Spatial Preference Rewarding for MLLMs Spatial Understanding

1. Details of Preference Data Construction

Region Query Construction. Before constructing random region-level queries, we first preprocess the annotations in the Objects365 dataset. Specifically, we leverage GroundingDINO to re-annotate objects labeled as "crowd" in the dataset and removed small objects with an area of less than 300. After obtaining valid annotations, we filtered out 10,000 images containing more than eight objects to generate random region-level questions. To construct random regions, we iteratively include the closest objects located outside the current region. We ensure that the aspect ratio of the generated random regions (short side to long side) are greater than 1/3. This restriction is imposed considering that most MLLMs preprocess input images into square shapes, and regions with extreme aspect ratios are too challenging for MLLMs.

Grounded Region Description Generation. We use cropped image regions and object indices to prompt MLLMs to generate descriptions of the target regions. When cropping the regions, we slightly expand both sides of the region by 1.5 times before performing the crop. In constructing the object indices, we only consider objects with at least 50% of their area falling within the queried region. For region descriptions generated using the cropped images, we remap their generated grounding coordinates back to the original image coordinates before further processing and scoring.

Preference Data Scoring and Refinement. For the semantic score, we leverage CLIP-L-Patch14-336 to compute the similarity between the region description and image semantic. For the similarity with whole image, For similarity based on the full image, we preprocess the image into a square shape and ensure that the queried region is fully contained within the square. (CLIP performs center cropping, which might cut off the queried region.). For similarity based on the region image, we crop the queried region with a 1.5 times expansion, then pad it into a square before feeding it into the CLIP model. For the localization score, we first use the cropped region and GroundingDINO to detect objects mentioned in the description within the region. We then calculate the localization score. When refining the preferred description, we adjust the coordinates of objects mentioned within the described region and retain the coordinates of objects outside the region mentioned in the description, provided their IoU with the object annotations exceeds 0.5.

2. Training Hyperparameter

We LORA employ the AdamW optimizer with a batch size of 2 and a peak learning rate of 1e-5/1e-5/2e-7 for Ferret/CogVLM/LLaVA-Onevision, respectively. The rank of LORA is set as 128 and alpha as 256 for all MLLMs. During training, we freeze all vision-related parameters, such as the visual encoder, projector, and embedding layer, and only add and update the LORA parameters in the language model part.