Supplementary Materials of Grounding Language Models for Visual Entity Recognition

Zilin Xiao[1], Ming Gong[2], Paola Cascante-Bonilla[1],
Xingyao Zhang[2], Jie Wu[2], and Vicente Ordonez[1]

[1] Department of Computer Science, Rice University, USA
{zilin, pc51, vicenteor}@rice.edu
[2] Microsoft STCA, China
{migon, xingyaozhang, jiewu1}@microsoft.com

A Experimental Details

Table 1: Hyperparameter settings for training of AutoVER.

Hyperparameter	Value
learning rate	2e-5
weight_decay	0
batch size per device	8
effective batch size	256
learning rate strategy	Cosine
optimizer	AdamW
gradient clipping	Disabled

We present the hyperparameter choice in Table 1. As the effectiveness of contrastive learning only reveals in a large batch size setting, we conduct all pieces of training in a batch size of 256 on 32 V100-SXM2-32GB GPUs. Utilizing DeepSpeed [2] for distributed training management, we train AutoVER-7B for 2 days and AutoVER-13B for 6.5 days with 2.5 million training samples for one epoch.

B Hard Negative Groups Construction Details

We first illustrate the process of constructing KB-HARD group construction. Specifically, we consider three types of relations in Wikidata for category hierarchy construction, "instance of (P31)", "parent taxon (P171)" and "subclass of (P279)". To obtain category hierarchy labels, we execute the SPARQL statement shown on the left of Figure 2 on Wikidata knowledge graph. To exclude generic categories, we limit the depth of traversal on the knowledge graph to 3, and manually filter generic categories[3]. On the right of Figure 2, we visualize the 3-hop

[3] ["Wikidata metaclass", "second-order class", "taxon", "classification scheme", "metaclass", "third-order class", "human", "direct anatomical metaclass", "organisms known

Fig. 1: Error analysis of two query image-question pairs. **Left:** two query image-question pairs. **Right:** some of retrieved candidates. While the correct entity answer is retrieved indicated in green font, the MLLM decides to produce the wrong entity identifier depicted in red font.

category hierarchy map for the entity ATR 42. Finally, we collect entities that share any parent nodes in the category hierarchy as knowledge-similar entities.

We use `ViTL/16-224px`[4] fine-tuned on ImageNet-1k as the visual classifier in VISION-HARD. While we collect the entities that share the same prediction label as vision-similar entities, we filter those entities with its image top-1 confidence score lower than 0.4, those entities without an infobox image, and those prediction labels with less than 10 entities. In Figure 3, we pick some entity images that share the image prediction "Alps" and "Goose".

C Error Analysis

We present two typical types of errors in Figure 1. Although the model is able to retrieve the correct entity, the MLLM predicts a too specific (but actually correct as it is indeed a great blue heron) entity in the upper sample. This is due to the misunderstood of intent by the MLLM, as the query does not ask for species but category. The lower sample has the MLLM make a wrong prediction on "Wetsuit" instead of "Diving suit".

by a particular common name", "Wikimedia disambiguation page", "variable-order class", "concept", "food", "artificial physical object", "artificial object", "physical object", "physical substance", "product", "object", "artificial physical structure", "architectural structure", "equipment", "tool", "physical location", "Wikimedia list article"]

[4] https://huggingface.co/google/vit-large-patch16-224

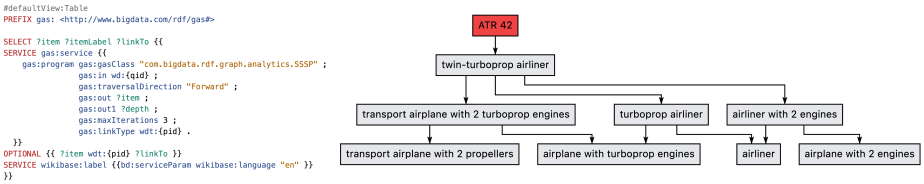

Fig. 2: KB-HARD group construction illustration. **Left:** The SPARQL query template that used for constructing the category hierarchy of entities. We only consider category labels within 3 hops to exclude general categories like "object", "food", etc. **Right:** Visualized 3-hop category hierarchy map for the entity ATR 42. Code highlight and category visualization are from Wikidata Graph Builder.

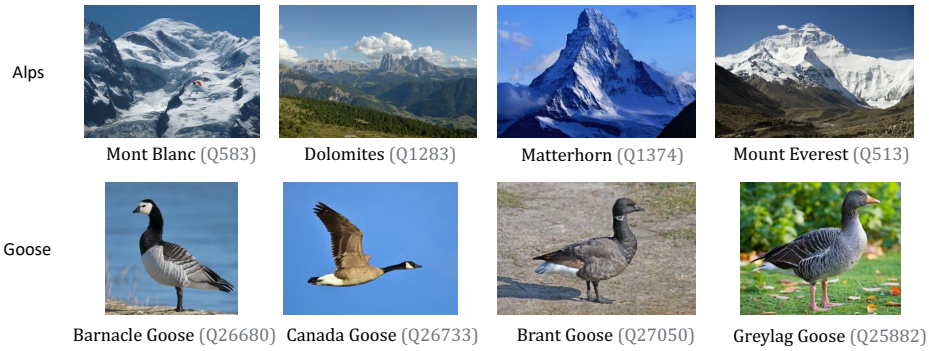

Fig. 3: Example VISION-HARD groups constructed with predictions from `ViTL/16-224px`.

D More Case Study

We present 7 cases for side-by-side comparison between GPT-4V and AUTOVER-7B in Figure 5. They include situations where AUTOVER-7B has an advantage (first 4 rows), GPT-4V outperforms (row 5), both make accurate predictions (row 6) and both commit mistakes (row 7). We also conclude the reasons behind GPT-4V's failures, such as its inability to ground accurately to entities, refusal to predict with insufficient clues, hallucinatory responses, struggles with disambiguating challenging entities and lack of specific domain knowledge.

E Qualitative Analysis on A-OKVQA-ENT Dataset

We present a qualitative analysis in Fig. 4 for zero-shot baselines reported in the main paper. For all baselines under **multi-choice** settings except AUTOVER, we append an explicit instruction[5] that is omitted in the figure. Still, LLaVA-v1-7B

[5] "Please strictly follow the options you are given. Only output the option number or the option itself."

and OpenFlamingo-9B fail to follow the instructions, leading to their extremely low evaluation results. Our AutoVER does not need any prompt engineering as it is guided by a prefix tree which never allows it to generate answers out of scope. We highlight the keywords where the model exhibits strong hallucinations in the analysis.

F Current Limitations

Although AutoVER demonstrates superior performance using fewer trainable parameters compared to PaLI-17B, the accuracy on the UNSEEN subset remains suboptimal, particularly when compared with the almost identical human performance on the SEEN and UNSEEN subsets. We expect future work that focuses on exploring visual entity recognition for emerging entities that never show in the training set, and even closing the gap between model and human performance. In addition, the autoregressive nature makes it challenging to predict through a single cheap vector retrieval like a bi-encoder. Instead, it requires multiple forward passes on a billion-size language model. We believe that efficient inference such as in [1] could be an interesting direction toward real-time autoregressive visual recognition systems.

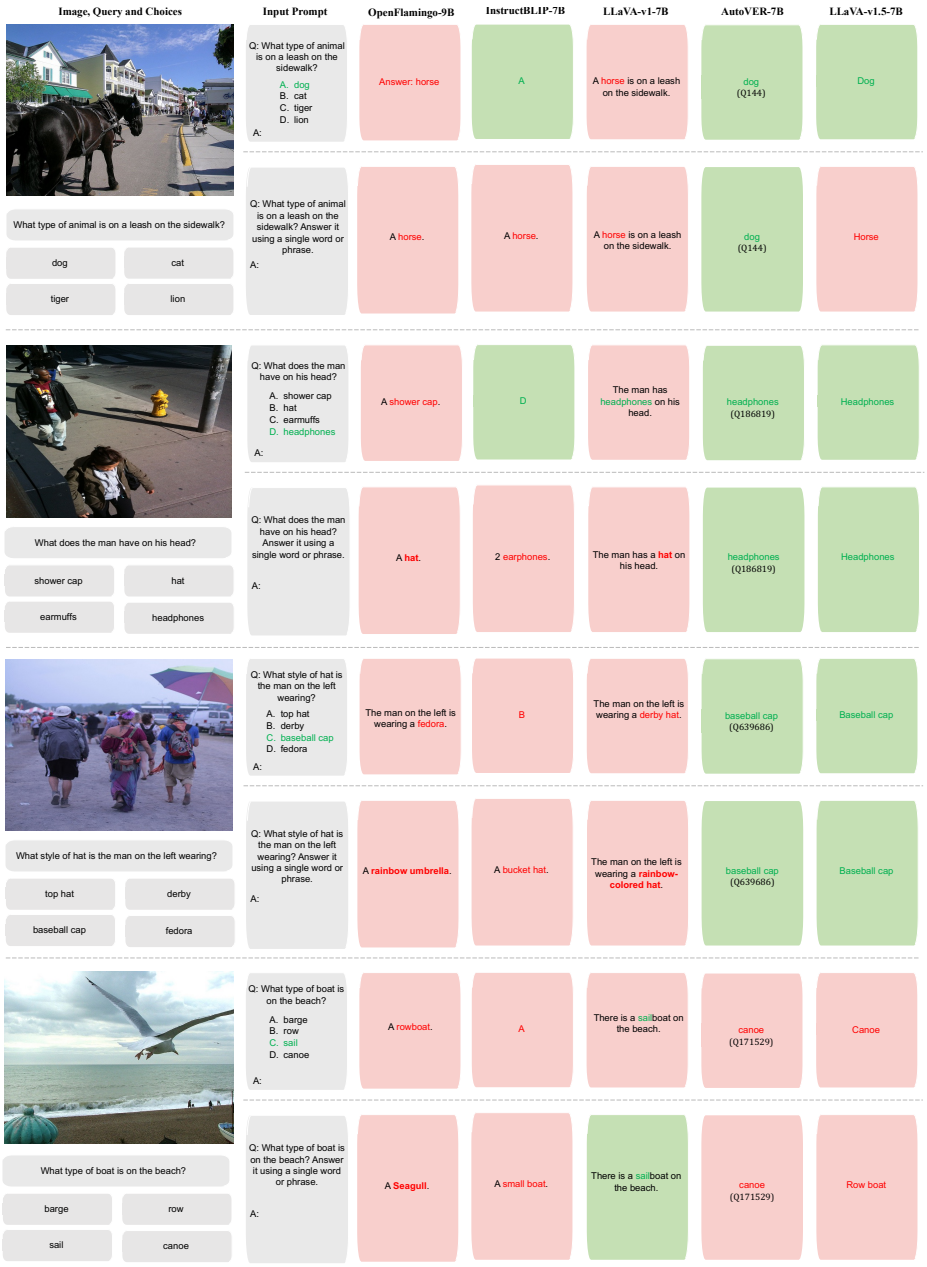

Fig. 4: Qualitative analysis on A-OKVQA-ENT dataset. For each triplet of image, query and choices, the upper row illustrates the input prompt and response from different baselines under **multi-choice** evaluation strategy, while the lower row is under **entity match**. Background color in red indicates an incorrect answer under the specific evaluation setting and that in green denotes a correct response. **Bold** highlights the model is heavily hallucinating.

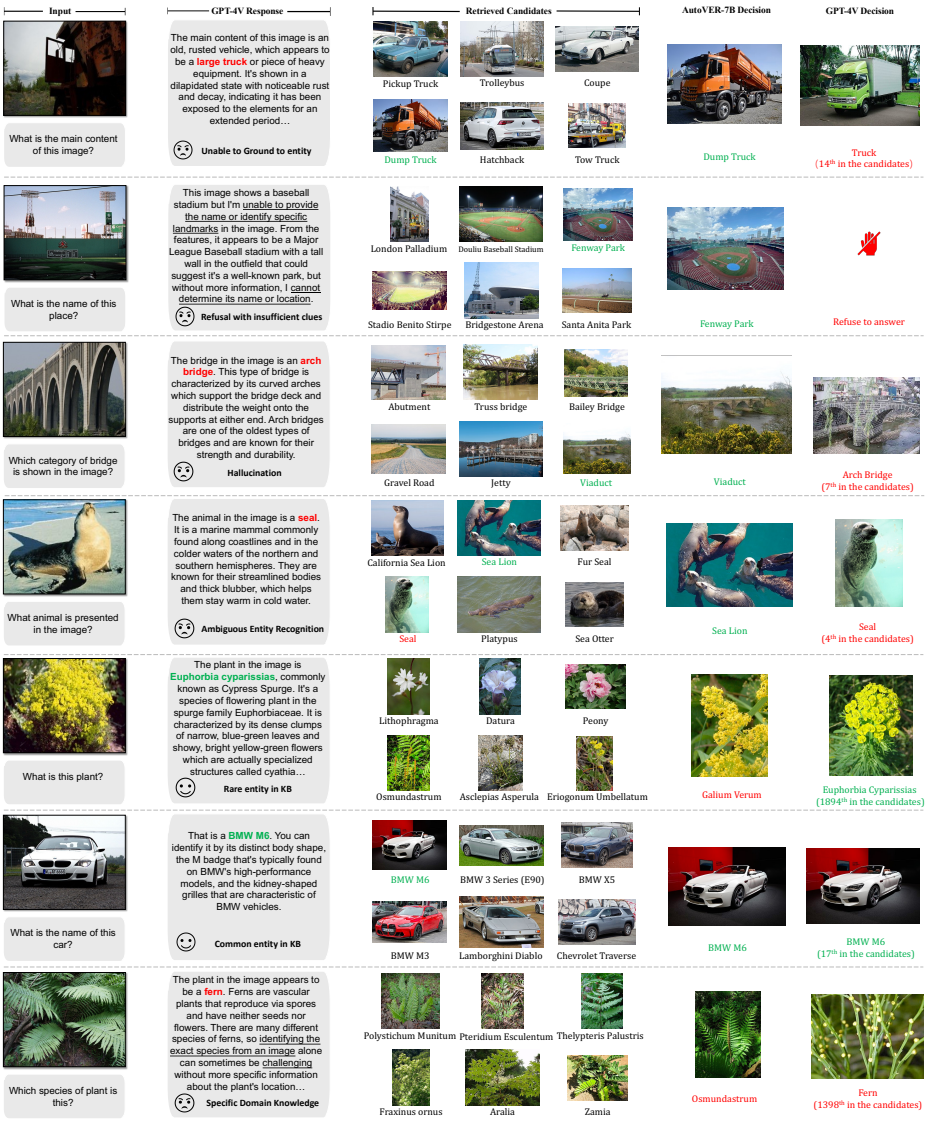

Fig. 5: Case study on comparison between the GPT-4V response and AutoVER-7B decision. **Bold** indicates the entity keyword. In the GPT-4V response, **red** indicates a false entity keyword and **green** indicates a correct one. In retrieved candidates and model decisions, red indicates a false entity prediction and green denotes a correct prediction. We also include the top-k position of the GPT-4V predicted entity for reference. The answer for the last query is *Matteuccia*.

References

1. He, Z., Zhong, Z., Cai, T., Lee, J.D., He, D.: Rest: Retrieval-based speculative decoding. arXiv preprint arXiv: 2311.08252 (2023)
2. Rasley, J., Rajbhandari, S., Ruwase, O., He, Y.: Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 3505–3506. KDD '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3394486.3406703, https://doi.org/10.1145/3394486.3406703