Controlling Multimodal LLMs via Reward-guided Decoding

Oscar Mañas[1,2,4], **Pierluca D'Oro**[1,2,4], **Koustuv Sinha**[4],
Adriana Romero-Soriano[1,3,4,5], **Michal Drozdzal**[4], **Aishwarya Agrawal**[1,2,5]
[1]Mila, [2]Université de Montréal, [3]McGill University, [4]Meta FAIR, [5]Canada CIFAR AI Chair
oscar.manas@mila.quebec

Abstract

As Multimodal Large Language Models (MLLMs) gain widespread applicability, it is becoming increasingly desirable to adapt them for diverse user needs. In this paper, we study the adaptation of MLLMs through controlled decoding. To achieve this, we introduce the first method for reward-guided decoding of MLLMs and demonstrate its application in improving their visual grounding. Our method involves learning a reward model for visual grounding and using it to guide the MLLM's decoding process. Our approach enables on-the-fly controllability of an MLLM's inference process in two ways: first, by giving control over the relative importance of reward and output likelihood during decoding, allowing a user to dynamically trade off object precision and recall in image captioning tasks; second, by giving control over the breadth of the search during decoding, allowing a user to trade off compute for output quality. We evaluate our method on standard object hallucination benchmarks, showing that it provides significant controllability over MLLM inference, while matching or outperforming existing visual grounding methods.

1 Introduction

Multimodal Large Language Models (MLLMs) have shown great potential for solving a wide range of visiolinguistic tasks, while offering a language interface to users [5, 8]. As adoption of MLLMs increases [1, 30, 12], a demand to easily control their behavior to satisfy diverse user needs is emerging.

Two needs, in particular, arise among the most important for users of MLLMs: control over the precision and thoroughness of their output, and control over the amount of compute spent to generate those outputs. For instance, a user with visual impairment using the system to understand their surroundings may want the MLLM to respond with thorough outputs that maximize object precision, while avoiding overly high latency on limited compute; instead, a user leveraging the MLLM to generate synthetic captions to train downstream models may value more the object recall of the model's output, while having more flexibility on spending more compute to obtain higher-quality results.

In this paper, we tackle this problem and propose a method for inference-time alignment of MLLMs. Our method is based on reward-guided decoding (RGD) with a reward function tailored for hallucination reduction [3]. Using the reward model as a heuristic for searching for better outputs, our method gives control over the two axes mentioned above: by giving the option to set a relative weight for reward and the MLLM's output likelihood, it allows to control the trade off between object precision and recall of the MLLM's outputs; by varying the breadth of the search, we can control the trade off between compute and output quality.

While other methods such as prompting [37], supervised fine-tuning [21] and RLHF fine-tuning [29, 34, 40] have been proposed to reduce hallucinations, reward-guided decoding enables on-the-fly granular controllability, which is typically hard to obtain with other techniques.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

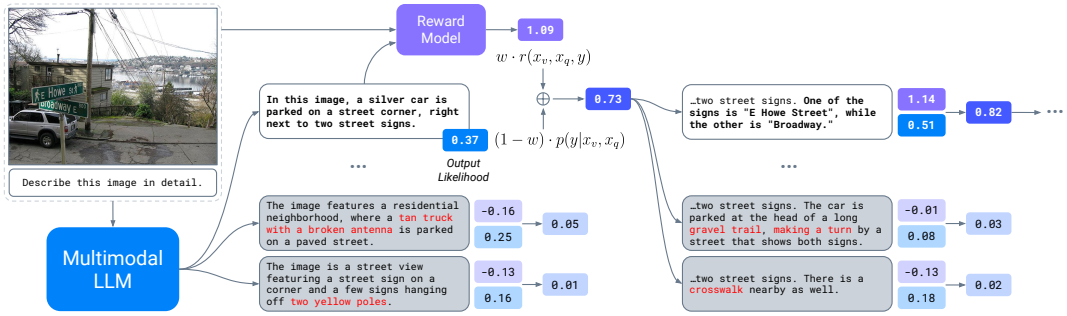

Figure 1: Illustration of reward-guided decoding (RGD) for MLLMs. At each iteration, k candidate completions to a partial response are sampled from the MLLM and evaluated according to a linear combination of their reward and their likelihood (the process is illustrated for the first selected completion and omitted elsewhere). The completion with largest score is selected and added to the context to generate the next k candidates, until the <EOS> token is encountered.

In summary, the main contributions of our paper are: (1) We propose the first approach for guided decoding for MLLMs, based on training a reward function for visual grounding and using it to guide the search for outputs at inference time. (2) We demonstrate on standard visual hallucination benchmarks that using guided decoding allows on-the-fly controllability of the balances between (a) object hallucination and output thoroughness, and (b) amount of compute spent to produce an output and its quality, while obtaining competitive performance to existing non-controllable approaches. (3) We analyze the object precision/recall of generated captions and compute trade-offs of reward-guided decoding for MLLMs, as well as the crucial properties of the proposed search process.

2 Method

2.1 Multimodal reward-guided decoding

Given an image x_v and a visual instruction x_q, an MLLM π generates a text response $y = \{y_1, ..., y_n\}$ autoregressively token-by-token, i.e., $y = \pi(x_v, x_q)$. To measure how well a response satisfies the preferences of a user, we use a reward function $r(x_v, x_q, y)$, which outputs a scalar score given the multimodal instruction $x = (x_v, x_q)$ and the generated response y.

Our goal is to guide the generation of an MLLM such that the generated response can be modulated using the reward function. We leverage *reward-guided decoding* (RGD) [23, 11, 17]: we search for a response by expanding a search tree of partial responses and deciding which partial response to complete depending on a reward-based selection criterion. At each iteration, we sample k candidate completions $\{y^j_{i..i+m}\}^k_{j=1}$ from a single partial response, with $(m < n)$, evaluate each of them with a score $s(x_v, x_q, y^j_{1..i+m})$, select the one with the maximum score, and add it to the context. We then iterate this process until the <EOS> token is generated.

To avoid potential instabilities resulting from the evaluation of outputs that are not well-formed, we evaluate partial outputs at the end of sentence, according to an *evaluation period* of T, i.e., evaluating the output of the MLLM every T sentences. As T grows, the reward model will evaluate longer and longer outputs. For $T \to \infty$, only complete outputs concluded with an <EOS> token are evaluated, and one complete output is selected among them: this strategy is usually referred to as *rejection sampling* or *best-of-k* in the literature [7, 28]. We leave for future work exploring alternative methods to detect semantically complete output segments within a sentence.

To give a user the possibility of choosing the level of reward guidance on-the-fly, we use as score the linear combination of the reward of a partial output and its likelihood under the MLLM: $s(x_v, x_q, y) = w \cdot r(x_v, x_q, y) + (1-w) \cdot p_\pi(y|x_v, x_q)$, where $p_\pi(y|x_v, x_q)$ is the response's conditional likelihood according to the base MLLM, and $w \in [0, 1]$ is a guidance strength hyperparameter chosen at inference time. A user can modulate the strength of the reward guidance by varying w. At the extremes, for $w = 1$, the best response is chosen entirely by following the reward function, while when $w = 0$ the reward function has no effect. Figure 1 provides a summary of our method.

2.2 Adapting a VLM as a multimodal reward model

The effectiveness of our guided decoding strategy hinges on the existence of a reward function capable of successfully evaluating how well a response satisfies a certain objective. Unlike for math or coding problems [7], there are no verifiers for the open-ended responses generated by MLLMs. Therefore, we approximate the reward function by learning a reward model r_θ from preference data.

Given a dataset of multimodal preference data $D = \{x_v, x_q, y^+, y^-\}_i$, where y^+ and y^- are the chosen and rejected responses respectively, we train a reward model as a classifier that predicts the preference probability following the Bradley-Terry model [6, 24]. To facilitate combining the reward with the MLLM's likelihood, it is desirable that $r_\theta(x_v, x_q, y) \in [0, 1]$. Therefore, we add a pair of mean-squared error loss terms to encourage $r_\theta(x_v, x_q, y^+)$ to be close to 1 and $r_\theta(x_v, x_q, y^-)$ to be close to 0, while simultaneously avoiding the gradient saturation pitfalls of squashing activation functions. Ultimately, this leads to the following loss function: $\mathcal{L}(\theta) = \mathbb{E}_{(x,y^+,y^-)\sim D}[-\log \sigma(r_\theta(x, y^+) - r_\theta(x, y^-)) + (r_\theta(x, y^+) - 1)^2 + r_\theta(x, y^-)^2]$, where $x = (x_v, x_q)$. We use PaliGemma [4] as the backbone of our reward model, and add to it a linear regression head.

3 Experiments

While our method can be applied to align MLLMs with any arbitrary objective function, in this paper we focus on improving the visual grounding to produce responses that are more factually grounded in the visual input. Hence, we evaluate the effectiveness of our multimodal reward-guided decoding strategy in mitigating object hallucinations in long captions.

3.1 Experimental setup

Training data. We train our reward model on a mixture of publicly available multimodal preference datasets focusing on visual hallucinations: LLaVA-RLHF [29] (9.4k), RLHF-V [34] (5.7k), POVID [40] (17k) and RLAIF-V [35] (83k). In addition, we repurpose SugarCrepe [15] (7.5k) as preference data for the instruction "`Describe this image`". We split each dataset into 80% for training and 20% for validation. To compensate for data imbalance, we make sure each minibatch has roughly the same number of examples from each dataset.

Implementation details. We initialize our reward model's backbone from PaliGemma (`google/paligemma-3b-pt-224`), train the regression head from scratch and finetune the backbone with LoRA [16]. We use an effective minibatch size of 256, warm up the learning rate from 0 to $1e^{-3}$ during the first 5% of an epoch and decay it to zero with a cosine schedule. We train the reward model for a single epoch. We use LLaVA-1.5_{7B} [22] (`llava-hf/llava-1.5-7b-hf`) as our base MLLM and caption images with the prompt "`Describe this image in detail`". For guided decoding, we use a sampling temperature of 1.0.

Evaluation setup. We evaluate our method on two standard object hallucination benchmarks, CHAIR [26] and AMBER [32], and report instance-level (C_i/CHAIR) and sentence-level (C_s/Hal.) hallucination rates (the inverse of object precision). We also report object recall (Rec.)/coverage (Cov.) and caption length (Len.) to ensure our method generates meaningful captions rather than degenerating into object-less outputs.

Table 1: Results on object hallucination benchmarks. RGD with $k = 30$ and $T = 1$, BS@k indicates beam search with k beams, $*$ indicates reported values from [27], † values are from [10].

Model	Method	COCO			AMBER			
		C_i (↓)	C_s (↓)	Rec. (↑)	Len.	CHAIR (↓)	Hal. (↓)	Cov. (↑)
LLaVA-1.5_{7B}	Greedy	15.05	48.94	81.30	90.12	7.6	31.8	49.3
LLaVA-1.5_{7B}	BS@10	15.80	52.94	81.48	96.31	10.9	39.7	46.0
HA-DPO* [39]	BS@5	11.0	38.2	-	91.0	6.7	30.9	49.8
EOS* [36]	Greedy	12.3	40.2	-	79.7	5.1	22.7	49.1
HALVA$^*_{7B}$ [27]	Greedy?	11.7	41.4	-	92.2	6.6	32.2	53.0
LLaVA-1.5_{7B}	CGD† [10]	8.1	29.7	79.03	76.66	-	-	-
LLaVA-1.5_{7B}	RGD ($w = 1$)	6.77	26.02	74.45	95.17	5.3	25.8	47.9
LLaVA-1.5_{7B}	RGD ($w = 0.5$)	7.43	27.43	76.80	92.91	5.1	23.3	47.5

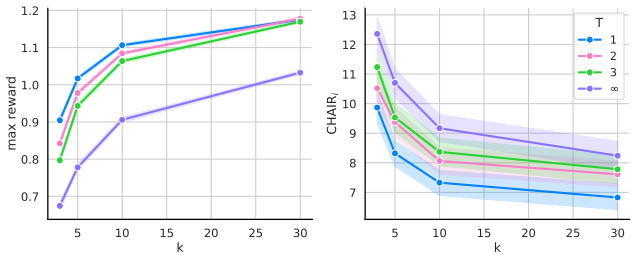

Figure 2: Reward value (left) and CHAIR$_i$ (right) on COCO varying k and T, for $w = 1.0$. Leveraging the reward model to guide the generation more often (lower T) improves sample-efficiency.

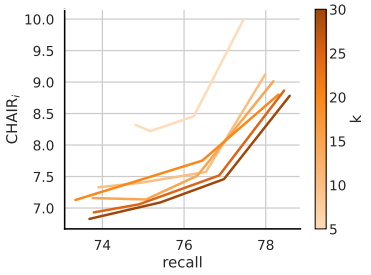

Figure 3: Object precision and recall on COCO, for $T = 1$. Each line represents a value of k, for varying w. Using more compute improves precision and recall, while varying w modulates the trade-off for a given level of compute.

3.2 Results

Reward model accuracy. We first evaluate the performance of our reward model in predicting preferences. We define accuracy as the percentage of times the reward model assigns a higher score to the chosen response, i.e. $r_\theta(x_v, x_q, y^+) > r_\theta(x_v, x_q, y^-)$. We obtain an average validation accuracy of 77.54%, which is in line with typical performance of reward models [18].

Downstream performance. In Table 1, we report the results of leveraging our reward model for guided decoding, and compare against multiple baselines. We observe our RGD method can considerably reduce object hallucinations at the expense of a few object recall/coverage points. For instance, on the COCO benchmark, CHAIR$_i$ is reduced by more than half (from 15.05 with greedy decoding to 6.77 with RGD and $w = 1.0$). By reducing the guidance strength to $w = 0.5$, recall is substantially increased without overly increasing the hallucination rate. Overall, RGD matches or surpasses the performance of existing methods for hallucination mitigation while enabling granular controllability of the MLLM's behavior at inference time.

Guided decoding is sample-efficient. Figure 2 shows how the maximum reward and hallucination rate (CHAIR$_i$) evolve as we increase the number of samples $k = \{3, 5, 10, 30\}$. As expected, we observe a lower hallucination rate when increasing k. More importantly, we see our RGD strategy (with $T = 1$) is considerably more sample-efficient than naive rejection sampling (with $T = \infty$). For instance, a similar hallucination rate is achieved with rejection sampling with $k = 30$ and RGD with $k = 5$, which makes RGD $\sim 6\times$ more sample-efficient than rejection sampling.

Inference-time reward-guidance modulation. Figure 3 shows the trade-off between hallucination rate (CHAIR$_i$) and object recall when varying the guidance strength $w = \{0.25, 0.5, 0.75, 1.0\}$. We observe a low w leads to high recall but also higher CHAIR$_i$, while a high w has the opposite effect. And this effect is more pronounced with higher k. Hence, we confirm the flexibility of our approach and its effectiveness in adapting to user needs at inference time.

4 Conclusion

In this paper, we presented a method for reward-guided decoding (RGD) of MLLMs, based on a multimodal reward model trained for visual grounding on a dataset of preferences. This reward model is then used in a search process, in which, at each iteration, several candidate responses are evaluated against a combination of their reward value and their likelihood. We show that, for the task of image captioning, this methodology affords on-the-fly controllability of a MLLM's output along two axes: first, it allows a user to trade off object precision and recall in a fine-grained way, by just changing the weight on each term of the search score; then, it allows to increase the amount of compute employed to generate a complete output of higher quality, by varying the breadth of the search. Our method provides significant controllability over MLLM inference while matching or surpassing the performance of existing visual grounding methods.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Elmira Amirloo, Jean-Philippe Fauconnier, Christoph Roesmann, Christian Kerl, Rinu Boney, Yusu Qian, Zirui Wang, Afshin Dehghan, Yinfei Yang, Zhe Gan, et al. Understanding alignment in multimodal llms: A comprehensive study. *arXiv preprint arXiv:2407.02477*, 2024.

[3] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

[4] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[5] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024.

[6] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[7] Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

[8] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The (r) evolution of multimodal large language models: A survey. *arXiv preprint arXiv:2402.12451*, 2024.

[9] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

[10] Ailin Deng, Zhirui Chen, and Bryan Hooi. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *arXiv preprint arXiv:2402.15300*, 2024.

[11] Haikang Deng and Colin Raffel. Reward-augmented decoding: Efficient controlled text generation with a unidirectional reward model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11781–11791, 2023.

[12] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[13] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312, 2024.

[14] Seungwook Han, Idan Shenfeld, Akash Srivastava, Yoon Kim, and Pulkit Agrawal. Value augmented sampling for language model alignment and personalization. *arXiv preprint arXiv:2405.06639*, 2024.

[15] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in neural information processing systems*, 36, 2024.

[16] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

[17] Maxim Khanov, Jirayu Burapacheep, and Yixuan Li. Args: Alignment as reward-guided search. In *The Twelfth International Conference on Learning Representations*.

[18] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

[19] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.

[20] Bolian Li, Yifan Wang, Ananth Grama, and Ruqi Zhang. Cascade reward sampling for efficient decoding-time alignment. In *ICML 2024 Next Generation of AI Safety Workshop*.

[21] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

[22] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[23] Sidharth Mudgal, Jong Lee, Harish Ganapathy, YaGuang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, et al. Controlled decoding from language models. In *Forty-first International Conference on Machine Learning*.

[24] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[25] Ahmad Rashid, Ruotian Wu, Julia Grosse, Agustinus Kristiadi, and Pascal Poupart. A critical look at tokenwise reward-guided text generation. *arXiv preprint arXiv:2406.07780*, 2024.

[26] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018.

[27] Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. Mitigating object hallucination via data augmented contrastive tuning. *arXiv preprint arXiv:2405.18654*, 2024.

[28] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

[29] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

[30] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[31] David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. Contrastive region guidance: Improving grounding in vision-language models without training. *arXiv preprint arXiv:2403.02325*, 2024.

[32] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

[33] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023.

[34] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024.

[35] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024.

[36] Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *arXiv preprint arXiv:2402.14545*, 2024.

[37] Zhuosheng Zhang, Aston Zhang, Mu Li, George Karypis, Alex Smola, et al. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*.

[38] Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via classifier-free guidance. *arXiv preprint arXiv:2402.08680*, 2024.

[39] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

[40] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

[41] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*.

[42] Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. Calibrated self-rewarding vision language models. *arXiv preprint arXiv:2405.14622*, 2024.

A Related work

Guided decoding of LLMs. In the text-only setting, several works have explored guiding LLMs with a reward model to control output features such as helpfulness and harmlessness, and summary quality [9, 23, 11, 17, 14, 25, 20]. Unlike existing methods, we train a *multimodal* reward model to evaluate responses to *multimodal* instructions, which additionally contain images, and focus on evaluating this class of methods on visual grounding tasks.

Mitigating hallucinations of MLLMs. Prior work on mitigating visual hallucinations of MLLMs has focused on prompting [37], supervised fine-tuning [21], RLHF/RLAIF fine-tuning [29, 39, 34, 40, 35, 42, 2, 27], post-hoc rectification [41, 33], or specialized decoding strategies [38, 31, 13, 19, 10]. RGD is more powerful than purely feed-forward methods, as the principles learned during fine-tuning or specified in (system) prompts are not guaranteed to be respected at generation time, while RGD directly optimizes the output. In addition, RGD can be combined with prompting or fine-tuning, and readily applied to many MLLMs without retraining.