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

Siming Yan⋆†1 Min Bai⋆2 Weifeng Chen2 Xiong Zhou2
Qixing Huang1 Li Erran Li2

1The University of Texas at Austin 2AWS AI

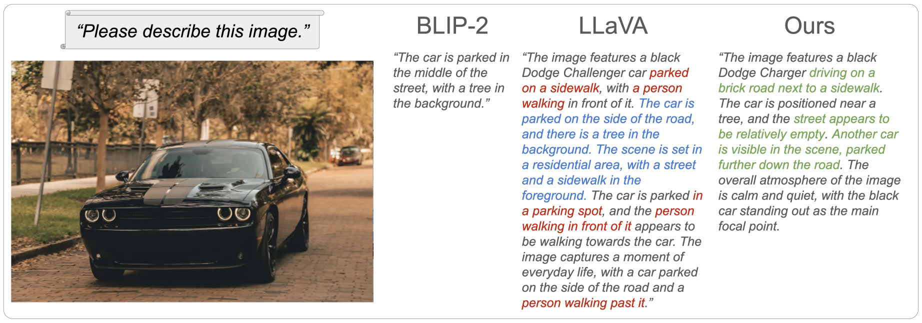
Abstract. By combining natural language understanding, generation capabilities, and breadth of knowledge of large language models with image perception, recent large vision language models (LVLMs) have shown unprecedented visual reasoning capabilities. However, the generated text often suffers from inaccurate grounding in the visual input, resulting in errors such as hallucination of nonexistent scene elements, missing significant parts of the scene, and inferring incorrect attributes of and relationships between objects. To address these issues, we introduce a novel framework, **ViGoR** (**Vi**sual **Gro**unding Through Fine-Grained **R**eward Modeling) that utilizes fine-grained reward modeling to significantly enhance the visual grounding of LVLMs over pre-trained baselines. This improvement is efficiently achieved using much cheaper human evaluations instead of full supervisions, as well as automated methods. We show the effectiveness of our approach through a variety of evaluation methods and benchmarks. Additionally, we released our human annotation (https://github.com/amazon-science/vigor) comprising 15,440 images and generated text pairs with fine-grained evaluations to contribute to related research in the community.

1 Introduction

Large language models (LLMs) have garnered intense interest throughout the research and academic communities. Typically, these models are pre-trained with tremendous amounts of automatically aggregated text data followed by further fine-tuning with specific user examples or evaluation feedback. This enables the models to follow human-provided prompts and retrieve useful relevant information or solve logical problems. Recently, numerous techniques [10, 16, 17, 21, 33, 34, 37] have enhanced these breakthroughs with the ability to understand visual information by further integrating image features into the prompt encoding process. Although these works have successfully aligned image features

[⋆] Equal contribution.

[†] Work done when interning at AWS AI. Contact email: siming@cs.utexas.edu.



"Please describe this image."

BLIP-2

"The car is parked in the middle of the street, with a tree in the background."

LLaVA

"The image features a black Dodge Challenger car parked on a sidewalk, with a person walking in front of it. The car is parked on the side of the road, and there is a tree in the background. The scene is set in a residential area, with a street and a sidewalk in the foreground. The car is parked in a parking spot, and the person walking in front of it appears to be walking towards the car. The image captures a moment of everyday life, with a car parked on the side of the road and a person walking past it."

Ours

"The image features a black Dodge Charger driving on a brick road next to a sidewalk. The car is positioned near a tree, and the street appears to be relatively empty. Another car is visible in the scene, parked further down the road. The overall atmosphere of the image is calm and quiet, with the black car standing out as the main focal point.

Fig. 1: An illustration of inaccurate visual grounding in Large Vision Language Models. BLIP-2 [17] failed to provide detailed image descriptions. LLaVA [21]'s more detailed description contains both correct sentences as well as sentences with hallucinations and inaccurate inference. In contrast, our model preserves the logical reasoning and creativity of LVLMs while exhibiting significantly better accuracy and detail.

into the large language model domain, they still exhibit significant problems. While a strong contextual grounding of language-only models can be learned from the enormous corpus of text data, paired training data for multimodal language/vision models is more limited, while the complexity of the task is arguably higher as the model must align two disparate modalities. Large, automatically compiled datasets, such as the LAION-5B dataset [27], tend to feature only simple images with very short text descriptions. Training with such data often leads large vision language models to fail in capturing the essential details of the image, returning only a short and coarse description (see BLIP-2 [17] output in Figure 1). Moreover, techniques such as InstructBLIP [10] and BLIP [17] primarily rely on high-quality paired language/image datasets (e.g. VQAv2 [13], VizWiz [14], TextCaps [28], etc.). However, these datasets are expensive to collect and challenging to adapt for broader coverage due to the need for manual text annotation. On the other hand, initiatives such as LLaVA [21] train on perception and simple caption datasets in conjunction with the reasoning capabilities of LLMs to semi-automatically generate synthetic conversational ground truth. Unfortunately, such outputs can also contain non-factual statements suffering from hallucinations, omissions, and inaccuracies in attribute or relational descriptions, as they are generated by text-only models based on sparse information about the actual image. Hence, the resulting trained model is still not ideal (see LLaVA's output in Figure 1).

Instead, we propose a novel and general framework using fine-grained reward modeling. It efficiently and substantially enhances the visual grounding of LVLMs beyond pre-trained baselines such as LLaVA, while preserving their conversational capabilities. Given a pre-trained LVLM (e.g., LLaVA), we input a set of images with prompts and generate multiple text outputs. Human annotators are asked to evaluate each image-text pair. As seen in the LLaVA output in Figure 1, a lengthy description generated by a competent LVLM can contain both correct and incorrect sentences. Attempting to assign a single holistic score is ambiguous for the annotator, as well as a subsequently trained reward model. In our design, annotators assign fine-grained, per-sentence evaluation scores. This results in a newly compiled dataset comprising image-text-evaluation trios. We train a reward model to also predict dense reward scores and use it to fine-tune the pre-trained LVLM (see Section 3). This fine-tuning process markedly improves the model's visual grounding capabilities with just 16K data samples, demonstrating the method's efficiency.

To further improve the performance of our system at negligible cost, we also develop a reward model scheme based on automatic methods *without additional human effort*, which is shown to be highly effective in improving the visual grounding capabilities of LVLMs.

Finally, we amalgamated the strengths of both reward models to develop a complete solution, which we refer to as **ViGoR** (**Vi**sual **Gro**unding Through Fine-Grained **R**eward Modeling) throughout the remainder of the paper. To assess the efficacy of ViGoR, we evaluate it against two commonly used benchmarks, POPE [19] and MME [12], where it demonstrated significant improvements over the baseline models. Finally, we compare the performance benefit of our approach to baselines for a variety of tasks that require accurate visual grounding. In summary, we make the following three major contributions.

- We introduce a novel framework that incorporates fine-grained reward modeling with easily implemented rejection sampling, substantially enhancing the visual grounding of LVLMs.
- We develop reward models that require little human effort while leveraging the impressive advances in powerful and robust visual perception models, demonstrating marked improvements in visual grounding efficiency.
- We create and release the human evaluation dataset comprising 15.4K pairs of images and generated results, as well as the fine-grained assessments of the latter by our annotators.

2 Related Work

Large Vision Language Models. Recent advances in extending large language models (LLM) such as GPT [5, 26], PaLM [9], BLOOM [32], LLaMA [30], and Vicuna [8] to the visual domain have been remarkable. Flamingo [1] and its open source counterpart OpenFlamingo [3], along with IDEFICS, have been pivotal in integrating LLMs with vision-language pretraining, using techniques like gated cross-attention dense blocks. PaLI's [7] research on the scaling of vision and language components across various tasks has been instrumental. As well, PaLM-E's [11] extension of LLM to the embodied domain, and BLIP-2's [17] introduction of the Querying Transformer to align image and language encoders mark significant progress. mPLUG-Owl [33] aligns visual characteristics before fine-tuning the language model using LoRA [15]. LLaVA directly injects visual tokens into a pre-trained LLM and finetunes the model with synthetic conversations generated by GPT-4 using metadata and short captions as input. While these techniques primarily focus on architectural innovation for aligning image features to the feature space of LLMs or leveraging extensive image-text data for instruction tuning, our work introduces a distinct and novel general framework designed to enhance the visual grounding capability of any LVLM.

Visual Perception Models. Recent foundational vision models such as CLIP [25] have shown remarkable proficiency in handling a variety of tasks in open world scenes. By reformulating object detection as a grounding problem, GLIP [18] achieves semantic alignment at both the phrase and region levels, leading to impressive open-set detection performance. GroundingDINO [22] combines the state-of-the-art object detector DINO [6] with language pretraining for open-set generalization. This model demonstrates a strong ability to determine the presence and count of objects in a scene. However, current LVLMs do not possess such advanced visual grounding abilities. As such, we devise an automatic method for building a reward model using these vision perception models, and distill their strong visual grounding capabilities directly into LVLMs.

Reward Modeling. Recent progress in training Large Language Models has increasingly emphasized the importance of reward modeling. This approach often incorporates human feedback and reinforcement learning optimization strategies, such as Proximal Policy Optimization. This approach is crucial in refining the accuracy and contextual relevance of model outputs. For example, Askell et al. [2, 24] highlighted the potential of using human feedback in the training of general language assistants, emphasizing the importance of aligning model responses with human standards and values. Recently, LLaMA-2 [31] introduced a novel rejection sampling strategy within reward modeling, claiming that it improves the generation of contextually appropriate high-quality responses.

However, in the realm of Large Vision Language Models, the application of reward modeling remains underexplored, with most existing work focusing predominantly on instruction tuning [21, 37]. One exception is LLaVA-RLHF which adapts the Reinforcement Learning from Human Feedback from the text domain to the task of vision-language alignment, where human annotators are asked to compare two responses and pin-point the more hallucinated one [29]. Inspired by LLaMA-2, we combine reward modeling with rejection sampling in the LVLM training framework. While LLaVA-RLHF uses a reward model that produces sparse signals, we leverage fine-grained reward models to improve the visual grounding capabilities of LVLMs, leading to more accurate and contextually relevant output in vision-language tasks.

3 ViGoR: Visual Grounding Improvement Framework

Our primary goal is to increase visual grounding and reduce hallucinations while keeping the strong intuitive reasoning and creative thought process of pre-trained LLMs and LVLMs. While it has been shown that high quality human annotations for supervised fine-tuning is a straightforward approach for significantly

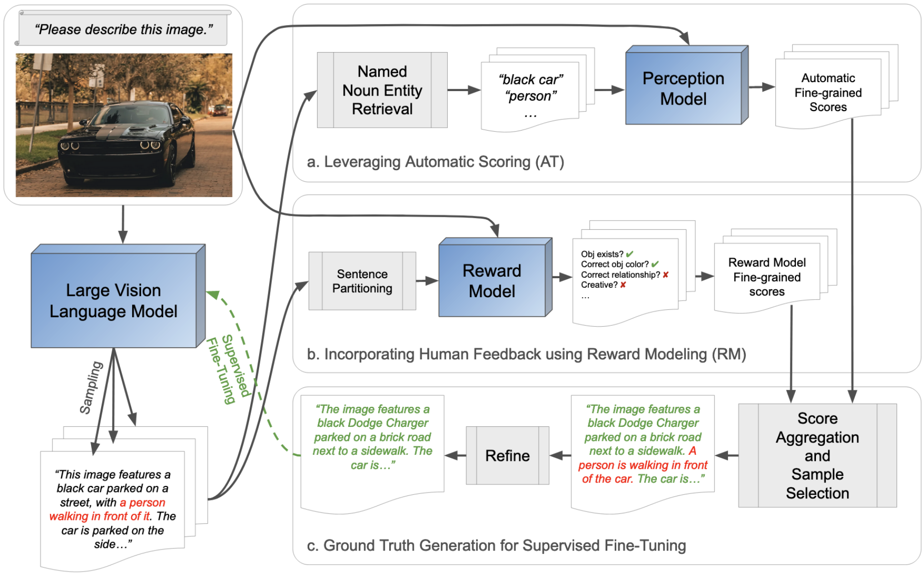

"Please describe this image."
Large Vision Language Model
Sampling
Supervised Fine-Tuning
"This image features a black car parked on a street, with a person walking in front of it. The car is parked on the side…"
Named Noun Entity Retrieval
"black car"
"person"
…
Perception Model
Automatic Fine-grained Scores
a. Leveraging Automatic Scoring (AT)
Sentence Partitioning
Reward Model
Obj exists? ✓
Correct obj color? ✓
Correct relationship? ✗
Creative? ✗
…
Reward Model Fine-grained scores
b. Incorporating Human Feedback using Reward Modeling (RM)
"The image features a black Dodge Charger parked on a brick road next to a sidewalk. The car is…"
Refine
"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. The car is…"
Score Aggregation and Sample Selection
c. Ground Truth Generation for Supervised Fine-Tuning

Fig. 2: Overview of our model training framework. Starting with an input image and a prompt to generate a detailed caption, we sample a number of responses from the LVLM. These responses are passed through two fine-grained reward signal generation branches (*a.* by leveraging state-of-the-art generalizable perception models and *b.* by using an LVLM-based reward model trained using annotator feedback). Finally, we *c.* combine the fine-grained assessment signals from both sources into a single reward score, and select the best sampled description. Finally, we use heuristics and byproducts from the automated scoring system to further refine this sample, and use it for supervised fine-tuning of the LVLM.

improving LVLMs, it is cost-prohibitive for many application scenarios. As such, we wish to efficiently leverage annotator time and the latest advances in direct visual perception models such as powerful open-set object detectors.

We construct a system to fine-tune a base LVLM with rejection sampling, similar to LLaMA-2 [31], allowing the model to improve through using intelligent ranking of its own sampled output. This requires a robust and perceptive scoring system for the generated text. In our work, we use two complementary solutions in parallel. In particular, we train a reward model to incorporate human annotator assessments of text outputs from the LVLM, and provide positive and negative assessments of the LVLM during training time with unlabeled examples. As well, we leverage an open-set object detector and heuristics to verify the existence or absence in the image of the named noun entities extracted from the generated descriptions. Finally, we combine these signals into a single reward score, and use it to select the best description among the initial samples. This sample undergoes additional refinement and is used for the supervised fine-tuning of the LVLM. We refer the reader to Figure 2 for a visualization.

3.1 Reward Modeling via Fine-Grained Human Feedback

Human Preference Data Collection. We design a system to incorporate human judgment and preference — considered the most reliable ground truth signal — into our model training. First, we select a pretrained LVLM's checkpoint state, and create image / caption pairs from the model using a nonzero temperature to balance factual rigor with creativity. Unlike other approaches that ask annotators to provide relatively sparse judgments for the LLM/LVLM's generated results, we ask crowd-workers to provide fine-grained feedback at the sentence level. For each sentence containing errors, the annotator selects the nature of the inaccuracy from a predefined list including object hallucinations, type of attribute error, incorrect relationships between multiple objects, error in the mentioned location or unreasonable conjectures about the image. Furthermore, the annotator provides a judgment about the creativity of the sentence. The creativity means if the sentence gives a thoughtful and reasonable interpretation or extrapolation of the image. Finally, the annotator provides a holistic assessment of the overall description's level of detail and identifies the missing elements in the scene. These requirements encapsulate our overall goal: to enhance the LVLM's visual grounding across the entire image while maintaining the insight and creativity inherent in the pre-trained language decoder.

Reward Model Training. Using the collected annotations, we fine-tune a dedicated LVLM as the reward model on the annotations using instruction tuning to judge the base LVLM's generated results during training time. The reward model is trained to output a sequence of text tokens which encodes the various scores given the underlying image and the LVLM's output as the input. While existing work [29] generates a single holistic score for the entire text output, this process can be ambiguous when the description contains both correct and incorrect components (see Figure 1). Instead, we train the model to produce sentence-level evaluations. This fine-grained approach reduces ambiguity, and increases the detailed visual grounding of the reward model. To provide the necessary context for scoring each sentence, we prepend the sentence with all preceding generated text and explicitly ask the reward model to score the last sentence (which is also the target sentence) in the given passage. A typical prompt is "*Assess the accuracy of the last sentence in the following description of this image, and return a single number.*" Due to the fine-grained feedback provided by the annotators, these prompts result in either a positive response (when no errors are found in the sentence), or a detailed negative response (where one of several error types are found in the sentence). More details are found in the Supplementary Materials.

As we show in the ablation studies, compared with holistic-based method, this fine-grained method significantly improves the reward model's capabilities to guide the fine-tuning of the LVLM. Furthermore, we see that the reward model trained with fine-grained feedback can better understand the link between errors in the descriptions and the image, resulting in superior fine-tuning performance.

3.2 Reward Modeling with Automatic Methods

While the preference annotations directly encapsulate human preferences and are cheaper than supervised fine-tuning annotations, the human effort is still non-negligible. This can limit the scale of potential datasets and subsequently the visual discerning capabilities of the learned reward model due to overfitting. To further improve the cost efficiency of the overall system, we leverage the advances in discriminative vision perception models to automatically score the grounding and fidelity of text generated by a LVLM on large quantities of unlabeled images. However, these discriminative models generally have structured input (such as images and semantic classes) and structured outputs (such as bounding boxes). On the other hand, LVLMs operate with unconstrained and unstructured input and output (free-form text). This gap prevents directly providing the LVLM's output to the discriminative models for scoring. As such, we carefully design the scoring system shown in part a of Figure 2.

Starting with a caption generated by the LVLM from an image, we identify the individual nouns mentioned using standard named entity recognition with NLTK [4]. Next, we prompt an open-set object detector with these nouns to verify the existence of the objects in the image. A correct identification is rewarded with a positive score, while hallucinated objects incur a penalty. We use Grounding DINO [22] for its strong performance across a wide variety of image domains. However, we note that while these detection models are adept at identifying the existence of objects, their ability to detect other types of errors is limited (e.g. object attributes and relationships). This limitation underscores the continued necessity for human-preference-based reward modeling.

3.3 Reward Score and Rejection Sampling

As both our reward model-based and automated methods provide fine-grained reward scores, we must design a strategy to combine them into a single score to enable rejection sampling at the description level. Note that this process embodies the same ambiguity faced by annotators in existing work, where they are asked to analyze complete text output in detail and rank them based on *holistic* preference. However, our solution combines the detailed analysis in a more principled and consistent fashion.

The sample with the best score is used as ground truth for supervised fine-tuning. As our focus is to reduce generation of erroneous descriptions, we use the signals that indicate errors as the primary selector. For each sample, we aggregate all negative signals from the two streams of description evaluation by normalizing their values with their respective variances and linearly combine them. We select the sample with the smallest penalty score as the best candidate, and use the positive scores as a tiebreaker when necessary.

3.4 Refinement Module

We design a simple refinement module to further polish the selected candidate description to use as a regression target (see part c of Figure 2). 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.

3.5 Model Training

The resulting description from the refinement module is used as ground truth in supervised fine-tuning with the standard autoregressive objective as in the original LLaVA [21]. As will be demonstrated by the ablation studies, two signal sources are complementary and provide better results than either one alone. This overall process is visualized in Figure 2.

4 Experiments and Results

We compare the effectiveness of our approach with competitive baselines, and delve into the contributions from each component in our design. As well, we provide visualizations to qualitatively compare the output of fine-tuned model with that of its initial state. In the following, "ViGoR-AT" refers to our approach utilizing only reward modeling with automatic methods. "ViGoR-RM" represents our method employing reward modeling with fine-grained human feedback. "ViGoR-All" and "ViGoR" denote our method with the combination of the reward modeling with automatic methods and fine-grained human feedback.

4.1 Testing Framework

We use the recently proposed LLaVA [21] as the base model to demonstrate the effectiveness of our fine-tuning strategies without modifications to its architecture. In particular, we select the variant of the LLaVA model with the pre-trained and frozen ViT-L / 14 @ 224px CLIP image encoder [25] and the pre-trained Vicuna v1.3 [8] with 7B parameters as the language model. Our method is not specific to any particular LVLM testing architecture or configuration. We use this configuration for both our base LVLM model and our learned reward model.

With computation efficiency in mind, we select the smallest variant of the Grounding DINO model with the Swin-T [23] backbone with the official checkpoint and the default box and text thresholds (0.25) for the open-set object detection task. Our proposed method can likely directly benefit from the larger and more computationally expensive variants or future advances in these models.

4.2 LVLM Model Fine-tuning

To support our experiment, we used 25,574 images from the ADE20K dataset [36] (training split) as the basis, as they cover a wide variety of relatively complex scenes. To preserve the generalization of our method, we disregard the original annotations of the dataset. We elicit responses from our base LVLM through a set of question prompts (e.g. "*Please provide a detailed description of the given image.*"), and sample 5 outputs for each input image. Subsequently, these outputs are fed into our reward model to obtain a corresponding reward score for each description. In addition to keeping the highest-scoring description following the common practice in rejection sampling, we enhance this description further by applying a refinement module to achieve additional quality improvements.

To reduce computational complexity, we choose an *offline* supervision generation strategy where we save the refined output as the ground truth for its respective image using the initial model state. However, we note that the rejection sampling process can take place *online* by refining an evolving model state. We believe that this has the potential to further improve the algorithm's effectiveness, but requires significantly more computation.

During the fine-tuning phase, we set the learning rate at 2×10^{-5} and train the model over two epochs with a batch size of 32. The entire process is executed on eight 40G A100 GPUs, taking 7 hours in total to complete.

4.3 Reward Model Training

Human feedback collection. To allow the reward model to be highly receptive to errors in the descriptions, we use a pre-trained checkpoint for LLaVA that is already fine-tuned with our automatic reward generation mechanism. Through hands-on experimentation, we observed that while the images from the ADE20K dataset exhibit excellent scene variability, the resulting cognitive load for human annotators is very high. As such, instead, we generate 15,440 detailed image captions using images selected from the somewhat simpler MS COCO [20] dataset. We enlist the services of 15 professional annotators to assist us in creating the evaluation data set for the training of the reward model using the process described in Section 3.1. We provide carefully designed and detailed annotation instructions (available in our Supplementary Materials) to our annotators, along with extensive sample annotations. The process took approximately 3 weeks.

Reward model. We initialize the same model architecture as our base LVLM with the same weights used to generate the annotation samples as the starting state of our reward model. We train this model for 5 epochs on the dataset with 15,440 samples with a batch size of 32 and an initial learning rate of 2×10^{-5}.

4.4 Quantitative Results

The types of natural interactions that a user can have with a LVLM are essentially unlimited in variety. Since the focus of our work is on improving visual

Method	HL	CA	AA	RA	RL	RS	DL	Average
LLaVA	3.28	3.27	3.28	3.27	3.27	3.27	3.01	3.24
ViGoR-AT	2.26	2.27	2.25	2.26	2.25	2.30	2.06	2.24
ViGoR-RM	2.48	2.50	2.47	2.49	2.49	2.46	2.96	2.55
ViGoR-All	**1.97**	**1.96**	**1.99**	**1.98**	**1.99**	**1.97**	**1.97**	**1.97**

Table 1: Detailed description generation evaluation. Scores are the *average preference rank* of the caption when compared within the 4 candidates in the table as judged by GPT-4Vision (ranging from 1 to 4; lower is better).
HL: Does the description have **hallucinations**?
CA: Does the description have good **counting accuracy** for objects?
AA: Does the description assign **accurate attributes** to objects?
RA: Does the description have **accurate relationships** between objects?
RL: Is the description **relevant** to the image as a whole?
RS: Does the description exhibit reasonable thought and **reasoning**?
DL: Does the description encompass all **details** in the image?

grounding, we primarily evaluate the benefits of our technique in three types of settings. First, we analyze the unstructured output of the LVLM when it is asked to generate a detailed and comprehensive caption for an image. This tests the model's ability to be perceptive to all regions and objects in the image. Second, we evaluate the ability of the model to form short sentences in response to targeted questions where some logical reasoning is needed. Lastly, we evaluate the model's performance on responding to highly specific questions for which a concise and unambiguous answer is required. This further tests the model's ability to use the input prompt to guide its visual understanding. We submit that this variety of tasks forms a comprehensive testing framework.

Detailed Description Generation Evaluation. Examining the factuality of detailed descriptions generated by a LVLM based on an input image allows provides highly comprehensive insight into visual grounding capabilities. However, objective quantitative measurement of free-form responses has been a major challenge in the LLM and LVLM space due to the inherent ambiguities in defining quality. Furthermore, it is unclear how structured information can be extracted from generated text so that a score can be assigned using traditional objective functions. As demonstrated by [35], the GPT family of models have very high levels of agreement with human annotators when asked to follow instructions to evaluate output of other models. This is likely due to the large quantities of proprietary data used for training. Therefore, we ask GPT-4Vision to rank the output of different LVLM candidates according to numerous criteria that signify the quality of the captions. This offers a detailed view of the visual foundation and text generation capabilities of the LVLMs under test.

For detailed insight, we evaluate model output in terms of hallucinations, relevance, reasoning, level of detail, as well as accuracy in counting, attributes, and relationships between objects. As it is difficult to grade lengthy text responses against an absolute scale, we instead use preference ranking. GPT-4Vision is asked to consider several candidate responses relative to one another against the sample input image and text query, and provide a ranked list over the responses for each metric. The rankings are averaged over all samples to obtain an overall ranking score for each model, where lower is better.

We evaluate using 300 randomly selected images from the COCO validation dataset and show the results in Table 1 along with an overview of the definitions of the metrics. We select the four most relevant configurations for comparison: 1) the original LLaVA baseline model that we use as the initialization; 2) ViGoR-AT: the ViGoR method using only reward modeling with automatic methods; 3) ViGoR-RM: the ViGoR method using only the guidance from the reward model trained using the annotator feedback, and lastly 3) ViGoR-All: the complete model. As is clearly evident, our complete model consistently achieves the best ranking within the comparison across all metrics. Furthermore, we observe that either source of supervision signals significantly outperforms the original baseline model. Further details are provided in the Supplementary Materials.

Short Answer MMHal-Bench Evaluation. Next, we showcase the performance of our system on the recently proposed MMHal-Bench [29], which involves 96 questions of a variety of types to which short answers are often appropriate, such as "*How would you describe the weather in the image?*". The corresponding images are hand-selected from OpenImages. The evaluation process sends ground truth annotations and the model's responses to GPT-4, and requests GPT-4 to judge the descriptiveness and level of hallucinations. This is combined into a composite score (higher is better).

Method	#Param	MMHal-Bench
LLaVA	7B	1.3
LLaVA-RLHF	7B	1.4 (+0.1)
ViGoR (ours)	7B	**1.6** (+0.3)

Table 2: Short answer evaluation on MMHal-Bench benchmark. For a fair comparison, we show the results of LLaVA-RLHF using solely RLHF optimization without supervised fine-tuning using additional data.

Table 2 shows the average score as reported by the MMHal-Bench. Our approach results in a much larger gain in performance (+0.3) compared to the baseline established by LLaVA-7B, which is more substantial than the improvement achieved through the RLHF scheme proposed by LLaVA-RLHF [29] without using supervised fine-tuning (+0.1). Note that our particular implementation of the ViGoR framework only uses prompts to elicit detailed descriptions of the entire scene, and does not cover additional types of questions. Therefore, the encouraging results on MMHal-Bench suggests that the improvements generalize well to different types of queries. The underlying images do not come from MS COCO or ADE20K (the data sources for our fine-tuning process), showing that our approach also generalizes to different images.

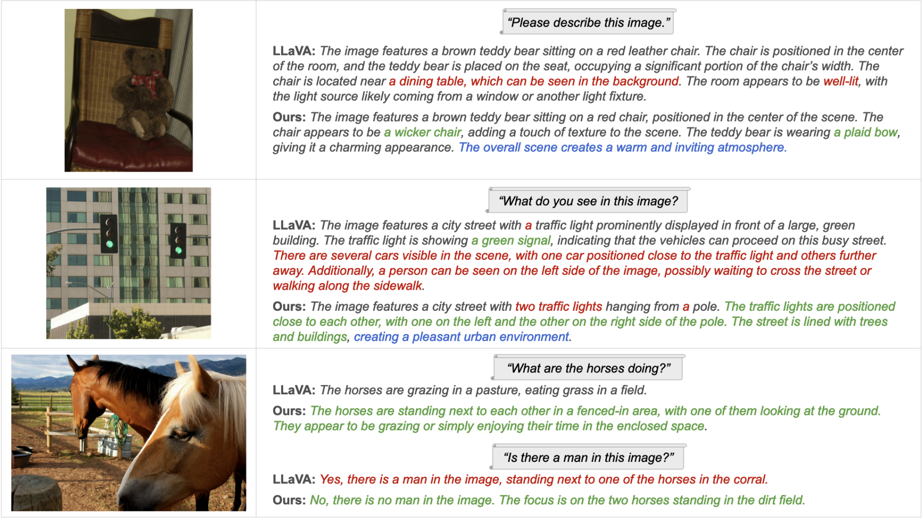
Method	#P	VQA	Random		Popular		Adversarial		Avg F1
			AC	PR	AC	PR	AC	PR	
mPLUG-Owl	7B	✗	53.3	51.7	50.6	50.3	50.7	50.3	67.2
MiniGPT-4	7B	✓	77.8	75.4	68.3	64.3	66.6	62.45	74.1
MM-GPT	9B	✓	50.0	50.0	50.0	50.0	50.0	50.0	70.1
InstructBLIP	13B	✓	**88.7**	85.0	81.4	75.1	74.4	67.7	83.7
LLaVA	7B	✗	54.4	52.3	52.4	51.3	50.8	50.4	67.8
LLaVA-RLHF	7B	✗	-	-	-	-	-	-	78.2
LLaVA-RLHF	7B	✓	-	-	-	-	-	-	82.7
ViGoR (ours)	7B	✗	85.1	**89.0**	**81.5**	**83.0**	**75.5**	**73.8**	**83.8**

Table 3: Quantitative results on POPE benchmark. POPE is comprised of three parts, each generated by different sampling strategies: Random, Popular, and Adversarial Sampling. We report the Accuracy(AC) and Precision(PR) for each part, alongside the average F1 score (Avg F1) across all three parts.

Method	#Param	Data	VQA	EX	CT	PO	CO	PT	CE	SC	LM	AT	OC	Overall
MiniGPT-4	7B	5K	✓	68.3	55.0	43.3	75.0	41.8	54.4	71.8	54.0	60.5	57.5	581.7
VPGTrans	7B	1.4M	✗	70.0	85.0	63.3	73.3	84.0	53.5	141.8	64.8	77.3	77.5	790.5
InstructBLIP	13B	1.2M	✓	185.0	143.3	66.7	153.3	123.8	101.2	153.0	79.8	134.3	72.5	1212.8
Cheetor	7B	500K	✓	180.0	96.7	80.0	116.7	147.3	164.1	156.0	145.7	113.5	100.0	1299.9
Muffin	13B	267K	✓	195.0	163.3	66.7	165.0	137.8	81.8	151.3	146.3	116.5	57.5	1281.0
LLaVA[†]	7B	158K	✗	158.3	83.3	51.7	85.0	94.2	85.0	145.8	125.8	74.3	57.5	960.1
ViGoR (ours)	7B	174K	✗	180.0	143.3	83.3	100.0	128.2	121.2	150.5	130.8	127.0	145.0	1309.3
Improvement				+21.7	+60.0	+31.6	+15.0	+34.0	+36.2	+4.7	+5.0	+52.7	+87.5	+349.2

Table 4: Quantitative results on MME benchmark. 'VQA' indicates whether the model is fine-tuned using the VQA datasets. Abbreviations explained: 'EX': Existence; 'CT': Count; 'PO': Position; 'CO': Color; 'PT': Poster; 'CE': Celebrity; 'SC': Scene; 'LM': Landmark; 'AT': Artwork; 'OC': OCR. For a fair comparison, we compare the Large Vision Language Model using the Vicuna backbone. [†] We report the number of LLaVA v1.0 reproduced by ourselves.

Short Answer Programmatic Benchmark Evaluation. Finally, we evaluate the models using benchmarks which consist of questions with concise and unambiguous one word answers such as "*Is there a red apple in the image?*" These benchmarks avoid the uncertainty and noise introduced from leveraging a third-party large language model for evaluation, and instead uses programmatic comparisons. Similar to MMHal-Bench, these benchmarks can measure both the visual grounding capabilities and the generalization of our method as the question types are outside the scope of the training data. We select the commonly used POPE [19] and MME [12] benchmarks. POPE probes the LVLM with 3000 *yes/no* questions targeting 500 images from COCO [20] to evaluate the LVLM's ability to determine the existence of specific objects in a scene. MME extends the questions to cover aspects such as numerical count, position, and color of objects, as well as other vision tasks such as scene/landmark identification and OCR from data sources other than MS COCO and ADE20K.



"Please describe this image."

LLaVA: *The image features a brown teddy bear sitting on a red leather chair. The chair is positioned in the center of the room, and the teddy bear is placed on the seat, occupying a significant portion of the chair's width. The chair is located near a dining table, which can be seen in the background. The room appears to be well-lit, with the light source likely coming from a window or another light fixture.*

Ours: *The image features a brown teddy bear sitting on a red chair, positioned in the center of the scene. The chair appears to be a wicker chair, adding a touch of texture to the scene. The teddy bear is wearing a plaid bow, giving it a charming appearance. The overall scene creates a warm and inviting atmosphere.*

"What do you see in this image?

LLaVA: *The image features a city street with a traffic light prominently displayed in front of a large, green building. The traffic light is showing a green signal, indicating that the vehicles can proceed on this busy street. There are several cars visible in the scene, with one car positioned close to the traffic light and others further away. Additionally, a person can be seen on the left side of the image, possibly waiting to cross the street or walking along the sidewalk.*

Ours: *The image features a city street with two traffic lights hanging from a pole. The traffic lights are positioned close to each other, with one on the left and the other on the right side of the pole. The street is lined with trees and buildings, creating a pleasant urban environment.*

"What are the horses doing?"

LLaVA: *The horses are grazing in a pasture, eating grass in a field.*

Ours: *The horses are standing next to each other in a fenced-in area, with one of them looking at the ground. They appear to be grazing or simply enjoying their time in the enclosed space.*

"Is there a man in this image?"

LLaVA: *Yes, there is a man in the image, standing next to one of the horses in the corral.*

Ours: *No, there is no man in the image. The focus is on the two horses standing in the dirt field.*

Fig. 3: Qualitative results. We compare examples of descriptions generated by our technique after using the fine-tuning scheme with the output from the original LLaVA [21]. Our greatly reduces 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.

As illustrated in Tables 3 and 4, our model shows consistent improvement across all categories evaluated in both datasets. Furthermore, while the competitive baseline methods [10, 16, 34, 37] use VQA datasets containing ground truth for these types of straightforward questions, our method is able to achieve comparable or superior results **without fine-tuning on such resources**. This suggests that the visual grounding abilities learned by the model through our process is general. When the model learns to generate accurate *comprehensive descriptions* for images, it is able to apply the capabilities to related *reasoning* tasks, suggesting that the logical deduction capabilities of the underlying pre-trained LLM is retained and leveraged. Furthermore, our 7B parameter model shows on par performance with some models with nearly double the number of parameters (13B). This underscores the efficiency and robustness of our model in handling diverse visual-language tasks. As with the case of MMHal-Bench, the MME benchmark draws imagery from sources other than MS COCO to encompass posters, celebrity images, scenes, landmarks, artwork, and OCR. Our method (using ADE20K and MS COCO in the pipeline) still shows significant improvements in these categories, further proving its generalizability.

Table 3 compares our method with LLaVA-RLHF [29], indicating that our method outperforms LLaVA-RLHF even when the latter is trained with additional VQA data. This further confirms the superior visual grounding capabilities of our model, attributed to our fine-grained reward modeling strategy.

4.5 Qualitative Results

Figure 3 compares our fine-tuned model with the base model (LLaVA) and highlights the enhancements our model has achieved, offering insights into the impact of our training strategies. More results are in the Supplementary Materials.

4.6 Fine-grained vs Holistic Evaluation by Reward Model

We compare the fine-grained analysis of sentences by the reward model with the more commonly used holistic evaluation of complete responses. This assessment is carried out in the context of reward modeling with human feedback. For the holistic reward model, we retrained a reward model using a singular, unified reward for each description. Specifically, holistic reward is determined based on the content of the entire description, which is negative if the description contains at least one erroneous sentence, and positive otherwise. Table 5 confirms that denser and more informative signals in the fine-grained approach allow the reward model to better discern the quality of the generated text by creating more direct links between the visual features in the image and the sentences describing them.

Reward model design	MME
none	960.1
holistic-based	1027.4
fine-grained-based	**1309.3**

Table 5: Fine-grained vs holistic evaluation by reward model.

5 Conclusion, Limitations, and Future Work

Our work takes a step toward improving visual grounding capabilities of LVLMs to reduce hallucination, errors in relational reasoning, counting, and so on. We developed a framework named **ViGoR** to combine fine-grained reward modeling of human preferences with powerful existing open-set visual perception models to efficiently improve LVLMs in these aspects. To validate our approach, we collect the first large dataset with fine-grained human annotation feedback for the LLaVA. We show that ViGoR significantly improves the LLaVA model's ability to generate accurate and relevant text given input images, while preserving its ability to creatively and intuitively reason about the scene.

Despite its successes, ViGoR has several limitations. The automated component of our reward generation relies on the perception model (i.e. GroundingDINO). The current choice is limited to objects suitable for detectors, and does not handle *stuff* regions, attributes, or layouts. As well, the human evaluation data used for training the reward model is specific to a particular LVLM's architecture and checkpoint with which the evaluated responses are generated. Additional annotation data may be needed for LVLMs with very different output or failure modes, which incurs further cost. We hope to address these limitations in future work. As well, we plan to apply RLHF training with our ViGoR framework, which may offer further improvements over our training based on rejection sampling. As well, we believe that explicitly linking visual entities with associated phrases in the generated text can further improve visual grounding.

References

1. Alayrac, J.B., et al.: Flamingo: a visual language model for few-shot learning. arXiv preprint arXiv:2204.14198 (2022)
2. Askell, A., et al.: A general language assistant as a laboratory for alignment. arXiv preprint arXiv:2112.00861 (2021)
3. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al.: Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023)
4. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media (2009)
5. Brown, T., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems **33**, 1877–1901 (2020)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
7. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., Kolesnikov, A., Puigcerver, J., Ding, N., Rong, K., Akbari, H., Mishra, G., Xue, L., Thapliyal, A.V., Bradbury, J., Kuo, W., Seyedhosseini, M., Jia, C., Ayan, B.K., Ruiz, C.R., Steiner, A.P., Angelova, A., Zhai, X., Houlsby, N., Soricut, R.: PaLI: A jointly-scaled multilingual language-image model. In: The Eleventh International Conference on Learning Representations (2023), `https://openreview.net/forum?id=mWVoBz4W0u`
8. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), `https://lmsys.org/blog/2023-03-30-vicuna/`
9. Chowdhery, A., et al.: Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311 (2022)
10. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)
11. Driess, D., et al.: Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023)
12. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., Li, K., Sun, X., Ji, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 (2023)
13. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6904–6913 (2017)
14. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3608–3617 (2018)
15. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), `https://openreview.net/forum?id=nZeVKeeFYf9`

16. Li, J., Pan, K., Ge, Z., Gao, M., Zhang, H., Ji, W., Zhang, W., Chua, T.S., Tang, S., Zhuang, Y.: Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. In: The Twelfth International Conference on Learning Representations (2023)
17. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
18. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022)
19. Li, Y., Du, Y., Kun Zhou, J.W., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: The 2023 Conference on Empirical Methods in Natural Language Processing (2023), `https://openreview.net/forum?id=xozJw0kZXF`
20. 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)
21. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
22. 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)
23. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
24. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 27730–27744. Curran Associates, Inc. (2022), `https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf`
25. Radford, A., Jong Wook Kim, C.H., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferrable visual models from natural language supervision. In: ICML (2021)
26. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
27. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
28. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioning with reading comprehension. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 742–758. Springer (2020)
29. Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L.Y., Wang, Y.X., Yang, Y., Keutzer, K., Darrell, T.: Aligning large multimodal models with factually augmented rlhf (2023)

30. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
31. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models (2023)
32. Workshop, B., Scao, T.L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A.S., Yvon, F., et al.: Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100 (2022)
33. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)
34. Yu, T., Hu, J., Yao, Y., Zhang, H., Zhao, Y., Wang, C., Wang, S., Pan, Y., Xue, J., Li, D., et al.: Reformulating vision-language foundation models and datasets towards universal multimodal assistants. arXiv preprint arXiv:2310.00653 (2023)
35. Zhang, Y., Mai, Y., Roberts, J.S.R., Bommasani, R., Dubois, Y., Liang, P.: Helm instruct: A multidimensional instruction following evaluation framework with absolute ratings, https://crfm.stanford.edu/2024/02/18/helm-instruct.html
36. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR (2017)
37. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)