ViGoR: Improving Visual Grounding of Large Vision Language Models with Fine-Grained Reward Modeling – Supplementary Material

Siming Yan[⋆†1] Min Bai[⋆2] Weifeng Chen[2] Xiong Zhou[2]
Qixing Huang[1] Li Erran Li[2]

[1]The University of Texas at Austin [2]AWS AI

In this document, we provide further details about our work, focusing on the data gathering and usage of our reward model, performance evaluation setup, and metrics. Finally, we present additional qualitative results to provide more anecdotal insight into the effectiveness of our model.

1 Visual Grounding Improvement Framework Details

1.1 Details of Reward Modeling via Fine-Grained Human Feedback

Human Annotation In this section, we give some clarifying details about our human annotation framework for collecting preference feedback.

Data source: To collect human annotation, we first select a group of images and sample a set of responses by the LVLM. In our case, we sample 7,200 images from the *train2017* split from the MS COCO dataset. Because of the expense of collecting human annotations, it is desirable for the reward model to learn more fine-grained and holistic signals that will allow it to further improve a relatively strong initial starting point of the LVLM. As shown in our main article, our automated signal generation approach can already significantly improve the visual grounding of the LLaVA model. Therefore, we generate responses using the checkpoint fine-tuned using only the automatic supervision approach. In particular, we generate 15,440 responses using the following prompts.

- Describe the given image in detail.
- Write a detailed description of the given image.
- Give a detailed description of the given image.
- Explain the visual content of the image in great detail.
- Analyze the image in a comprehensive and detailed manner.

⋆ Equal contribution.
† Work done when interning at AWS AI. Contact email: siming@cs.utexas.edu.

Annotation Instructions: For each image-description pair, we ask annotators to carefully consider each sentence of the description in two aspects: *accuracy* and *creativity*. *Accuracy* refers to whether the sentence is factually correct. The annotator is asked to select one of the following categories for each sentence.

- Accurate - all details mentioned in this sentence are true with respect to the input image.
- Hallucinated object - the sentence mentions an object that does not exist in the image.
- Incorrect object color - the sentence describes an object that exists in the image with the wrong color.
- Incorrect object quantity - the sentence describes an object that exists in the image with the wrong count.
- Incorrect object material - the sentence describes an object that exists in the image to be made of the wrong material.
- Incorrect object shape - the sentence describes an object that exists in the image to be the wrong shape.
- Incorrect object relationship - the sentence describes the relationship between two or more objects incorrectly. For example, the sentence mentions that *a person is riding a bicycle*, while the person is actually pushing the bicycle in the image.
- Incorrect object location - the sentence mentions objects with the wrong localization in the image.
- Incorrect reasoning: the sentence describes an illogical interpretation of the image. For example, the image shows an empty and dilapidated street, but the text describes it to be *lively and cheerful*.
- Other - the sentence assigns wrong descriptions in the image in a different way than the above.

Furthermore, the *creativity* score is a binary assessment of whether the sentence attempts to provide a reasonable interpretation of the extrapolation of the image. For example, a sentence such as *"The potted plants contribute to a pleasant and calming atmosphere in the room.* is considered creative, while *"There is a potted plant on the table."* is not.

Lastly, we ask annotators to perform a holistic assessment of the overall description's *level of detail* (as a score between 1 and 7) with the following rubric.

Ignore the accuracy, how comprehensively does the description capture the image's key elements?

- *1 = Extremely Lacking: The description omits all vital information.*
- *2 = Lacking: The description only has a summary without any description of the objects.*
- *3 = Somewhat Lacking: The description only has a summary but misses most detailed description of the objects in the image.*
- *4 = Neutral: The description only contains the details of few objects in the image.*
- *5 = Somewhat detailed: The description contains the details of most of the objects but misses two or three key elements in the image.*

- *6 = Detailed: The description covers nearly all the objects but misses one key element in the image.*
- *7 = Extremely Detailed: The description covers all essential details present in the image.*

Algorithm 1 Human Feedback Reward Model Score

Input: $Image$, $\{Descriptions\}$, Trained Reward Model R_H
for each S_i in $\{Descriptions\}$ **do**
  Divide S_i into a set of sentences $\{s_{i,1}, s_{i,2}, \ldots, s_{i,m}\}$
  $P_i^h \leftarrow 0$
  $N_i^h \leftarrow 0$
  for each sentence $s_{i,j}$ in $\{s_{i,1}, s_{i,2}, \ldots, s_{i,m}\}$ **do**
    if $R_H(s_{i,j}, \{s_{i,1}, s_{i,2}, \ldots, s_{i,j-1}\}, Image) = 0$ **then**
      $P_i^h \leftarrow P_i^h + 1$
    else
      $N_i^h \leftarrow N_i^h + 1$
    end if
  end for
  The final scores for S_i are (N_i^h, P_i^h)
end for

Reward Model Training After collecting human annotation data, we proceed to train a reward model. In our implementation, we adopt the same architecture as the LLaVA model for the reward model. We initialize the weights of the reward model using the pre-trained LLaVA model and then fine-tune it using the human annotation data. As mentioned in Section 1.1, our reward model predicts the *accuracy* score, which is a discrete integer ranging from 0 to 9, with the following categories: 0: Accurate, 1: Hallucinated object, 2: Incorrect object color, 3: Incorrect object quantity, 4: Incorrect object material, 5: Incorrect object shape, 6: Incorrect object relationship, 7: Incorrect object location, 8: Incorrect reasoning, 9: Others.

Inspired by [1], unlike the traditional approach of adding a linear layer for direct score regression, we treat the task as a language modeling problem, leveraging the inherent capabilities of the reward model. We observe that it has shown enhanced stability compared to direct score regression. To effectively gauge the accuracy, we have designed a specific prompt. It guides the reward model to evaluate the last sentence in a given image description, assigning a numerical value based on various criteria.

Assess the accuracy of the last sentence in the following description of this image and return a single number: 0 for accurate, 1 for hallucination of objects, 2 for incorrect object color, 3 for incorrect number of objects, 4 for incorrect object material, 5 for incorrect object shape, 6 for incorrect object relationship, 7 for incorrect object location, 8 for flawed reasoning and 9 for other types of inaccuracies.

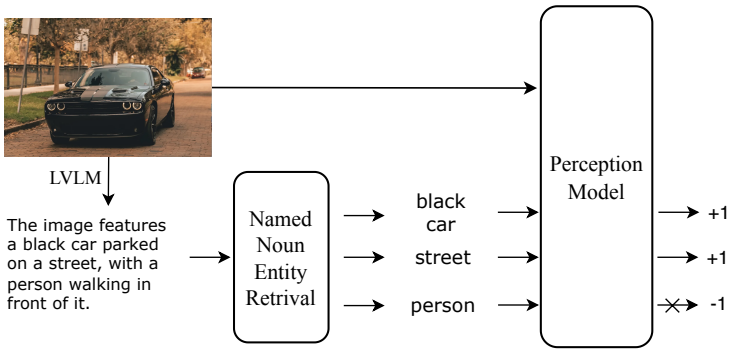

Fig. 1: Illustration of reward modeling with automatic methods.

Description: ...

The model is trained to predict the numerical score in text format. This method not only streamlines the evaluation process but also aligns with the intrinsic strengths of the language model, facilitating more nuanced and stable scoring.

In practice, we find that the *creativity*, *level of detail* and *missing objects* judgment for the holistic description by our annotators is unfortunately somewhat difficult to combine into our rejection sampling scheme. Therefore, we opted to ignore these annotated attributes in the models presented. However, we believe that they will be beneficial for future work.

Reward Model Scoring Upon completing the training of the reward model, it is employed to assign rewards to each description generated by LVLM. As the model provides sentence-level scores, these must be consolidated to derive an overall score for each description. To this end, we compute two distinct scores for each description: a positive score, denoted as P^h, and a negative score, denoted as N^h. The scoring process is as follows: for each sentence that the reward model evaluates as accurate (returning a value of 0), the positive score P^h is incremented by 1. Conversely, for each sentence deemed inaccurate, the negative score N^h is incremented by 1. This scoring method is elaborated upon in Algorithm 1.

1.2 Details of Reward Modeling with Automatic Methods

In this section, we show more details on the implementation of reward modeling with automatic methods. As shown in Figure 1, given a description generated by LVLM from an image, we first identify the individual nouns mentioned using standard recognition of the entity of nouns.

Following the identification of noun phrases, each is paired with the original image and analyzed using a perception model. In our implementation, we employ

Algorithm 2 Automatic Method Reward Model Score

Input: *Image*, {*Descriptions*}, Perception Model R_A
for each S_i in {*Descriptions*} **do**
    $P_i^a \leftarrow 0$
    $N_i^a \leftarrow 0$
    for each noun phrase $n_{i,j}$ in S_i **do**
        if $n_{i,j}$ is detected by R_A **then**
            $P_i^a \leftarrow P_i^a + 1$
        else
            $N_i^a \leftarrow N_i^a + 1$
        end if
    end for
    The final scores for S_i are (N_i^a, P_i^a)
end for

an open-set detector, Grounding DINO [5], for this purpose. The core function of this model is to verify the presence of the identified noun phrases within the image.

Similar to the approach of reward modeling that leverages detailed human feedback, the scoring process of the reward model for automatic methods also yields two distinct scores for each description: a positive score P^a and a negative score N^a. A positive score (+1) is assigned if the noun phrase is successfully detected in the image, for example, 'black car' and 'street' in this case. Conversely, if the perception model fails to locate the noun phrase in the image (e.g., 'person'), a negative score (-1) is attributed. This scoring method is elaborated upon in Algorithm 2.

1.3 Details of Reward Score Combination and Rejection Sampling

In this section, we provide a detailed exposition and clarification of the process for combining reward scores, as well as the subsequent rejection sampling procedure.

Reward Score Combination For each description, we obtain fine-grained reward scores from two sources: reward modeling via detailed human feedback and automated methods, denoted as (N^h, P^h) and (N^a, P^a), respectively. To compute the final score for each description, we sum the negative and positive scores separately, resulting in a composite score: $(N, P) = (N^h + N^a, P^h + P^a)$.

Rejection Sampling Procedure A general procedure of rejection sampling for LVLM is depicted in Figure 2. The process begins with the generation of N samples from LVLM. Each of these samples is evaluated by the reward modeling module, which assigns a specific reward score to each sample. The rejection sampling module selects the best sample according to the reward score.

In our case, the best sample is the one with the smallest N. If there are multiple descriptions with the same smallest N, then the one with the largest P is chosen as a tiebreaker. This best sample is used as the regression target in the supervised fine-tuning setting to improve the LVLM.

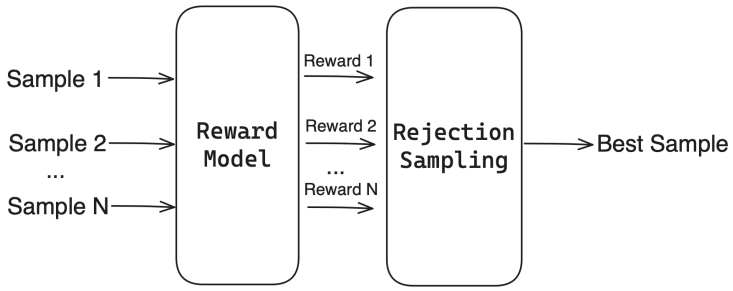

Fig. 2: Illustration of the rejection sampling procedure.

1.4 Details of Refinement Module

In this section, we provide a detailed overview of the refinement module. The core of our strategy lies in a straightforward refinement approach applied to the sample descriptions.

For each sentence within the description, we assess the presence of noun phrases deemed nonexistent by the reward modeling module. Should a sentence contain any such noun phrases, we eliminate that sentence entirely from the description. For instance, consider the description:*"The image features a black Dodge Charger parked on a brick road next to a sidewalk. A person is walking in front of the car."* If the reward modeling module identifies *'person'* as a non-existent noun phrase, the sentence *"A person is walking in front of the car."* is removed.

This refinement process effectively reduces the hallucination problem by removing inaccurate elements, thereby leading to notable improvements in the model's performance.

2 Performance Evaluation Details

2.1 Detailed Description Generation Evaluation

Objective and consistent evaluation of free-form text output has traditionally been a difficult problem. Previously, the research community has often relied on handcrafted metrics such as BLEU [6] and ROUGE [3] to evaluate the similarity between text. However, these generally only reward regurgitation of the exact

same passage as the ground truth, but disregard high-level meaning and alternative but equally valid expression. We elect to use the powerful proprietary GPT-4Vision (with impressive prompt follow-up and visual recognition capabilities [7]) as a tool to give us an imperfect yet still valuable method to evaluate the descriptions generated by various settings of our LVLM.

As our task is fundamentally to compare the quality of generated results of different model settings, we select a set of images and generate a group of descriptions using each model setting to be compared. We then provide a single instruction to GPT-4Vision using the following template along with the underlying image.

Fig. 3: Additional qualitative results. We show examples of descriptions generated by our technique after using the fine-tuning scheme, and compare them with the output from the original LLaVA [4]. Our approach is able to greatly reduce the amount of hallucinations and invalid observations while increasing the amount of detailed visual descriptions. Furthermore, the model retains the plausible intuitive reasoning capabilities of the underlying LLM. Please see Sec. 3 for further discussions.

Please rank the following four descriptions of the provided image according to the following criteria.

Existence accuracy (EA): Evaluate whether the objects mentioned in the description exist in the image.

Counting accuracy (CA): Assess the accuracy in the number of objects described.

Accuracy of attributes (AA): Gauge the accuracy of the attributes (color, material, shape, etc.) assigned to the objects in the description.

Relation accuracy (RA): Consider how accurately the description captures the spatial or relational aspects of objects in the image (e.g., an object being on top of, next to, or inside another object).

Relevance (RL): Assess whether the description is relevant and relevant to the content of the image.

Reasoning (RS): Evaluate the description for reasonable interpretations or extrapolations about the image (e.g., 1. The general atmosphere of the bathroom appears to be in the process of being renovated or updated. 2. The overall atmosphere of the kitchen appears warm and welcoming. 3. It indicates that they are enjoying the time.)

Detail Level (DL): Measure the level of detail provided in the entire description, considering both the quantity and the quality of the details.

After completing the ranking, please use the following template to report your results, where the first item represents the best and the last the worst.

For example: EA: [0,2,3,1]; CA: [1,3,2,0]; AA: [3,2,1,0]; RA: [1,0,2,3]; RL: [0,3,2,1]; RS: [3,1,0,2], DL: [1,3,0,2]. DO NOT return any explanation.

Description 1: ...
Description 2: ...
Description 3: ...
Description 4: ...

Upon retrieving the responses from GPT-4Vision, we aggregate the ranking information for each image's description set by computing the average rank within each metric type across the image samples. As such, the best average rank that a model can achieve within a particular metric is 1.0, while the worst is equal to the total number of competitor models, which is 4.0 in the above case.

3 Additional Qualitative Results

In Figure 3, we provide additional samples of conversations that contrast the behavior of the LVLM before and after the fine-tuning using ViGoR. Through these varied examples, it is evident that our fine-tuning scheme focusing on improving visual grounding in creating detailed descriptions has strong generalization capability to other types of queries. In the first, third, and last rows, we see further cases in which our approach is able to produce more accurate descriptions with fewer incidents of hallucinations that are present in the responses of the original model. The second and fourth rows demonstrate that our approach is able to maintain the powerful abstract reasoning capabilities of the original LLM, but further refine its attention and visual grounding capacities.

4 Potential Negative Societal Impact

In our experiments, we train and debug ViGoR using eight 40G A100 GPUs for approximately 100 hours. The total emission is estimated to be 60.00 kgCO2eq, equivalent to 242.6 km driven by an average car. This estimation is conducted using the Machine Learning Impact calculator presented in [2].

References

1. Chen, T., Saxena, S., Li, L., Fleet, D.J., Hinton, G.: Pix2seq: A language modeling framework for object detection. arXiv preprint arXiv:2109.10852 (2021)
2. Lacoste, A., Luccioni, A., Schmidt, V., Dandres, T.: Quantifying the carbon emissions of machine learning. arXiv preprint arXiv:1910.09700 (2019)
3. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013
4. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
5. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
6. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002). https://doi.org/10.3115/1073083.1073135, https://aclanthology.org/P02-1040
7. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.C., Liu, Z., Wang, L.: The dawn of lmms: Preliminary explorations with gpt-4v(ision) (2023)