Revisiting Referring Expression Comprehension Evaluation in the Era of Large Multimodal Models

Supplementary Material

6. Labeling Errors in Existing Benchmarks

In the REC task, a referring expression should uniquely describe an instance, which is represented by an accurate bounding box. We have identified and visualized three common types of labeling errors in the RefCOCO, RefCOCO+, and RefCOCOg benchmarks: 1) non-unique referring expressions (Fig. 7), which refer to multiple instances within the same image; 2) inaccurate bounding boxes (Fig. 8); and 3) misalignment between target instances and their referring expressions (Fig. 9), where the referring expressions are either ambiguous or do not refer to any instance in the image.

7. Examples of Our Ref-L4 Benchmark

In Fig. 10 and Fig. 11, we present various examples of our Ref-L4 benchmark, illustrating a broad spectrum of scene diversity, instance scales, image aspect ratios, and expression lengths.

8. Prompts

8.1. Prompt for Context-Independent Description Generation

Briefly describe the [*Category Name*] in one sentence. Begin your description with the object name, including adjectives if appropriate to describe its color or shape. Focus only on visible features and avoid mentioning blurriness.

Input image: [*Cropped Image*].

8.2. Prompt for Context-Aware Description Generation

You are a sophisticated referring expression generator. Your task is to generate a clear and specific description for the target instance highlighted by a red circle in the provided image, based on a given hint and the following criteria:

Criteria 1: The description should enable individuals to understand and accurately identify the specified region within the image.

Criteria 2: The description may should various attributes such as category, shape, size, color, visibility, exposure, texture, orientation, absolute position, relative position, facial features, clothing, accessories, gestures, context, semantic attributes, emotions, age, gender, posture, action, and especially interactions with other instances. The selection of features should be relevant to the particular region and the image context.

Criteria 3: The red circle is solely for highlighting the region of interest. Do not refer to it in your descriptions.

Criteria 4: Avoid using unnecessary words like "look for", "spot", "observe", "find", "notice", "identify", "outline", "target" and "question".

Criteria 5: Ensure that the subject of each sentence matches the subject given in the hints. Do not incorrectly use the subject as the object.

Criteria 6: Use the correct singular or plural form when referring to the target, which may be a single object, a pair of objects, or a group of objects.

Criteria 7: Integrate all relevant information from the hints, noting that some hints may be redundant or contain errors.

Input image: [*Raw Image*].

Hint: [*Context-Independent Description*].

8.3. Prompt for Rephrasing Referring Expressions

Rewrite the subsequent description while preserving the main information. Utilize varied expressions and reorganize the sentences if necessary. Begin each sentence with the same subject being referred to.

Description: [*The Referring Expression to be Rephrased*].

8.4. Prompt for GPT4-V Evaluation

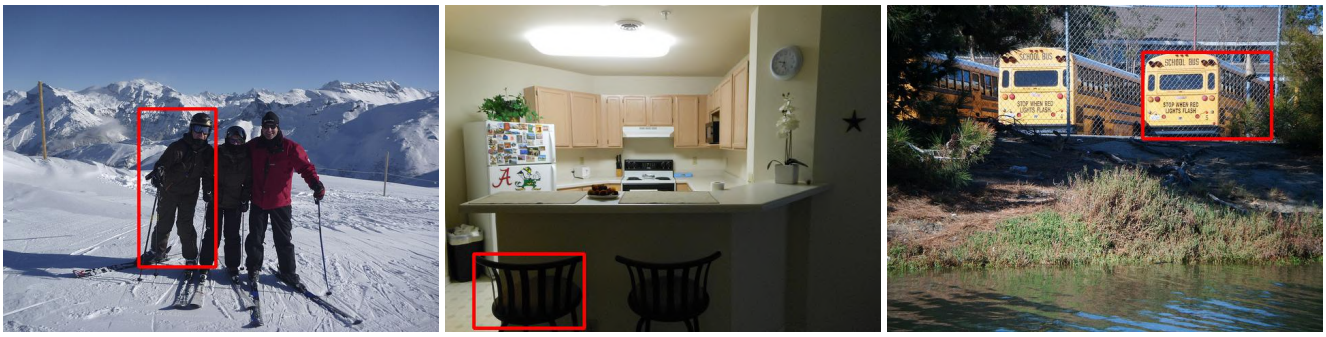
You are an expert in referring expression comprehension and localization. Your task is to locate the object in the image based on the provided expression. The coordinates range from the top left (0, 0) to the bottom right ([*Image Width*], [*Image Height*]). Please provide the bounding box in the format (x_0, y_0, x_1, y_1), where (x_0, y_0) represents the top-left corner and (x_1, y_1) represents the bottom-right corner.

Expression: [*The Referring Expression*].

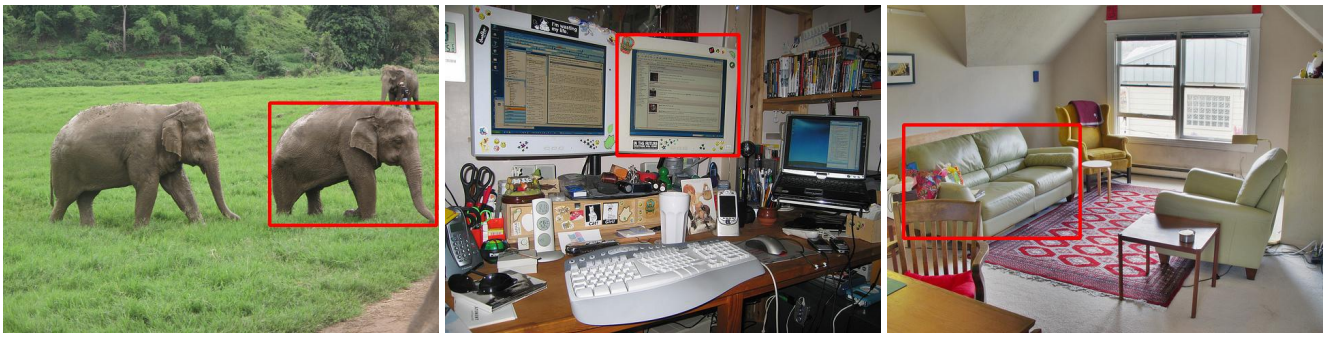
9. More Experiments

9.1. Category-Wise Performance.

Fig. 5 presents the per-category performance of the top four models. In Fig. 12 and Fig. 13, we show the performance for all 24 models on a per-category basis, with mAcc serving as the metric, along with the average performance for each model across all categories.



(a) man with glasses on (b) a black colur chair (c) bus in fence frame 1



(d) an elephant walking in the grass (e) a white computer screen (f) white couch

Figure 7. Visualization of labeling errors, where a referring expression refers to multiple instances within the same image. For each sub-figure, we display the original bounding box annotation with a red rectangle and include the corresponding referring expression in the caption.

9.2. Evaluation on Diverse Data Sources.

Fig. 6 illustrates the performance of six models across three subsets, namely "COCO", "O365-P1" and "O365-P2". In Fig. 14, the comprehensive results of 24 models across the same three subsets are displayed.

10. Limitations and Broad Impacts

Ref-L4 provides a more comprehensive and detailed evaluation of REC capabilities, helping to better understand and improve the performance of large multimodal models capable of handling the REC task. The public availability of Ref-L4 and its evaluation code encourages further research and collaboration, driving innovation and advancements in the field of REC and beyond. While Ref-L4 aims to cover a wide range of scenarios, it may still miss out on specific edge cases or unique contexts that could be encountered in real-world applications. The detailed and lengthy referring expressions might pose a challenge for current models, requiring significant advancements in natural language processing and comprehension capabilities.

11. Author Statement

The authors of the Ref-L4 benchmark accept full accountability for any rights violations, such as copyright infringement or other legal breaches. They emphasize that all data included in the Ref-L4 dataset adheres to the licensing agreements of the original source datasets. The Ref-L4 benchmark is made available under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license. Meticulous attention has been paid to ensure that the dataset upholds the highest legal and ethical standards. The authors are committed to addressing any issues arising from the use of this dataset and stand prepared to take necessary actions to resolve them.

12. Maintenance and Long Term Preservation

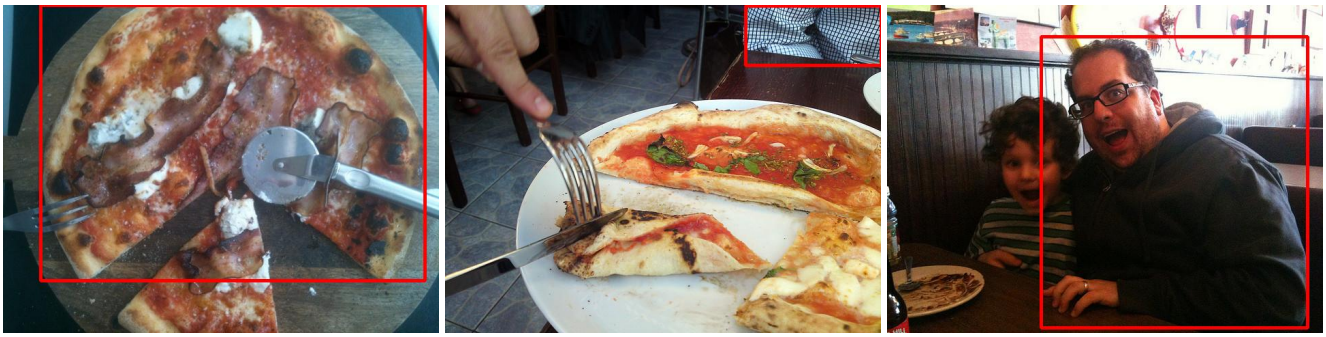
To ensure the benchmark remains relevant and useful for evaluating REC models, we will establish a protocol for regular updates. This includes the addition of new image sets and text annotations that reflect current trends and challenges in the field. A version control system will be implemented to track changes and updates to the benchmark. Each version will be documented with detailed notes on the modifications, including the addition of new data, changes to annotation guidelines, and improvements based on user feedback. We will utilize reliable cloud storage solutions with multiple redundancy mechanisms to safeguard against data loss.

(a) tail of elephant (b) red jacket (c) name on oven

(d) pumpkin (e) a knife cutting a cake (f) white hair

Figure 8. Visualization of labeling errors, where the bounding box annotations are inaccurate. For each sub-figure, we display the original bounding box annotation with a red rectangle and include the corresponding referring expression in the caption.

(a) left (b) next to him (c) yep

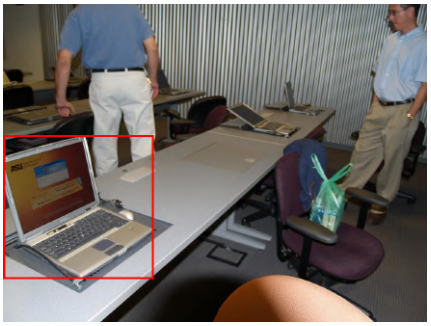


(d) not the slice (e) why is the game doing this checker shirt (f) last hotdog mess thing

Figure 9. Visualization of labeling errors, where the referring expressions are either ambiguous or do not refer to any instance in the image. For each sub-figure, we display the original bounding box annotation with a red rectangle and include the corresponding referring expression in the caption.

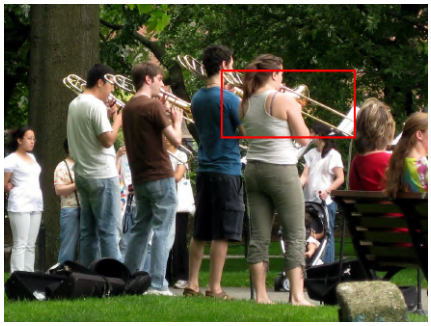
(a) The object with alternating black and white rings, imitating a target, is positioned on the left of a stack of books and adjacent to a basketball.

(b) A modest assortment of fragile flowers in a transparent container, positioned on the lengthy wooden cupboard below the attached-to-wall looking glass.



(c) A mobile computer featuring a grey keyboard, trackpad, and apparent side ports is positioned on the left side of the picture, with its display facing the observer.

(d) A decorative baseball with a unique red and gold color scheme, situated amongst various baseball memorabilia.



(e) The brass instrument with a long slide mechanism being played by the person at the rightmost of the group.

(f) Two clear plastic bags filled with swirled, cream-colored confections suspended beside various colorful toys on a mobile stall.

(g) The yellow electronic gadget situated on the music stand before the musician.

(h) The ridged potato chips situated between the two people enjoying their outdoor meal.

(i) The bright yellow fire extinguisher resting on the stone ledge near the concrete counter.

Figure 10. Examples from our Ref-L4 benchmark. For each sub-figure, we display the original bounding box annotation with a red rectangle and include the corresponding referring expression in the caption.

(a) The person wearing blue is a bystander located in the backdrop, beyond the row of jumping performers.

(b) This table, covered in black, features a spread of Korean delicacies, paired with informative materials about the Chuseok festival.

(c) The lighting fixture mounted on the ceiling is situated between the fan blades.

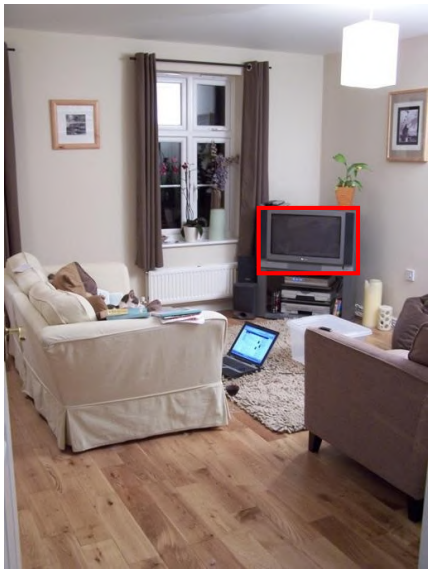
(d) The mop, featuring a uniquely twisted loop pattern in various hues, stands out against the neighboring vivid red and orange mops, each exhibiting a more even and consistent head appearance.

(e) A laundry appliance situated under a wall-mounted mirror.

(f) A stool with a red top is positioned to the right of the wooden coffee table, near the right bed.

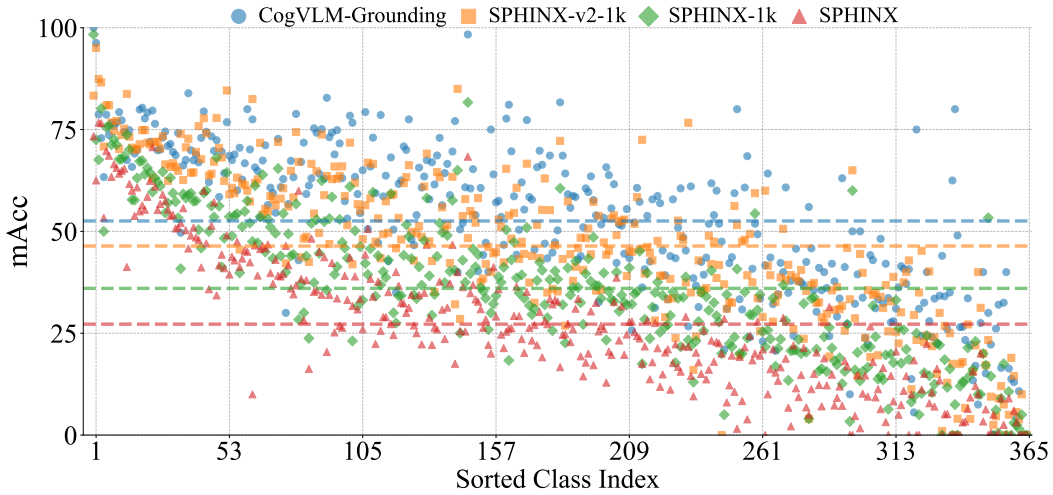
(g) A black dress shoe is visible on the left foot of the man with crossed legs, located at the bottom edge of the picture.

(h) Four spring rolls, located near the upper right corner.

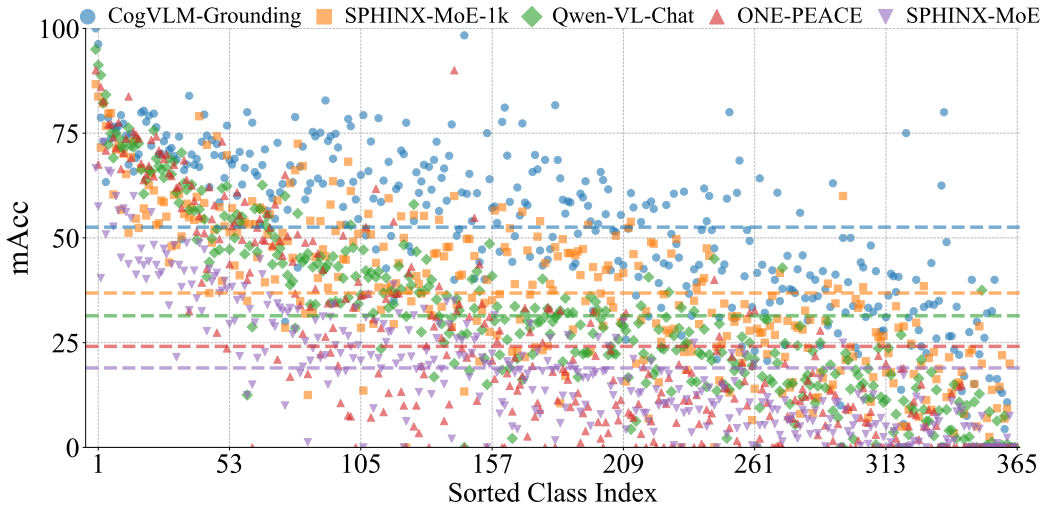


(i) The electronic device mounted above a DVD player and beneath a green plant on the stand in the living area.

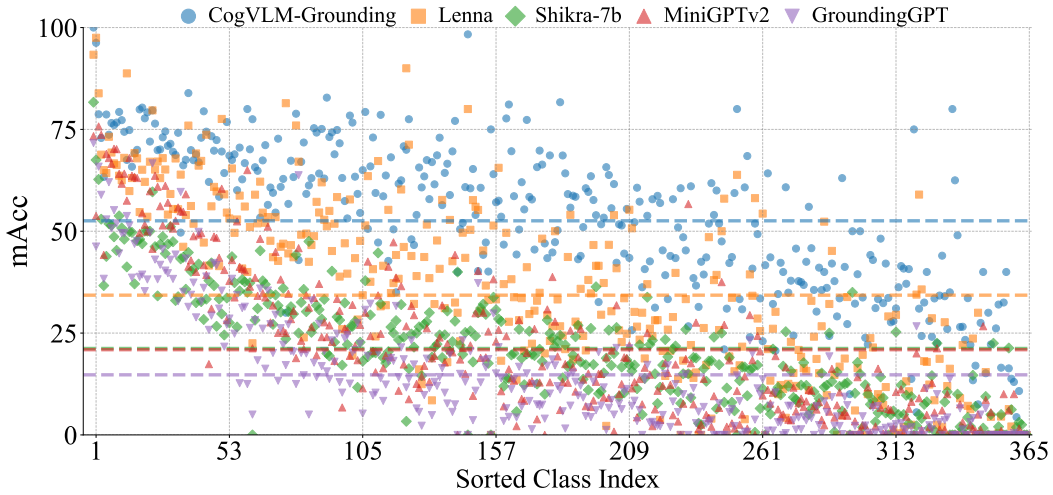
Figure 11. Examples from our Ref-L4 benchmark (continued in Fig. 10). For each sub-figure, we display the original bounding box annotation with a red rectangle and include the corresponding referring expression in the caption.



(a) The average performance across all categories (dot lines) for CogVLM-Grounding [63], SPHINX-v2-1k [29], SPHINX-1k [29], and SPHINX1 [29] are 52.56, 46.40, 36.01, and 26.95, respectively.

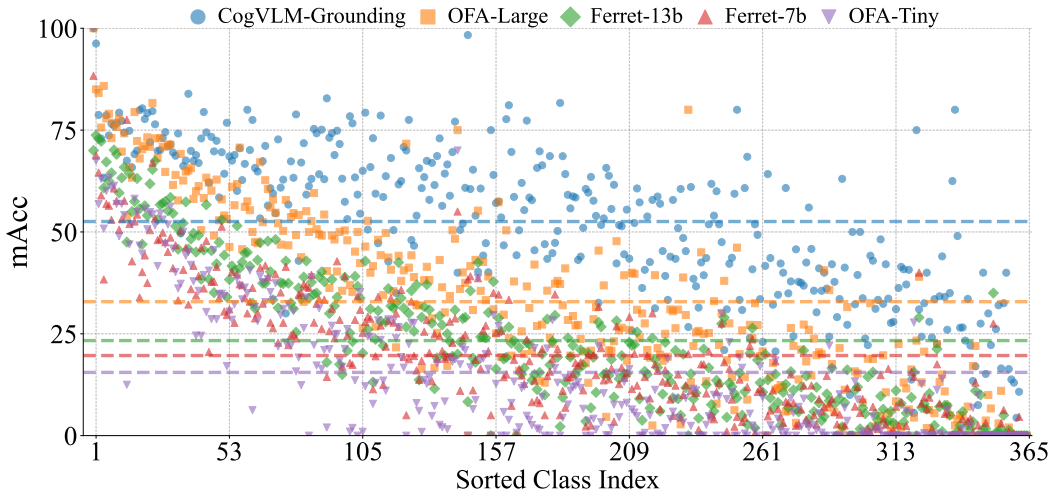


(b) The average performance across all categories (dot lines) for SPHINX-MoE-1k [14], Qwen-VL-Chat [1], ONE-PEACE [61], and SPHINX-MoE [14] are 36.84, 31.41, 24.11, and 18.77, respectively.

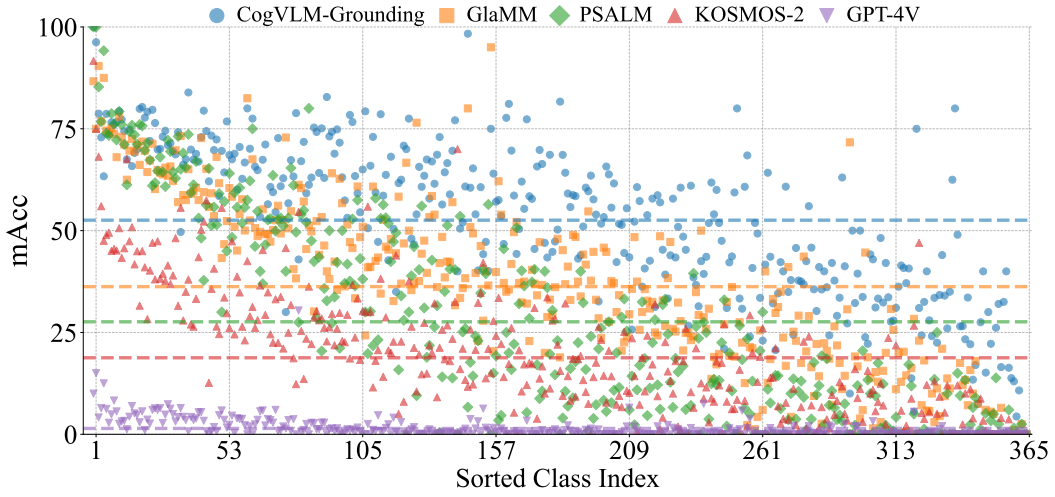


(c) The average performance across all categories (dot lines) for Lenna [67], Shikra-7b [6], MiniGPTv2 [5], and GroundingGPT [27] are 34.30, 21.22, 21.13, and 14.60, respectively.

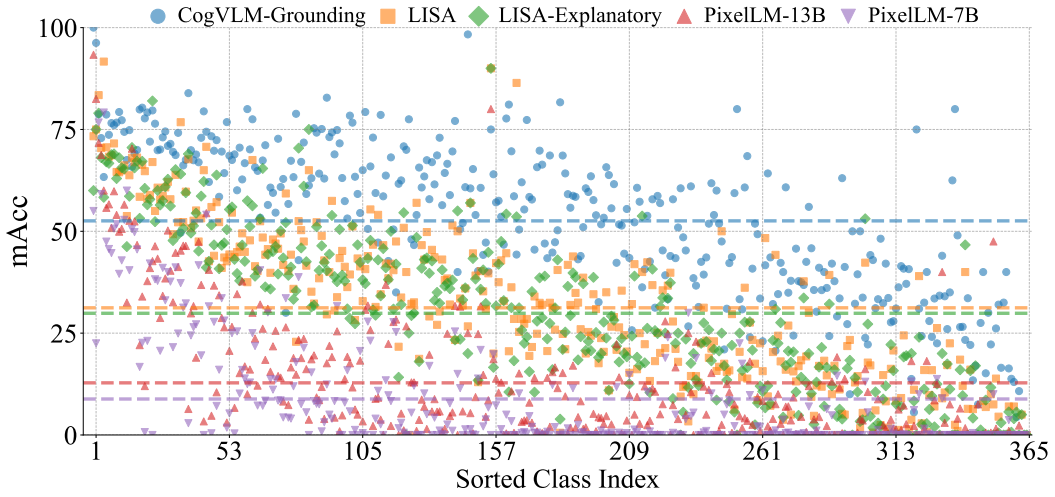
Figure 12. Category-wise performance of 24 models (part-1), sorted in the same order as in Fig. 5. We use CogVLM-Grounding as a reference for comparison in each sub-figure.



(a) The average performance across all categories (dot lines) for OFA-Large [60], Ferret-13b [72], Ferret-7b [72] and OFA-Tiny [60] are 32.88, 23.33, 20.27, and 15.37, respectively.



(b) The average performance across all categories (dot lines) for GlaMM [48], PSALM [81], KOSMOS-2 [41] and GPT-4V [38–40] are 36.25, 27.62, 19.37, and 1.42, respectively.



(c) The average performance across all categories (dot lines) for LISA [24], LISA-Explanatory [24], PixelLM-13B [51] and PixelLM-7B [51] are 31.22, 29.87, 13.19, and 8.74, respectively.

Figure 13. Category-wise performance of 24 models (part-2), sorted in the same order as in Fig. 5. We use CogVLM-Grounding as a reference for comparison in each sub-figure.

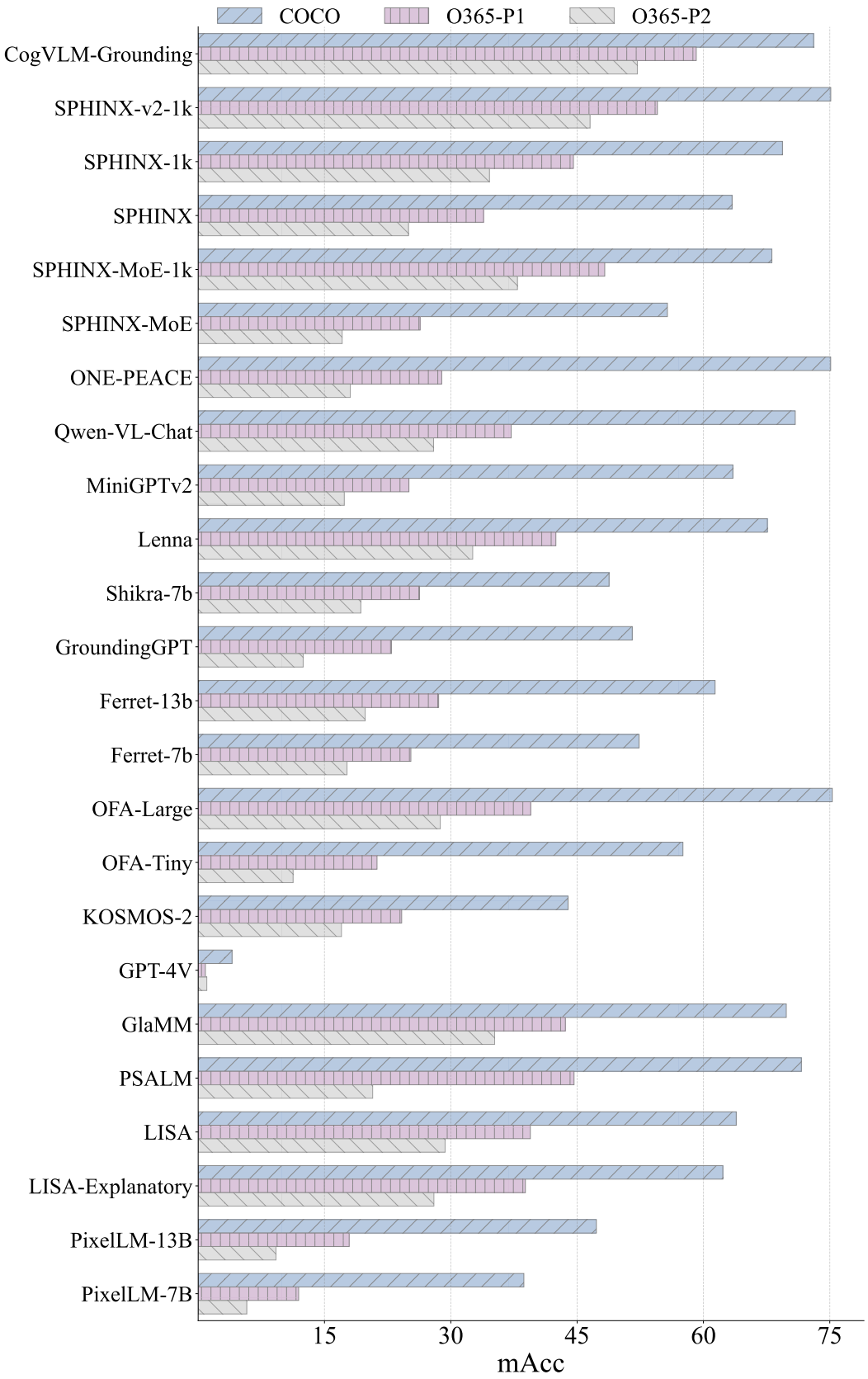


Figure 14. Evaluation of 24 models on various data sources, with mAcc acting as the metric.

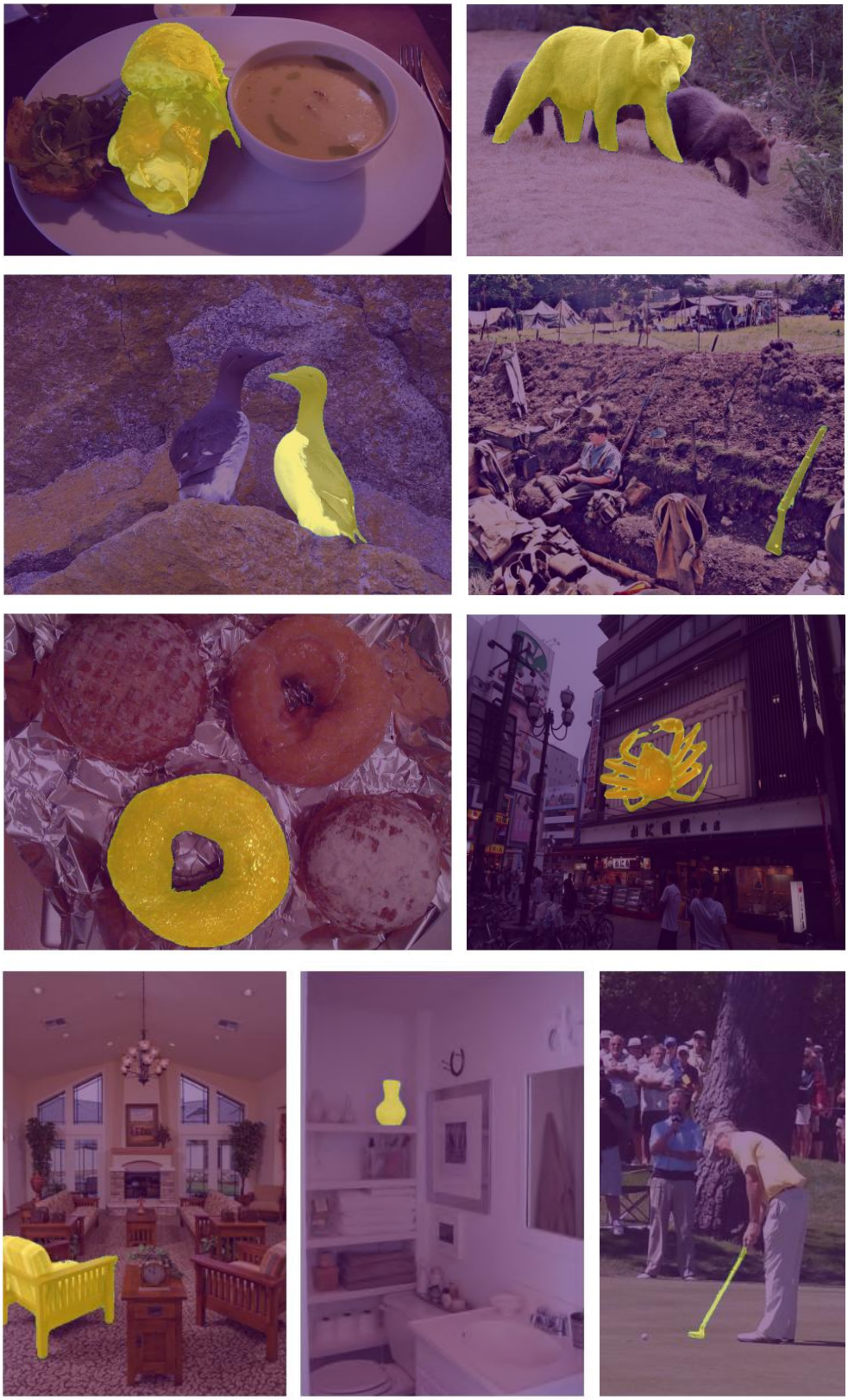


Figure 15. We provide visualizations of nine randomly selected segmentation annotations from various categories within our benchmark. The annotations are highlighted in yellow.