi.org/10.1145/3544548.3580710

[64] Annika Muehlbradt and Shaun K. Kane. 2022. What's in an ALT Tag? Exploring Caption Content Priorities through Collaborative Captioning. *ACM Transactions on Accessible Computing (TACCESS)* 15, 1, Article 6 (mar 2022), 32 pages. https://doi.org/10.1145/3507659

[65] Vishnu Nair, Hanxiu 'Hazel' Zhu, and Brian A. Smith. 2023. ImageAssist: Tools for Enhancing Touchscreen-Based Image Exploration Systems for Blind and Low Vision Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. ACM, Article 76, 17 pages. https://doi.org/10.1145/3544548.3581302

[66] Royal National Institute of the Blind (RNIB) and Vocaleyes. 2003. *The Talking Images Guide. Museums,galleries and heritage sites: Improving access for blind and partially sighted people.* Retrieved December 9, 2023 from https://www.familyarts.co.uk/wp-content/uploads/2014/12/Talking_Images_Guide_-_PDF_File.pdf

[67] Jennifer O'Meara and Cáit Murphy. 2023. Aberrant AI creations: co-creating surrealist body horror using the DALL-E Mini text-to-image generator. *Convergence* 29, 4 (2023), 1070–1096. https://doi.org/10.1177/13548565231185865

[68] 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. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[69] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. *Language Models are Unsupervised Multitask Learners.* Technical Report. https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf

[70] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[71] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv* (2022). https://arxiv.org/abs/2204.06125

[72] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8821–8831. https://proceedings.mlr.press/v139/ramesh21a.html

[73] Kyle Rector, Keith Salmon, Dan Thornton, Neel Joshi, and Meredith Ringel Morris. 2017. Eyes-Free Art: Exploring Proxemic Audio Interfaces For Blind and Low Vision Art Engagement. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3, Article 93 (sep 2017), 21 pages. https://doi.org/10.1145/3130958

[74] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

[75] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. 2022. *Imagen: Unprecedented photorealism × deep level of language understanding.* Retrieved September 9, 2023 from https://imagen.research.google/

[76] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=08Yk-n5l2Al

[77] Elliot Salisbury, Ece Kamar, and Meredith Morris. 2017. Toward Scalable Social Alt Text: Conversational Crowdsourcing as a Tool for Refining Vision-to-Language Technology for the Blind. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 5, 1 (Sep. 2017), 147–156. https://doi.org/10.1609/hcomp.v5i1.13301

[78] Filipo Sharevski and Aziz Zeidieh. 2021. "I Just Didn't Notice It:" Experiences with Misinformation Warnings on Social Media amongst Users Who Are Low Vision or Blind. https://acal.cdm.depaul.edu/wp-content/uploads/2023/09/Paper-2-I-Just-Didnt-Notice-It-NSPW-2023.pdf

[79] Filipo Sharevski and Aziz Zeidieh. 2023. Designing and Conducting Usability Research on Social Media Misinformation with Low Vision or Blind Users. In *Proceedings of the 16th Cyber Security Experimentation and Test Workshop* (Marina del Rey, CA, USA) *(CSET '23)*. ACM, 75–81. https://doi.org/10.1145/3607505.3607525

[80] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. ACM, 1–13. https://doi.org/10.1145/3313831.3376404

[81] 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)*. ACM, Article 16, 15 pages. https://doi.org/10.1145/3441852.3471233

[82] Abigale J. Stangl, Esha Kothari, Suyog D. Jain, Tom Yeh, Kristen Grauman, and Danna Gurari. 2018. BrowseWithMe: An Online Clothes Shopping Assistant for People with Visual Impairments. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (Galway, Ireland) *(ASSETS '18)*. ACM, 107–118. https://doi.org/10.1145/3234695.3236337

[83] Alina Valyaeva. 2023. *AI Has Already Created As Many Images As Photographers Have Taken in 150 Years. Statistics for 2023.* Retrieved September 9, 2023 from https://journal.everypixel.com/ai-image-statistics

[84] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[85] Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman (Eds.). 56 – 61. https://doi.org/10.25080/Majora-92bf1922-00a

[86] Lucas Whittaker, Tim C. Kietzmann, Jan Kietzmann, and Amir Dabirian. 2020. "All Around Me Are Synthetic Faces": The Mad World of AI-Generated Media. *IT Professional* 22, 5 (2020), 90–99. https://doi.org/10.1109/MITP.2020.2985492

[87] Candace Williams, Lilian de Greef, Ed Harris, Leah Findlater, Amy Pavel, and Cynthia Bennett. 2022. Toward Supporting Quality Alt Text in Computing Publications. In *Proceedings of the 19th International Web for All Conference* (Lyon, France) *(W4A '22)*. ACM, Article 20, 12 pages. https://doi.org/10.1145/3493612.3520449

[88] 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)*. ACM, 1180–1192. https://doi.org/10.1145/2998181.2998364

[89] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=AFDcYJKhND

[90] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. 2020. WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. ACM, 2382–2390. https://doi.org/10.1145/3394171.3413769

A APPENDIX

Received 14 September 2023; revised 12 December 2023; accepted 19 January 2024

Table 5: Participants' demographic information shown on an aggregate level to maintain anonymity.

Category		Creators (count)	SRUs (count)
Gender	Male	11	8
	Female	5	8
Age (years)	18–24	0	2
	25–34	6	4
	35–44	8	5
	45–54	2	5

Table 6: Comparative summary and examples showing how alt text from different sources presented important information related to AI images, as described by participants. The 'Creator' column reports information about both creator-original and creator-ideal alt texts to reduce redundancy. The count values are based on the 32 images that all participants—creators and SRUs—evaluated.

Factor	Prompt	V2L	Expert	Creator	SRU-ideal
Provenance	Not described but jargon (e.g., cybernetic, –upbeta) and unusual sentence structure implied AI generation.	Not described	Described watermark if present i.e., in 27/32 images. Example: "Model_name is written near the bottom right corner."	Described watermark in 5 'original' and 16 'ideal' versions out of 32 images (27 with watermark). Example: For I9, C9 added, "At the bottom right, there are 5 colored squares, the signature for Model_name."	Described AI generation in plain language in 23/32 images; all 23 had visible watermark described by other alt texts. Example: S31 described I7, I17, and I25 as "Image made by model_name."
Aberrant and uncanny content	Not described	Not described	Described in 12/32 images among which 5 explicitly called those out as possible "aberrations of the AI model." Example: I13's expert alt text said, "a birthday-style multi-colored banner that says "KIRRY ARIHOA" with some illegible letters, a possible aberration of AI models."	Described in 10/32 images (both 'original' and 'ideal' versions); only 4 'ideal' versions called these out as aberrations. Example: In I13's 'original' alt text, C13 said, "Some of the edges of the drawing are smudged or slightly stretched out of shape… A letter banner above the mouse reads "DIRRY ARIHIOA"." In the 'ideal' version, they added, "a possible aberration of AI model used to generate this image." Creators sometimes considered aberration information as unnecessary and irrelevant to the images' key point.	Described in 3 alt texts, although none called them out as aberrations. Example: S29's alt text for I13 mirrors the description from C13's 'original' version, not the ideal version which called out aberrations explicitly. SRUs felt that aberrations and uncanny content are difficult to visualize, require a longer description, and are unnecessary information except in certain contexts (e.g., if the image is created or shared by the SRU for their own work, or used as a funny example of T2I hallucinations.)
Visual uncertainty	Not described	Not described	Used qualifying language to describe visually uncertain content. Example: Experts described I25's animal to have an "otter-like body" and "parrot-like short curved gray beak."	Sometimes used qualifying language for visually uncertain content. Example: C11 described I25's animal to be a "hamster-like creature with sharp beak."	Preferred accurate description (even if generic) over specific but potentially inaccurate interpretations. Example: In I29, S23 chose 'weapon' instead of 'flamethrower' or 'bat' to describe an object, because they preferred *"to say that it's a weapon if it's not explicitly sure that it's a flamethrower."* SRUs also sometimes used qualifying language. Example: S25 described I27 as showing "sign language-like gestures" to imply that *"this kind of looks like sign language but I don't know if they're actually saying something in sign language."*
Creator intent	Indicated creator intent. Example: I9 prompt says, "Draw a professional black and white graphic logo of a llama standing with a unicorn horn."	Not described	Not described	Indicated creator intent, which sometimes was a determining factor in choosing between multiple interpretations of visually uncertain content. Example: In I9, C9 considered the experts' interpretation of alpaca to be incorrect, explaining that *"I asked for a llama. So I'm thinking that this is supposed to be a llama, not an alpaca."*	Not described. Some SRUs valued knowing creator intent if available. Example: S23 said, *"I would lean more toward the intent of what the author wanted to do."*

Table 7: Continued from Table 6. Comparative summary and examples showing how alt text from different sources presented important information related to AI images, as described by participants. The 'Creator' column reports information about both creator-original and creator-ideal alt texts to reduce redundancy. The count values are based on the 32 images that all participants—creators and SRUs—evaluated.

Factor	Prompt	V2L	Expert	Creator	SRU-ideal
Image medium	Mentioned in 16/32 images. Example: photo, infographic, poster, pixel art, oil painting, pencil sketch, charcoal drawing.	Mentioned in 18/32 images. Example: painting drawing, photos, pictures.	Mentioned in 21/32 images. Example: painting, pixel art, clip art, charcoal/graphite pencil drawing, photograph, studio shot, illustration.	Mentioned in 17 'original' versions and 24 'ideal' versions out of 32 images. Example: drawing, painting, pencil sketch, illustration, photo, pixel art	Mentioned in 19/32 images. Example phrases: drawing, painting, pencil sketch, photograph, illustration, studio shot, pixel art, charcoal/graphite pencil drawing, oil painting
Style	Mentioned in 20/32 images. Example: arcane style, psychedelic style, Van Gogh style, cinematic, photorealistic, anthropomorphic, renaissance, graphic logo, black-and-white, high resolution, high quality scan, highly detailed, isometric render, octane render, insular art, Christian iconography	Mentioned in 5/32 images. Example: cartoon image, black-and-white.	Mentioned in 9/32 images. Example: black-and-white, book-style, abstract, cartoon style, closeup.	Mentioned in 9 'original' and 11 'ideal' versions out of 32 images. Example: style of anime, movie poster-style, renaissance style, style of the 13th century, style similar to Rembrandt, black-and-white, abstract, photorealistic, graphic logo, closeup, impressionist.	Mentioned in 11/32 images. Example phrases: detailed, closeup, Van Gogh style, abstract, black-and-white, style similar to Rembrandt, renaissance portrait, movie poster-style, anime style, book-style, photorealistic. Some SRUs were initially against image medium and style descriptors but appreciated this information upon learning more about T2I variations.
Ambience	Conveyed ambience and subjective emotions in several images. Example: In I19 "sad person", in I4 "happy monkey", in I18 "happy and friendly sloth", and in I22 "beautiful landing page."	Did not describe ambience or emotions, but on a few occasions identified the subjects' facial expressions that conveyed emotions. Example: "a smiling monkey" in I4.	Described overall ambiance on a few occasions. Example: In I46 "relaxed vibe" and in I33 "dystopian" cityscape. In several cases, described the subjects' facial expressions or body language that indicated emotions. Example: In I4 "laughing monkey" or in I18 "arms raised skywards as if in celebration."	Described overall ambience and emotion in several images, in addition to describing explicit facial or bodily expressions. Example: In I3, C3 described, "the entire bird seems to glow with fiery energy. The bird appears very regal and majestic." In I17, C3 added, "The image gives off the feeling that it is a scary movie." In I19, C5's ideal version described, the person's "posture suggests sadness or grieving." In I4, C4 said, "a happy monkey".	Generally preferred more objective descriptions instead of interpretive ambience. Example: In I17, both S17 and S31 added that the letters were "shaky" instead of explicitly mentioning that the image gives a "scary" vibe. SRUs also included explicit facial or bodily expressions. Example: In I14, S30 added, "smiling" turtles. In I4, S20 added, "hands up in a celebratory position." On a few rare occasions, SRUs included subjective expressions. Example: In I3, S18 mirrored the creator's emotive expression, "the entire bird seems to glow with fiery energy. The bird appears very regal and majestic." In I4, S28 said, "happy monkey".