Enhancing Multimodal Learning with Knowledge Graphs

Ananya Kulshrestha Ethan Harbaugh Julianna Schneider

Abstract

Mulitmodal models hold great potential to address tasks in accessibility, where certain modalities may be inaccessible to users due to health conditions. One such task is generating scene descriptions for images, where capturing situational understanding and relationships between objects is critical. However, standard neural-network-based captioning systems have no guarantees of developing an understanding of such relationships. We propose the integration of knowledge graphs (KG), networks whose edges describe relationships between the entities represented as nodes, to provide this prior knowledge. We present a language model that leverages image embeddings, KG embeddings, and a recurrent neural network to generate image captions. We find that it outperforms the control model along certain benchmarks and provides a proof of concept for the utility of KG-based inductive bias in pursuit of contextually enriched scene descriptions.

1. Introduction

Multimodal models—a growing sub-group of natural language processing (NLP) models that process data from multiple formats, including text and images—have demonstrated increasing potential for solving pressing societal issues that require systems to operate across diverse information types. One such areas is accessibility, where multimodal models can perform tasks to bridge gaps in areas difficult for individuals with sensory disabilities, like image description, navigation assistance, and language translation/captioning [11, 18]. However, these tasks require models to encode implicit relationships between entities across different modalities, for example, identifying characteristics that make a walking path inaccessible. The populations which are likely to be best served by these models are also those most vulnerable to the gaps in performance caused by insufficient relational knowledge.

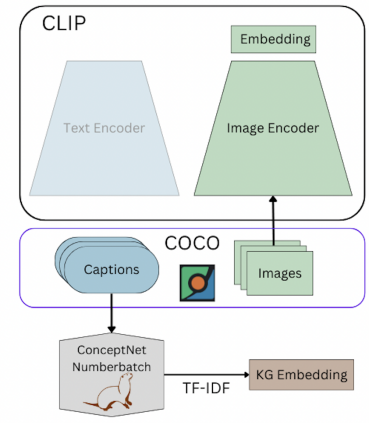
In pursuit of improved performance in this domain, we propose integrating knowledge graphs (KGs) into multimodal models. Knowledge graphs offer structured, relational data with contextual knowledge about objects, their relationships, and their functions in various scenarios. Models equipped with these structures can be trained to utilize them to avoid the aforementioned errors due to a lack of situational context [16]. We investigate the effects of incorporating KGs in multimodal models in constrained situations and evaluate our model on a task that assessess its performance in an accessibility use case: captioning images of everyday objects and scenes for visually impaired users.

2. Related Works

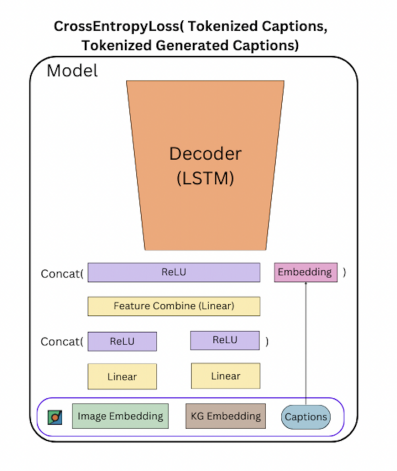
One such multimodal model is CLIP, a state-of-the-art model developed by OpenAI that excels in co-training image and text embeddings, optimizing a learning objective across both modalities [10]. This design enables CLIP to perform a variety of tasks, from image recognition and search to zero-shot classification, and, indirectly, to assist in generating descriptive labels for images.

Recent research has built towards applying multimodal image-captioning models to accessibility tasks through the development of evaluation datasets curated to the domain. One such example is VizWiz, which introduced an innovative dataset specifically designed for accessibility applications [4]. This dataset contains images taken by visually impaired users along with questions they would ask about the visual content, illustrating the unique challenges in accessibility contexts, such as atypical object placements, nonstandard angles, and complex backgrounds. Gurari et al. (2020) expanded on this with the VizWiz-Captions dataset [5], which consists of over 39,000 images, each with multiple descriptive captions. This dataset highlights the critical need for image captioning systems that can produce context-aware, detailed descriptions that convey not just the objects in a scene but also their relationships and the overall context.

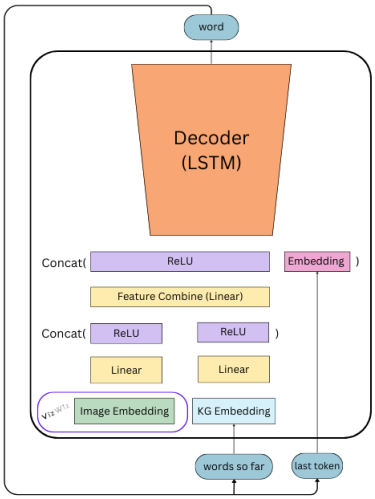
One promising approach to enhancing image captioning models for accessibility applications is the integration of knowledge graphs (KGs). Knowledge graphs offer structured, relational data that can supplement multimodal models with contextual knowledge about objects, their relationships, and their functions in various scenarios. In text-only NLP, KGs have shown great promise in improving model performance on tasks that require relational understanding [15, 19]. By integrating KGs with multimodal models, we investigate whether this semantic information can enable models to generate captions that are not only accurate but also enriched with background knowledge, thereby provid-



(a) Embedding stage showing extraction of embeddings from images, captions, and knowledge graph (KG) features.



(b) Feature combination stage where embeddings are projected, concatenated, and passed through a linear layer with ReLU activation, and later concatenated with caption embeddings.



(c) Illustration of the model in autoregressive inference, using previous tokens to generate the next word.

Figure 1. Network architecture showing the embedding extraction, feature combination, and inference stages.

ing visually impaired users with more informative descriptions that go beyond simple object identification.

In practical applications, systems that generate relationally aware and contextually enriched captions would offer substantial benefits for visually impaired users. Microsoft's Seeing AI project [7] provides a prime example of a learning-based system that would benefit from such enhancements. The system only provides real-time descriptions of objects and scenes but its output is limited to straightforward object labels, often missing the nuanced relational context that would make captions more situationally meaningful. Knowledge graphs could bridge this gap by enabling models to produce captions that encapsulate both the presence of objects and their interactions, such as "a lamp on a bedside table next to a book," providing a clearer mental image for the user.

Recent work in image captioning also reflects a growing trend towards emphasizing models' abilities to display relational understanding. For example, the Object Relation Transformer [6] explicitly models relationships between objects, allowing for more contextually relevant captions. The approach in Sec. 3 differs in that we provide access to external knowledge sources of relational knowledge, rather than attempting to learn these relations without any priors.

Building on this host of work across the diverse fields of KGs, image captioning, and LLMs, we perform a benchmark of one approach to integrating learnings from each domain. By incorporating KGs, we hope to further enrich these models, allowing them to leverage structured background knowledge to generate more detailed and contextually appropriate descriptions, particularly in complex or ambiguous scenes.

3. Methods

We present a language model that uses CLIP image embeddings to generate image captions via a recurrent neural network (NN). Our accessibility case study of interest is captioning images taken by visually impaired users to provide contextually relevant scene descriptions. This motivates our use of a lightweight predictive model to reflect the limits of the low-compute hardware that is most broadly available to the population we aim to serve. To model the breadth of data quality that accessibility captioning systems encounter in real-world scenarios, we train on a state-of-the-art dataset, COCO [3], and test on a dataset collected by visually impaired users and captioned by crowdsourced annotators, VizWiz Captions [4].

Our model consists of three major components: embedding (see Fig. 1a), language modeling (see Fig. 1b), and decoding. Together, these three stages enable our model to integrate visual and semantic knowledge about the captioning task with the goal of generating contextually-enriched captions.

3.1. Embedding

Image Embedding. We use CLIP's [10] image encoder to project our images of interest to a high-dimensional vector embedding E_{image}. These embeddings capture fine-grained visual details and are semantically aligned with textual descriptions, providing a meaningful representation of the image content. We choose to embed our images, rather

than provide them to the language model directly, to enable a shared latent space between our image and KG embeddings (see Sec. 3.2).

KG Embedding. We use the pre-trained ConceptNet Numberbatch embeddings to represent our target captions in a form that reflects the relationships modeled in the ConceptNet KG [12]. These embeddings encode the relationships between entities in the ConceptNet KG in a word2vec-style embedding. Therefore, each word in the target caption can be represented as a ConceptNet Numberbatch vector $E_{\text{kg\_word}}(w)$. We then determine the relative importance of these words using the Term Frequency-Inverse Document Frequency (TF-IDF) scheme, which computes the relevance of a word based on its frequency in the caption and its overall occurrence across captions in the dataset (see Eqn. 1).

$$\text{TF-IDF}(w) = \text{TF}(w) \cdot \log\left(\frac{N}{\text{DF}(w)+1}\right) \tag{1}$$

where $\text{TF}(w)$ represents the term frequency of word w in the caption, $\text{DF}(w)$ denotes the document frequency of w (the number of captions containing w), and N is the total number of captions. By applying TF-IDF weighting, rare but meaningful words are given more weight in the final representation. The caption-level KG embedding $E_{\text{kg\_caption}}$ is computed by aggregating the word embeddings, weighted by their respective TF-IDF values (see Eqn. 2).

$$E_{\text{kg\_caption}} = \frac{\sum_{w \in \text{caption}} \text{TF-IDF}(w) \cdot E_{\text{kg\_word}}(w)}{\sum_{w \in \text{caption}} \text{TF-IDF}(w)} \tag{2}$$

This approach ensures that semantically important words contribute more significantly to the overall caption embedding, creating a representation that aligns with the meaning of the caption. Figure 2 illustrates how embeddings of important words like "stairs" and "outline" dominate the caption's representation, compared to less meaningful words like "which" or "a".

3.2. Language Modeling

To move from embeddings to probability distributions over sequences of words, we train a custom model that first projects our image and caption embeddings into a shared latent space then applies a Long-Short-Term-Memory (LSTM) based recurrent NN [13] in an autoregressive fashion (see Fig. 1b).

First, we reconcile the image and caption-level KG embeddings' original dimensions, 512 and 300, respectively, by projecting them into a shared latent space of dimension 256. This transformation is achieved through learnable linear layers followed by ReLU activations (Eqn. 3).

$$\begin{aligned} E_{\text{proj\_img}} &= \text{ReLU}(W_{\text{image}} \cdot E_{\text{image}} + b_{\text{image}}) \\ E_{\text{proj\_kg}} &= \text{ReLU}(W_{\text{kg}} \cdot E_{\text{kg\_caption}} + b_{\text{kg}}) \end{aligned} \tag{3}$$

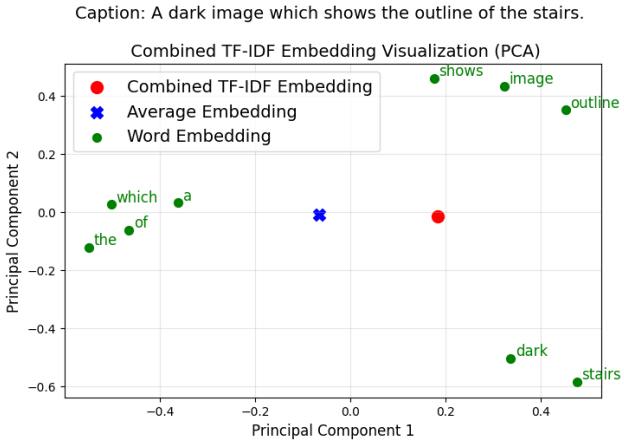


Figure 2. PCA visualization highlighting the distribution of word embeddings (green), the unweighted caption embedding (blue), and the TF-IDF weighted caption embedding (red). The TF-IDF embedding is closer to semantically important words, like 'stairs,' 'outline,' and 'dark', than the unweighted embedding.

The projected embeddings $E_{\text{proj\_img}}$ and $E_{\text{proj\_kg}}$ are then concatenated and passed through an additional linear layer, resulting in a unified feature vector (Eqn. 4).

$$\begin{aligned} E_{\text{linear}} &= W_{\text{comb.}} \cdot \text{concat}(E_{\text{proj\_img}}, E_{\text{proj\_kg}}) + b_{\text{comb.}} \\ E_{\text{comb.}} &= \text{ReLU}(E_{\text{linear}}) \end{aligned} \tag{4}$$

To integrate the last word (either target, for training, or predicted, for inference) into the feature representation, the last word is passed through an embedding layer, producing E_{caption}. This embedding is then concatenated with the combined image and KG features to form the final representation that is used as input to a one-layer LSTM (Eqn. 5).

$$E_{\text{input}} = \text{concat}(E_{\text{comb.}}, E_{\text{caption}}) \tag{5}$$

The model is trained to generate captions by minimizing the cross-entropy loss between the predicted and ground truth captions. This objective ensures that the predicted sequence aligns with the target sequence.

For comparison, we train a model with the same architecture as but omit the KG embeddings. This is achieved by performing only the first of the projections in Eqn. 3 and setting $E_{\text{comb.}} = \text{ReLU}(E_{\text{proj\_img}})$ in Eqn. 5. All other aspects of the training and inferencing pipeline remain identical. From here forth, we refer to the model with KG embeddings as "With KG" and the control as "Without KG".

3.3. Decoding

We generate captions at inference time using a modified version beam search. We allow for a maximum caption length of 20 words, and pad shorter predicted captions

as needed. Our beam search modifications were motivated by empirical observation that greedy decoding and standard beam search were notably susceptible to repetition. We implement three main heuristics: length normalization, repetition penalty, and trigram blocking. The length normalization heuristic re-weights the candidate beams' scores so as to avoid unfair favoring of shorter predictions over longer ones. The scaling coefficients seen in Eqn. 6 were informed by Wu et al.'s work [17], where s denotes a candidate beam's score.

$$p_{\text{length}} = \left(\frac{5 + \text{len}(\texttt{sequence})}{6}\right)^{0.65}$$
$$s_{\text{norm}} = \frac{s_{\text{original}}}{p_{\text{length}}} \tag{6}$$

The repetition penalty is applied as a 0.5 scaling to the logits that have appeared in the sub-sentence preceding them. Finally, in the trigram blocking step, if a group of three consecutive words is repeated within a candidate sequence, that sequence is discarded. Similarly, if any word is repeated more than twice, the sequence it belongs to is discarded as well. To ensure sufficient sequences are maintained, up to six beams are considered as candidates – once three valid beams (the desired beam width) make it through the heuristic checks, the top-scoring candidate is selected.

3.4. Training & Inference

Datasets. For training, one of the five reference captions provided by COCO for each image is randomly chosen to create unique image-caption pairs, leading to a sample size of 17.9K. For inference, all five of the crowdsourced reference captions provided by VizWiz-Captions for each image is used to compute the evaluation metrics in Sec. 4. Images where one or more captions described "quality issues too severe" to be captioned were discarded, leading to a sample size of 5k.

Embeddings. During training, we compute both the image and KG embeddings offline before training our language model (see Sec. 3.2). During testing, we mix offline and online computation, pre-computing the image embeddings and computing the KG embedding that corresponds to the caption we've predicted so far at inference time (see Fig. 1c).

4. Results

To evaluate the quality of our model's captions and their similarity to the crowdsourced VizWiz-Captions targets, we examined its BLEU-1 [9] and ROUGE-L [8] scores and compare it to our control. We select BLEU-1 to quantify the quality of our model's word choice and ROUGE-L to measure its sentence-level performance with respect to the target captions. As seen in Table 1 and Fig. 3, the model with KG inductive bias has a higher mean score across both metrics. To compute a two-tailed 95% confidence interval for each score, we bootstrapped the BLEU-1 scores for each of our test images then looked to the 2.5th and 97.5th percentiles. Since the models' 95% confidence intervals (CI) for their respective BLEU-1 scores do not overlap, we can conclude that the model with KG's superior performance on the BLEU-1 metric is statistically significant. Meanwhile, the overlapping 95% CIs on the ROUGE-L metric indicate no statistically significant conclusions can be drawn.

	BLEU-1		ROUGE-L	
	Mean	95% CI	Mean	95% CI
With KG	**28.34**	[28.11, 28.52]	**14.38**	[5.56, 26.67]
Without KG	22.88	[22.61, 23.16]	12.51	[00.00, 25.00]

Table 1. The model with knowledge graph inductive bias scores higher and has lower spread than the control along both metrics, which measure word-for-word matching and similarities in overall sentence structure with the target captions.

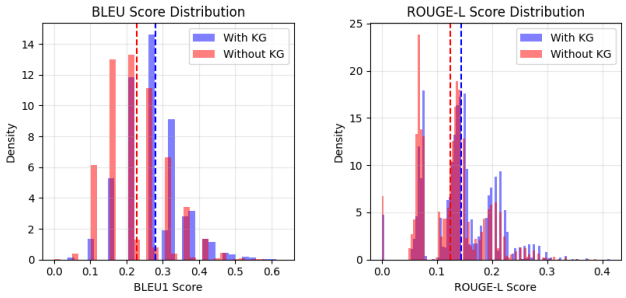


Figure 3. Histogram of BLEU-1 and ROUGE-L scores across the test set. The model with KG (blue) has a higher density of high-scoring scores compared to the model without KG (red).

To further characterize the differences between our model with KG and without KG, we compute each model's CIDEr [14] and SPICE [1] scores. The former combines an n-gram evaluation and TF-IDF weighting to score the overall match between a predicted caption and its references, while the latter turns the predicted and target captions into scene descriptions that capture relationships between objects. As seen in Table 2, the control outperforms the model with KG in terms of CIDEr score, but the model with KG does better in terms of SPICE. Though both models' scores are generally low, it is noteworthy that the version of our model with KG embeddings outperforms the control in terms of SPICE score.

Table 3 provides an example image from the Viz Wiz test set and the captions produced by our model with the KG, without KG, and one of the five target captions. Qual-

	CIDEr	SPICE
With KG	1.58	**2.84**
Without KG	**2.47**	1.82

Table 2. The control outperforms the model with inductive bias along the CIDEr metric, which measures a scaled version of n-gram similarity to the reference captions, but the model with inductive bias outperforms the control along the SPICE metric, which measures accuracy of scene description.

itatively, we see that the model with KG captures relations present in the scene, such as the person being "on" the image, while the control does not. Similarly, the model with KG's word choice shows better categorical accuracy than the control's. It refers to "people", "image", and "man", all of which are present in the scene description, and even its less accurate word choice, like "woman", "guy", and "boy", are still somewhat relevant to the objects in the scene it's captioning. In contrast, the control uses words like "black dog", "orange juice", and "blender", none of which have categorical similarity with the depicted scene. However, we empirically observe that, overall, the control demonstrates greater diversity in its word choice than the model with KG. We hypothesize that this may reflect how the prevalence of certain entities or relationships in the KG, such as people, may bias the model's output towards referring to the these common entities and relationships – this warrants further investigation in future work.

5. Summary & Outlook

In conclusion, we present a language model that leverages image embeddings and KG embeddings to generate captions that describe scenes for visually impaired users. Our model leverages latent spaces, recurrent NNs, and a modified beam search decoding algorithm in pursuit of contextually enriched and relevant image captions. We compare our method against a control and find that its mean score outperforms the control on three out of the four benchmarks, one of which is guaranteed to be a statistically significant improvement in score by a 95% CI. Considering that we train on a high-quality dataset, COCO, and evaluate on a notably challenging one, VizWiz-Captions, this is noteworthy.

Our method has room for improvement in the overall strength of its generative captions and score magnitude. For reference, a state-of-the-art attention-based model [2], when pre-trained on COCO and evaluated on Viz Wiz, scored 52.8 on BLEU-1, 35.8 on ROUGE, 18.9 on CIDEr, and 5.8 on SPICE [4]. We believe the score discrepancy between this reference and our models reflects the limited expressive power of our architecture. Specifically, Anderson et al. combine a bottom-up and top-down attention to

Model	Caption
Without KG	A black dog lying next to pick up close shot of orange juice carton on the blender with bat
With KG	A person on an image of someone is another man holding a woman guy one boy in black
Target	A view of a building with a picture of a man on it with a couple big trees in front of it

Table 3. Predicted and target captions for a VizWiz-Captions test set image. The model with inductive bias shows a qualitative better match in terms of entities referenced and the relationships between them.

identify meaningful regions of an image at different scales. Due to using image embeddings, which obscure direct access to the original non-vector representation, and the nature of LSTM's internal operations, our model cannot learn detailed features in the same way. Furthermore, as a result of the user group capturing the reference images, they tend to be not focused, cropped in non-standard ways, and have objects obscuring one another. Our networks have not seen similarly low-quality data when training on COCO and are likely ill-equipped to model and perform through the noise that comprises such a large train-test gap.

Nonetheless, our work serves as a proof of concept for the utility of KGs in multimodal learning. The increase in standard image captioning scores and qualitative advantages of the model with KG compared to the model without KG provides evidence that even a simple embedding integration scheme, like our shallow section of the model that projects embeddings to a shared latent space, has potential for future work. We believe interesting avenues for further study include exploring the KG embeddings we use in a stronger generative framework, such as one with attention,

and experimenting with more complicated architectures to embed the image and KG data in a shared latent space.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation, 2016.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[3] Lin et al. Coco: Common objects in context, 2017. https://cocodataset.org/home.

[4] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people, 2018.

[5] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind, 2020.

[6] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words, 2019.

[7] Sherlock Huang, Salman Gadit, Antony Deepak Thomas, Anirudh Koul, Meher Kasam, Serge-Eric Tremblay, Eren Song, Wes Sularz, Mary Bellard, Anne Taylor, Abhinav Shrivastava, Margaret Mitchell, Ross Girshick, Kartik Sawhney, Ishan Misra, Gaurang Prajapati, Saqib Shaikh, and Microsoft Research Team. Seeing ai, 2016.

[8] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[11] Hasib-Al Rashid, Argho Sarkar, Aryya Gangopadhyay, Maryam Rahnemoonfar, and Tinoosh Mohsenin. Tinyvqa: Compact multimodal deep neural network for visual question answering on resource-constrained devices, 2024.

[12] Joshua Chin Robyn Speer and Catherine Havasi. Conceptnet 5.5: An open, multilingual knowledge graph, 2017.

[13] Hasim Sak, Andrew W Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. 2014.

[14] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015.

[15] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. Kepler: A unified model for knowledge embedding and pre-trained language representation, 2020.

[16] Sean Williams and James Huckle. Easy problems that llms get wrong, 2024.

[17] Yonghui Wu. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[18] Bufang Yang, Lixing He, Kaiwei Liu, and Zhenyu Yan. Viassist: Adapting multi-modal large language models for users with visual impairments, 2024.

[19] Liang Yao, Chengsheng Mao, and Yuan Luo. Kg-bert: Bert for knowledge graph completion, 2019.

Impact Statement

Our project revolves around challenges faced by people in need of accessibility accommodations. We specifically focus on solving the task of scene description for visually impaired people in need of assistance in navigating their environment. Through a technological solution, we aim to help members of this population interact with their environment more safely and confidently. While none of our team members belong to this group, we thoughtfully designed our methodology to ensure adequate representation of disabled people's experiences in our development and evaluation processes. Specifically, we evaluated our system only on images collected by visually impaired users as they navigate everyday scenes. This helps reflect the real-world challenges faced by visually impaired users and helps validate that the contributions are equitable, focused on the audience, and addresses the needs of underserved communities.

While this work seeks, and takes active steps, to serve underprivileged populations, we recognize the potential harms posed by deploying these technologies without proper testing and design. For example, once a user heavily relies on this tool to perform daily functions, they are susceptible to malfunctions, such as irrelevant or inaccurate captioning. Our current results constitute a strong proof of concept, but future work is required to improve generation quality and work towards a production-ready solution. Furthermore, every individual's needs are different, and the metrics we used as a proxy for improving accessibility are only some of many important factors. Similarly, the single dataset we used to evaluate our model may not be representative of the captioning tasks encountered by individuals whose disability or life experiences differ from the VizWiz-Captions population. In summary, this research is designed to explore methods for improving prospective technologies and the possible impacts of implementation should be considered with levity to ensure ethical use, especially when designing solutions for vulnerable populations.