

Supplemental: Probing Image-Language Transformers for Verb Understanding

Lisa Anne Hendricks Aida Nematzadeh

DeepMind

{lmh, nematzadeh} @google.com

1 The Annotation Pipeline

We provide figures depicting the interfaces of our three annotation tasks.



- ☐ This image won't load. No need to answer any more questions; move onto the next hit.
- ☐ This image is a cartoon. No need to answer any more questions; move onto the next hit.

Q1: Can you see the verb **play** in the image?

- ☐ Yes
- ☐ No
- ☐ Unsure

Q2: Can you see a/an **activist** and a/an **guitar** in the image?

- ☐ Yes
- ☐ No
- ☐ Unsure

Figure 1: The interface of our first task for collecting images.

Describe the image below with a short grammatical sentence using the three bolded words:

actor, play, cat?

Example: If the words are person, eat, apple, a possible sentence would be: "A person eating an apple."



Describe the image with a short sentence using the three words above...

☐ Cannot write a caption which describes the image accurately with three provided words

Figure 2: The interface of our second task for writing sentences for images.

Which image matches the sentence **The girl will go down the footpath.**?

- ☐ Both
- ☐ Image 1
- ☐ Image 2
- ☐ Neither
- ☐ Unsure

Image 1



Image 2



Figure 3: The interface of our third task for confirming images.