Supplementary Material for Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation

Visual Context				
Textual Context	Oh my gosh, i'm so buying this shirt.	I found a cawaii bird.	Stocking up!!	Only reason I come to carnival.
Question	Where did you see this for sale?	Are you going to collect some feathers?	Ayee! what the prices looking like?	Oh my God. How the hell do you even eat that?
Response	Midwest sports	There are so many crows here I'd be surprised if I never found one.	Only like 10-20% off..I think I'm gonna wait a little longer.	They are the greatest things ever chan. I could eat 5!

Table 1: Example conversations in $\text{IGC}_{\text{Twitter}}$.

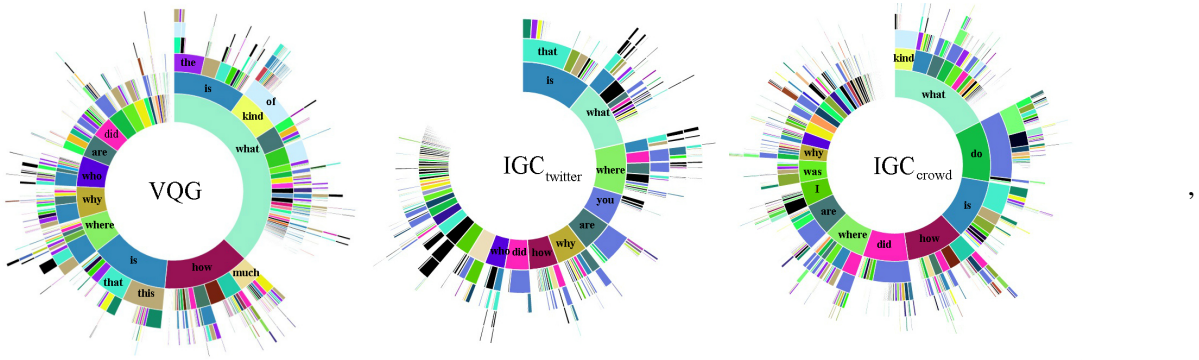

Figure 1: Sunburst visualization of distributions of n-gram sequences (with $n \leq 6$) in questions in VQG, $\text{IGC}_{\text{Twitter}}$, and $\text{IGC}_{\text{Crowd}}$. $\text{IGC}_{\text{Twitter}}$ is the most diverse set, with the lighter-colored part of the circle indicating sequences with less than 0.1% representation in the dataset.

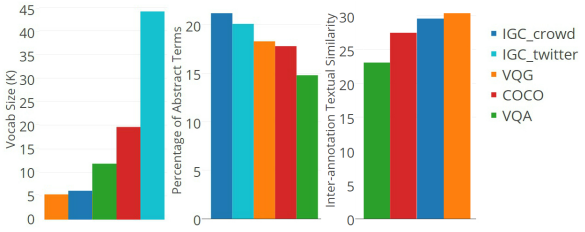

Figure 2: Comparison of IGC questions with VQG (Mostafazadeh et al., 2016) and VQA (Antol et al., 2015) questions in terms of vocabulary size, percentage of abstract terms, and inter-annotation textual similarity. The COCO (Lin et al., 2014) image captioning dataset is also included as a point of reference. The $\text{IGC}_{\text{Twitter}}$ dataset has by far the largest vocabulary, making it a more challenging dataset for training purposes. The $\text{IGC}_{\text{Crowd}}$, followed in order by $\text{IGC}_{\text{Twitter}}$, exhibit the highest ratio of abstract to concrete terms. Broadly, abstract terms refer to intangibles, such as concepts, qualities, and feelings, whereas concrete terms refer to things that can be experienced with the five senses. Conversational content may often involve more abstract concepts than captions or questions directly targeting visible image content.

The right-hand plot in compares the inter-annotation textual similarity of our $\text{IGC}_{\text{Crowd}}$ questions using a smoothed BLEU metric (Lin and Och, 2004). Contextually grounded questions of $\text{IGC}_{\text{Crowd}}$ are competitive with VQG in inter-annotation similarity.

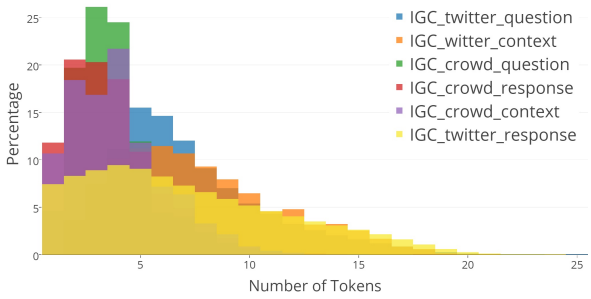

Figure 3: Distribution of the number of tokens across subsections of the datasets. On average, $\text{IGC}_{\text{Twitter}}$ has longer sentences.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proc. ICCV*.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proc. ACL*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proc. ECCV*.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proc. ACL*.