A Task-Specified Instruction Templates

In complementary to discussion on training datasets in Sec. 3.3, we list a few instruction templates used to convert task-specified datasets to instruction following format in Tab. 1. Specifically, we convert the detection dataset COCO to multiple objects grounding data in a similar format as REC data.

Table 1: Instruction templates. We randomly select three templates from each task for illustration.

Task	Template
Image captioning	What is this photo about? Describe the following image. Analyze the image in a comprehensive and detailed manner.
Region captioning	What is <region>? Please briefly describe <region>. Give a concise description of <region>.
Referring expression comprehension	Locate {expression} in the image. Which region matches {expression} ? Identify the region that corresponds to {expression} .
Multiple objects grounding	Locate all {object class} in this image. Find out all instances of {object class} in the image. Detect and list each {object class} that appears in the picture.
Grounded image captioning	[grounding] Give me a short description of the image. [grounding] Succinctly summarize what you see in the image. [grounding] Please summarize the content of this image in brief.
Grounded chat	[grounding] {Free-form user instructions}.

B LVIS-Ground Benchmark

Current MLLMs typically do not support detecting multiple categories of objects at the same time. Therefore, to customize the LVIS [1] detection benchmark for MLLM evaluation, each time we only select one object class that is included in the image to ground. For instance, the grounding query can be formulated as "Locate all {object class name} in this image". However, this 'one-by-one' evaluation strategy unavoidably leads to low efficiency. To save time and maintain class balance, we randomly sample at most 5 images for each object category[1] from the LVIS validation set to construct LVIS-Ground.

There are often multiple ground-truth boxes for a query in LVIS-Ground. In such cases, traditional methods either adopt the ANY-Protocol or MERGED-BOXES-Protocol to evaluate performance [2]. To be specific, the ANY-Protocol considers recall to be 100% if the prediction matches any of the ground-truth boxes (*e.g.*, with IoU > 0.5), which fails to truly reflect the model's capability

[1] Some categories have fewer than 5 samples in the original LVIS validation set.

in finding out all object instances. On the other hand, the MERGED-BOXES-Protocol merges all ground-truth boxes into a smallest enclosing box as the ultimate ground-truth box. However, this protocol ignores the atomicity of individual boxes, and is not well-suited for instance-level prediction evaluation.

To better evaluate recall for multiple ground-truths, we propose a new protocol termed AS-MANY-Protocol. This protocol selects the top-k predicted boxes (where k is the number of ground-truth boxes) and measures recall over all ground-truth boxes. For example, if there are 3 out of 5 ground-truth boxes hit by the top-5 predicted boxes, the recall is 60%. Besides, we follow common practice in detection [3] to calculate average recall over 10 IoU thresholds (ranging from 0.5 to 0.95) as the primary metric on LVIS-Ground.

C More Implementation Details

Table 2 lists the detailed hyper-parameter configuration used for Groma training. It takes roughly 5/2.5/0.5 days to finish stage 1/2/3 training on 8 A100 GPUs. For some large-scale datasets, we merely sample a subset from them during training. The total number of training samples in one epoch is given in Tab. 2.

Table 2: Training details. RP, RE, and VLP stand for region proposer, region encoder, and vision-language projector (an MLP), respectively.

Configuration	Detection pretrain	Alignment pretrain	Instruction finetune
optimizer	AdamW	AdamW	AdamW
epochs	12	2	1
batch size	64	128	128
learning rate	2e-4	1e-4	1e-5
weight decay	1e-4	0	0
resolution	448p	448p	448p
training samples	5.7m	3.2m	857k
trainable param.	RP	RE, VLP	RE, VLP, LLM

References

1. Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5356–5364 (2019) 1
2. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1780–1790 (2021) 1
3. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 2