# **Evaluation of Image Descriptions**

Hello, thanks for being part of our research study. Our goal is to accurately characterize the performance of vision-language models (i.e., AI systems that can describe images). By doing so, you'll help us gauge how well such systems would perform in consequential settings such as the automatic generation of accessibility text for people who are blind or have low vision.

In our annotation interface, you'll see 1) an image, 2) a **CORRECT** description of the image and 3) two **GENERATED** descriptions of the image. Your task is to provide **relative grades** of the **GENERATED** descriptions across *three dimensions*: **mistakes**, **missing details** and **overall quality**.

**Mistakes** in **GENERATED** descriptions are incorrectly added details or identifications of nouns, their descriptors or their relationships that are not true of the image. **Missing details** are details in the **CORRECT** description that are not accounted for in the **GENERATED** descriptions **after correcting their mistakes**. **Overall quality** is more subjective – we want you to grade the generations by which one is the best stand-in for the **CORRECT** description.

For each task, please follow the instructions below:

- 1) Look at the image. Get a quick sense of any relevant people or objects, their actions and their broader setting.
- 2) Read the **CORRECT** description of the image.
- 3) Read the two **GENERATED** descriptions of the image.
- 4) Provide relative grades of the **GENERATED** descriptions by their **mistakes**, their **missing details** or their **overall quality** as specified in the interface. When providing these relative grades, please consider the **significance** of mistakes and missing details, not just their numerical quantity.
- 5) For any individual relative grade, if you are unsure between "Slightly Better" and "Much Better", please choose "Slightly Better". If you are unsure between "Slightly Better" and "Equal", please choose "Equal". Across all three dimensions, "better" indicates that a description is actually better i.e., has less significant mistakes, less significant missing details or higher overall quality.
- 6) After providing relative grades across all three dimensions, please hit SUBMIT to move on to the next task.

Please note that as you progress through your assigned tasks, you'll see the same image and **CORRECT** description twice (one after another), each time paired with two different

**GENERATED** descriptions. We provide more details on what constitutes a **mistake** and a **missing detail** below.

#### **Mistakes**



*Incorrectly Added Details* include anything in the **GENERATED** description that is not actually true in the image.

CORRECT: A dog on the beach.

**GENERATED:** A dog playing fetch on the beach.

*Incorrect Nouns* include misidentifications in the **GENERATED** description of nouns in the image. Specifically, they should be wrong. If below, instead of *otter* the generated description used *animal* (vague but correct), it would **not be** a mistake.

**CORRECT:** A dog on the beach. **GENERATED:** An otter on the beach.

*Incorrect Descriptors* include misidentifications in the **GENERATED** description of descriptors in the image. Additionally, they include errors in descriptor attachment (i.e., if the descriptor is true of a different noun in the image than specified in the description).

**CORRECT**: A dog on a sandy beach.

**GENERATED:** A black dog on a sandy beach.

*Incorrect Relationships* include misidentifications in the **GENERATED** description of relationships (verbal, positional, etc.) between two nouns in the image. Additionally, they include errors in relationship attachment (i.e., if the relationship is true of a different noun in the image than specified in the description).

**CORRECT**: A dog by the ocean. **GENERATED**: A dog in the ocean.

# **Missing Details**

Before judging missing details in a generation, please remember to account for the mistakes. You should be comparing the **CORRECT** description to a version of each

**GENERATED** description where all of its *mistakes* have been corrected. We include examples of *missing details* below.



Below, we highlight *sandy* as it's a detail that's not reflected in the **GENERATED** description.

**CORRECT**: A dog on a sandy beach. **GENERATED**: A dog on a beach.

Below, we **do not** highlight *dog* as we've already penalized the generation for mistakenly identifying an *otter* in its place.

CORRECT: A dog on a beach.
GENERATED: An otter on a beach.

Below, we highlight dog as the generation's use of animal is lacking in specificity.

CORRECT: A dog on a beach.

**GENERATED**: An animal on a beach.

## **Overall Quality**

When providing relative grades of *overall quality* (i.e., as stand-ins for the **CORRECT** description), please rely on your own preference and judgment.

### **Other Notes**

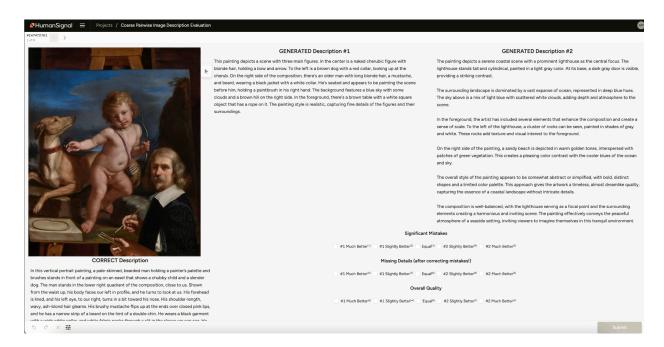
 on occasion, generated descriptions will include language that is not explicitly visually grounded but instead interprets the painting; you can ignore these cases when considering mistakes

#### **Annotation Interface**

You should see the annotations that you've been assigned here: https://app.heartex.com/projects/?page=1.

When you begin your annotations, you'll see the annotation interface below. Remember, first look at the image, then read the **CORRECT** description and then read the two **GENERATED** descriptions. Then, provide relative grades for mistakes, missing details and overall quality by selecting the appropriate checkboxes. After grading all three, please hit SUBMIT at the bottom.

As some generations are quite large, you might find it helpful to 1) close the panel on the right hand side and 2) zoom out (Command + Shift + [MINUS]) so you can consider the image and text simultaneously.



We have assigned you a number of tasks that should be completeable in the hour(s) you've been allocated. We expect each task to take no longer than 6 minutes (in our internal pilot, each task took 2.5 minutes on average).

We will be randomly sampling a subset of your annotations to verify their quality.

Thank you so much for your participation!