Granular 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.

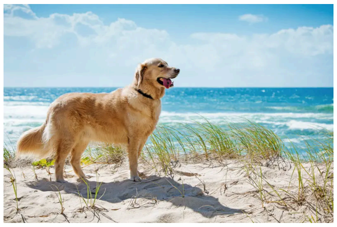
In our annotation interface, you'll see 1) an image, 2) a **CORRECT** description of the image and 3) a **GENERATED** description of the image. Your task is to first identify **minimal** spans in the **GENERATED** description that are ***mistakes*** (e.g. incorrectly added details that are not true of the image; identifications of nouns, their descriptors or their relationships that are not true of the image) and then identify **minimal** spans in the **CORRECT** description that are ***missing*** (e.g. details not reflected in the **GENERATED** 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 **GENERATED** description of the image.

4) Read the **GENERATED** description of the image again. As you encounter ***mistakes*** (e.g., incorrectly added details that are not true of the image or nouns, their descriptors or their relationships that are not true of the image), click & drag your cursor to select minimal spans of text to highlight the error. If you are not sure if something is a ***mistake***, please fall back to the information included in the **CORRECT** description. We include further information on different kinds of ***mistakes*** below.



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.

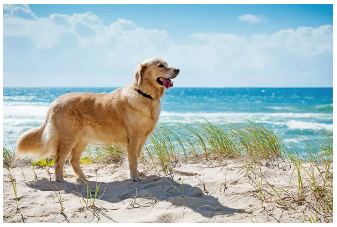
Incorrect Descriptors include misidentifications in the **GENERATED** description of descriptors in the image. Additionally, they include errors in descriptor attachment (i.e., if the descriptor is true of a different noun in the image than specified in the description).

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.

5) Read the **CORRECT** description of the image again. As you encounter ***missing*** details **that you did not mark as *mistakes* in the GENERATED description**, click and drag your cursor to select minimal spans of text in the **CORRECT** description to highlight what's missing. Mentally, you should be comparing the **CORRECT** description to a version of the **GENERATED** description where all the ***mistakes*** you previously identified have been corrected. We include examples 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 correct but lacking in specificity.

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

If at any point you want to undo the selection of a span, you can select it again with your mouse. After you're done selecting all *mistakes* and *missing details*, please hit **Submit** to save your progress and continue to the next task.

Other Notes

1) On occasion, generated descriptions will include language that is not explicitly visually grounded but instead attempts to interpret the painting; you can ignore these cases when marking mistakes.

2) When marking mistakes / missing details, please try to mark as little text as is required to narrowly identify the mistake or missing detail. For example, if a figure's "left hand is resting on a bench" but in the correct description their left hand is held aloft, please only mark "resting on a bench". This will help us greatly in targeting areas to improve the quality of assistive text.

3) If a generation is overly vague / general but correct, please mark the specifics that it is missing as missing details. For example, if a generation describes a "figure" but a correct description specifies a "child" or a generation describes "ornamentation" but a correct description specifies "pearls", please mark "child" and "pearls" in the correct description as missing details but do not mark "figure" / "ornamentation" as mistakes.

4) When judging mistakes, there is a reasonable question about strictness. We ask that you credit generations when they are reasonably close to a generation but still correct (i.e. a reasonable person with some familiarity with paintings would produce a similar description). Obviously there's some subjectivity here which is totally fine. The important thing to note is that it needn't be an exact match to the generation.

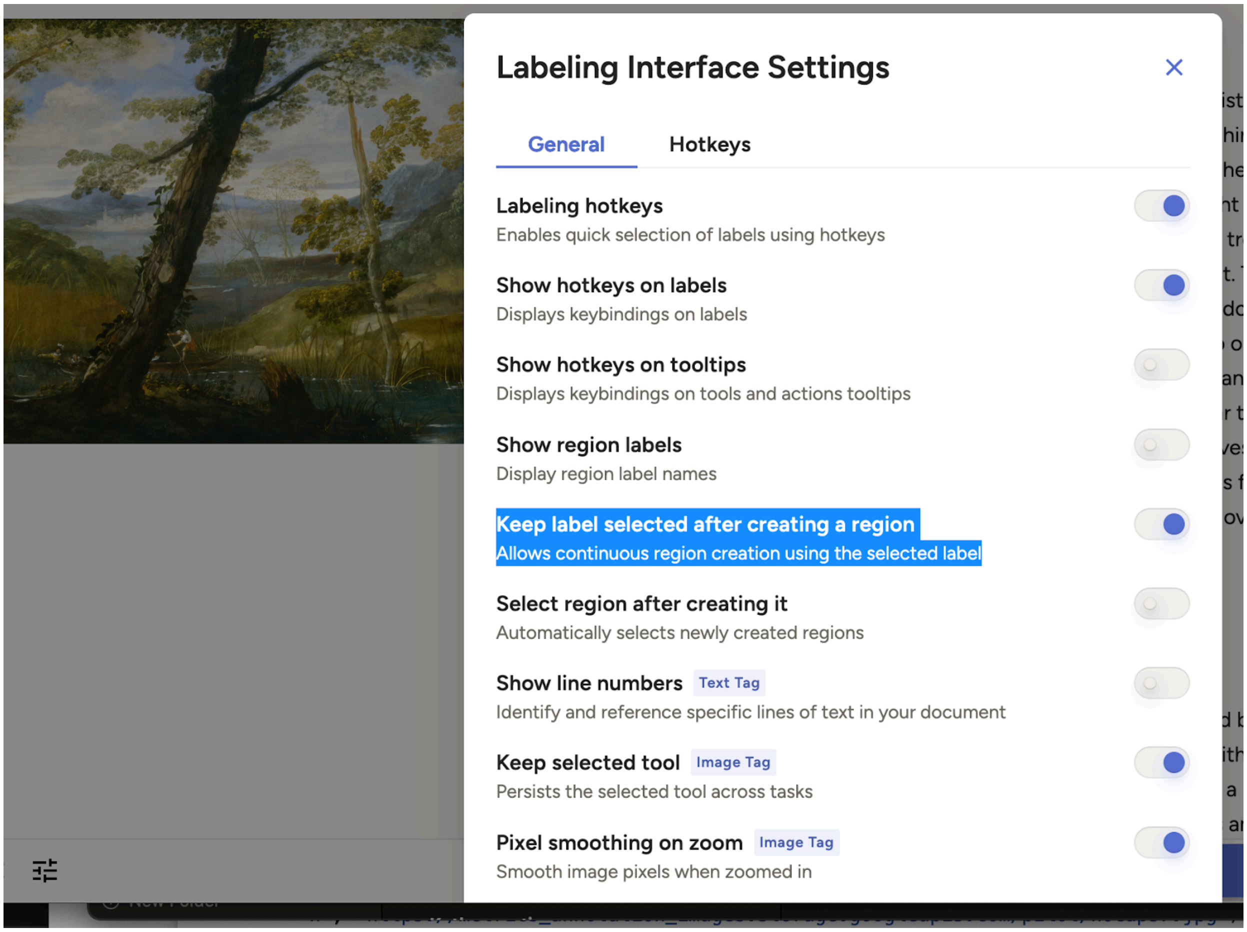
5) Finally, when judging missing details, please do try to mentally correct mistakes you've identified first. For example, if a generation specifies "three women" but the correct description says "four women", please only mark "three" as a mistake and leave "four" unmarked if possible. Again, this is not straightforward to do in all cases so don't stress too much if it's non obvious how to accommodate. (The goal here is to avoid double penalizing generations -- when it gets something wrong, that's an issue with its precision

in description but when it misses a detail, that's an issue with its coverage in description -- fixing each involves different techniques so it's helpful to disambiguate the former from the latter).

Annotation Interface

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

Once you open your annotations, click the settings button in the bottom left and enable "Keep label selected after creating a region". This will make the annotation process simpler.



Labeling Interface Settings

Labeling hotkeys ●
Enables quick selection of labels using hotkeys

Show hotkeys on labels ●
Displays keybindings on labels

Show hotkeys on tooltips ○
Displays keybindings on tools and actions tooltips

Show region labels ○
Display region label names

Keep label selected after creating a region ●
Allows continuous region creation using the selected label

Select region after creating it ○
Automatically selects newly created regions

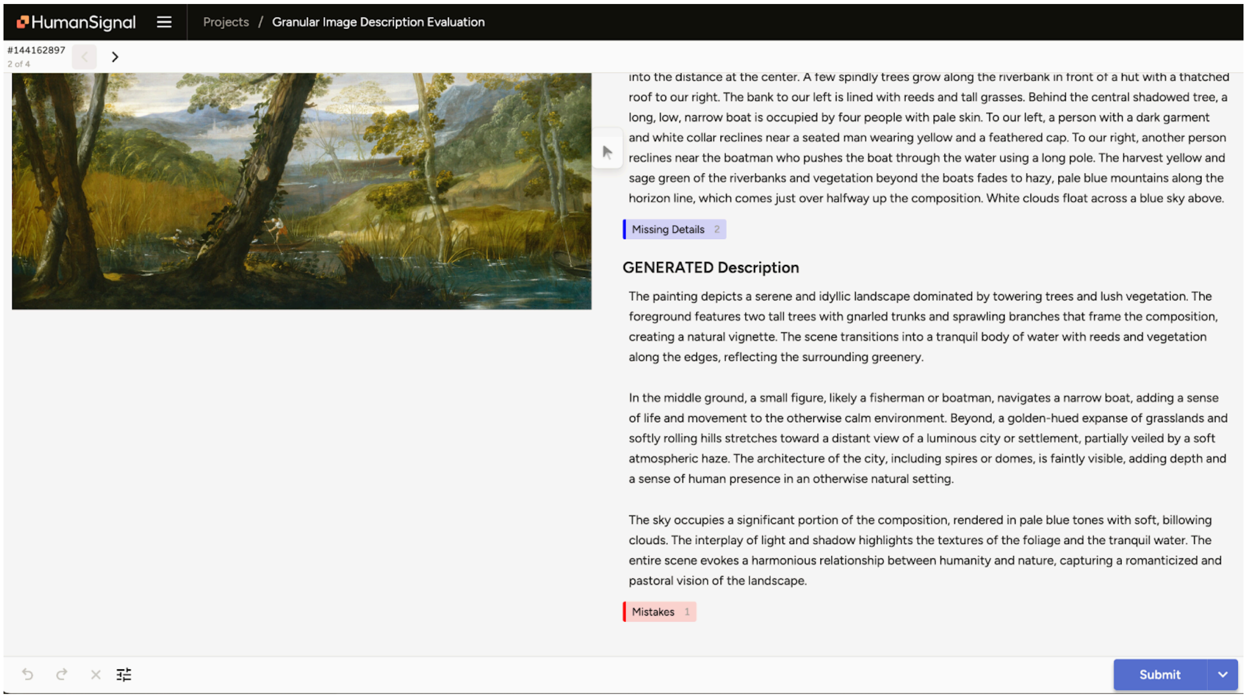
Show line numbers Text Tag ○
Identify and reference specific lines of text in your document

Keep selected tool Image Tag ●
Persists the selected tool across tasks

Pixel smoothing on zoom Image Tag ●
Smooth image pixels when zoomed in

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 **GENERATED** description. Press "1" and begin by drag-selecting minimal spans of text that identify *mistakes* in the **GENERATED** description. Text you select should appear highlighted in red. When complete, press "2" and drag-select minimal spans of text that identify *missing details* in the **CORRECT** description (after mentally correcting all *mistakes*). Text you select should appear highlighted in blue. Once satisfied with your annotation, hit SUBMIT.

If you would like to delete a selection, you can simply click on the highlighted region and hit "DELETE".



into the distance at the center. A few spindly trees grow along the riverbank in front of a hut with a thatched roof to our right. The bank to our left is lined with reeds and tall grasses. Behind the central shadowed tree, a long, low, narrow boat is occupied by four people with pale skin. To our left, a person with a dark garment and white collar reclines near a seated man wearing yellow and a feathered cap. To our right, another person reclines near the boatman who pushes the boat through the water using a long pole. The harvest yellow and sage green of the riverbanks and vegetation beyond the boats fades to hazy, pale blue mountains along the horizon line, which comes just over halfway up the composition. White clouds float across a blue sky above.

Missing Details 2

GENERATED Description

The painting depicts a serene and idyllic landscape dominated by towering trees and lush vegetation. The foreground features two tall trees with gnarled trunks and sprawling branches that frame the composition, creating a natural vignette. The scene transitions into a tranquil body of water with reeds and vegetation along the edges, reflecting the surrounding greenery.

In the middle ground, a small figure, likely a fisherman or boatman, navigates a narrow boat, adding a sense of life and movement to the otherwise calm environment. Beyond, a golden-hued expanse of grasslands and softly rolling hills stretches toward a distant view of a luminous city or settlement, partially veiled by a soft atmospheric haze. The architecture of the city, including spires or domes, is faintly visible, adding depth and a sense of human presence in an otherwise natural setting.

The sky occupies a significant portion of the composition, rendered in pale blue tones with soft, billowing clouds. The interplay of light and shadow highlights the textures of the foliage and the tranquil water. The entire scene evokes a harmonious relationship between humanity and nature, capturing a romanticized and pastoral vision of the landscape.

Mistakes 1

We have assigned you a number of tasks that should be completeable in the hour(s) you've been allocated. You will see each image and its **CORRECT** description paired with **4** different **GENERATED** descriptions. We expect each task to take no longer than 10 minutes (in our internal pilot, each task took 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!