ATTENTION-AWARE DPO FOR REDUCING HALLUCINATIONS IN MULTI-IMAGE QA

Shaswat Patel, Harsh Sutaria, Jeet Patel, Vishvesh Trivedi
Department of Computer Science
New York University

ABSTRACT

Large Language Vision Models(LLVMs) have illustrated significant improvements across various multimodal tasks. To enhance the usability of LLVMs, preference alignment has become a standard technique. The Direct Preference Optimization (DPO) has emerged as a de-facto preference alignment algorithm which generally improves LLVMs single image performance, and LLVMs as a direct consequence struggle in multi-image scenarios. Previous studies have illustrated that LLVMs hallucinate when prompt contains containing multiple images and reference such as "In Image1". The misalignment can be improved using three broad solutions: (i) generating scalable dataset generation pipelines; (ii) improving alignment loss; and (iii) improving alignment at inference time. In this work, we present, Attention-aware Multi-Image Augmented Direct Preference Optimization, a preference alignment approach to handle multi-image inputs. Our improved alignment loss has shown promising results, with an improvement of 8.5% in terms of accuracy over the base model. Lastly, we also tackle improving alignment at inference time, by extending the previous studies on adaptive attention scaling at inference time to multi-image inputs and see an improvement of 10% over the base model. We make code publicly available for research purposes at https://github.com/harsh-sutariya/AA-DPO

1 INTRODUCTION

Large Vision-Language Models (LVLMs) have emerged as a powerful reasoners, trained on diverse web-scale data. LVLMs excel at single image queries, as both pretraining and post-training processes are primarily designed around single-image inputs. However, in real-world applications, multi-image understanding and reasoning is crucial to increase usability across industries. While prosperity models such as GPT-4o excel at multi-image queries, such performance gains are missing from open-sourced models.

Recent works have addressed the shortcoming of models for multi-image understanding by incorporating multi-image examples during pre-training (Su et al. (2023)), post-training (Liu et al. (2025)), and benchmarking dataset (Meng et al. (2025)). One of the drawbacks of the recent developments has been the loss performance on single-image tasks post multi-image finetuning or pre-training. Liu et al. (2025) proposed MIA-DPO, a highly automatic scalable data generation pipeline for Direct Preference Optimization(DPO) which has shown to improve the multi-image performance with little to no performance degradation on single-image benchmarks. The data generation pipeline uses existing single-image benchmarks by randomly sampling multiple images for a single query. Furthermore, for generating (chosen, rejected) pairs, model's attention ratio on target image was also used which acts as a soft loss during DPO training.

We propose a modification to the DPO loss by introducing an attention-based penalty that discourages the model from incorrectly attending to the irrelevant images. This modification is motivated by observations in Liu et al. (2025), which show that the model often allocates attention to irrelevant images, despite explicit references to image indices in the query. Our DPO loss modification is model agnostic and scalable to different multi-image QA settings.

Figure 1: **Example of Multi-Image hallucination**. When presented with a multi-image query, the LLaVA1.5-7B model allocates only 29.43% of its attention to the target image, resulting in a hallucinated response. After fine-tuning with our proposed Attention-aware DPO loss, the model correctly answers the query, increasing the attention ratio on the target image to 33.93%.

Finally, we extend AdaptVis (Chen et al. (2025))—which adaptively scales attention scores based on the model's confidence to the multi-image setting. A known limitation of AdaptVis is that if the model is confidently incorrect, scaling the attention scores can further degrade the quality of the generated output. By integrating our model, trained with the proposed modified loss, into the AdaptVis pipeline, we demonstrate that enhancing the base model's capability for multi-image reasoning can be further enhanced using AdaptVis pipeline. Figure 2 illustrates our proposed pipeline.

In summary, our key contributions are as follows:

1. We contribute to the understanding of various forms of multi-image hallucinations and propose using attention patterns as a signal for mitigating hallucinations.

2. We further extend the inference time optimizations proposed in AdaptViz to multi-image hallucinations setting.

2 Related Work

Hallucination in Large Vision-Language models is a significant issue. Strong language prior from the large-scale web data used for pre-training is one of the main reasons for this phenomenon (Min et al. (2025)). Work on mitigating hallucinations mainly focuses on post-training methods such as preference learning and inference-time strategies such as contrastive decoding. Below we briefly cover related works in both of these areas.

2.1 VISUAL PREFERENCE LEARNING

Visual preference learning involves aligning vision-language models with human preferences based on chosen-rejected data pairs. Specifically, the chosen sample reflect the ground truth while rejected samples contain hallucinated output. Training is carried out via policy learning methods like DPO (Rafailov et al. (2024)) and PPO (Schulman et al. (2017)). Yu et al. (2024a) introduce dense DPO to incorporate detailed feedback from human annotators, giving more weightage to correct samples in the loss function. Zhao et al. (2024) construct synthetic data pairs for the chosen-rejected dataset for use in DPO. Further, Wang et al. (2024a) optimize for image preference by proposing a multi-modal DPO objective to improve preference learning. Moreover, Yu et al. (2024b) incorporate AI feedback from open-source LVLMs to generate high-quality preference data and also introduce iterative feedback learning to limit distribution shift. Although a lot of methods have been employed for the single-image scenarios, none of the techniques involve using explicit training signal for the multi-image case.

2.2 SPECIAL DECODING METHODS

In this body of work, hallucination is tackled through specially adapted decoding techniques. Traditional inference-time decoding is vulnerable to producing hallucinatory output due to the inherent probabilistic nature. Contrastive decoding methods such as Leng et al. (2023), Wang et al. (2024b), Suo et al. (2025) aim to reduce the dependence on linguistic prior using disturbances as inputs. Further, as generation progresses LVLMs increasingly rely on language information while neglecting visual tokens Min et al. (2025). Chen et al. (2025) introduce AdaptVis - a technique which uses confidence score-based thresholding to adaptively scale attention onto image tokens in order to steer the LVLM response. Recently, Lyu et al. (2025) combine contrastive decoding and DPO training to effectively alleviate object hallucinations. However all these methods focus exclusively on single-image cases while we extend the idea to multi-image settings.

3 METHOD

3.1 ATTENTION AWARE DPO

Direct Preference Optimization (DPO) training uses pairs of model responses to adjust the model's probabilities in favor of the better answer. Formally, given a question content x, a chosen answer y^+ and a rejected answer y^-, the DPO objective Rafailov et al. (2023) maximizes the difference in log-likelihood between the two:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\log \sigma \left(\beta \left(\log \pi_\theta(y^+ \mid x) - \log \pi_\theta(y^- \mid x) \right) \right), \tag{1}$$

where σ is the sigmoid function and β is the scaling hyperparameter. Intuitively, this loss is low (good) when the model assigns higher probability to the chosen answer than to the rejected answer, and high when the model mistakenly prefers the wrong answer. DPO thus trains θ to align with the preference order without needing a separate reward model Liu et al. (2025).

In our attention aware DPO, we augment this loss with an additional term that penalizes mis-focused attention. We want the model to not only rank the correct answer higher, but alse to have attended to the correct image when producing that answer. We obtain the attention from the cross-attention weights of the decoder over image tokens, averaged across decoding steps for the answer Liu et al. (2025); Chen et al. (2025). The attention ratio is the proportion of attention allocated to the correct image. In a well grounded scenario, we expect the attention ratio to be high (the model looked at the right image for the correct answer) and low attention ratio signifies the model is hallucinating (the model was looking elsewhere, which is often why it went wrong). However standard DPO does not directly account for this internal behavior, it only considers the output probabilities. To inject attention alignment into training, we define an attention penalty term:

$$\mathcal{L}_{\text{attn}}(\theta) = \max \left\{ 0, \left(\delta - \frac{a_{corect\_image}}{a_{all\_images}} \right) \right\}, \tag{2}$$

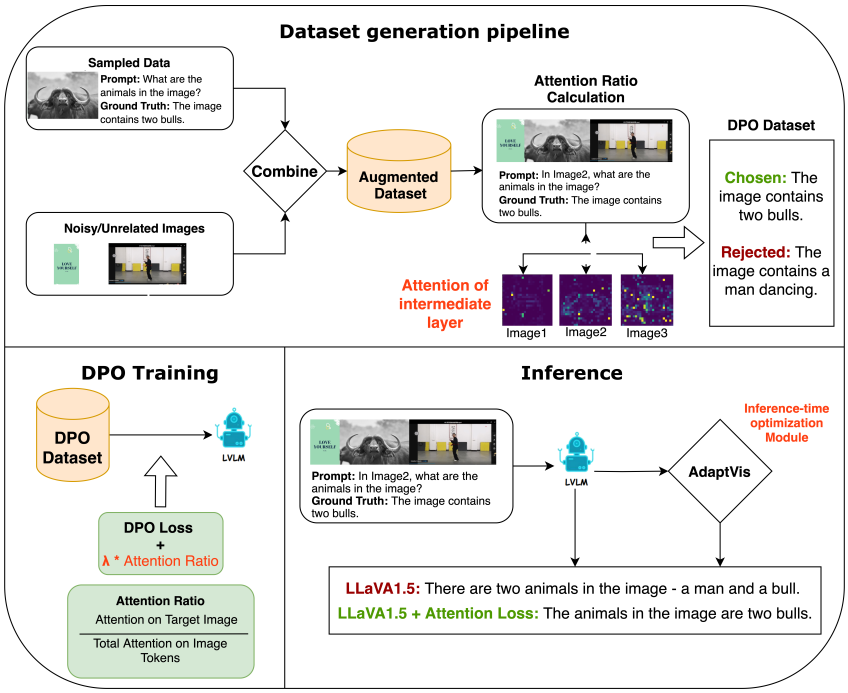

Figure 2: **Overview of our proposed method.** First, we augment single-image datasets such as LLaVA665k to multi-image setting. Using attention-aware sampling, we generate a (chosen, rejected) pair for DPO finetuning. We then finetune the model with our attention-aware DPO loss. Lastly, we couple the benefits of our finetuned model and AdaptVis for higher quality output generation.

where δ is a threshold hyperparameter (e.g. $\delta = 0.75$). This term is 0 when the attention ratio is greater than at least δ, which is a desirable situation. If, however, the chosen answer did not have sufficiently higher focus on the target image, then $\mathcal{L}_{\text{attn}}$ becomes positive, growing linearly with the excess of non-target attention. We found the hinge style formulation effective: it gently penalizes cases where the model isn't distinctly focusing on the right image. This encourages a wider separation in attention patterns Cortes & Vapnik (1995).

Our final loss combines the preference loss and the attention penalty:

$$\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{DPO}}(\theta) + \lambda\, \mathcal{L}_{\text{attn}}(\theta) \tag{3}$$

where λ controls the strength of the attention term. In practice, we tune λ to balance the two objective (alignment of output choice vs. alignment of attention). Our method is architecture agnostic: as long as the model yields attention weights, we can apply this loss. During training time, we utilize layers from 14 to 22 for calculation of attention ratio loss, as illustrated in Figure 3, these layers consistently allocate highest attention on the correct target image. Averaging the attention loss over these layers helps stabilize training, leading to more consistent optimization of the model.

3.2 Inference Time Optimization

Previous work (Azaria & Mitchell (2023), Kadavath et al. (2022), Chen et al. (2024)) has highlighted how the model's generation confidence itself can be used as an indicator of its output trustworthiness. Adapt-Vis (Chen et al. (2025)) borrows this concept to assign the threshold for scaling attention logits through which we can modify the attention distribution of the LVLM at inference time. Since image tokens receive a small fraction of the total attention as compared to text tokens even though they form a good chunk of the total tokens - we dynamically scale the raw attention logits to sharpen or smoothen the attention distribution. The decision boundaries for when to perform scaling is determined by a confidence threshold (β) - which is the output probability of the first token. If the

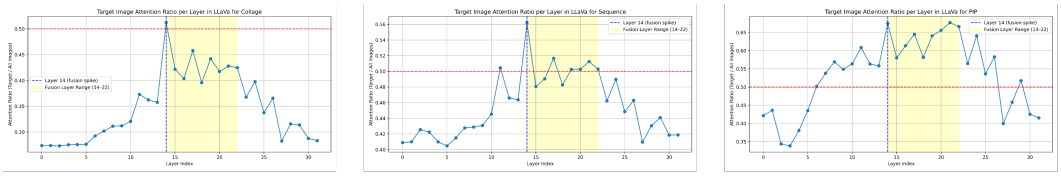

Figure 3: **Layer-wise analysis of attention ratio.** We perform a layer-wise analysis of attention ratio of LLaVA1.5 and observe that layers 14 to 22 consistently allocate highest attention to correct image target across all three input formats.

output probability is higher than a certain threshold - indicating that the model is confident in its response - then we sharpen the attention distribution by multiplying the raw logits with a constant factor ($\alpha > 1$). This in turn forces the model to pay more attention to the visual tokens, thus reinforcing correctness. On the other hand if the output probability is lower then the threshold then we smoothen the attention distribution by down-scaling the raw logits by a constant factor ($\alpha < 1$). This forces the model to not look too closely at a few areas and instead spread out its focus to the other areas where there could be possible clues.

More specifically, this scaling of the raw logits is done only for the attention of the final input token to the image tokens - allowing us in a way to steer the LVLM's generation. Mathematically, this intervention is expressed as:

$$\mathbf{A}_{n,j}^{(l,h)} = \begin{cases} \alpha_1 \mathbf{A}_{n,j}^{(l,h)} & \text{if } j \in \mathcal{I} \\ \mathbf{A}_{n,j}^{(l,h)} & \text{otherwise} \end{cases} \quad \text{if } \mathcal{C} < \beta \tag{4}$$

$$\mathbf{A}_{n,j}^{(l,h)} = \begin{cases} \alpha_2 \mathbf{A}_{n,j}^{(l,h)} & \text{if } j \in \mathcal{I} \\ \mathbf{A}_{n,j}^{(l,h)} & \text{otherwise} \end{cases} \quad \text{if } \mathcal{C} > \beta \tag{5}$$

3.3 Dataset

We evaluate the models using the PixMo datasetDeitke et al. (2024) specifically designed to simulate various multi image scenarios: Sequence (multiple unrelated images presented sequentially), Grid Collage (multiple images combined into a labeled composite), and Pic-in-Pic (one image overlaid onto another). Each format aims to provoke specific hallucination types, such as sequence confusion or element interference. The test set consists of 500 test questions per format which ensures diverse coverage of typical hallucination scenarios encountered in multi image question answering tasks. The number of images per query are also equally split between the range of 2 to 8.

4 Results

We present the experimental outcomes evaluating our proposed method Attention Aware DPO and it's inference time optimization variant (based on AdaptVis Chen et al. (2025)) against baseline models including the base LLaVA model Liu et al. (2024) and MIA-DPO aligned model Liu et al. (2025). We use the same hyper-parameter values for $\alpha_1 = 2.0, \alpha_2 = 0.5, \beta = 0.3$ as Chen et al. (2025) and the same learning rate, lora rank, and dropout from Liu et al. (2025). Further we use $\lambda = 0.9, \delta = 0.75$ and train using our loss for 1 Epoch on 4xRTX8000 for 72 hours.

We report comprehensive quantitative results across multiple multi image QA setups: Sequence, Grid Collage, and Pic-in-Pic, using PixMo evaluation dataset. Metrics include accuracy, relevancy, clarity, and completeness as judged by an LLM, and attention ratios indicating model focus correctness. We selected Gemini Team (2025) as our LLM-as-a-Judge model, owing to its strong performance across a wide range of multimodal benchmarks. We also provide the evaluation prompt in Figure A1. Lastly, we also provide qualitative results in Figure A2.

For evaluation we define the following metrics:

1. **Relevance**: Assesses whether the predicted answer directly addresses the question, taking into account the information provided in the accompanying caption. The model is asked to score between 1 to 5, 5 being the highest.
2. **Accuracy**: Measures the factual correctness of the prediction by comparing it to the ground truth. The answer should faithfully reflect the ground truth without introducing inaccuracies. The model is asked to score between 1 to 5, 5 being the highest.
3. **Clarity**: Evaluates the readability and coherence of the response. Predictions should be free from repetition, ambiguity, or logical inconsistencies that could impede understanding. The model is asked to score between 1 to 5, 5 being the highest.
4. **Completeness**: Determines whether the prediction fully captures the scope of the ground truth answer. The answer should not omit essential information and should include all necessary details to be considered comprehensive. The model is asked to score between 1 to 5, 5 being the highest.

Model	**Sequence**	**Collage**	**Pic-in-Pic**
LLaVA-1.5-7B	3.99	3.96	4.08
+ MIA-DPO	4.05	4.05	4.15
+ Ours	4.08	**4.07**	4.18
+ Ours $w/$ Adapt-Vis	**4.10**	4.06	**4.19**

Table 1: Combined rubric score (mean of relevance, accuracy, clarity, and completeness) across datasets. Highest per column is **bolded**.

Model	**Sequence**	**Collage**	**Pic-in-Pic**
LLaVA-1.5-7B	2.95	2.88	3.17
+ MIA-DPO	3.12	3.09	3.31
+ Ours	3.20	3.18	3.40
+ Ours $w/$ Adapt-Vis	**3.25**	**3.22**	**3.46**

Table 2: Average **accuracy** scores for each model across three datasets. Best model per dataset in **bold**.

Model	**Sequence**	**Collage**	**Pic-in-Pic**
LLaVA-1.5-7B	**0.3987**	**0.4009**	0.6157
+ MIA-DPO	0.3899	0.3921	0.6084
+ Ours	0.3927	0.4002	0.6179
+ Ours $w/$ Adapt-Vis	0.3933	0.4007	**0.6181**

Table 3: Average **attention ratio** for each model across datasets. Highest attention per dataset in **bold**.

Our Attention Aware DPO improves the multi image QA performance. Specifically, it achieved an average accuracy score of 3.20, as observed from Table 2, surpassing the base LLaVA model (2.95) Liu et al. (2024) and MIA-DPO (3.12) Liu et al. (2025). Importantly, our model exhibited higher relevancy scores, as observed from Table 1 which combines relevance, accuracy, clarity and completeness. Our method has outperformed the previous methods across all three input settings, this illustrates the effectiveness of using explicit attention penalty term, as it helps in discouraging attention drift to irrelevant images. This improved internal attention focus resulted in an average attention ratio of 0.393, compared to 0.389 for MIA-DPO (Table 3). The base model exhibits an average attention ratio of 0.399 on the target image. Our approach might surpass this performance, with increase in number of training epochs. Such enhancement demonstrates that incorporating attention signals directly into the training objective yields superior alignment and reduces hallucinations.

Incorporating our inference time attention adaption (based on AdaptVis Chen et al. (2025)) yielded additional performance improvements. When confidence was high, sharpening attention distributions helped the model consolidate correct visual references which boosted accuracy scored by an additional 0.05 points in ambiguous cases(Table 2). Conversely, broadening attention distributions in uncertain situations prevented premature fixation on incorrect visual cues, particularly beneficial in Grid Collage and Pic-in-Pic scenarios with improvements of 0.04 and 0.06 points respectively.

5 Conclusion and Future work

In this paper, we introduce a scalable pipeline to improve the alignment of Large Language and Vision Models (LLVMs) for multi-image queries. Our approach requires no human annotations and extends existing single-image datasets into multi-image formats. Quantitative results demonstrate that our method enhances the model's ability to accurately allocate attention to the correct target image.

Future work can explore applying this pipeline to more complex multi-image reasoning tasks, such as those found in the MANTIS datasetJiang et al. (2024), where understanding multiple images is necessary to answer a query. Additionally, the impact of various attention-based loss functions (e.g., log-loss) on model performance could be further investigated.

We also show that combining inference-time optimization techniques like AdaptVis with our fine-tuned model yields promising results. Future research may delve deeper into the comparative effectiveness of different inference-time strategies in mitigating hallucinations.

Finally, we observe that hallucinations are more frequent when the input includes visually similar images—particularly when multiple images contain the same object of interest. To facilitate further study, a benchmark dataset could be developed using our pipeline, leveraging CLIP embeddings to sample similar images instead of random sampling. This benchmark could feature varying levels of difficulty, directly correlated with the number of semantically related images in the query.

6 Challenges and Learnings

Throughout the course of this work, we encountered several challenges and gained valuable insights:

1. **Distributed training:** This was the first time we have used distrusted training using DeepSpeed. We got a chance to understand the different zero configurations, their impact on memory usage and performance.
2. **Extending DPOTrainer:** To incorporate attention loss, we modified the DPOTrainer by huggingface. This required a deep dive into the Hugging Face codebase, through which we significantly improved our familiarity with its internal workings. Although our initial implementation contained several bugs, careful debugging and iteration led to a functional and effective solution.
3. **AdaptVis:** The AdaptVis and MIA-DPO codebases were structurally quite different. Our first attempt to combine them involved extensive monkey patching, which introduced numerous integration issues. Eventually, by selectively modifying key components within MIA-DPO with AdaptVis code, we avoided unnecessary complexity and streamlined the integration process.
4. **Inference:** The inference module in the MIA-DPO codebase had multiple issues, requiring substantial debugging and refactoring. Extending it to support evaluation across our experimental settings was time-intensive.
5. **Computation graph:** Incorporating attention loss also taught us a valuable lesson of debugging. Always check the computation graph after modifying anything! Otherwise modifications can cause computation graph to break.

7 Individual Contributions

Below we list individual contributions:

1. **Literature survey and ideation**: Everyone contributed equally.
2. **Dataset generation**: Vishvesh Trivedi & Jeet Patel
3. **Integrating attention loss**: Shaswat Patel, Harsh Sutaria & Vishvesh Trivedi
4. **AdaptVis integration**: Shaswat Patel, Harsh Sutaria & Jeet Patel
5. **Evaluation pipeline**: Shaswat Patel, Jeet Patel & Harsh Sutaria
6. **Report writing**: Everyone contributed equally.

REFERENCES

Amos Azaria and Tom Mitchell. The internal state of an llm knows when it's lying, 2023. URL https://arxiv.org/abs/2304.13734.

Shiqi Chen, Miao Xiong, Junteng Liu, Zhengxuan Wu, Teng Xiao, Siyang Gao, and Junxian He. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7553–7567. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chen24av.html.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas, 2025. URL https://arxiv.org/abs/2503.01773.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. URL https://doi.org/10.1007/BF00994018.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. URL https://arxiv.org/abs/2409.17146.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning, 2024. URL https://arxiv.org/abs/2405.01483.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

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, 2023. URL https://arxiv.org/abs/2311.16922.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. URL https://arxiv.org/abs/2310.03744.

Ziyu Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Conghui He, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. MIA-DPO: Multi-image augmented direct preference optimization for large vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=f7WBRSuf9l.

Xinyu Lyu, Beitao Chen, Lianli Gao, Jingkuan Song, and Heng Tao Shen. Alleviating hallucinations in large vision-language models through hallucination-induced optimization, 2025. URL https://arxiv.org/abs/2405.15356.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2408.02718.

Kyungmin Min, Minbeom Kim, Kang il Lee, Dongryeol Lee, and Kyomin Jung. Mitigating hallucinations in large vision-language models via summary-guided decoding, 2025. URL https://arxiv.org/abs/2410.13321.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023. URL https://arxiv.org/abs/2305.18290.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Weijie Su, Xizhou Zhu, Chenxin Tao, Lewei Lu, Bin Li, Gao Huang, Yu Qiao, Xiaogang Wang, Jie Zhou, and Jifeng Dai. Towards all-in-one pre-training via maximizing multi-modal mutual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15888–15899, 2023. URL https://arxiv.org/abs/2211.09807.

Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. Octopus: Alleviating hallucination via dynamic contrastive decoding, 2025. URL https://arxiv.org/abs/2503.00361.

Gemini Team. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models, 2024a. URL https://arxiv.org/abs/2406.11839.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding, 2024b. URL https://arxiv.org/abs/2403.18715.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback, 2024a. URL https://arxiv.org/abs/2312.00849.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness, 2024b. URL https://arxiv.org/abs/2405.17220.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization, 2024. URL https://arxiv.org/abs/2311.16839.

.

A Appendix

LLM-as-a-Judge Prompt Format

Your task is to act as an impartial and objective assessor of answers generated by a Large Multimodal Model (LMM) for questions based on images. You will be given an image (optionally), a posed question, and the model's provided answer. Your evaluation should focus on the following aspects:
- **Relevance**: Does the predicted answer directly address the question posed, considering the information provided in the caption?
- **Accuracy**: Compare the predicted answer to the ground truth answer. Does the prediction accurately reflect the information given in the ground truth answer without introducing factual inaccuracies?
- **Clarity**: Assess the clarity of the predicted answer. Look for issues such as repetition, unclear descriptions, or any logical errors that could hinder understanding.
- **Completeness**: Determine if the predicted answer fully covers the scope of the ground truth answer. Does it leave out critical information or does it include all necessary details?

Input: # Question: {question} \n # Model Predicted Answer: {base_prediction} \n # **Output Format**: # ```python # {{ # "llm_judge_response": "<brief judgement of prediction>", # "model": {{ # "relevance": <score between 1 to 5>, # "accuracy": <score between 1 to 5>, # "clarity": <score between 1 to 5>, # "completeness": <score between 1 to 5>, # "hallucination": <score between 1 to 5> # }} # }} # ``` # Do not provide any conversation and make sure you adhere to the output format. Ensure your response is valid JSON inside the python code block. # '"

Figure A1: **LLM-as-a-Judge Evaluation Prompt**

Image1

Image2

Prompt: In Image2, How many fingernails can we see?

LLaVA: In Image2, we can see two fingernails.

MIA-DPO: In Image2, we can see two fingernails.

Our Model: In the image, we can see three fingernails.

Image1

Image2

Prompt: In Image2, what are the uniform numbers of the two nearest defending players?

LLaVA: In Image2, the uniform numbers of the two nearest defending players are 11 and 10.

MIA-DPO: In Image2, the uniform numbers of the two nearest defending players are 11 and 17.

Our Model: In the image, the two nearest defending players are wearing uniform numbers 11 and 14.

Figure A2: **Qualitative Comparison** between various model outputs