Underspecification in Scene Description-to-Depiction Tasks

Ben Hutchinson
Google Research, Australia
benhutch@google.com

Jason Baldridge
Google Research, USA
jasonbaldridge@google.com

Vinodkumar Prabhakaran
Google Research, USA
vinodkpg@google.com

Abstract

Questions regarding implicitness, ambiguity and underspecification are crucial for understanding the task validity and ethical concerns of multimodal image+text systems, yet have received little attention to date. This position paper maps out a conceptual framework to address this gap, focusing on systems which generate images depicting scenes from scene descriptions. In doing so, we account for how texts and images convey meaning differently. We outline a set of core challenges concerning textual and visual ambiguity, as well as risks that may be amplified by ambiguous and underspecified elements. We propose and discuss strategies for addressing these challenges, including generating visually ambiguous images, and generating a set of diverse images.

1 Introduction

The classic Grounding Problem in AI asks *how is it that language can be interpreted as referring to things in the world?* It has been argued that demonstrating natural language understanding requires mapping text to something that is non-text and that functions as a model of meaning (e.g., Bender and Koller, 2020). In this view, multimodal models that relate images and language have an important role in pursuing contextualized language understanding. Indeed, joint modeling of linguistic and visual signals has been argued to play a critical role in progress towards this ultimate goal, as precursors to modeling relationships between language and the social and physical worlds (Bisk et al., 2020).

Recent text-to-image *generation* systems have demonstrated impressive capabilities (Zhang et al., 2021; Ramesh et al., 2021; Ding et al., 2021; Nichol et al., 2021; Gafni et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Ramesh et al., 2022; Yu et al., 2022). These employ deep learning methods such as generative adversarial networks (Goodfellow et al., 2014), neural discrete representation learning

Figure 1: Generated depictions of the scene "A robot and its pet in a tree." Many elements are underspecified in the text, e.g., pet type, perspective, and visual style.

(van den Oord et al., 2017) combined with autoregressive models (Brown et al., 2020), and diffusion models (Sohl-Dickstein et al., 2015), trained on large datasets of images and aligned texts (Radford et al., 2021; Jia et al., 2021).

With such developments in multimodal modeling and further aspirations towards contextualized language understanding, it is import to better understand both task validity and construct validity in text-to-image systems (Raji et al., 2021). Ethical questions concerning bias, safety and misinformation are increasingly recognized (Saharia et al., 2022; Cho et al., 2022); nevertheless, understanding which system behaviors are desirable requires a vocabulary and framework for understanding the diverse and quickly expanding capabilities of these systems. This position paper addresses these issues by focusing on classic problems (in both linguistic theory and NLP) of ambiguity and underspecification (e.g., Poesio, 1994; Copestake et al., 2005; Frisson, 2009). Little previous work has looked into how underspecification impacts multimodal systems, or what challenges and risks they pose.

This position paper presents a model of task formulation in text-to-image tasks by considering the relationships between images and texts. We use this foundation to identify challenges and risks when generating images of scenes from text descriptions, and discuss possible mitigations and strategies for addressing them.

2 Background

2.1 Image meanings

Like texts, images are used in communicative contexts to convey concepts. Images often convey meaning via resemblance, whereas the correspondence between language and meaning is largely conventional ("icons" vs "symbols" in the vocabulary of semiotics (e.g. de Saussure, [1916] 1983; Hartshorne et al., 1958; Jappy, 2013; Chandler, 2007)). For example, both the English word "cat" or images of a cat—including photographs, sketches, etc—can signify the concept of a cat. Furthermore they each can be used in contexts to represent either the general concept of cats, or a specific instance of a cat. That is, images can have both i) concepts/senses, as well as ii) objects/referents in the world. As such, both images and text can direct the mind of the viewer/reader towards objects and affairs in the world (also known as "intentionality" in the philosophy of language (e.g., Searle, 1995)), albeit in different ways. Despite the adage that a picture is worth a thousand words, even relatively simple diagrams may not be reducible to textual descriptions (Griesemer, 1991).

Like texts, images can also indirectly convey meaning about the agent who produced the image, or about the technology used to create or transmit it (cf. the model of communication of Jakobson and Sebeok, 1960). Also like language, the meanings of images can be at least partly conventional and cultural, e.g., logos, iconography, tattoos, crests, hand gestures, etc. can each convey meaning despite having no visual resemblance to the concept or thing being denoted. Shatford (1986) describes this in terms of images being *Of* one thing yet potentially *About* another thing. Such "aboutness" is not limited to iconography, for photographic imagery can convey cultural meanings too—Barthes (1977) uses the example of a photograph of a red chequered tablecloth and fresh produce conveying the idea of Italianicity.

2.2 Text-image relationships

A variety of relationships between text and image are possible, and have been widely discussed in creative and cultural fields (e.g., Barthes, 1977; Berger, 2008). The Cooper Hewitt Design Museum has, for example, published extensive guidelines on accessible image descriptions.[1] These make a fundamental distinction between image *descriptions*, which provide visual information about what is depicted in the image, and *captions*, which explain the image or provide additional information. For example, the following texts could apply to the same image, while serving these different purposes:

- **description**: "Portrait of former First Lady Michelle Obama seated looking directly at us."
- **caption**: "Michelle LaVaughn Robinson Obama, born 1964, Chicago, Illinois."

This distinction is closely related to that between *conceptual descriptions* and *non-visual descriptions* made by Hodosh et al. (2013), building on prior work on image indexing (Jaimes and Chang, 2000). Hodosh et al. subdivide conceptual descriptions into *concrete* or *abstract* according to whether they describe the scene and its entities or the overall mood, and also further differentiate a category of *perceptual descriptions* which concern the visual properties of the image itself such as color and shape. van Miltenburg (2019, Chapter 2) has a more detailed review of these distinctions.

As images have meanings (see §2.1), describing an image often involves a degree of interpretation (van Miltenburg, 2020). Although often presented as neutral labels, captions on photographs commonly tell us how visual elements "ought to be read" (Hall, 2019, p. 229). Literary theorist Barthes distinguishes two relationships between texts and images: *anchorage* and *relay*. With anchorage, the text guides the viewer towards certain interpretations of the image, whereas for relay, the text and image complement each other (Barthes, 1977, pp. 38–41). McCloud's theory of comics elaborates on this to distinguish four flavours of word-image combinations (McCloud, 1993): (1) the image supplements the text, (2) the text supplements the image, (3) the text and image contribute the same information, (4) the text and image operate in parallel without their meanings intersecting. Since language is interpreted contextually, these image-accompanying texts might depend on the multimodal discourse context, the writer, and the intended audience. The strong dependence on the writer, in particular, highlights the socially and culturally subjective nature of image descriptions (van Miltenburg et al., 2017; Bhargava and Forsyth, 2019). This subjectivity can result in speculation (or abductive inference), for example when people describing images fill in missing details (van Miltenburg, 2020), in human reporting biases regard-

[1] https://www.cooperhewitt.org/cooper-hewitt-guidelines-for-image-descriptions/

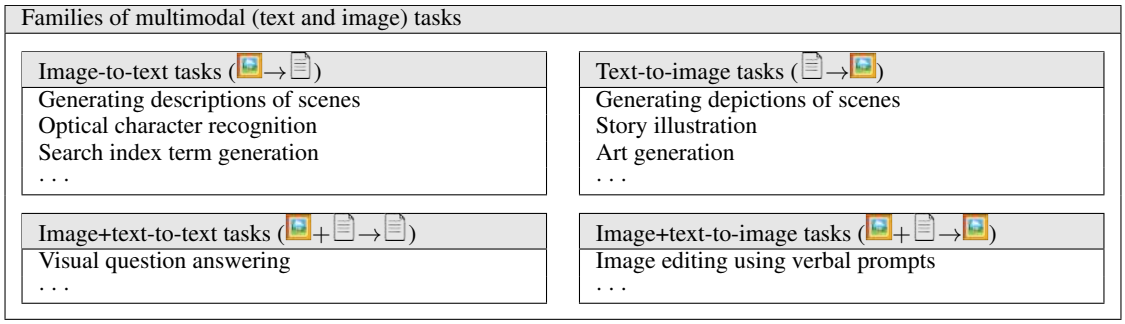

Figure 2: Sketch of a taxonomy of text+image tasks. The taxonomy has gaps which suggest novel tasks, e.g., "optical character generation" (generating images of texts), or querying text collections using images.

ing what is considered noteworthy or unexpected (Van Miltenburg et al., 2016; Misra et al., 2016), in social and cultural stereotyping (van Miltenburg, 2016; Zhao et al., 2017; Otterbacher et al., 2019), and in derogatory and offensive image associations (Birhane et al., 2021; Crawford and Paglen, 2019).

Despite the frequently stated motivation of ML-based multimodal image+text technologies as assisting the visually impaired, the distinction between captions and descriptions—relevant to accessibility—is mostly ignored in the text-to-image literature (van Miltenburg, 2019, 2020). It is common for systems that generate image descriptions to be described as "image-captioning" (e.g., Nie et al., 2020; Agrawal et al., 2019; Srinivasan et al., 2021; Lin et al., 2014; Sharma et al., 2018), without making a distinction between captions and descriptions. An exception is a recent paper explicitly aimed at addressing image accessibility (Kreiss et al., 2021). Other NLP work uses "caption" to denote characterizations of image content, using "depiction" for more general relations between texts and images (Alikhani and Stone, 2019).

Within multimodal NLP, building on annotation efforts, Alikhani et al. have distinguished five types of coherence relationships in aligned images and texts (of which multiple can hold concurrently) (Alikhani et al., 2020, 2019): (1) the text presents what is depicted in the image, (2) the text describes the speaker's reaction to the image, (3) the text describes a bigger event of which the image captures only a moment, (4) the text describes background info or other circumstances relevant to the image, and (5) the text concerns the production and presentation of the image itself.

Finally, we also note the case where the image is of (or contains) text itself. Not only is this relevant to OCR tasks, but also to visual analysis of web pages (e.g., Mei et al., 2016), memes (e.g., Kiela et al., 2020), advertising imagery (e.g., Lim-Fei et al., 2017), as well as a challenging aspect of image generation when the image is desired to have embedded text (for example on a book cover). (Prior to movable type printing, the distinction between texts and images-of-texts was likely less culturally important (Ong, 2013; Sproat, 2010).)

2.3 Text-to-image tasks

Figure 2 situates the family of text-to-image tasks within the greater family of multimodal (text and image) tasks. One of the important factors distinguishing different flavors of text-to-image tasks is the semantic and pragmatic relationship between the input text and the output image. Although commonly used as if it describes a single task, we posit that "text-to-image" describes a family of tasks, since it only denotes a structural relationship: a text goes in and an image comes out. Although some relationship between input and output is perhaps implied, it is just as implicit as if one were to speak of a "text-to-text" task without mentioning whether the task involves translation, paraphrase, summarization, etc. It is important to emphasise that tasks and models are typically not in a 1:1 relationship: even without multi-head architectures, a model may be used for many tasks (e.g., Raffel et al., 2020; Chen et al., 2022), while many (single-task) NLP architectures employ multiple models in sequence. As van Miltenburg (2020) argues, the dataset annotations which often act as extensional definitions of the task of interest (Schlangen, 2021) are often produced via underspecified crowdsourcing tasks that do not pay full attention to the rich space of possible text-image relationships described above. Similarly, text-image pairs repurposed from the web often have poorly specified relationships: although the Web Content Accessibility Guidelines recommend that "alt" tags "convey the same function or purpose as the image" (Chisholm et al., 2001) (for a survey of guidelines,

see Craven (2006)), real-world usage may deviate considerably (see, e.g., (Petrie et al., 2005) and the discussion in (Muehlbradt and Kane, 2022)).

Recent literature on text-to-image modeling has been characterized by simplified task formulations. For example, despite the impressive outputs of recent models—e.g., unCLIP (a.k.a., DALL-E 2) (Ramesh et al., 2022), Imagen (Saharia et al., 2022), and Parti (Yu et al., 2022)—the papers introducing these models rely on the broadest task formulation, wherein the model takes a textual prompt of any kind and produces an image of any kind. While they discuss terms such as *diversity*, *caption similarity*, *high fidelity*, and *high quality* to discuss properties of model outputs, these are not precisely defined, nor are they fully operationalized in current evaluation metrics. Similarly, the XMC-GAN paper asserts that systems should produce "coherent, clear, photo-realistic scenes" yet the authors fail to either justify or clarify these objectives (Zhang et al., 2021). In fact, this objective seems to be at least partly a by-product of the fact that the model training and evaluation was on photographs from the MS-COCO dataset. Setting photo-realistic imagery as the ideal raises questions about both justification (why not other styles of images?) and correspondence (e.g., how does photography construct relationships between images and reality?).

3 Task Formulation

Underspecification in task formulation is a major challenge for machine learning and artificial intelligence disciplines as a whole (D'Amour et al., 2022; Raji et al., 2021). Clarity around task formulation helps system designers navigate ambiguous inputs; for example, given a prompt such as "a painting of a horse", should the system create an image whose style resembles a painting, or an image of a scene containing a painting, including the frame and other plausible contextual details? This paper postulates that accounts of *image meaning* and *text-image relationships* are of central relevance to formulating task definitions in text-to-image systems generally. Such accounts are thus important for characterizing underspecification in such systems.

We take the notion of *world* to be important too, for two reasons. Like texts, images can reference objects in the world, and in doing so are human-mediated representations of the observable world that involve selection and filtering processes. Also, the notion of possible worlds has played an important role in theories of semantics (e.g., Kratzer and Heim (1998)). Therefore, two questions that we believe should be central to an account of underspecification in text-to-image tasks are:

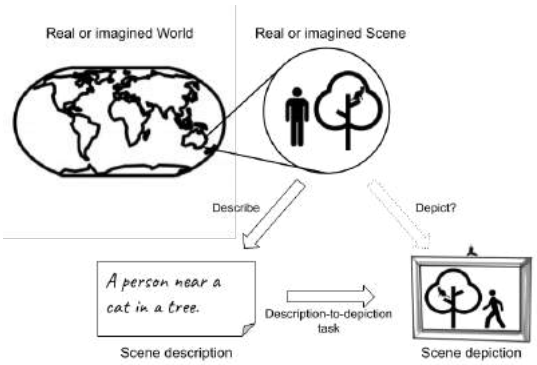

Figure 3: Scene depictions and descriptions are communicative acts conveying information (or misinformation) about a real or imagined scene in the world.

1. What are the two-way relationships between images-text pairs and (real or imagined) worlds?
2. What is the three-way relationship between the images, texts and the world?

We do not attempt here to unify or rebut the many theories of image meanings and text–image relationships, but instead highlight what we see as essential considerations for scene depiction tasks:

1. We use *scene* to mean a small fragment of a (real or imagined) world. A scene can be *described* in texts, and can also be *depicted* in images. Both descriptions and depictions can thus convey information about a scene.[2]
2. The production and sharing of descriptions and depictions both constitute *communicative acts*. These acts are interpreted within social contexts, and can have locutionary (what is said/shown) as well as perlocutionary dimensions (effects on the viewer/reader such as scaring, offending or prompting action) and illocutionary dimensions such as connotations.
3. Descriptions and depictions necessarily convey *incomplete* information about all but the most trivial scene. The two modalities necessarily *underspecify* different types of information, both due to intra-modal constraints and assumptions of extra-modal contextual information.

We propose two components, *coherence* and *style*, for the formulation of the family of text-to-

[2]We use "depiction" in the sense of "to show visually", rather than the definition by Alikhani and Stone (2019).

image tasks. We argue in the following section that both are relevant to underspecification.

- **Coherence**: Any valid semantic and/or pragmatic relationship between a static image-text pair, e.g., those listed in §2.2, is a potentially valid semantic relationship for a given flavor of text-to-image task. For example, one can meaningfully speak of a description-to-depiction task or an event-to-image-moment task.
- **Style**: Valid text-to-image tasks can encompass a multitude of visual styles. That is, text-to-image is not constrained to photo-realism but rather can involve styles resembling cartoons, paintings, woodcut prints, etc, and even to specific genres such as manga, impressionist, or ukiyo-e.

Given this conceptual framework, one natural challenge that presents itself is that visual and linguistic information often serve to complement each other in multimodal texts. Indeed this can be utilized for skilled effect leading to greater engagement with readers/viewers by requiring that they mentally fill in the missing information (McCloud, 1993; Iyyer et al., 2017).

4 Challenges in Description to Depiction

Having laid out the relevant considerations of meaning and reference in text-to-image systems in §3, we now focus specifically on systems that produce an image depiction of a scene from a description of that scene. We distinguish challenges from three sources: linguistic ambiguity in descriptions, underspecification in descriptions, and underspecification of desired depictions. Our use of the term *underspecification* here reflects how it has been used in NLP literature, referring both to ambiguity in the objects of study (e.g., linguistic forms (Bender and Lascarides, 2019, p. 29)), as well as to properties of the technical apparatus used to model meaning (e.g., Bender and Lascarides 2019, p. 30).

4.1 Linguistic Ambiguity in Descriptions

Many if not all forms of linguistic ambiguities are likely to occur in scene descriptions. However we call out a few of notable importance.

- *Syntactic ambiguities* including locative PP attachment can present ambiguities concerning spatial relationships. For instance, in the input "A cat chasing a mouse on as skateboard", is the cat or the mouse—or both—on the skateboard? See Figure 4a.
- *Word sense ambiguities* (including metonymy) and ontological vagueness present challenges as to how objects should be depicted; e.g., for "The man picked up the bat", is the bat a flying mammal or a sports implement? Visualizing ambiguous words is also a challenge for verbs: "riding a bus" and "riding a horse" are very different actions (consider that "riding a bus in the way one would normally ride a horse" is easier to imagine than the converse) (Gella et al., 2017).
- *Anaphoric ambiguities* including pronouns can also cause challenges, e.g., what is the toy beside in "a book on a chair and a toy beside it"?
- *Quantifier scope ambiguities* also arise, e.g., how many books are there in "three people holding a large book"?

4.2 Underspecification in Descriptions

Finite and reasonable-length linguistic descriptions of real-world or realistic scenes will by necessity omit a great deal of visual information. Within NLP, underspecification in descriptions has perhaps been discussed most often in the context of generating referring expressions for objects (see Krahmer and Van Deemter (2012) for a survey). However, underspecification in input texts also causes major challenges in description to depiction tasks.

- *Unmarked defaults* can lead to potentially unbounded amounts of underspecified information (e.g., should people be depicted as clothed even if clothing is not mentioned, as is the social norm in images?) (Misra et al., 2016). Visual details such as lighting, color and texture may be omitted from texts: What does a carpet's surface look like? Where is the light source and do shadows need to be depicted?). See Figure 4b.
- *Ontological vagueness* may also present challenges as to what types of objects should be depicted: for "a tall dark-skinned person with a toy", what type of toy? See also Figure 4b. Scalars typically often present underspecification (e.g., how tall is "tall person"?; how dark is "dark-skinned"?), and points of reference are often underspecified (cf. "tall" and "dark-skinned" in Japan vs South Africa). Ontological specificity in texts depends at least partly on which categories are considered to be basic (e.g. Rosch et al., 1976; Ordonez et al., 2015).
- *Geo-cultural context* of input descriptions is often left unspecified. For instance, in "a woman eating breakfast beside her pet", the types of

(a) Outputs for "A cat chasing a mouse on a skateboard." The number of boards and which animal is on any given board is ambiguous.

(b) Outputs for "A ball on a rug." The types and visual details of balls and rugs are unspecified.

(c) Outputs for "A monkey cutting a cake." The cutting instrument is unspecified, as is the style.

(d) Outputs for "Two cats looking out of a space shuttle window. DSLR photograph." Perspective is unspecified.

Figure 4: Example treatments of underspecified inputs. These examples and those elsewhere in this paper were generated using Parti (Yu et al., 2022) followed by the super-resolution third stage of Imagen (Saharia et al., 2022).

things that count as breakfast and pets are culturally subjective. In many cases, object forms are institutionally regulated, e.g., for "a man counting money in a car", the physical appearance of money and license plates, and the positioning of the steering wheel (left vs. right), are institutionally regulated and only implicit in the text.

- *Implied objects* that are part of many events or states are often not specified in corresponding descriptions. For example "a monkey cutting a cake" implies a cutting instrument (see Figure 4c); "a wedding" has many implied objects, but at a minimum seems to imply two people.

While description to depiction models often generate images that fills in such implied details or objects, such extrapolations run the risk of perpetuating social stereotypes (§5).

4.3 Underspecification of Desired Depictions

The underspecification challenges in the linguistic inputs to text-to-image systems are complemented by a different set of challenges in the output generation concerning precise visual details.

- *Style*. Text inputs often do not specify a desired visual style of depiction, e.g., photo-realism, cartoons, paintings, woodcut prints, etc., or genres such as manga, impressionist, and ukiyo-e. While this is a question relevant also for task formulation (see §3), this ambiguity need to be resolved for text-to-image systems capable of generating multiple styles of images. It is also possible to imagine and create new styles using these tools. This is a fascinating use case, but it also raises questions about how to evaluate whether a model has succeeded—for example, when mashing together multiple style specifications, e.g. "The New York City skyline in ukiyo-e style by van Gogh."
- *Technical*. Goals of photo(graphic)-realism raise questions about what sort of photographic technologies are implied, including implicit lens, implicit depth of focus and implicit exposure time, each of which produce different visual artefacts.
- *Perspective*. Many image styles, including but not limited to photographic ones, have an implied perspective, and an implied frame or shot (Chandler, 2007, p. 89), including not just an implied eye but also an implied angle or tilt. The choice of perspective can have socio-cultural connotations. A perspective closer to the ground may represent that of a child, and low viewing angles are used by filmmakers to make subjects appear powerful or convey vulnerability.[3] Such low-shots might also impact subject credibility (Mandell and Shaw, 1973). Different social groups may have proclivities for different angles (Aguera y Arcas et al., 2017) or perspectives (e.g., Green 2009, discussed in Cohn 2013).
- *Spatial orientations* with respect to the implied viewer (see Figure 4d) are not typically men-

[3] https://www.nfi.edu/low-angle-shot/

(a) Outputs for "Wedding attire displayed on a mannequin" may show gender and Western cultural biases.

(b) Outputs for "Graffiti on the New York Public library. DSLR photo." might cause offence to bibliophiles.

(c) Outputs for "A photo of a famous city with opera house" may spread misinformation.

(d) Outputs for "A photo of a non-venomous Australian spider" may have safety risks for animal lovers.

Figure 5: Example of risks in scene description-to-depiction.

tioned in the image descriptions upon which models are trained. For example, it is common in a portrait for the subject to be oriented so their face is visible, however such orientation towards the viewer is often not made explicit.

Finally, we note that linguistic ambiguities can interact with underspecified perspectives. An example provided by Levelt (1999) is the congruity of an image with the text "a house with a tree to the left of it" depends not just on the perspective taken in framing the image, but also whether "to the left of" is with respect to the viewer's orientation (facing the house) or to the house's orientation (e.g., facing the viewer, if the front of the house is depicted).

5 Risks and Concerns

Some datasets used for training multimodal systems have previously been shown to contain biases, stereotypes and pornography (Birhane et al., 2021; van Miltenburg, 2016). We now discuss potential concerns in applications employing scene description-to-generation tasks, including how underspecification challenges can exacerbate them.

Bias: As in image-to-text (Bennett et al., 2021), there are risks of text-to-image amplifying societal biases including those concerning gender, race, and disability. Since English-language texts do not grammatically require specification of gender identities of people mentioned in a scene, there is a great potential for systems to reproduce existing societal biases. For example, the prompt "a boss addressing workers" might produce an image of a boss with masculine phenotypes. Similar outcomes are likely to be obtained with respect to other social roles, social groups and stereotypical phenotypes. Cultural biases are expected to be prevalent in any text-to-image systems, since what events and artefacts look like vary wildly around the world—e.g., weddings, bank notes, places of worship, breakfast dishes, etc. When a prompt is ambiguous or underspecified, an ML model is likely to revert to correlations in its training data for deciding details about objects and their appearances. Thus underspecification leads to a greater risk of stereotyping biases, which can cause offense and representational harm especially to marginalized groups with a history of being stereotyped. See Figure 5a.

Harmful, taboo and offensive content: Images depicting violent scenes may have a greater impact on the viewer than corresponding text descriptions. Similarly, pornographic images can be more shocking or culturally taboo than texts. Some societies, such as indigenous Australian ones, may have taboos on visual depictions of the recently deceased (Australian Special Broadcasting Service, 2018, p. 20). This exemplifies potential dangers of non-taboo inputs (permissible referring expressions) producing taboo outputs. Attempts to predict image offensiveness within the context of an input text are likely to encounter challenges when inputs are underspecified. See Figure 5b.

Mis/dis-information: For text-to-image systems which aspire to realism, important ethical concerns arise concerning the deliberate or accidental misleading of viewers' beliefs about the world. Misinformation can lead to adopting addictive habits, belief in pseudoscience or in dangerous health or crisis response information, and other harms (see, e.g., (Neumann et al., 2022)). This is especially risky when systems output photorealistic images, and viewers may be more prone to believe fake photorealistic images than readers are to view fake texts. Identifying mis/dis-information concerns in scene description-to-depiction requires comparing the depicted scene with a model of reality in order to identify misalignments and classify them according to risk of harm. However an underspecified input to a scene description-to-depiction

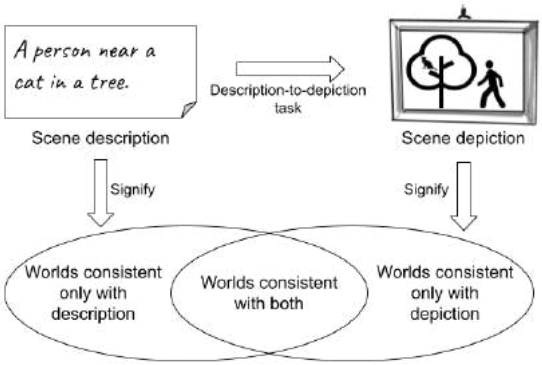

Figure 6: Visual scene depictions and textual scene descriptions may be consistent with different worlds.

system may have one interpretation which is consistent with reality and alternative interpretations which are not. Underspecification hence risks inadvertent misinterpretation of innocuous inputs, potentially leading to misinformation. See Figure 5c.

Safety: Since images can convey meaning (§2.1), they can mislead with potentially harmful consequences. Instruction manuals, road signs, labels, gestures and facial expressions, and many other forms of visual information can lead viewers to take actions in the world which would potentially lead to harm in inappropriate contexts. As with the misinformation risks concerning underspecification outlined above, there is a risk that inadvertent misinterpretation of innocuous inputs could potentially leading to unsafe images in high-risk scenarios. See Figure 5d.

In summary, challenges around input ambiguity seem to exacerbate the risks of many potential concerns around text-to-image systems

6 Paths Forward

6.1 Approaches to input ambiguity

It is impossible to avoid ambiguous inputs. We describe two possible approaches to managing underspecification in scene-description-to-depiction tasks, which we call *Ambiguity In, Ambiguity Out* (AIAO) and *Ambiguity In, Diversity Out* (AIDO).

The AIAO approach posits that a generated image is a model of the intent of the user inputting the text. As such, this approach proposes that generated depictions should underspecify as much as the input does. Given the framework in §3 whereby scene depictions and descriptions both signify concepts about a (real or imaginary) world fragment, we can consider a depiction I algorithmically generated from a description T. A reasonable assumption regarding image quality is that (all else being equal) the depiction I is better if it is consistent with all and only the same world fragments that T is consistent with. This objective of preserving ambiguity suggests a range of strategies. Deliberate visual blurring of non-foreground elements (akin to camera lens and/or exposure effects) can reduce the specificity of objects not mentioned in the text. Some visual styles reproduce social stereotypes less than others, for example a stick figure drawing style could minimize depictions of phenotypes associated with specific social groups. Orientation choices can be manipulated to obscure information not present in the input text, for example if a figure is facing away from the viewer there may be less need to generate specific facial characteristics.

In contrast, the AIDO approach acknowledges that since text and image communicate meaning in different ways, it is often extremely challenging or impossible to translate linguistic ambiguities into visual ambiguities (especially discrete structural ambiguities such as PP attachment or word sense ambiguities). This approach instead advocates for systems which output sets of images, such that the diversity of the output set captures the space of interpretations of the input. When asked to depict "a boss", the AIDO approach would aim to show many diverse people. Some challenges that arise include how to measure image diversity in a socially appropriate way (Mitchell et al., 2020), as well as what space of possibilities should be represented at all.

Due to the challenges in translating ambiguities between mediums, the AIDO approach is likely to generally be more tractable and operationalizable in application systems that permit multiple outputs. However in practice the two approaches are not exclusive and it is possible to combine them. For example, a system generating images for "a boss" may both generate a set of images that includes both diverse faces (AIDO) as well as stick figures and images with obscured facial features (AIAO). Also, the two approaches agree that what is specified in the input should also be specified in the output(s). For example, if asked to depict "eight tall buildings" then the system should aim to generate an image that provides both perspective and spatial configurations that allow the count of eight buildings to be verified using the image alone.

6.2 Clarifying tasks and capabilities

When people collaborate to produce comics, an "important ingredient is the writer's understand-

ing of the artist's style and capabilities" (Eisner, 2008)—and the same is true of human-machine text-to-image collaborations. Just as the Bender Rule advocates for explicitly naming the languages of NLP systems (Bender, 2019), developers of multimodal systems should aim to understand and communicate the "visual language" capabilities of their systems. Understanding and documenting a deployed text-to-image system's interpretive and generative capabilities—including what visual styles it produces and which text-to-image tasks (§3) it can perform—is therefore important for managing user expectations, aiding users in interpreting system behaviours, and mitigating risks of misuse (§5). Understanding the landscape of visual capabilities (and also non-capabilities, i.e., both the range and the codomain of the model) will require engaging with experts in visual disciplines, such as photographers, artists, designers, and curators. We propose that care should be taken when handling training and test data in order to distinguish the semantic and pragmatic relationships between aligned text-image pairs (§2.2), using relationships which make sense for the tasks and applications at hand.

6.3 Risk mitigation

We recommend adopting clear principles of desirable and undesirable system behaviors, especially with regards to biases, offensive and taboo topics, safety, and misinformation risks (§5). Robust stress testing with an adversarial mindset can help to detect corner cases which might trigger undesirable model behaviors, and a culturally diverse pool of stress testers broadens the space of issues which are likely to be detected. Communicating application-specific uses cases of a text-to-image system (see Mitchell et al., 2019) can help to mitigate risk since specific applications come with specific user expectations (e.g., applications for entertainment may not have expectations of truthfulness).

A description-to-depiction system should take into account the potential effects on viewers concerning sensitive and taboo topics. One simple mitigation strategy is for a system to refuse to generate images which are (predicted to be) harmful or offensive, e.g., based on the offensiveness of the input or analysis of the output. However, even if an image or a text are inoffensive alone, an image can nevertheless be offensive if generated in response to the text; for example neither a portrait of a black woman nor the text "an angry person" is offensive on their own, yet the former may reproduce the "angry black woman" stereotype (Walley-Jean, 2009) if generated in response to the latter.

Derczynski et al. (2022) present recommendations for handling harmful text that are relevant to images. These include using overlays to convey that the contents or associations of the harmful image is not condoned, being transparent about why the image is being used within some context (e.g., as an example of something problematic), stating that the harmful image is harmful, or using cropping, blurring or other visual obfuscation techniques (as adopted, e.g., by Birhane et al. (2021)).

7 Conclusion

We have motivated greater consideration of task formulation and underspecification in text-to-image tasks. We laid out the conceptual elements required for this, including greater clarity around the formulation of the space of tasks, as well as consideration of how texts and images each convey concepts. Echoing van Miltenburg (2019), our goal in connecting state-of-the art technologies to theories of cultural and social studies is both to promote deeper understanding of these technologies, and also to foster dialogue across disciplines. We outlined some of the primary challenges concerning textual and visual specification and proposed that systems should consider both reproducing visually the vagueness and ambiguities of the input and producing a diversity of images which convey the breadth of text interpretations. We encourage more work on measuring visual objectives discussed in cultural fields—such as clarity, aesthetics, etc.—and on task-specific utility of generated images (cf. Fisch et al., 2020; Zhao et al., 2019).

Limitations Any position paper at least somewhat reflects the backgrounds and standpoints of its authors. The authors have backgrounds in NLP, computational social science, and AI ethics. Although we call for greater engagement with creative disciplines, we do not represent those disciplines. Although we raise culturally sensitive questions, we have first-hand lived experiences in only Australia, India, the UK and the USA.

Acknowledgements

We would like to thank Emily Denton, Kieran Browne, and the anonymous reviewers for their suggestions and feedback.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8948–8957.

Blaise Aguera y Arcas, Margaret Mitchell, and Alexander Todorov. 2017. Physiognomy's new clothes. *Medium (6 May 2017), online:< https://medium. com/@ blaisea/physiognomys-new-clothesf2d4b59fdd6a*.

Malihe Alikhani, Sreyasi Nag Chowdhury, Gerard de Melo, and Matthew Stone. 2019. CITE: A corpus of image-text discourse relations. In *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2019*, pages 570–575. Association for Computational Linguistics (ACL).

Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. 2020. Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535.

Malihe Alikhani and Matthew Stone. 2019. "Caption" as a coherence relation: Evidence and implications. In *Proceedings of the Second Workshop on Shortcomings in Vision and Language*, pages 58–67.

Australian Special Broadcasting Service. 2018. *The greater perspective: Protocol and guidelines for the production of film and television on Aboriginal and Torres Strait Islander Communities (Supplementary Guidelines)*.

Roland Barthes. 1977. *Image-music-text*. Macmillan.

Emily Bender. 2019. The #Benderrule: On naming the languages we study and why it matters. *The Gradient*, 14.

Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198.

Emily M Bender and Alex Lascarides. 2019. Linguistic fundamentals for natural language processing II: 100 essentials from semantics and pragmatics. *Synthesis Lectures on Human Language Technologies*, 12(3):1–268.

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*, pages 1–19.

John Berger. 2008. *Ways of seeing*. Penguin UK.

Shruti Bhargava and David Forsyth. 2019. Exposing and correcting the gender bias in image captioning datasets and models. *arXiv preprint arXiv:1912.00578*.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv e-prints*, pages arXiv–2110.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Daniel Chandler. 2007. *Semiotics: the basics*. Routledge.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Wendy Chisholm, Gregg Vanderheiden, and Ian Jacobs. 2001. Web content accessibility guidelines 1.0. *Interactions*, 8(4):35–54.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2022. DALL-Eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *arXiv preprint arXiv:2202.04053*.

Neil Cohn. 2013. *The Visual Language of Comics: Introduction to the Structure and Cognition of Sequential Images.* A&C Black.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A Sag. 2005. Minimal recursion semantics: An introduction. *Research on language and computation*, 3(2):281–332.

Timothy C Craven. 2006. Some features of "alt" texts associated with images in web pages. *Information Research: An International Electronic Journal*, 11(2):n2.

Kate Crawford and Trevor Paglen. 2019. Excavating AI: The politics of images in machine learning training sets. *AI and Society*.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2022. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23:1–61.

Ferdinand de Saussure. [1916] 1983. *Course in General Linguistics*. Duckworth, London. (trans. Roy Harris).

Leon Derczynski, Hannah Rose Kirk, Abeba Birhane, and Bertie Vidgen. 2022. Handling and presenting harmful text. *arXiv preprint arXiv:2204.14256*.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. 2021. CogView: Mastering text-to-image generation via transformers. In *Advances in Neural Information Processing Systems*.

Will Eisner. 2008. *Comics and sequential art: Principles and practices from the legendary cartoonist*. WW Norton & Company.

Adam Fisch, Kenton Lee, Ming-Wei Chang, Jonathan H Clark, and Regina Barzilay. 2020. Capwap: Image captioning with a purpose. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8755–8768.

Steven Frisson. 2009. Semantic underspecification in language processing. *Language and Linguistics Compass*, 3(1):111–127.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*.

Spandana Gella, Frank Keller, and Mirella Lapata. 2017. Disambiguating visual verbs. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):311–322.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NeurIPS*.

Jennifer Anne Green. 2009. *Between the earth and the air: Multimodality in Arandic sand stories*. Ph.D. thesis.

James R Griesemer. 1991. Must scientific diagrams be eliminable?: The case of path analysis. *Biology and Philosophy*, 6(2):155–180.

Stuart Hall. 2019. The determinations of news photographs (1973). In *Crime and Media*, pages 123–134. Routledge.

Charles Hartshorne, Paul Weiss, Arthur W Burks, et al. 1958. Collected papers of Charles Sanders Peirce.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan Boyd-Graber, Hal Daume, and Larry S Davis. 2017. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 7186–7195.

Alejandro Jaimes and Shih-Fu Chang. 2000. A conceptual framework for indexing visual information at multiple levels. In *SPIE proceedings series*, volume 3964, pages 2–15.

Roman Jakobson and Thomas A Sebeok. 1960. Closing statement: Linguistics and poetics. *Semiotics: An introductory anthology*, pages 147–175.

Tony Jappy. 2013. *Introduction to Peircean visual semiotics*. A&C Black.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624.

Emiel Krahmer and Kees Van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Angelika Kratzer and Irene Heim. 1998. *Semantics in generative grammar*, volume 1185. Blackwell Oxford.

Elisa Kreiss, Noah D Goodman, and Christopher Potts. 2021. Concadia: Tackling image accessibility with context. *CORR*, abs/2104.08376.

W Levelt. 1999. Producing spoken language. *The neurocognition of language*, pages 83–122.

Victor Lim-Fei, KYS Tan, and K Yin. 2017. Multimodal translational research: Teaching visual texts. *New studies in multimodality: Conceptual and methodological elaborations*, 175.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Lee M Mandell and Donald L Shaw. 1973. Judging people in the news—unconsciously: Effect of camera angle and bodily activity. *Journal of broadcasting & electronic media*, 17(3):353–362.

Scott McCloud. 1993. Understanding comics: The invisible art. *Northampton, Mass*.

Tao Mei, Lusong Li, Xinmei Tian, Dacheng Tao, and Chong-Wah Ngo. 2016. PageSense: Toward stylewise contextual advertising via visual analysis of web pages. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(1):254–266.

Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2939.

Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. 2020. Diversity and inclusion metrics in subset selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 117–123.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

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):1–32.

Terrence Neumann, Maria De-Arteaga, and Sina Fazelpour. 2022. Justice in misinformation detection systems: An analysis of algorithms, stakeholders, and potential harms. In *Proceedings of the conference on fairness, accountability, and transparency*.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv e-prints*, pages arXiv–2112.

Allen Nie, Reuben Cohn-Gordon, and Christopher Potts. 2020. Pragmatic issue-sensitive image captioning. In *EMNLP (Findings)*.

Walter J Ong. 2013. *Orality and literacy*. Routledge.

Vicente Ordonez, Wei Liu, Jia Deng, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2015. Predicting entry-level categories. *International Journal of Computer Vision*, 115(1):29–43.

Jahna Otterbacher, Pınar Barlas, Styliani Kleanthous, and Kyriakos Kyriakou. 2019. How do we talk about other people? group (un) fairness in natural language image descriptions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 106–114.

Helen Petrie, Chandra Harrison, and Sundeep Dev. 2005. Describing images on the web: a survey of current practice and prospects for the future. *Proceedings of Human Computer Interaction International (HCII)*, 71(2).

Massimo Poesio. 1994. Ambiguity, underspecification and discourse interpretation. In *Proceedings of the First International Workshop on Computational Semantics*, pages 151–160.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

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*, 21:1–67.

Inioluwa Deborah Raji, Emily Denton, Emily M Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the everything in the whole wide world benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *arXiv e-prints*, pages arXiv–2204.

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 *ICML*.

Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439.

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, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv e-prints*, pages arXiv–2205.

David Schlangen. 2021. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674.

John R Searle. 1995. *The construction of social reality*. Simon and Schuster.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Sara Shatford. 1986. Analyzing the subject of a picture: a theoretical approach. *Cataloging & classification quarterly*, 6(3):39–62.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France. PMLR.

Richard Sproat. 2010. *Language, technology, and society*. Oxford University Press.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *NeurIPS*.

CWJ van Miltenburg. 2016. Stereotyping and bias in the flickr30k dataset. In *11th workshop on multimodal corpora: computer vision and language processing*.

CWJ van Miltenburg. 2019. *Pragmatic factors in (automatic) image description*. Ph.D. thesis, SIKS, the Dutch Research School for Information and Knowledge Systems.

Emiel van Miltenburg. 2020. On the use of human reference data for evaluating automatic image descriptions. In *2020 VizWiz Grand Challenge Workshop*.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2017. Cross-linguistic differences and similarities in image descriptions. In *10th International Conference on Natural Language Generation*, pages 21–30. Association for Computational Linguistics.

Emiel Van Miltenburg, Roser Morante, and Desmond Elliott. 2016. Pragmatic factors in image description: The case of negations. In *Proceedings of the 5th Workshop on Vision and Language*, pages 54–59.

J Celeste Walley-Jean. 2009. Debunking the myth of the "angry Black woman": An exploration of anger in young African American women. *Black Women, Gender & Families*, 3(2):68–86.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*.

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 833–842.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989.

Sanqiang Zhao, Piyush Sharma, Tomer Levinboim, and Radu Soricut. 2019. Informative image captioning with external sources of information. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6485–6494.