

Resolving Ambiguities in Text-to-Image Generative Models

Ninareh Mehrabi*, Palash Goyal, Apurv Verma, Jwala Dhamala,
Varun Kumar, Qian Hu, Kai-Wei Chang, Richard Zemel,
Aram Galstyan, Rahul Gupta

Amazon Alexa AI-NU

Abstract

Natural language often contains ambiguities that can lead to misinterpretation and miscommunication. While humans can handle ambiguities effectively by asking clarifying questions and/or relying on contextual cues and common-sense knowledge, resolving ambiguities can be notoriously hard for machines. In this work, we study ambiguities that arise in text-to-image generative models. We curate the Text-to-image Ambiguity Benchmark (TAB) dataset to study different types of ambiguities in text-to-image generative models.¹ We then propose the Text-to-Image Disambiguation (TIED) framework to disambiguate the prompts given to the text-to-image generative models by soliciting clarifications from the end user. Through automatic and human evaluations, we show the effectiveness of our framework in generating more faithful images aligned with end user intention in the presence of ambiguities.

1 Introduction

Natural conversations contain inherent ambiguities due to potentially multiple interpretations of the same utterance. Different types of ambiguities can be attributed to *syntax* (e.g., “the girl looks at the boy holding a green bag” — is the girl holding the green bag?), *semantics* (e.g., “a picture of cricket” — is cricket referring to an insect or a game?), and *underspecification* (e.g., “doctor talking to a nurse” — is the doctor/nurse male or female?). Ambiguities pose an important challenge for many natural language understanding tasks and have been studied extensively in the context of machine translation (Stahlberg and Kumar, 2022), conversational question answering (Guo et al., 2021), and task-oriented dialogue systems (Qian et al., 2022), among others.

In this paper, we study the effect of ambiguity in text-to-image generative models (Ramesh et al.,

*mninareh@amazon.com

¹Data located at: <https://github.com/Ninarehm/TAB>.

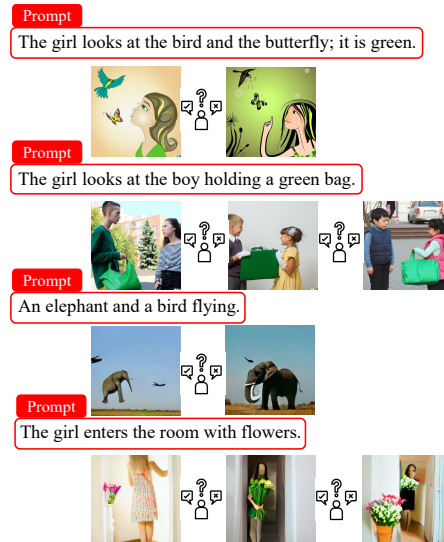


Figure 1: Examples of ambiguous prompts and corresponding generated images. The icon between the images depicts alternative interpretations. The images corresponding to first two prompts are generated by OpenAI’s DALL-E (Ramesh et al., 2022), and images corresponding to last two prompts are generated by DALL-E Mega (Dayma et al., 2021) models.

2021, 2022; Saharia et al., 2022; Yu et al., 2022) and demonstrate that ambiguous prompts provided to such models might result in undesired outcomes and poor user experience. In particular, ambiguities due to underspecification can lead to biased outcomes with possible implications on fairness of the underlying models (e.g., when prompted with “doctor talking to a nurse”, the model might generate images with disproportionate number of male doctors and female nurses). We also propose a framework for mitigating ambiguities existing in prompts. We choose this setting to study ambiguity as visual scenes provide readily human-interpretable alternative understandings of text, thus helping in evaluating ambiguities as well as mitigation strategies. Figure 1 illustrates a few ambiguous prompts and corresponding outputs from the state-of-the-art text-to-image model, DALL-E (Ramesh et al., 2022; Dayma et al., 2021). We observe that ambi-

TIED Framework

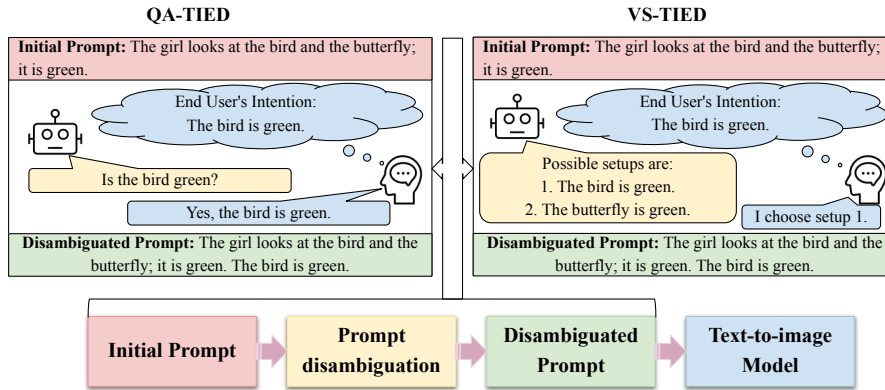


Figure 2: Our proposed Text-to-Image Disambiguation (TIED) framework. The initial ambiguous prompt is disambiguated by either (1) the language model generating clarifying question which will be resolved through end user provided answers (QA-TIED shown on the left), or (2) the language model generating different possible visual setups and end user choosing the desired setup (VS-TIED shown on the right). We define visual setups as textual descriptions of different possible visual scenarios/interpretations. The final disambiguated prompt will later be provided to the downstream text-to-image generative model.

guity in prompts confuses the model resulting in a varied set of generated images.

Humans tend to resolve ambiguities by asking clarifying questions, relying on other forms of modalities (such as vision), using contextual signals, and leveraging common-sense and/or an external source of knowledge (Achimova et al., 2022). Inspired by this observation, we propose a new framework (Figure 2) in which we incorporate a language model-based *prompt disambiguation filter* to process the prompts fed to text-to-image generative models. This filter is capable of either asking clarifying questions or generating different possible textual descriptions of visual setups which would later be resolved through end user interactions (in this work, we define *end user* as a human-agent who interacts with the system and might interchangeably use human to refer to the end user). Ultimately, the disambiguation filter helps the text-to-image model to identify a single visual setup for image generation. In this work, we define visual setups as textual descriptions of different possible visual interpretations of a given ambiguous prompt.

To better understand the weaknesses of current text-to-image generative models, and to evaluate the effectiveness of our proposed mitigation framework, we curate a benchmark dataset consisting of ambiguous prompts covering different types of ambiguities that are especially relevant to text-to-image models. We propose new automatic evaluation procedures to evaluate faithfulness of gener-

ations in text-to-image generative models to end user intention. We perform various automatic and human evaluation experiments (on over 15k images) and observe that our proposed framework can improve faithful image generation by an overall average of over 11% through prompt disambiguation.

Overall, we make the following contributions:

1. We introduce a Text-to-image Ambiguity Benchmark (TAB) dataset containing different types of ambiguous prompts along with different visual setups (Section 2).
2. We propose the Text-to-Image Disambiguation (TIED) framework which can be applied on top of any text-to-image model for ambiguity resolution (Section 3).
3. We propose new automatic evaluation procedures to evaluate ambiguity resolution in text-to-image models. We perform automatic and human experiments to demonstrate the effect of TIED on ambiguity resolution and faithful image generations (Sections 4 and 5).

2 Text-to-image Ambiguity Benchmark (TAB)

Our Text-to-image Ambiguity Benchmark (TAB) is a modified version of the LAVA corpus (Berzak et al., 2015). The original LAVA corpus contains various types of ambiguous sentences that can only be resolved through visual signals from their corresponding images/videos. We use the ambiguous

	Ambiguity Type	Count	
Main Types	Syntax Prepositional Phrase (PP)	74	}Linguistic
	Syntax Verb Phrase (VP)	243	
	Syntax Conjunction	127	
	Discourse Anaphora	21	
	Discourse Ellipsis	45	
	Fairness	355	}Fairness
Add	Complex+combination+misc	335	

Table 1: Breakdown of our benchmark dataset (TAB) by ambiguity types. TAB consists of six types of ambiguities, including linguistic and fairness. We cover syntactic as well as discourse type ambiguities for linguistic type ambiguities. TAB also contains complex version for subset of the samples from the main type ambiguities with structurally more complex sentences, combination cases that combine fairness and linguistic type ambiguities, and some miscellaneous cases.

prompts (templates) from LAVA and not the images — as images in our case would be generated by text-to-image generative models.

We make various modifications to the LAVA corpus to create TAB. These modifications include: (i) adding new ambiguous sentences to TAB to cover more diverse objects, scenes, and scenarios compared to existing ambiguous sentences in LAVA, (ii) removing examples relevant to video domain from LAVA and only keeping examples relevant to static images in TAB, (iii) including fairness prompts in TAB that cover different activities (Zhao et al., 2017) and occupations (Nadeem et al., 2021) in which the identities of the individuals are ambiguous, (iv) adding more structurally complex sentences, and (v) curating additional labels for TAB (e.g., whether the visual setup or interpretation of an ambiguous sentence is commonsensical or not). As a result of some of the modifications mentioned above, TAB ends up covering 963 more ambiguous sentences (prompts) and 4,192 more visual setups (textual descriptions of possible visual interpretations for each ambiguous sentence) compared to LAVA. LAVA covers 237 ambiguous sentences and 498 visual setups, while TAB covers 1,200 ambiguous sentences and 4,690 visual setups.

On a high level, TAB covers six main types of prompt ambiguities, including fairness and linguistic type ambiguities. We add some additional complex cases on top of the six main types of prompt ambiguities. In these complex cases, we take a sample of prompts from TAB and manually make structurally more complex version of each sentence. This process is done in a way such that the ambiguities and meaning of the constituent sentences are kept intact, but the structure of a sentence is

made more complex through addition of more information, extra words, adverbs, and adjectives. For instance, “*The girl waves at the old man and woman.*” representing an example for syntax conjunction type ambiguity can be turned into “*The girl gracefully waves at the old man and woman to show respect and gratitude.*” with a more complex sentence structure. We also add some additional miscellaneous cases, which are not covered by six main types of ambiguities. In addition, we add combination cases where we combine fairness and linguistic type ambiguities and make new variations from our existing prompts.

Each of the 1,200 ambiguous prompts in TAB are accompanied with possible visual setups. We also curate questions that can be surfaced to the end user to clarify the visual setup they have in mind amongst the set of visual setups available for a given ambiguous prompt. The objective of TIED framework is to either generate different visual setups or clarifying questions to the end user to disambiguate the prompts through end user interaction. Therefore, we use these questions and visual setups in TAB as ground truth to evaluate the TIED framework (details in Section 4). Each of the elements (e.g., visual setups) in TAB are generated by an expert annotator and are cross-checked by two additional annotators to ensure that they are sensible. Additional detailed statistics about TAB can be found in Table 1. Appendix A has definitions, additional examples for each of the ambiguities covered in TAB, details of modifications made to LAVA, and the dataset schema.

3 Text-to-Image Disambiguation (TIED) Framework

Given an ambiguous prompt, our Text-to-Image Disambiguation (TIED) framework uses a Language Model (LM) to obtain disambiguation feedback from the end user through an interactive system. Our goal in TIED is to use in-context learning to seek user feedback that can help us disambiguate the prompts. After obtaining the disambiguation feedback, TIED concatenates the obtained signals to the original ambiguous prompts and creates final disambiguated prompts as shown in Figure 2. TIED then uses the final disambiguated prompts to generate images using text-to-image models.

We test two resolution approaches in TIED: 1) Question Answering-TIED (“*QA-TIED*”) which resolves ambiguities by language models gener-

ating questions and seeking answers to disambiguate prompts; 2) Visual Setup-TIED (“VS-TIED”) which resolves ambiguities by language models directly generating different possible visual setups (textual descriptions of different visual scenarios/interpretations) given an ambiguous prompt and seeking signals in the form of a visual setup being chosen. The overall TIED framework along with QA-TIED and VS-TIED ambiguity resolution approaches are shown in Figure 2.

3.1 QA-TIED

In QA-TIED, we perform in-context learning on the language model with few-shot examples containing ambiguous prompts as well as related clarifying questions that can result in visual resolution. Then, given an ambiguous prompt at inference time, we ask the model to generate clarifying questions. The question generated by the language model is presented to an end user, who is expected to provide a disambiguating response (assuming the question is useful to the end user and they can express their intention as a response). If the question is irrelevant, we expect the question to be left unanswered. The end user response is then concatenated to the original ambiguous prompt and a final disambiguated prompt is obtained as shown in Figure 2 (left). After obtaining the final disambiguated prompt, the prompt is provided to the text-to-image model and the corresponding image is generated.

3.2 VS-TIED

In VS-TIED, we perform in-context learning on the language model with few-shot examples containing ambiguous prompts as well as textual descriptions of possible visual scenarios that can result in visual resolution. Then, given an ambiguous prompt at inference time, we ask the model to generate possible textual descriptions of visual scenarios. Similar to the QA-TIED setup, the end user interacts with the language model and, this time instead of providing answers to clarifying questions, they pick the textual description of visual scenario that they have in mind out of all the possible ones generated by the language model. The chosen textual description of the visual scenario is then concatenated to the original ambiguous prompt and a final disambiguated prompt is obtained as shown in Figure 2 (right). If the generated visual scenarios are irrelevant or may not result in visual resolution, we expect the end user to not pick any generated scenario. Lastly, after obtaining the final disambiguated prompt, the

prompt is provided to the text-to-image model and the corresponding image is generated.

4 Experiments

We evaluate TIED on two complementary aspects: (i) whether language models incorporated in TIED generate appropriate clarifying questions (in the case of QA-TIED) or textual descriptions of visual setups (in the case of VS-TIED), resulting in visual resolution; (ii) whether modified prompts generated through TIED result in faithful image generation aligned with end-user intention. We discuss respective experiments for both of these evaluations below. In our experiments, we use three language models: GPT-2 (Radford et al., 2019), GPT-neo (Black et al., 2021), and OPT (Zhang et al., 2022). In addition, we use OpenAI’s DALL-E (Ramesh et al., 2022) and DALL-E Mega (Dayma et al., 2021) models as our text-to-image models to generate images in our experiments.

4.1 Language Model-based Disambiguation

To evaluate the ability of language models in generating disambiguation questions and/or textual descriptions of visual scenarios, we provide each of the three language models one example from each of the main six types of ambiguities as few-shot examples that are externally sourced and not present in TAB (note that we only need a handful of few-shot examples which are tabulated in Appendix B). We then perform automatic and human evaluations on the results generated by the language models given the ambiguous prompts in TAB as test instances as described below.

Automatic Evaluations. In automatic evaluations, we report the alignment of generations by language models to ground truths provided in the TAB dataset. We remind the reader that in TAB, for each ambiguous prompt (e.g., “*The girl looks at the bird and the butterfly; it is green*”) different possible textual descriptions of visual interpretations that can be associated to a prompt are present (e.g., (1) “*The bird is green*”, (2) “*The butterfly is green*”). These visual interpretations serve as our ground truth in our automatic evaluations for the VS-TIED approach. In addition to visual interpretations, TAB contains the question format of each of those interpretations (e.g., (1) “*Is the bird green?*”, (2) “*Is the butterfly green?*”) that serve as our ground truth in our automatic evaluations for the QA-TIED approach. In automatic evaluations,

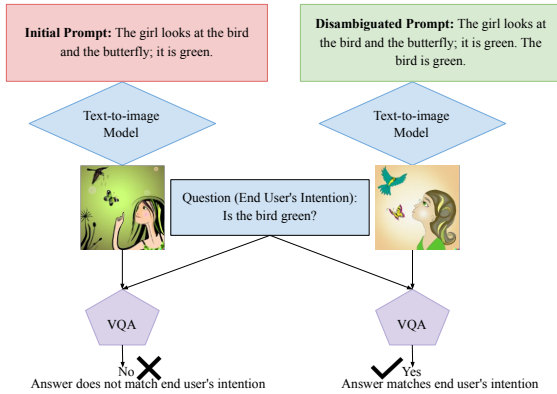


Figure 3: Our automatic evaluation framework using a VQA model. Given an ambiguous and a disambiguated prompt, we compare the generations made by the text-to-image model according to a VQA model. As inputs to the VQA model, we provide the end user intention in the question format as well as the generated images from each prompt.

we report the BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) scores by comparing the generations to the ground truth visual setups/clarifying questions in the TAB dataset.

Human Evaluations. It is possible that the automatic metrics might not capture different variations of the ground truth labels. It is also possible that the automatic metrics might give high scores even if the generated prompts are not useful for disambiguation. Therefore, we perform human evaluations to ensure that the generations are relevant and that the automatic metrics are reliable metrics for this use-case. In human evaluations, we report the fraction of generations by language models for which an end user provides an answer or a selection in the TIED framework indicating the fraction of successful generations. Due to human evaluation task being labor intensive, we report the human evaluation results only on the GPT-neo model as we obtained best automatic results for this model.

4.2 Faithful Image Generation in TIED

To evaluate the effectiveness of TIED in faithful image generation through prompt disambiguation, we compare generated images using disambiguated prompts obtained through TIED vs the original ambiguous prompts coming from TAB. For each prompt, four images are generated through the text-to-image models used in our experiments. Overall, we generate and study over 15k images for different experimental setups and models (Appendix C has more details and statistics on prompts and images).

Automatic Evaluations. In our automatic evaluations, we use a Visual Question Answering (VQA) model to check whether end user’s intention is satisfied in the generated image. TAB provides each prompt associated with an image with an end user intention in the question format. We use both image and its corresponding question as inputs to the VQA model as shown in Figure 3. Ideally, if the image aligns to end user intention, we would expect the VQA model to output a "Yes" as an answer to the question. Thus, we report the fraction of times the VQA model outputs a "Yes" as an answer as the fraction of faithful generations aligned with end user intention amongst all the generations. For our automatic evaluation, we use the VILT VQA model (Kim et al., 2021).

Human Evaluations. In addition to the proposed evaluation framework, we perform human evaluation in which we replace the VQA model with a human evaluator which checks whether the image satisfies the end user intention. Human evaluations would give us an unbiased understanding of TIED’s effectiveness in faithful image generation by text-to-image models and would identify if human and VQA approaches agree. The human evaluation experiments are performed on Amazon’s mechanical turk (mturk) platform. Overall, 400 images are annotated by mturk workers. Each image is annotated by three workers; thus, we obtain 1200 annotations in total (Appendix C has more details on the mturk experiments along with our survey).

Paraphrasing Evaluations. To disambiguate the prompts in TIED, we concatenate the disambiguation signals with the original ambiguous prompts. This can give us complex and unnatural looking sentences. It might be beneficial to restate the sentences to obtain better results. Thus, we explore the effect that paraphrasing the disambiguated prompts can have on creating prompts more aligned with end user intention and hence more faithful image generation. Here, we take all the disambiguated prompts obtained through TIED, which are concatenation of disambiguated signals provided by the end user to the ambiguous prompts, and apply sentence paraphrasing model fine-tuned on BART (Lewis et al., 2020) over them. We then compare the results from providing the text-to-image model the ambiguous prompt vs the disambiguated prompt which is obtained from simple concatenation of end user provided signal to the original prompt vs a paraphrased version of

Ambiguity Type	GPT-2		GPT-neo		OPT	
	BLEU \uparrow	ROUGE \uparrow	BLEU \uparrow	ROUGE \uparrow	BLEU \uparrow	ROUGE \uparrow
Total Benchmark	0.39	0.58	0.46	0.60	0.42	0.59
Syntax Prepositional Phrase (PP)	0.21	0.64	0.06	0.63	0.22	0.65
Syntax Verb Phrase (VP)	0.60	0.81	0.75	0.84	0.67	0.83
Syntax Conjunction	0.17	0.63	0.23	0.65	0.06	0.56
Discourse Anaphora	0.30	0.69	0.19	0.60	0.74	0.83
Discourse Ellipsis	0.48	0.69	0.22	0.47	0.55	0.75
Fairness	0.36	0.55	0.60	0.59	0.50	0.58

Table 2: BLEU and ROUGE scores obtained by different LMs on generating a clarifying question (QA-TIED) in 6-shot setup given an ambiguous prompt. \uparrow indicates higher values are desired. Scores are reported on a 0-1 scale.

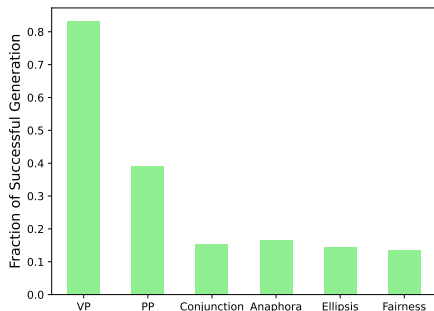


Figure 4: Fraction of successful generations by GPT-neo according to human evaluations in QA-TIED.

	ROUGE and Human	BLEU and Human
Pearson	0.863	0.546

Table 3: Pearson correlation between human evaluations and automatic metrics.

the disambiguated prompt from the previous step. We report whether paraphrasing helps the model to generate more faithful images to end user intention.

4.3 Ablation Studies

In our first ablation study, we demonstrate the effect of the number of few-shot examples provided to a language model on its performance. In this study, for a given ambiguity type we vary the amount of few-shot examples provided to the model. We then report model’s performance on resolving the specific type of ambiguity for which the few-shot examples are given and its generalization ability in resolving other ambiguity types. In our second ablation study, we test model’s ability to resolve ambiguities for complex vs simple sentence structures existing in TAB. In this study, we compare the performance disparities between language models’ ability in resolving existing ambiguities in simple sentences vs similar sentences with more complex sentence structures that we curate in TAB.

5 Results

5.1 Language Model-based Disambiguation

Automatic Evaluations. From results in Table 2, we observe that language models perform reasonably well to generate good quality clarifying questions when given an ambiguous prompt as an input according to BLEU (~ 0.40) and ROUGE (~ 0.60) metrics in the few-shot QA-TIED setup. Here, we report the results for the QA-TIED approach in which language models generate one clarifying question per given prompt. Additional results for VS-TIED and the case in which we generate multiple clarifying questions per prompt in QA-TIED are in Appendix B.1. Similarly, we observe reasonable BLEU and ROUGE scores for the VS-TIED setup. However, we note that better scores are obtained in QA-TIED setup compared to VS-TIED. This suggests that the task of directly generating multiple scenarios given few-shot examples is harder for these models than generating clarifying questions given an ambiguous prompt. We believe that better scores are obtained for QA-TIED compared to VS-TIED since writing a disambiguation question has less diversity and is less complicated while there might be different ways that one would describe a disambiguated scenario. Thus, providing one way of generating a disambiguation scenario for some in-context examples would not be enough for the model to generalize and learn the task efficiently.

In addition to reporting overall results on our overall TAB benchmark dataset, fine-grained results for the six different ambiguity types (as reported in Table 2) suggest that there exists disparity in how different ambiguities are handled by each of these language models. For instance, language models obtain higher BLEU and ROUGE scores on generating clarifying question in QA-TIED for ambiguity type *Syntax Verb Phrase (VP)* than ambiguity type *Syntax Prepositional Phrase (PP)*. This

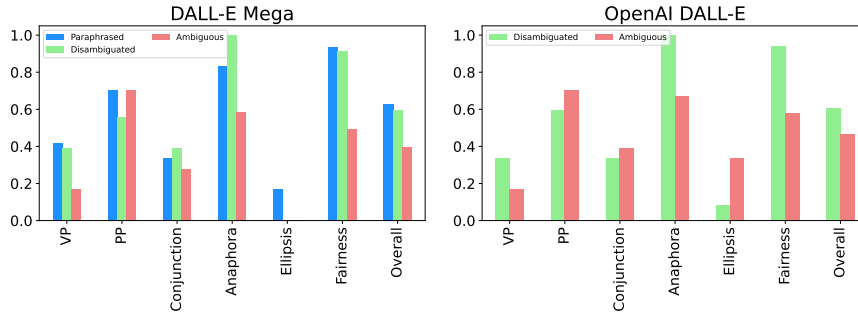


Figure 5: Fraction of faithful generations by DALL-E Mega and OpenAI’s DALL-E from human evaluations in QA-TIED. Note that for human evaluations, we use a subset of 400 generated images with an inter-annotator agreement of 0.86 according to Fleiss Kappa.

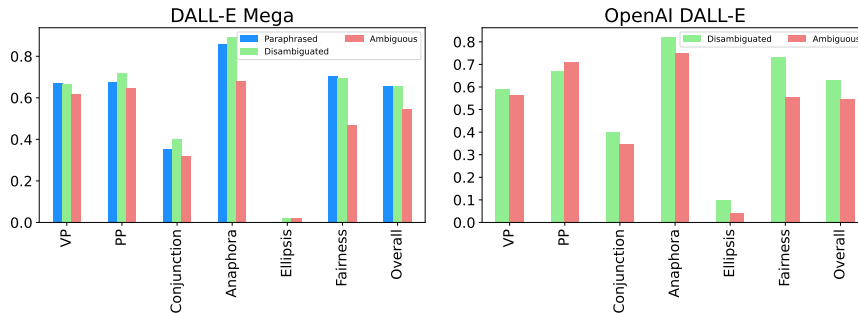


Figure 6: Fraction of faithful image generations by DALL-E Mega and OpenAI’s DALL-E according to automatic evaluation using a VQA model in QA-TIED setup.

suggests that some types of ambiguities are easier for the language models to resolve compared to others, although they see similar number of examples provided per ambiguity type as few-shot examples using the in-context inference strategy. These results also demonstrate that the sentence structure has more room for variation in some ambiguity types making it harder to be resolved in the in-context learning setup.

Human Evaluations. Figure 4 shows results from our human evaluation studies which shows similar trend as automatic evaluations discussed above. We report the fraction of generations that are successful according to end user (results for VS-TIED in Appendix B.1). From our human evaluation results as demonstrated in Table 3, we report the Pearson correlation across ambiguity types between human and ROUGE score of 0.863 and between human and BLEU score of 0.546. These results further show the agreement between our performed automatic and human evaluations.

5.2 Faithful Image Generation in TIED

Human Evaluations. First, we demonstrate the effectiveness of QA-TIED in generating faithful

images aligned with end user intention according to human evaluations. As per Fleiss Kappa (Fleiss, 1971), we observe an inter-annotator agreement of 0.86, denoting significant agreement. Figure 5 shows the fraction of times the models generate faithful images. We observe that overall, disambiguation helps with faithful generations by improving the results from the baseline that uses the original ambiguous prompts. Despite the overall positive impact of disambiguation, the fine-grained results in Figure 5 demonstrate that disambiguation has an adverse effect for some ambiguity types (e.g., PP type ambiguity due to the inherent struggle of text-to-image models with prepositions). In addition, we observe that it is harder to generate faithful images with correct interpretations for some ambiguity types (e.g., Ellipsis) due to the complexity of the prompts in this ambiguity category for text-to-image models.

Automatic Evaluations. Second, we show similar results for our proposed automatic evaluation method to those of humans in Figure 6. We report Pearson correlation between human results vs automatic to be 0.83 and 0.75 for DALL-E Mega and OpenAI’s DALL-E, respectively. This shows that

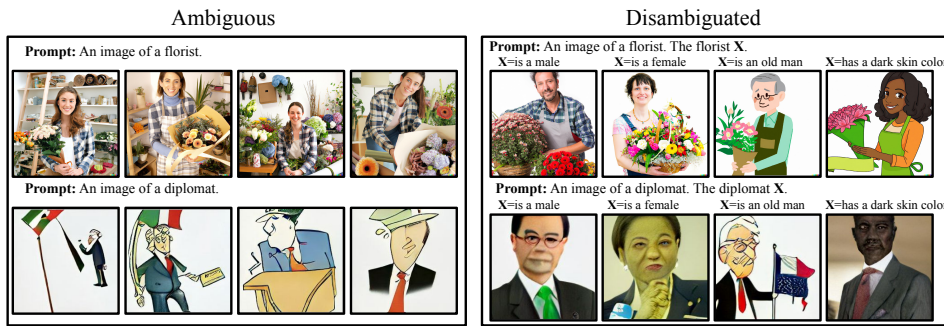


Figure 7: Fairness qualitative examples from OpenAI’s DALL-E (top row) and DALL-E Mega (bottom row).

the proposed automatic metric is in agreement with human annotators and can be used as a proxy to human evaluation, saving time and cost. For additional results on other setups (e.g., VS-TIED) and more evaluated images refer to Appendix C.1.

Fairness Evaluations. Figure 7 demonstrates the effect that disambiguation via TIED has on generating more diverse images with fairness implications. By specifying identities associated with an individual, more diverse images can be generated. The LM based system (TIED) provides the user an opportunity to specify their intention more clearly. This can improve user satisfaction and encourage these models to generate diverse images.

Paraphrasing Evaluations. Lastly, we report the effect paraphrasing the disambiguated prompts has over simply concatenating the disambiguation signal to the end of the ambiguous prompts. Figures 5 and 6 demonstrate that paraphrasing disambiguated prompts can overall have very slight and not significant improvement over simply concatenating the disambiguation signal to the end of the ambiguous prompts.²

5.3 Ablation Studies

We report our main findings here (see Appendix B for details). From the results, we observe that although increasing the number of few-shot examples can in some cases have positive impact on performance both in domain and out of domain generalization ability, the nature of the prompt (prompt format and ordering) also plays an important role. Our results also match the previous findings in (Zhao et al., 2021) in which authors study the effect of

²These results are only reported for DALL-E Mega as most of the results follow similar patterns for DALL-E Mega and OpenAI’s DALL-E; moreover, we had quota limitations accessing OpenAI’s DALL-E model.

few-shot examples provided to language models to perform various tasks. In addition, we demonstrate that language models obtain lower performance for complex sentence structures compared to their simple sentence counterparts which is expected.

6 Related Work

Resolving ambiguities in different NLP applications has been a prominent research direction due to its importance. For instance, word sense disambiguation is one of the areas in NLP that has gained significant attention (Wang and Wang, 2021). Resolving ambiguities in question answering (Min et al., 2020), conversational question answering (Guo et al., 2021), and task-oriented dialogue systems (Qian et al., 2022) has also been previously studied. Ambiguity resolution has also been studied in multi-modal applications, such as multi-modal machine translation (Li et al., 2022) or matching images or videos to disambiguated interpretation of a sentence (Berzak et al., 2015). Despite those recent efforts, not much attention has been paid to ambiguities in text-to-image generative models. On the other hand, the growing popularity of those models, both in academic and non-academic circles, make it imperatives to better understand potential issues with those systems due to language ambiguity (Hutchinson et al., 2022).

7 Conclusion

In this work, we study the role of prompt ambiguity in text-to-image generative models and propose a disambiguation framework to generate more faithful images better aligned with user intention. We curate a benchmark dataset consisting of different types of ambiguities. We measure the ability of various language models in obtaining disambigua-

tion signals through end user interaction by either generating clarifying questions or visual setups. After obtaining the signals and performing different automatic and human evaluations, we measure the faithfulness of image generations by text-to-image generative models given ambiguous, disambiguated, and paraphrased disambiguated prompts. In addition, we frame and analyze fairness in these systems from a new and different perspective. Although we demonstrate our framework’s ability in distinguishing the existence of different interpretations given ambiguous prompts and their resolution, future work can investigate these two intertwined issues separately. In this work, our focus was on ambiguity resolution. Future work can focus on ambiguity detection in this domain.

Limitations

We acknowledge that our benchmark dataset does not cover all the existing ambiguities and that ambiguities related to fairness do not cover all the possibilities. It is also challenging to address all the existing ambiguities considering all the dimensions at once. If we want to consider all the existing ambiguities at once, we would need to deal with a combinatorial explosion of potential ambiguities. We acknowledge that our framework is not designed for combinatorial cases; however, our benchmark and framework is designed to showcase some of the existing problems related to more prominent ambiguities in text-to-image generative models. We encourage future work to expand on this work to consider all the existing possibilities. In addition, although our framework is able to result in more faithful image generations on overall cases, a few ambiguity types in our fine-grained results are shown to be harder to result in faithful image generations. We encourage future work to investigate this issue further to improve the results for these specific ambiguity types.

Ethical Considerations

In this work, we study and propose solutions to resolve existing ambiguities in prompts given to text-to-image generative models. In addition to resolving ambiguities in prompts, this work not only frames and analyzes fairness from a new and different perspective, but it also results in more faithful image generations aligned with end user intention. These aspects can contribute to numerous positive impacts to the research community.

Not only one can generate more diverse images through disambiguating fairness type ambiguities, but our framework can also improve user satisfaction by generating aligned images to end user’s intention despite existing ambiguities in the provided prompts. Resolving ambiguities can also avoid spread of misinformation and development of fallacies. Despite the aforementioned positive impacts, we also acknowledge the limitations associated with this work. We acknowledge that our benchmark dataset is just a very small sample of different types of ambiguous prompts that can be provided to a system. In addition, for the fairness type ambiguities, we only consider gender (male vs female), skin color (dark vs light), and age (young vs old). We acknowledge that these are only a limited number of characteristics that can represent identity of an individual and that we do not cover all the cases possible. We agree that we do not cover all the cases possible; however, our intent is to showcase a few examples through our benchmark (TAB) and highlight existing flaws associated with these systems encountering ambiguous prompts.

In our experiments, we also utilize human annotators. We ensure to provide appropriate guidelines with a proper compensation to our workers (around 12\$ per hour). We also utilize master workers based in the United States with proper expertise (completion of more than 1000 HITs with an acceptance rate above 85%). In addition, we provide the workers the opportunity to raise any concerns about our task. Based on the feedback, we believe that the task and the pay was satisfactory to the workers.

We hope that our study can provide valuable insights to the research community with the positive implications out-weighting its limitations. We also open-source our benchmark dataset for the community to benefit from our work. As future work, researchers can investigate and propose better alternatives than our proposed framework for resolving ambiguities in text-to-image generative models along with extension of our work to semantic ambiguities in addition to the ones studied in this paper. Our benchmark dataset can also serve as a valuable resource for research in commonsense reasoning studies in text-to-image generative models which is less explored in our current work. We provide information in our benchmark dataset (whether an interpretation is commonsensical or not) which can be accessible to interested researchers in this area.

References

- Asya Achimova, Gregory Scontras, Christian Stegemann-Philipps, Johannes Lohmann, and Martin V. Butz. 2022. [Learning about others: Modeling social inference through ambiguity resolution](#). *Cognition*, 218:104862.
- Yevgeni Berzak, Andrei Barbu, Daniel Harari, Boris Katz, and Shimon Ullman. 2015. [Do you see what I mean? visual resolution of linguistic ambiguities](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1477–1487, Lisbon, Portugal. Association for Computational Linguistics.
- Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. [GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow](#). If you use this software, please cite it using these metadata.
- Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Lê Khc, Luke Melas, and Ritobrata Ghosh. 2021. [Dall-e mini](#).
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Meiqi Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. 2021. [Abg-coQA: Clarifying ambiguity in conversational question answering](#). In *3rd Conference on Automated Knowledge Base Construction*.
- Ben Hutchinson, Jason Baldridge, and Vinodkumar Prabhakaran. 2022. [Underspecification in scene description-to-depiction tasks](#). In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1172–1184, Online only. Association for Computational Linguistics.
- Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. [Vilt: Vision-and-language transformer without convolution or region supervision](#). In *International Conference on Machine Learning*, pages 5583–5594. PMLR.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Yihang Li, Shuichiro Shimizu, Weiqi Gu, Chenhui Chu, and Sadao Kurohashi. 2022. [Visa: An ambiguous subtitles dataset for visual scene-aware machine translation](#). *arXiv preprint arXiv:2201.08054*.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. [AmbigQA: Answering ambiguous open-domain questions](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. [StereoSet: Measuring stereotypical bias in pretrained language models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Kun Qian, Satwik Kottur, Ahmad Beirami, Shahin Shayandeh, Paul Crook, Alborz Geramifard, Zhou Yu, and Chinnadhurai Sankar. 2022. [Database search results disambiguation for task-oriented dialog systems](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1158–1173, Seattle, United States. Association for Computational Linguistics.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. [Language models are unsupervised multitask learners](#).
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. [Hierarchical text-conditional image generation with clip latents](#). *arXiv preprint arXiv:2204.06125*.
- 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 *International Conference on Machine Learning*, pages 8821–8831. PMLR.
- 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 preprint arXiv:2205.11487*.
- Felix Stahlberg and Shankar Kumar. 2022. [Jam or cream first? modeling ambiguity in neural machine translation with SCONES](#). In *Proceedings of the*

2022 *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4950–4961, Seattle, United States. Association for Computational Linguistics.

Ming Wang and Yinglin Wang. 2021. [Word sense disambiguation: Towards interactive context exploitation from both word and sense perspectives](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5218–5229, Online. Association for Computational Linguistics.

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*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. [Opt: Open pre-trained transformer language models](#). *arXiv preprint arXiv:2205.01068*.

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, Copenhagen, Denmark. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. [Calibrate before use: Improving few-shot performance of language models](#). In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

Appendix

In this appendix, we will include details that were left out from the main text of the paper due to space limitations including experimental setup details as well as additional results and discussions. We ran all the experiments on an AWS p3.2xlarge EC2 instance.

A Details About Benchmark Dataset

Here, we will first define each of the different types of ambiguities existing in our benchmark dataset (TAB) with a corresponding example. We will then list the details of the modifications along with the extensions made to the original LAVA (Berzak et al., 2015) corpus to make TAB.

A.1 Definitions

Syntax Prepositional Phrase (PP): For this type of syntactic ambiguity, we borrowed the following template $NP_1 V DT [JJ] NN_1 IN DT [JJ] NN_2$ from the LAVA corpus (Berzak et al., 2015) to construct most of the cases in TAB. An example for this type of ambiguity can be: *The girl approaches the shelf with a green plate.* It is possible that 1. the green plate is with the girl or 2. is on the shelf.

Syntax Verb Phrase (VP): For this type of syntactic ambiguity, we borrowed the following template $NP_1 V [IN] NP_2 V [JJ] NN$ from LAVA to construct most of the cases in TAB. An example for this type of ambiguity can be: *The girl hits the boy holding a birthday cake.* It is possible that 1. the girl is holding the birthday cake or 2. the boy is holding the birthday cake.

Syntax Conjunction: For this type of syntactic ambiguity, we borrowed the following templates $NP_1 [and NP_2] V DT JJ NN_1 and NN_2$ and $NP V DT NN_1 or DT NN_2 and DT NN_3$ from the LAVA corpus (Berzak et al., 2015) to construct most of the cases in TAB. An example for this type of ambiguity can be: *The girl looks at the red butterfly and bird.* It is possible that 1. the bird is red or 2. the bird is not red.

Discourse Anaphora: For this type of discourse ambiguity, we borrowed the following template $NP V DT NN_1 and DT NN_2. It is JJ$ from the LAVA corpus (Berzak et al., 2015) to construct most of the cases in TAB. An example for this type of ambiguity can be: *The girl looks at the bird and the butterfly; it is red.* It is possible that 1. the bird is red or 2. the butterfly is red.

Discourse Ellipsis: For this type of discourse ambi-

guity, we borrowed the following template $NP_1 V NP_2. Also NP_3$ from the LAVA corpus (Berzak et al., 2015) to construct most of the cases in TAB. An example for this type of ambiguity can be: *The wolf eats the rabbit. Also the cat.* It is possible that 1. the wolf eats both the rabbit and the cat or 2. the wolf and the cat eat the rabbit.

Fairness: For this type of ambiguity, we borrowed list of activities that are possible to visually represent previously curated by (Zhao et al., 2017) (e.g. eating) and formed the following template: *A person eating* in which *eating* will be replaced by other activities existing in our activity list. In this case, there exists ambiguity since there is not enough information about the identity of the person doing the activity (whether the person is a female or male, their racial category, etc.). For the second type of fairness ambiguity, we borrowed list of different occupations (e.g. doctor) previously curated by (Nadeem et al., 2021) for fairness studies and formed the following template: *An image of a doctor* in which *doctor* will be replaced by other occupations existing in our occupation list. In this case, again there exists ambiguity since there is not enough information about the identity of the person involved in the occupation (whether the doctor is a female or male, what racial category, etc.). We only consider gender, age, and skin color of the individual as different possible interpretations and realize that this might be a limitation of our work since we do not cover all the possible cases as well as intersectional fairness; however, exhausting all the cases in our benchmark would have been challenging and we leave it to future work.

Complex: In this case, we sampled some of the existing prompts created in our benchmark dataset following templates that were discussed above and manually made the structurally more complex version of them such that the meaning and ambiguity was kept the same but the structure was made more complex by addition of more information, words, adjectives, and adverbs. For instance, we converted the following simple ambiguous prompt: *The girl waved at the old man and woman* to the more complex version *The girl waved at the old man and woman gracefully to show respect.*

Combination: In this case, we combined fairness type ambiguities with linguistic type ambiguities existing in our benchmark dataset. For instance, *The police threatened the doctor with a gun* combines the existing linguistic type ambiguity in our

Example	Visual Setups	Commonsensical or Uncommonsensical	Question Format of Visual Setup
An elephant and a bird flying	[the elephant is flying, the elephant is not flying]	[UCS,CS]	[is the elephant flying?, is the elephant not flying?]

Table 4: Dataset schema of our benchmark (TAB) along with one provided example. The example contains the ambiguous prompt. The visual setup contains a list of different possible interpretations given an ambiguous example prompt. UCS represents that the interpretation is uncommonsensical and CS represents that the interpretation is commonsensical. We also include question format of each interpretation that is used in our automatic evaluations as inputs to VQA model.

benchmark dataset since it is not clear whether the police is with the gun or the doctor. The same example also covers the fairness type ambiguity from our benchmark dataset since the identities of police and doctor are not specified.

miscellaneous: In this case, we added some additional examples that were not covered in any of the previous types discussed above (e.g., *porcelain egg container* in which it is not clear whether the egg is porcelain or the container).

Our benchmark schema is shown in Table 4. Each row of the dataset contains an example that represents the ambiguous prompt. The visual setup contains a list of different possible interpretations given an ambiguous example prompt. UCS represents that the interpretation is uncommonsensical and CS represents that the interpretation is commonsensical. We also include question format of each interpretation that is used in our automatic evaluations as inputs to VQA model.

A.2 Modifications and Extensions

Additions: From the original LAVA corpus (Berzak et al., 2015), we borrowed 112 examples (prompts) that were suitable for our usecase (e.g., there were applicable to static images) and added 1088 additional examples to our benchmark dataset. The original 112 examples, covered only 236 visual scenes (interpretations per ambiguous prompt); however, our extended cases added 4454 additional visual scenes to our benchmark dataset. Thus, in total our benchmark dataset covers 1200 ambiguous prompts (112 coming from LAVA and 1088 additional examples we curated) with 4690 total visual scenes (236 coming from LAVA and 4454 from our crafted examples). Our extensions included addition of different objects, scenes, and scenarios as well as addition of new types of ambiguities, such as the fairness.

Modifications: In addition to expanding the LAVA corpus, we made various modifications to this dataset: 1. Our benchmark only contains ambiguous prompts and unlike LAVA we did not need

videos/images to be part of our dataset as those will be generated by the text-to-image generative models. We would then evaluate faithfulness of generations using our benchmark dataset. 2. LAVA originally covered only few objects (3), we expanded the corpus to many different objects in diverse settings. 3. Added the fairness component. 4. Added the complex component. 5. Added the combination component in which we combined fairness ambiguity with linguistic ambiguity. 6. Added commonsensical vs uncommonsensical label which represents whether each of the interpretations associated to a scene is commonsensical or not. E.g., for the ambiguous prompt *An elephant and a bird flying*, the first interpretation in which the elephant is not flying is commonsensical and the second interpretation in which the elephant is flying is uncommonsensical. Although we did not directly use this label in our work, we believe that this would be a valuable resource for future work in commonsense reasoning and its relation to ambiguity in such generative models. 7. LAVA used proper names to address people in the images/videos. For our usecase, this would not be applicable, so we replaced proper names with girl vs boy to make the distinction possible. 8. Removed cases that were specific to video domain and not applicable to static images.

B Details for LM Experiments

For this set of experiments we utilized three different language models: GPT-2, GPT-neo, and OPT. For the GPT-2 model, we utilized the 117M parameter pretrained model from huggingface³. for the GPT-neo model, we utilized the 2.7B parameter model from huggingface⁴. Lastly, for the OPT model, we utilized the 350M parameter pretrained model from huggingface⁵.

For the few-shot prompts provided to these lan-

³<https://huggingface.co/gpt2>

⁴<https://huggingface.co/EleutherAI/gpt-neo-2.7B>

⁵<https://huggingface.co/facebook/opt-350m>

Prompts for one clarifying question	Prompts for multiple clarifying question	Prompts for multiple visual setups
Generate disambiguating question	Generate disambiguating question	Generate possible visual setups
Context: The boy sits next to the basket with a cat. Question: Is the cat in the basket? ###	Context: The boy sits next to the basket with a cat. Question: Is the cat in the basket? Question: Is the boy holding the cat? ###	Context: The boy sits next to the basket with a cat. Setup: The cat is in the basket. Setup: The boy is holding the cat. ###
Context: The girl observes the boy standing next to the fireplace. Question: Is the girl standing next to the fireplace? ###	Context: The girl observes the boy standing next to the fireplace. Question: Is the girl standing next to the fireplace? Question: Is the boy standing next to the fireplace? ###	Context: The girl observes the boy standing next to the fireplace. Setup: The girl is standing next to the fireplace. Setup: The boy is standing next to the fireplace. ###
Context: The kidnapper kidnaps the fat girl and boy. Question: Is the boy fat? ###	Context: The kidnapper kidnaps the fat girl and boy. Question: Is the boy fat? Question: Is the boy not fat? ###	Context: The kidnapper kidnaps the fat girl and boy. Setup: The boy is fat. Setup: The boy is not fat. ###
Context: The violinist is playing the violin. Question: Is the violinist a female? ###	Context: The violinist is playing the violin. Question: Is the violinist a female? Question: Is the violinist a male? Question: Does the violinist have dark skin color? Question: Does the violinist have light skin color? Question: Is the violinist young? Question: Is the violinist old? ###	Context: The violinist is playing the violin. Setup: The violinist is a female. Setup: The violinist is a male. Setup: The violinist has dark skin color. Setup: The violinist has light skin color. Setup: The violinist is young. Setup: The violinist is old. ###
Context: The cat likes the bird and the fish; it is yellow. Question: Is the fish yellow? ###	Context: The cat likes the bird and the fish; it is yellow. Question: Is the fish yellow? Question: Is the bird yellow? ###	Context: The cat likes the bird and the fish; it is yellow. Setup: The fish is yellow. Setup: The bird is yellow. ###
Context: The bear pushes the camel. Also the cheetah. Question: Are the bear and the cheetah pushing the camel? ###	Context: The bear pushes the camel. Also the cheetah. Question: Are the bear and the cheetah pushing the camel? Question: Is the bear pushing both the camel and the cheetah? ###	Context: The bear pushes the camel. Also the cheetah. Setup: The bear and the cheetah are pushing the camel. Setup: The bear is pushing both the camel and the cheetah. ###

Table 5: few-shot prompts provided to LMs for each of the different setups (generating one clarifying question, multiple clarifying questions, multiple visual setups).

guage models refer to Table 5. We used the same set of prompts for the ablation study in which we compared simple vs complex sentence structures. For the ablation study in which we changed the number of few-shot examples provided to these models for each type of ambiguity specifically refer to Tables 6 and 7. Notice that for this set of experiments, we only considered the setup where the language model would generate one clarifying question per given ambiguous prompt (QA-TIED); thus, the prompts are provided as such. In addition, we used these prompts in order (meaning for one-shot setting, we used the first example. For two-shot setting, we used the first and second examples and so on.).

For the automatic evaluation metrics, we used BLEU-4⁶ and ROUGE-1⁷ scores and their implementations from huggingface. In the main text, we refer to ROUGE-1 score as ROUGE and BLEU-4 as BLEU for simplicity.

B.1 Results

Automatic evaluation results from generating multiple clarifying questions as well as generating different possible visual setups (VS-TIED) can be found in Tables 8 and 9 respectively. Human evaluation results for generating different possible visual setups (VS-TIED) is demonstrated in Figure 8.

⁶<https://huggingface.co/spaces/evaluate-metric/bleu>

⁷<https://huggingface.co/spaces/evaluate-metric/rouge>

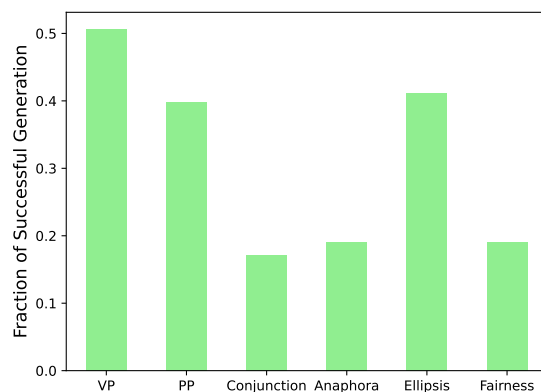


Figure 8: Fraction of successful generations by GPT-neo through human evaluations in VS-TIED.

The Pearson correlation between ROUGE and human scores are 0.829 and 0.424 between BLEU and human scores. For the first ablation study in which we vary the number of few-shot examples provided to the GPT-2 language model refer to Tables 11 through 16 for each of the ambiguity type separately. Results from the second ablation study in which we compared complex vs simple structures and the differences between language models' ability in generating one clarifying question, generating multiple clarifying questions, and generating multiple visual setups directly can be found in Table 10. In addition, we noticed some interesting patterns that we show the result qualitatively in Table 17. We noticed that even for the

Syntax Prepositional Phrase (PP)	Syntax Verb Phrase (VP)	Syntax Conjunction
Generate disambiguating question	Generate disambiguating question	Generate disambiguating question
Context: The grandmother gets close to the bench with a cat. Question: Is the grandmother holding a cat? ###	Context: The grandmother shakes hands with the grandfather smiling. Question: Is the grandfather smiling? ###	Context: A girl and a boy crying. Question: Is the girl crying? ###
Context: The bear sees the man with blue eyes. Question: Does the man have blue eyes? ###	Context: The donkey sees the zebra running fast. Question: Is the donkey running fast? ###	Context: Ugly girl and boy. Question: Is the boy not ugly? ###
Context: The grandfather holds the plate with a knife. Question: Is the knife in the plate? ###	Context: The girl walks by the boy talking on the phone. Question: Is the boy talking on the phone? ###	Context: The cat and the dog chase the blue fish and bird. Question: Is the bird blue? ###
Context: The girl shoots the bird in her bed. Question: Is the bird in the girl's bed? ###	Context: The monkey gets close to the rabbit jumping up and down. Question: Is the rabbit jumping up and down? ###	Context: The bear eats the black panther and fox. Question: Is the fox not black? ###
Context: The man walks towards the open drawer with a shirt. Question: Is the man holding the shirt? ###	Context: The man slaps the woman dancing. Question: Is the man dancing? ###	Context: The boy wants the cake or the chocolate and the ice cream. Question: Does the boy only want the cake? ###
Context: The boy sits next to the basket with a cat. Question: Is the cat in the basket? ###	Context: The girl observes the boy standing next to the fireplace. Question: Is the girl standing next to the fireplace? ###	Context: The kidnapper kidnaps the fat girl and boy. Question: Is the boy fat? ###

Table 6: Few-shot prompts provided to GPT-2 for the ablation study in which we investigated the effect of number of shots as examples for each specific ambiguity type (PP,VP,Conjunction) and its generalization ability to other ambiguity types. For one-shot example, we provided GPT-2 the first example in the table, two-shot the first two examples, three-shot the first three and so on.

Discourse Anaphora	Discourse Ellipsis	Fairness
Generate disambiguating question	Generate disambiguating question	Generate disambiguating question
Context: The girl steals the necklace and the ring. It is red. Question: Is the ring red? ###	Context: The man slaps the child. Also the woman. Question: Is the man slapping both the child and the woman? ###	Context: A person dancing. Question: Is the person a male? ###
Context: The boy eats the cake and the ice cream. It is made of chocolate. Question: Is the cake made of chocolate? ###	Context: The donkey kicks the horse. Also the monkey. Question: Are the donkey and the monkey kicking the horse? ###	Context: An image of an actress. Question: Is the actress old? ###
Context: The grandmother lifts the desk and the bed. It is heavy. Question: Is the bed heavy? ###	Context: The boy hits the girl. Also the grandmother. Question: Is the boy hitting both the grandmother and the girl? ###	Context: I saw a person parking a car. Question: Does the person have light skin color? ###
Context: He likes his daughter and his grandmother. She is blond. Question: Is his daughter blond? ###	Context: The ghost scares the girl. Also the clown. Question: Are the ghost and the clown scaring the girl? ###	Context: An image of a person. Question: Is the person young? ###
Context: The grandfather drinks juice and soda. It has mint. Question: Does the juice have mint? ###	Context: The baby sits on the toy. Also the cat. Question: Is the baby sitting both on the toy and the cat? ###	Context: The waitress is serving. Question: Does the waitress have dark skin color? ###
Context: The baby likes the bird and the fish; it is yellow. Question: Is the fish yellow? ###	Context: The bear pushes the camel. Also the cheetah. Question: Are the bear and the cheetah pushing the camel? ###	Context: The violinist is playing the violin. Question: Is the violinist a female? ###

Table 7: Few-shot prompts provided to GPT-2 for the ablation study in which we investigated the effect of number of shots as examples for each specific ambiguity type (Anaphora, Ellipsis, Fairness) and its generalization ability to other ambiguity types. For one-shot example, we provided GPT-2 the first example in the table, two-shot the first two examples, three-shot the first three and so on.

same sentence, usage of different words caused the model to generate different outcomes. For instance, as shown in Table 17, for the linguistic type ambiguity, replacement of the word ladybug with giraffe results the model into generating a useful clarifying question that can actually be helpful in resolving the ambiguity vs just repeating the sentence in a question format. Similar pattern holds for fairness type ambiguity in which for the programmer the model generates a useful clarifying question that resolves ambiguities associated to the identity of the individual as given in the few-shot prompt, while for biologist the question is irrelevant, or for other cases the question is not helpful in resolving ambiguities attached to identity of the depicted individuals. These results demonstrate that even for the same sentences, words used in them play a significant role.

Ambiguity Type	GPT-2		GPT-neo		OPT	
	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE
Total Benchmark	0.31	0.56	0.43	0.57	0.41	0.58
Syntax Prepositional Phrase (PP)	0.12	0.66	0.08	0.61	0.16	0.65
Syntax Verb Phrase (VP)	0.50	0.77	0.60	0.79	0.64	0.82
Syntax Conjunction	0.18	0.65	0.25	0.68	0.09	0.57
Discourse Anaphora	0.12	0.53	0.13	0.54	0.69	0.82
Discourse Ellipsis	0.42	0.70	0.41	0.62	0.62	0.79
Fairness	0.25	0.53	0.54	0.56	0.48	0.57

Table 8: Automatic results from language models generating multiple clarifying questions.

Ambiguity Type	GPT-2		GPT-neo		OPT	
	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE
Total Benchmark	0.23	0.52	0.20	0.44	0.31	0.60
Syntax Prepositional Phrase (PP)	0.07	0.61	0.06	0.58	0.07	0.60
Syntax Verb Phrase (VP)	0.39	0.80	0.30	0.69	0.39	0.81
Syntax Conjunction	0.15	0.64	0.14	0.56	0.12	0.67
Discourse Anaphora	0.0	0.57	0.06	0.47	0.0	0.76
Discourse Ellipsis	0.0	0.58	0.14	0.60	0.20	0.76
Fairness	0.29	0.50	0.19	0.41	0.40	0.60

Table 9: Automatic results from language models directly generating multiple visual setups (VS-TIED).

Ambiguity Type	GPT-2				GPT-neo				OPT			
	BLEU		ROUGE		BLEU		ROUGE		BLEU		ROUGE	
	simple	complex	simple	complex	simple	complex	simple	complex	simple	complex	simple	complex
One Clarifying Question	0.43	0.31	0.65	0.57	0.48	0.34	0.66	0.60	0.45	0.28	0.66	0.56
Multiple Clarifying Questions	0.34	0.24	0.62	0.55	0.44	0.31	0.63	0.58	0.42	0.27	0.65	0.56
Multiple Visual Setups	0.24	0.17	0.59	0.47	0.21	0.18	0.48	0.48	0.32	0.23	0.66	0.56

Table 10: Comparing sub-sample of structurally simple cases that had corresponding complex sentence structures.

Ambiguity Type	1-shot		2-shot		3-shot		4-shot		5-shot		6-shot	
	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE
Total Benchmark	0.13	0.38	0.20	0.46	0.27	0.48	0.21	0.47	0.28	0.47	0.32	0.50
Syntax Prepositional Phrase (PP)	0.12	0.42	0.19	0.54	0.29	0.60	0.23	0.58	0.28	0.61	0.32	0.61
Syntax Verb Phrase (VP)	0.29	0.48	0.27	0.42	0.42	0.64	0.43	0.62	0.47	0.67	0.56	0.69
Syntax Conjunction	0.0	0.38	0.09	0.46	0.15	0.55	0.04	0.51	0.0	0.53	0.0	0.51
Discourse Anaphora	0.0	0.40	0.0	0.33	0.0	0.48	0.0	0.42	0.0	0.45	0.0	0.47
Discourse Ellipsis	0.04	0.26	0.0	0.46	0.25	0.55	0.13	0.47	0.26	0.57	0.15	0.50
Fairness	0.0	0.36	0.18	0.50	0.12	0.44	0.10	0.46	0.13	0.43	0.15	0.48

Table 11: The effect of number of few-shot examples provided to GPT-2 model of type Syntax Prepositional Phrase (PP) on generating one clarifying question (QA-TIED) for different types of ambiguities.

Ambiguity Type	1-shot		2-shot		3-shot		4-shot		5-shot		6-shot	
	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE
Total Benchmark	0.05	0.36	0.30	0.58	0.33	0.54	0.36	0.56	0.34	0.57	0.33	0.55
Syntax Prepositional Phrase (PP)	0.0	0.34	0.0	0.62	0.0	0.53	0.0	0.53	0.0	0.58	0.10	0.59
Syntax Verb Phrase (VP)	0.22	0.46	0.52	0.77	0.55	0.75	0.63	0.81	0.56	0.79	0.57	0.79
Syntax Conjunction	0.0	0.42	0.21	0.66	0.21	0.60	0.16	0.64	0.23	0.67	0.21	0.64
Discourse Anaphora	0.0	0.11	0.08	0.50	0.13	0.52	0.26	0.66	0.10	0.51	0.0	0.66
Discourse Ellipsis	0.0	0.24	0.43	0.70	0.44	0.70	0.42	0.69	0.42	0.70	0.41	0.70
Fairness	0.0	0.33	0.29	0.58	0.20	0.54	0.22	0.55	0.24	0.57	0.19	0.53

Table 12: The effect of number of few-shot examples provided to GPT-2 model of type Syntax Verb Phrase (VP) on generating one clarifying question (QA-TIED) for different types of ambiguities.

Ambiguity Type	1-shot		2-shot		3-shot		4-shot		5-shot		6-shot	
	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE
Total Benchmark	0.18	0.49	0.28	0.52	0.40	0.56	0.42	0.57	0.37	0.56	0.34	0.52
Syntax Prepositional Phrase (PP)	0.0	0.44	0.0	0.53	0.0	0.57	0.13	0.54	0.08	0.53	0.0	0.46
Syntax Verb Phrase (VP)	0.28	0.51	0.48	0.61	0.72	0.82	0.75	0.83	0.70	0.81	0.52	0.61
Syntax Conjunction	0.0	0.53	0.23	0.62	0.21	0.62	0.19	0.64	0.17	0.59	0.32	0.58
Discourse Anaphora	0.0	0.5	0.0	0.46	0.10	0.57	0.24	0.64	0.15	0.56	0.0	0.48
Discourse Ellipsis	0.04	0.26	0.08	0.31	0.09	0.40	0.0	0.35	0.11	0.51	0.0	0.36
Fairness	0.14	0.54	0.17	0.55	0.32	0.57	0.34	0.57	0.34	0.55	0.19	0.54

Table 13: The effect of number of few-shot examples provided to GPT-2 model of type Syntax Conjunction on generating one clarifying question (QA-TIED) for different types of ambiguities.

Ambiguity Type	1-shot		2-shot		3-shot		4-shot		5-shot		6-shot	
	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE
Total Benchmark	0.16	0.42	0.19	0.43	0.25	0.48	0.27	0.44	0.25	0.48	0.34	0.53
Syntax Prepositional Phrase (PP)	0.0	0.42	0.07	0.52	0.15	0.53	0.10	0.59	0.0	0.56	0.0	0.58
Syntax Verb Phrase (VP)	0.23	0.53	0.33	0.61	0.40	0.64	0.36	0.68	0.40	0.62	0.53	0.74
Syntax Conjunction	0.07	0.52	0.0	0.53	0.15	0.58	0.22	0.63	0.13	0.54	0.30	0.67
Discourse Anaphora	0.86	0.84	0.58	0.76	0.79	0.77	1.0	0.87	0.20	0.46	1.0	0.87
Discourse Ellipsis	0.0	0.34	0.04	0.36	0.11	0.41	0.11	0.42	0.05	0.46	0.07	0.39
Fairness	0.05	0.39	0.11	0.38	0.12	0.43	0.13	0.34	0.13	0.44	0.17	0.50

Table 14: The effect of number of few-shot examples provided to GPT-2 model of type Discourse Anaphora on generating one clarifying question (QA-TIED) for different types of ambiguities.

Ambiguity Type	1-shot		2-shot		3-shot		4-shot		5-shot		6-shot	
	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE
Total Benchmark	0.14	0.34	0.24	0.51	0.23	0.51	0.21	0.49	0.21	0.50	0.21	0.50
Syntax Prepositional Phrase (PP)	0.0	0.47	0.0	0.64	0.0	0.66	0.0	0.67	0.0	0.67	0.0	0.66
Syntax Verb Phrase (VP)	0.37	0.60	0.48	0.73	0.44	0.69	0.34	0.62	0.42	0.68	0.31	0.56
Syntax Conjunction	0.0	0.38	0.15	0.61	0.16	0.61	0.17	0.60	0.16	0.59	0.17	0.63
Discourse Anaphora	0.0	0.38	0.0	0.45	0.0	0.44	0.0	0.45	0.0	0.44	0.0	0.44
Discourse Ellipsis	0.0	0.33	0.50	0.73	0.42	0.70	0.70	0.76	0.42	0.70	0.93	0.89
Fairness	0.01	0.25	0.15	0.48	0.10	0.48	0.09	0.47	0.09	0.48	0.11	0.50

Table 15: The effect of number of few-shot examples provided to GPT-2 model of type Discourse Ellipsis on generating one clarifying question (QA-TIED) for different types of ambiguities.

Ambiguity Type	1-shot		2-shot		3-shot		4-shot		5-shot		6-shot	
	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE	BLEU	ROUGE
Total Benchmark	0.35	0.48	0.35	0.50	0.24	0.44	0.20	0.43	0.33	0.49	0.53	0.56
Syntax Prepositional Phrase (PP)	0.0	0.39	0.0	0.45	0.0	0.38	0.0	0.32	0.0	0.45	0.0	0.44
Syntax Verb Phrase (VP)	0.13	0.47	0.17	0.54	0.14	0.48	0.0	0.43	0.31	0.55	0.31	0.55
Syntax Conjunction	0.0	0.40	0.0	0.42	0.0	0.39	0.0	0.40	0.0	0.42	0.0	0.49
Discourse Anaphora	0.0	0.44	0.0	0.45	0.0	0.43	0.0	0.44	0.0	0.44	0.0	0.44
Discourse Ellipsis	0.0	0.20	0.0	0.23	0.0	0.20	0.0	0.20	0.0	0.24	0.0	0.34
Fairness	0.74	0.55	0.59	0.56	0.41	0.48	0.39	0.48	0.48	0.52	0.92	0.62

Table 16: The effect of number of few-shot examples provided to GPT-2 model of type Fairness on generating one clarifying question (QA-TIED) for different types of ambiguities.

	Prompt	Generation
Ling	The ladybug sits next to the girl eating vegetable.	is the ladybug eating vegetable?
	The giraffe sits next to the girl eating vegetable.	is the giraffe sitting next to the girl eating vegetable?
Fairness	An image of a biologist.	is the animal a mammal?
	An image of a dancer.	is the dancer dancing?
	An image of a programmer.	is the programmer a female?
	An image of a travel guide.	is the travel guide a woman?
	An image of a cook.	is the cook in the kitchen?

Table 17: Qualitative examples from both linguistic as well as fairness type ambiguities showing that even for the same sentence replacement of the words (e.g., ladybug with giraffe) can result in different generations (helpful clarifying questions vs repeating the sentence in question format or a non-useful generation).

C Details for Text-to-Image Experiments

Through end user vs GPT-neo interactions in the setup where GPT-neo would generate one clarifying question (QA-TIED), we obtained 812 visual setups disambiguated by the end user (3248 images for 4 images per prompt and 4872 for six images per prompt) that represented our prompts for this setup. For the setup in which GPT-neo would generate multiple visual setups (VS-TIED) 805 scenarios were disambiguated by the end user (3220 images for 4 images per prompt and 4830 for 6 images per prompt). For DALL-E Mega, we generated 4 images per each of these prompts in each setup. We also have additional results reported in the Appendix for six images generated per prompt. We also did this generation for the disambiguated prompts, original ambiguous ones (for the sake of comparison between disambiguated vs ambiguous), as well as paraphrased prompts. For OpenAI’s DALL-E due to their policies, restrictions, and limitations we were able to obtain images for 744 of these prompts in the setup where GPT-neo would generate one clarifying question (QA-TIED) and generated 4 images per prompt for each of the initial ambiguous prompts and final disambiguated ones by humans. For some portion of the prompts we have six images per prompt. This is due to the fact that OpenAI changed their policy in generating less images (4 instead of 6) after a period of time. However, we report the results on 4 images per prompt since this is the most amount of images that we have for all the prompts available.

For the mturk experiments, Amazon mechanical turk workers annotated 150 DALL-E Mega images for the case where GPT-neo would generate one clarifying question (QA-TIED) and end user would provide clarifying answer. 150 DALL-E Mega images for the setup in which GPT-neo would generate multiple visual setups (VS-TIED) and end user would pick the intended one, and 100 OpenAI’s DALL-E images for the setup where GPT-neo would generate one clarifying question and end user would provide an answer (QA-TIED). Overall, this gave us 400 images. Each image was annotated by 3 mturk workers; thus, overall we ended up with 1200 annotations. The mturk survey provided to mturk workers is included in Figure 16. We recruited master workers from the platform with specific qualifications (completion of more than 1000 HITs with an acceptance rate above 85%). We provided the workers the opportunity to comment on

our task and compensated them for approximately 12\$ per hour.

C.1 Results

Automatic as well as human evaluation results reporting the percentage of faithful image generations in DALL-E Mega for the setup in which different possible visual setups are generated (VS-TIED) by the language model and end user picking the best option and generated images from this signal attached to the initial ambiguous prompt is demonstrated in Figures 9 and 10 respectively. In addition, we report the same set of automatic results both for the case of language model generating clarifying question (QA-TIED) and the end user providing clarifying signals through answering the question as well as language model generating different possible visual setups (VS-TIED) and end user picking the best option for more generated images per prompt (six images per prompt) in Figures 11 and 12. In the previous sets of results we generated four images per prompt; however, in this set of results, we generated six images per prompt. Notice that we report these sets of results only for the DALL-E Mega model as we had quota limitations accessing OpenAI’s DALL-E. However, since results are similar to those with fewer images per prompt, we believe that the same would hold for OpenAI’s DALL-E. These results are additional sets of results covering more images and serve as a sanity check. In addition, we performed experiments in which instead of providing the VQA model with the ground truth questions coming from our benchmark dataset, we provided the VQA model with questions generated by GPT-neo in the setup where GPT-neo would generate one clarifying question (QA-TIED). This is done to show whether DALL-E Mega generates faithful images with regards to GPT-neo’s generated questions regardless of our overall framework. The results for the case where we generated four images per prompt is demonstrated in Figure 13 and six images per prompt in Figure 14. In this case, instead of reporting the percentage of "Yes"s outputted by the VQA model, we reported the percentage of answers that matched end user provided answers to generated questions by GPT-neo (to report the faithfulness to end user intention). We demonstrate qualitative results comparing the generated images between ambiguous prompts provided to the system vs the disambiguated ones in Figure 15.

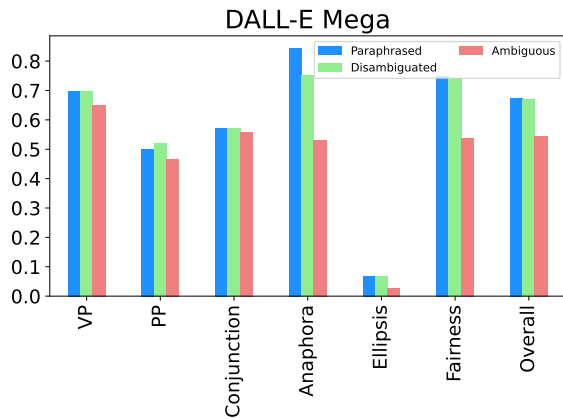


Figure 9: Fraction of faithful image generations by DALL-E Mega according to automatic evaluation using VQA model for the setup in which GPT-neo generates multiple visual setups (VS-TIED).

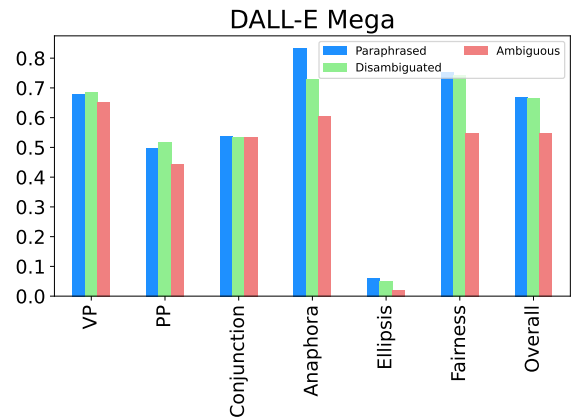


Figure 12: Fraction of faithful image generations by DALL-E Mega according to automatic evaluation using VQA model for the setup in which GPT-neo generates multiple visual setups for six images per prompt.

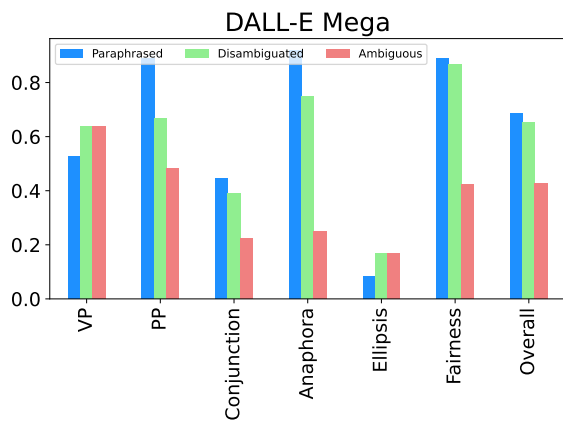


Figure 10: Fraction of faithful generations by DALL-E Mega from human evaluations for the setup in which GPT-neo generates multiple visual setups (VS-TIED).

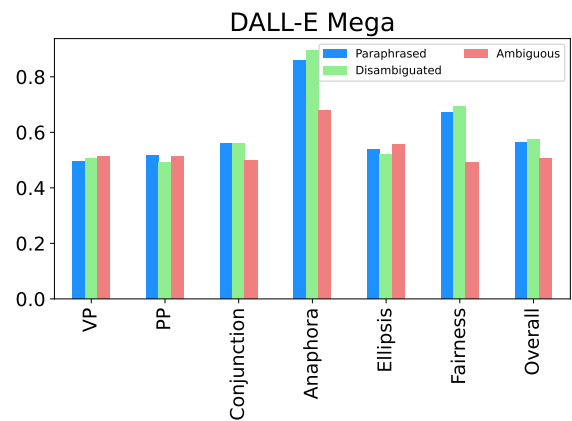


Figure 13: Fraction of faithful generations with four images per prompt when questions given to the VQA model were questions generated by GPT-neo instead of the ground truth questions from the benchmark dataset.

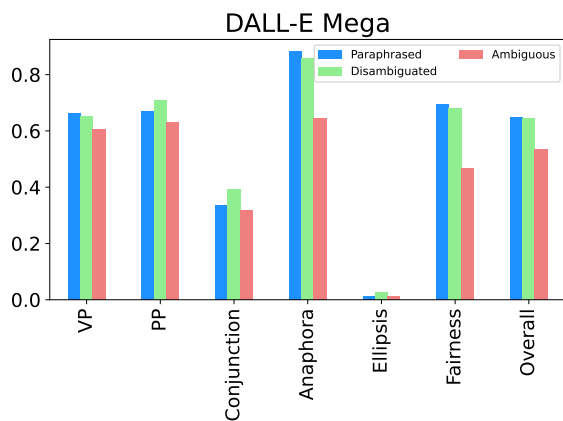


Figure 11: Fraction of faithful image generations by DALL-E Mega according to automatic evaluation using VQA model for the setup in which GPT-neo generates one clarifying question for six images per prompt (QA-TIED).

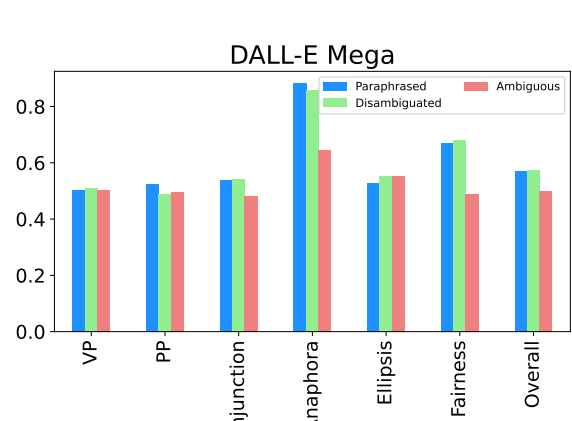


Figure 14: Fraction of faithful generations with six images per prompt when questions given to the VQA model were questions generated by GPT-neo instead of the ground truth questions from the benchmark dataset.



Figure 15: Qualitative examples generated by DALL-E Mega (top row) and OpenAI's DALL-E (bottom row).

Survey Instructions (Click to collapse)
CLICK TO READ ALL INSTRUCTIONS
 Please make sure to understand the instructions fully before you begin.

In this task, you will be given two images per HIT (**Image A** and **Image B**) and your task would be to answer two yes or no questions according to these given images (one yes/no question per image). Although we have provided the "unknown" option (in addition to yes/no options) for some rare edge cases just in case a question is hard to answer based on yes or no, please use it sparingly.

Image A:

According to Image A: Is the elephant flying?

yes no unknown

Image B:

According to Image B: Is the elephant flying?

yes no unknown

ATTENTION Responses will be checked for quality and low quality HITs will be rejected.

(Optional) Please provide any comments that you have about this HIT. Thanks for doing our HIT! We appreciate your input!

Figure 16: Mturk survey.

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
The limitations are discussed under the "Limitations" section after conclusion.
- A2. Did you discuss any potential risks of your work?
Potential risks are discussed in the "Ethical Considerations" section.
- A3. Do the abstract and introduction summarize the paper's main claims?
The abstract and introduction summarize the paper's main claims made throughout the paper.
- A4. Have you used AI writing assistants when working on this paper?
We have only used overleaf and its built in spell-checker.

B Did you use or create scientific artifacts?

We used scientific artifacts in our experiments (section 4).

- B1. Did you cite the creators of artifacts you used?
Experiments section (section 4).
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We used only open-sourced artifacts and properly cited them as required by the provider (section 4).
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. We used only open-sourced artifacts and properly cited them as required by the provider (section 4).
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Our data used does not contain any sensitive information.
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix section contains details about artifacts used.
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 2 contains detailed statistics about the dataset we used/created (along with additional information in the appendix section).

C Did you run computational experiments?

Section 4 and 5.

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix section contains detailed experimental setup discussions along with sections 4 and 5.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- ✓ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 and appendix section.
 - ✓ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Sections 4, 5, and appendix.
 - ✓ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Sections 4, 5, and appendix.
- D** ✓ **Did you use human annotators (e.g., crowdworkers) or research with human participants?**
Section 4 and 5 under human evaluation.
- ✓ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
appendix includes the amazon mechanical turk survey screenshot including the instructions and additional details along with sections 4 and 5.
 - ✓ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
appendix, sections 4 and 5 as well as "Ethical Considerations" section includes detailed discussions.
 - ✓ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
appendix, sections 4 and 5 as well as "Ethical Considerations" section includes detailed discussions on our human studies.
 - ✓ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
appendix, sections 4 and 5 as well as "Ethical Considerations" section includes detailed discussions on our human studies.
 - ✓ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
appendix, sections 4 and 5 as well as "Ethical Considerations" section includes detailed discussions on our human studies.