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.

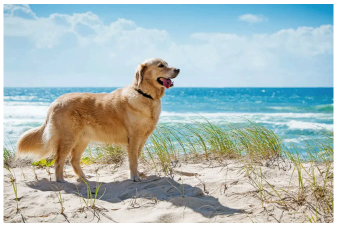
5) For any individual relative grade, if you are unsure between "Slightly Better" and "Much Better", please choose "Slightly Better". If you are unsure between "Slightly Better" and "Equal", please choose "Equal". Across all three dimensions, "better" indicates that a description is actually better – 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.

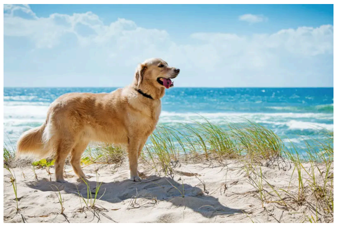
Incorrect Relationships include misidentifications in the **GENERATED** description of relationships (verbal, positional, etc.) between two nouns in the image. Additionally, they include errors in relationship attachment (i.e., if the relationship is true of a different noun in the image than specified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

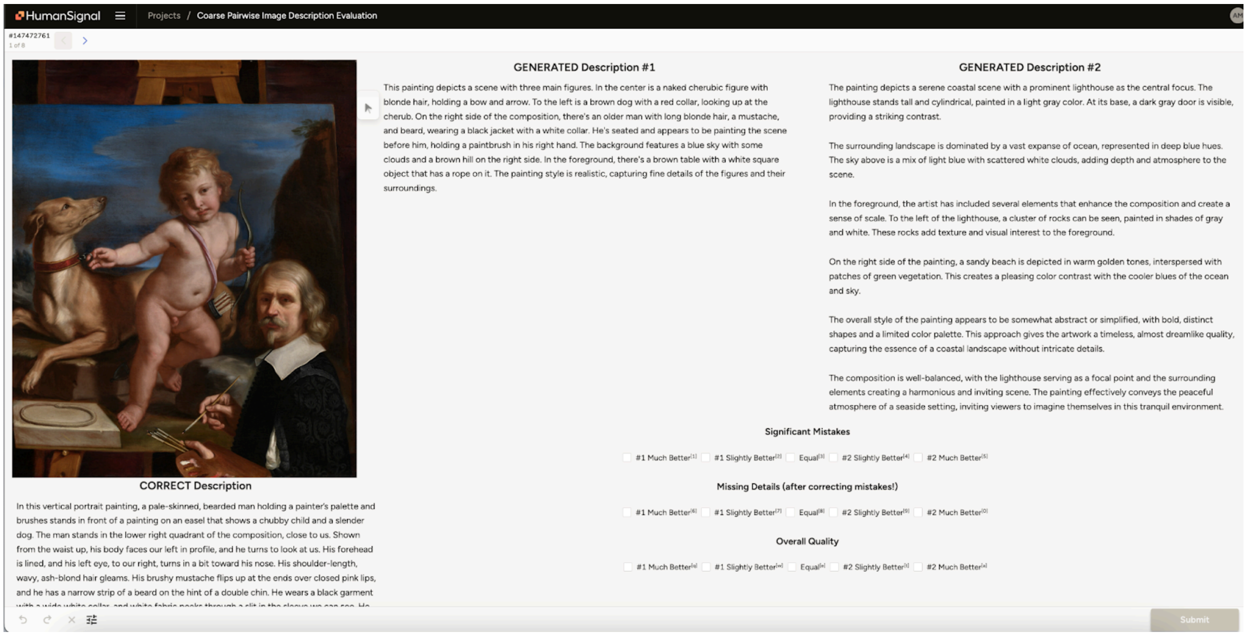
- on occasion, generated descriptions will include language that is not explicitly visually grounded but instead interprets the painting; you can ignore these cases when considering mistakes

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.



GENERATED Description #1

This painting depicts a scene with three main figures. In the center is a naked cherubic figure with blonde hair, holding a bow and arrow. To the left is a brown dog with a red collar, looking up at the cherub. On the right side of the composition, there's an older man with long blonde hair, a mustache, and beard, wearing a black jacket with a white collar. He's seated and appears to be painting the scene before him, holding a paintbrush in his right hand. The background features a blue sky with some clouds and a brown hill on the right side. In the foreground, there's a brown table with a white square object that has a rope on it. The painting style is realistic, capturing fine details of the figures and their surroundings.

GENERATED Description #2

The painting depicts a serene coastal scene with a prominent lighthouse as the central focus. The lighthouse stands tall and cylindrical, painted in a light gray color. At its base, a dark gray door is visible, providing a striking contrast.

The surrounding landscape is dominated by a vast expanse of ocean, represented in deep blue hues. The sky above is a mix of light blue with scattered white clouds, adding depth and atmosphere to the scene.

In the foreground, the artist has included several elements that enhance the composition and create a sense of scale. To the left of the lighthouse, a cluster of rocks can be seen, painted in shades of gray and white. These rocks add texture and visual interest to the foreground.

On the right side of the painting, a sandy beach is depicted in warm golden tones, interspersed with patches of green vegetation. This creates a pleasing color contrast with the cooler blues of the ocean and sky.

The overall style of the painting appears to be somewhat abstract or simplified, with bold, distinct shapes and a limited color palette. This approach gives the artwork a timeless, almost dreamlike quality, capturing the essence of a coastal landscape without intricate details.

The composition is well-balanced, with the lighthouse serving as a focal point and the surrounding elements creating a harmonious and inviting scene. The painting effectively conveys the peaceful atmosphere of a seaside setting, inviting viewers to imagine themselves in this tranquil environment.

Significant Mistakes

□ #1 Much Better[1] □ #1 Slightly Better[2] □ Equal[3] □ #2 Slightly Better[4] □ #2 Much Better[5]

Missing Details (after correcting mistakes!)

□ #1 Much Better[6] □ #1 Slightly Better[7] □ Equal[8] □ #2 Slightly Better[9] □ #2 Much Better[0]

Overall Quality

□ #1 Much Better[q] □ #1 Slightly Better[w] □ Equal[e] □ #2 Slightly Better[r] □ #2 Much Better[t]

CORRECT Description

In this vertical portrait painting, a pale-skinned, bearded man holding a painter's palette and brushes stands in front of a painting on an easel that shows a chubby child and a slender dog. The man stands in the lower right quadrant of the composition, close to us. Shown from the waist up, his body faces our left in profile, and he turns to look at us. His forehead is lined, and his left eye, to our right, turns in a bit toward his nose. His shoulder-length, wavy, ash-blond hair gleams. His brushy mustache flips up at the ends over closed pink lips, and he has a narrow strip of a beard on the hint of a double chin. He wears a black garment with a wide white collar, and white fabric peeks through a slit in the sleeve we can see. He

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!