GroundingSuite: Measuring Complex Multi-Granular Pixel Grounding

Supplementary Material

A. Dataset Details

Fig. 6 shows the word cloud visualization of our benchmark's textual descriptions, highlighting the linguistic diversity and domain coverage of GSEval.

Fig. 10 illustrates additional samples from our GSEval. The six images on the left represent the "stuff" class category, while the six images on the right demonstrate part-level segmentation examples.

Fig. 11 further showcases the diversity of our dataset, with the left panel displaying multi-object instances and the right panel presenting single-object examples.

In addition, we analyze mask size distribution (Fig. 7) and position distribution (Fig. 8) towards understanding mask size and position biases, which demonstrate GSTrain (Sampled from SA-1B) and GSEval (Sampled from unlabeled COCO) have significantly different distributions. We add a detailed breakdown of performance for different object classes. Partial statistics are presented in Fig. 9.

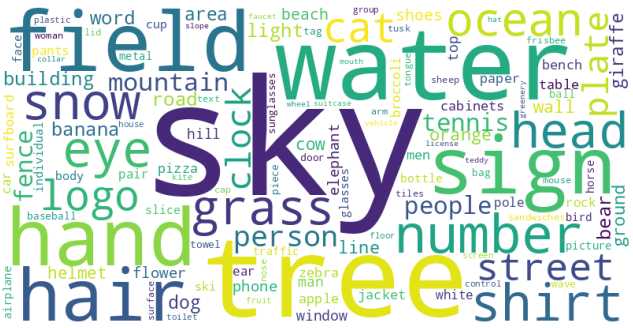

Figure 6. The word cloud of GSEval.

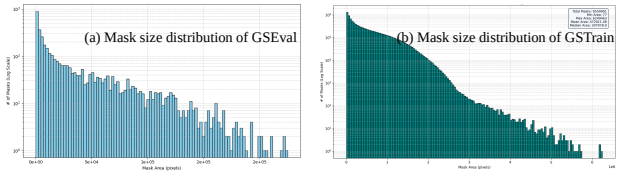

Figure 7. Distribution of mask sizes of GroundingSuite (zoom in for more detailed information).

B. Analysis of long expressions and filter types

B.1. Analysis of long expressions

Our quantitative results (Tab. 7) demonstrate that training the EVF-SAM model with our GSTrain-10M dataset yields a significant 1.1 cIoU point improvement on RefCOCOg for expressions of 16 words or more. We attribute this to the textual diversity and complexity of GSTrain-10M. More samples are visualized in revision.

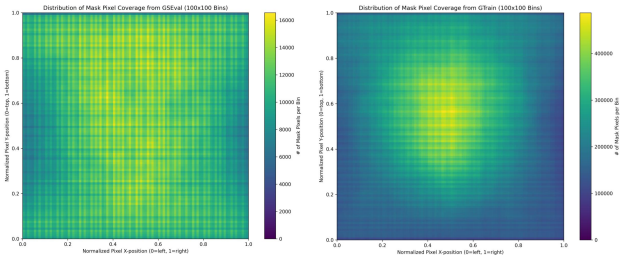

Figure 8. Distribution of mask positions of GroundingSuite (zoom in for more detailed information).

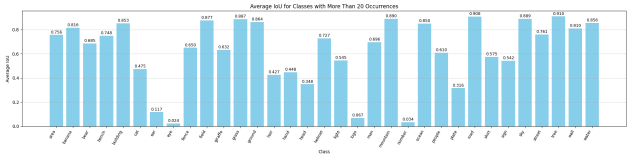

Figure 9. Partial statistics of fine-grained class IoU (zoom in for more detailed information).

Expression Length	≤10	11 - 15	≥16
w/o GSTrain-10M	79.2	77.4	81.7
w/ GSTrain-10M	79.7 +0.5	78.0 +0.6	82.8 +1.1

Table 7. Performance comparison on RefCOCOg across different referring expression lengths.

Filter Type	RefCOCO	gRefCOCO	RefCOCOm
No Filter	36.9	26.2	25.9
CLIP-based Filter	60.7	28.0	31.5
IoU-based Filter (*Ours.*)	63.5 +2.8	28.9 +0.9	32.7 +1.2

Table 8. Ablation study on the effectiveness of different filter types.

B.2. Analysis of different filter types

We compared different filtering methods: no filter (baseline for unfiltered data), a CLIP-based filter (GranD), and our proposed IoU-based filter. The results in Tab. 8 indicate that filtering is essential for generating high-quality data. Significantly, our IoU-based filter demonstrates superior performance over the CLIP-based approach, achieving a more substantial reduction in labeling noise and enhancing data

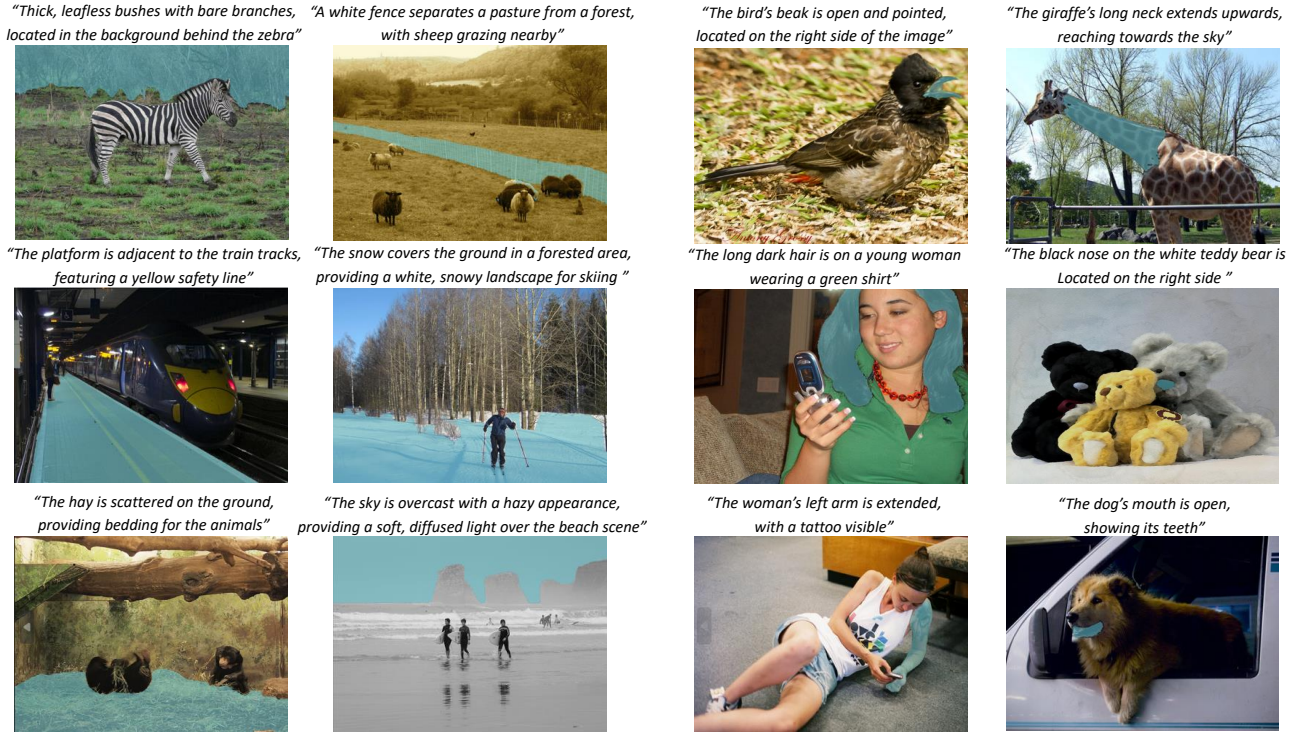

Figure 10. **More selected samples from our GSEval.** Stuff class and part level

precision.

C. Prompt

C.1. Global caption generation

We utilize the InternVL2.5-78B model to generate comprehensive global captions for the images. The specifically designed prompt template employed for this purpose is presented in Figure 12.

C.2. Grounding text generation

We employ specialized prompt templates with InternVL2.5 to generate unambiguous references that emphasize spatial relationships and distinctive visual features. The detailed prompt template is illustrated in Figure 12.

C.3. Noise filtering

During the noise filtering stage, we employ a two-step approach: first prompting the Vision-Language Model (VLM) to assess referring expression accuracy, then using specialized prompts to classify the referring expressions by category. Both prompt templates are illustrated in Figure 13.

C.4. Prompt for different grounding models

For different grounding models, we apply customized prompt templates to generate bounding box coordinates:

- **Gemini-1.5-Pro:**
  Return a bounding box for [***Referring***] in this image in [xmin, ymin, xmax, ymax] format.
- **GPT-4o and Claude-3.7-sonnet:**
  In this image, please locate the object described as: '[***Referring***]'. Provide the bounding box coordinates in the format [x_min, y_min, x_max, y_max]. You can use either absolute pixel coordinates or normalized coordinates (0-1 range).
- **Doubao-1.5-vision-pro:**
  Please provide the bounding box coordinate of the region this sentence describes: [***Referring***]
- **InternVL2.5:**
  Please provide the bounding box coordinate of the region this sentence describes: "<ref>[***Referring***]</ref>"
- **Qwen2.5-VL:**
  Please provide the bounding box coordinate of the region this sentence describes: <|object_ref_start|>[***Referring***]<|object_ref_end|>
- **Deepseek-VL-2:**
  <image><|ref|>[***Referring***]<|/ref|>.

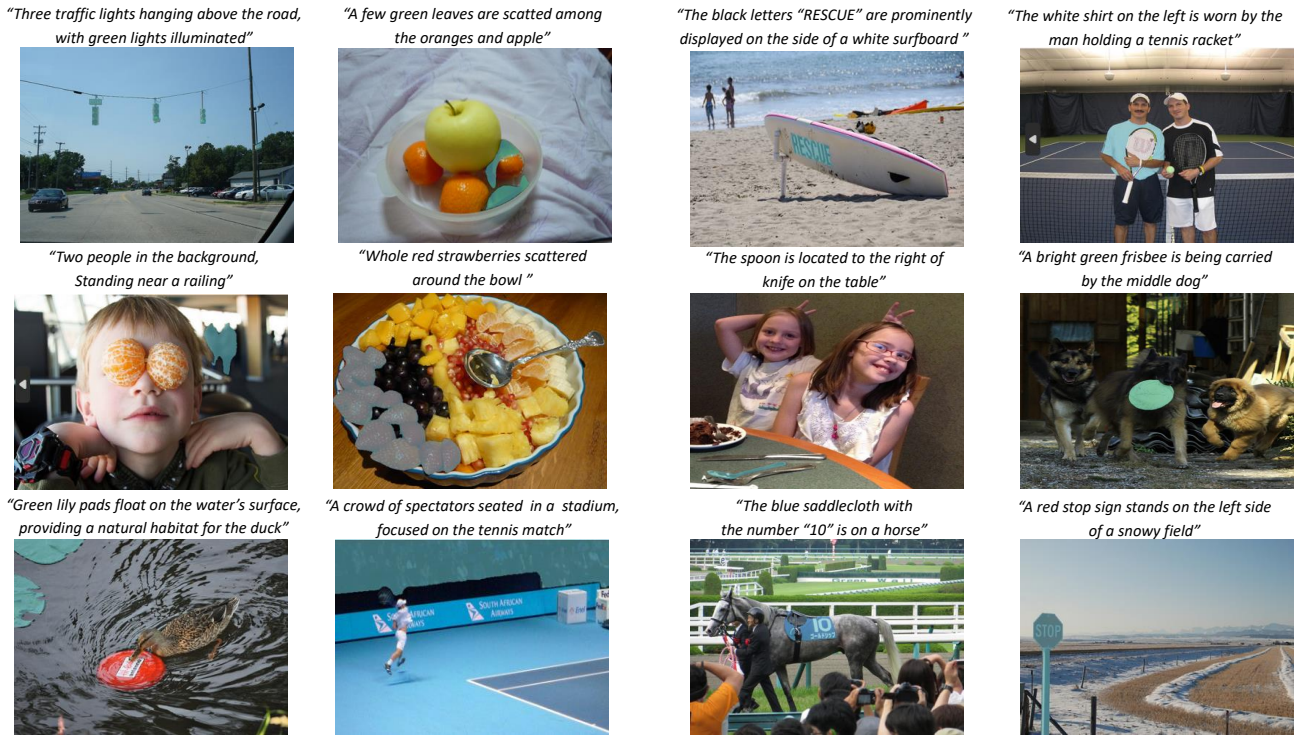

Figure 11. **More selected samples from our GSEval.** Multi object and single object

Prompt for Global Caption

Generate an accurate, single-paragraph description based on the given image. Do not use multiple paragraphs or line breaks. Avoid generating speculative content. Ensure that the description is based on clearly visible information in the image and avoid any over-speculation.

Prompt for Grounding Text Generation

Please give me a short description of [*Category Name*] [*x1,y1,x2,y2*]
Notice the following:
1: Ensure that this description distinguishes the [*Category Name*] from other [*Category Name*] by adding a relative position or unique features
2: If the image contains multiple images, specify the location of the image where the object is located.
3: If there are multiple objects in the region, describe them all in one sentence.
4: If the class is not correct, describe it in your own words.
5: Do not mention the coordinates. Using a short phrase.

Figure 12. Prompt for global caption and grounding text generation

Prompt for VLM Judger

Please review if the red box mask correctly annotates [*Referring*].
The accuracy standards are:
1.The object in the red box is consistent with the text meaning [*Referring*]
2.No object is missed or over-annotated; if other similar objects exist in the image, the annotation is inaccurate
3.No repeated or redundant red object boxes; if redundant red boxes exist in the image, the annotation is inaccurate.

Prompt for Prompt for VLM Classifier

Please classify according to the following categories.
Your final output should be only one number from 1-4, with no detailed analysis:
1. Stuff class description (materials, textures, backgrounds, natural elements like sky, water, grass, etc.)
2. Part level description (human or animal facial features, body parts, etc.)
3. Multi object description (mask contains multiple objects)
4. Single object description (one distinct item, person, animal, vehicle, or other countable entity)

Figure 13. Prompt for noise filtering