The Effect of Data Corruption on Multimodal Long Form Responses

Daniel Z Kaplan [*1,2] **Alexis Roger** [*1,3,4] **Mohamed Osman** [*5] **Irina Rish** [1,3,4]

Abstract

Despite significant progress, Vision-Language Models (VLMs) still struggle with hallucinations, especially in long-form responses. Existing strategies have had limited successes in specific cases, and long-form generation remains problematic. In this work we attempt to establish the link between the data used to train the model and the hallucinations in the model's output. To this end, we examine hallucinations through data corruption. We develop a method to corrupt training data and then train models with this data to see the effect on performance. We will show that corrupting only a small portion of the long-form training data significantly impairs the performance of the model on long-form tasks, while leaving simpler tasks like visual question-answering and multiple choice relatively intact. All training code and models are released for reproducibility and future research.

1. Introduction

Foundation models (FMs), including Vision-Language Models (VLMs), have revolutionized the field of artificial intelligence by enabling advanced applications such as image captioning, visual question answering, and cross-modal retrieval (Radford et al., 2021; Alayrac et al., 2022). These models hold immense potential for real-world applications, from enhancing accessibility in education to supporting critical decision-making in healthcare and finance. However, their deployment in practical, in-the-wild, scenarios presents significant challenges, particularly concerning their reliability and ethical implications.

One of the most pressing issues in the deployment of VLMs is hallucinations — instances where the model generates outputs that are factually incorrect or inconsistent with the given visual input (Rohrbach et al., 2019). This problem is especially pronounced in long-form responses and detailed image descriptions, which are critical for applications in domains like clinical health and education as well as for downstream machine learning applications (Betker et al., 2023; Chen et al., 2023; Hammoud et al., 2024). Ensuring the factual accuracy and coherence of VLM outputs is essential for their responsible and effective use.

It is well known that for pure language models, the likelihood of hallucinations increases as the length of the language model's output grows (Holtzman et al., 2020). Similarly, in VLMs, longer sequences provide more opportunities for the model to deviate from the visual grounding and generate text based on spurious correlations or statistical patterns in the training data. (Zhou et al., 2024)

In this paper, we aim to explore the link between the quality of the training data and the prevalence of hallucinations in VLM outputs. We propose a method to systematically corrupt training data and train models on this corrupted data to study its impact on performance.

Our contributions include the development of a data corruption pipeline, the creation of corrupted datasets, and the training of models on varying levels of data corruption. These resources are made publicly available to support reproducibility and further research. By addressing the issue of hallucinations through the lens of data quality, our work contributes to the broader goal of deploying foundation models in the wild with higher reliability, while reducing concerns surrounding their use.

2. Related Work

Different strategies have been brought up to try and reduce hallucinations, such as rewriting-based approaches, guidance-based approaches and data based approaches. We decide to focus on this last category of data-driven strategies. These focus on improving the quality and robustness of the model at training. Training based methods typically change something about the actual process in an online manner. (Liu et al., 2024) developed LRV-Instruction and GAVIE for robust visual instruction tuning to reduce hallucinations. (Yu et al., 2024) proposed HalluciDoctor, which uses a consistency cross-checking paradigm to detect and

[*]Equal contribution [1]CERC-AAI [2]Realiz.ai [3]Mila – Quebec AI Institute [4]Université de Montréal [5]Virginia Commonwealth University. Correspondence to: Daniel Z Kaplan <daniel.z.kaplan@realiz.ai>.

Published at ICML 2024 Workshop on Foundation Models in the Wild.

eliminate hallucinations in visual instruction data. (Zhao et al., 2024) proposed Hallucination-Aware Direct Preference Optimization (HA-DPO) to mitigate hallucinations during inference. (Kim et al., 2023) creates synthetic data aimed to remove spurious object correlations. (Ben-Kish et al., 2024) introduced MOCHa, a framework for optimizing image captioning models to reduce open-vocabulary hallucinations while preserving caption quality. However, no comprehensive study has been conducted on how the data quality, across a spread of qualities, affects the downstream model's performance.

3. Corrupting Data

To conduct our experiments on long-form responses, we required a dataset that contained both reliable prompts and systematically corrupted prompts. However, we were unable to locate any existing datasets that comprised prompts alongside a corrupted or altered version. Consequently, we decided to create our own dataset. We based our dataset on the Detailed Caption dataset (Li et al., 2024), which consists of 213k long captions of images. These captions served as our ground truth for the experiments.

The procedure to corrupt the prompts involved iterating over the list of prompts and selectively masking words of each prompts. Each word is masked with a probability of 0.5, which consists of replacing the word with "[MASK]". After masking a prompt, we utilized the Gemini LLM (Team et al., 2023) to repair it. We prompted the model with the instruction: "Replace the [MASK] with a suitable word". No information about the image was passed to the Gemini LLM.

An example provided in Appendix A illustrates how this method introduces hallucinations, as the most likely word often replaces the original word. For instance, in the example, the laptop in the image was replaced with a book by the Gemini LLM, and a "black dress" was erroneously changed to "white dress". However, semantic integrity was well preserved by the Gemini LLM.

For training, we would start with the original list of uncorrupted prompts and select an amount n of prompts that we wished to have corrupted. We would then replace the n first prompts with their corrupted version. The prompts were then scrambled before beginning training.

4. Training Corrupted Models Efficiently

Our model training protocol was guided by methodologies outlined in the Robin paper (Kaplan et al., 2023), while the data utilized originated from the Monkey paper (Li et al., 2024). To this end, the VLM itself is built using the OpenHermes instruction tuning of the Mistral 7B model (Nous Research, 2023; Jiang et al., 2023) combined with the ViT SigLip vision encoder (Zhai et al., 2023) with the LLaVA architecture (Liu et al., 2023). This was chosen as it had shown the most interesting results in the Robin paper (Kaplan et al., 2023). The composition and breakdown of the dataset used for training is shown in Table 3. The specificity in our training is the execution of two distinct steps: VLM pretraining and VLM finetuning. For reproducibility, all hyperparameters used are detailed in Table 4.

In the pretraining phase, the model was exposed to the comprehensive dataset shown in Table 3, excluding the Detailed Caption dataset. This phase encompassed a total of 1.23 million samples covering many tasks such as short form image captioning, general Visual Question Answering (VQA), scientific VQA and document-oriented VQA. The pretraining process was resource-intensive, requiring 8 hours of compute time on 8 NVIDIA H100 GPUs. This step ensured that the model developed a robust foundational understanding before being exposed to the specific nuances of the Detailed Caption data.

The subsequent finetuning phase focused exclusively on the Detailed Caption dataset, which had been systematically corrupted using the method described in Section 3. This finetuning allowed us to efficiently and economically train models across varying levels of corruption. The compute cost for this phase was significantly lower, demanding only 1 hour on the same hardware setup of 8 NVIDIA H100 GPUs. This two-step approach facilitated rapid experimentation and enabled us to train a wide range of corrupted models.

5. Results

We present the results of our experiments across three key areas: General Visual Question Answering (VQA), Scene Text-centric VQA, and AI assisted evaluations.

General VQA tasks necessitate the model ability to understand and integrate visual and textual information. This involves a comprehensive grasp of how these modalities interrelate. We validate our model using four benchmarks: ScienceQA (Lu et al., 2022), GQA (Hudson & Manning, 2019), VQAv2 (Goyal et al., 2017), and POPE (Li et al., 2023). These benchmarks provide a broad assessment of the model capabilities in general visual question answering scenarios.

Scene Text-centric VQA tasks consist of text within images, as it is prevalent in real-world environments. This highlights the ability to address questions about such text, which is a critical component of VQA tasks. For evaluating our model's performance in this area, we utilized the TextVQA (Singh et al., 2019) benchmark. This dataset specifically test the model's proficiency in interpreting and responding to queries involving text found within images.

Table 1. Performance of the VLMs on standard benchmarks. The model is pretained on all the data *except* for the 213k samples from the Detailed Caption, and is then finetuned only on these Detailed Captions with varying level of corruption. The mean and standard deviation are calculated only on the finetuned models.

Corruption	SQA Text	SQA Image	GQA	VQA v2	POPE	TextVQA	MM-VET	LLaVA Bench
pretrained	82.55%	**82.35%**	**52.81%**	**70.19%**	77.10%	**47.58%**	21.9%	27.5%
0	82.53%	82.00%	51.95%	68.93%	**78.57%**	46.75%	26.3%	**46.3%**
10k	82.55%	82.15%	51.55%	68.97%	77.57%	46.84%	**26.9%**	46.2%
20k	82.36%	82.05%	52.00%	68.87%	77.57%	46.58%	25.4%	41.5%
30k	82.60%	82.05%	51.87%	68.87%	77.70%	46.24%	25.7%	39.9%
40k	82.50%	82.05%	51.96%	68.81%	77.87%	46.21%	25.4%	38.6%
50k	82.39%	81.76%	51.85%	68.79%	77.20%	46.33%	24.4%	38.1%
70k	82.46%	81.76%	51.92%	68.68%	77.27%	46.18%	24.1%	36.8%
100k	82.60%	81.80%	51.73%	68.55%	76.47%	45.85%	23.7%	37.3%
150k	82.65%	82.00%	51.55%	68.12%	76.20%	44.89%	25.4%	35.1%
213k	**82.79%**	81.85%	51.33%	67.09%	75.70%	44.67%	24.1%	37.5%
Mean	82.54%	81.95%	51.77%	68.57%	77.21%	46.05%	25.14%	39.73%
STD	0.13%	0.14%	0.22%	0.58%	0.86%	0.73%	1.04%	3.84%

AI-based Evaluations are used to further gauge our model's performance on long-form prompts. To this end, we employed the MM-Vet (Yu et al., 2023) and LLaVA-Bench (Liu et al., 2023) evaluation frameworks. These comprehensive evaluations provide additional insights into the model's effectiveness in handling detailed and complex textual queries.

We then proceeded to evaluate both the pretrained model and all of the finetuned models on these benchmarks. The complete table of results can be seen in Table 1. The first result that we notice is that all our automated benchmarks, both general and scene text-centric VQA, have a very tight grouping of results, with the Detailed Caption finetuning seeming to have very little effect on the model's performance, regardless of the amount of prompt corruption that it present. This is quantified with the small standard deviation in these results.

These results can be explained quite simply by the tasks performed in these benchmarks. ScienceQA Text represents questions asked only on a textual input, with no image, so it is obvious why it is the least affected by our corruption attempts on captioning prompts. Following this trend, we note that ScienceQA Image also has a very low variation. As this is a multiple choice benchmark, this stops the model from improvising and having too many hallucinations. GQA, VQA and TextVQA are all performed with the instruction "Answer the question using a single word or phrase." which again contrains the model. We do see that the standard deviation creeps up on TextVQA, as the prompts tend to be more image focused by the scene-centric focus of the benchmark. This requires a better understanding of the image, which may have been impeded in training runs with corrupted data. POPE shows the most variance out of the automated benchmarks. This is because the POPE benchmark is a series of yes/no questions, regarding specific objects and if they are present or not in the image. How corruption can damage this is made clear in Appendix A, where a computer became a book. However, this effect remains rather weak as many prompts would have to be corrupted in a similar manner to make a meaningful impact.

Figure 2 illustrates the above results by showing the gentle decline of the automated metrics as the training data becomes more corrupted. We can also clearly distinguish the TextVQA and POPE benchmarks which degrade faster than the other benchmarks. However, this loss in performance remains rather small compared to the AI benchmarks. Both Table 1 and Figure 1 also show that there is very little change in the model's performance over the automated benchmarks, whether it has followed long form finetuning or not. In fact, the model may have become more verbose during this finetuning steps, as even with uncorrupted data the finetuned model underperforms, compared to the pretrained one. A notable exception to this trend is on the POPE benchmark, where the model trained on the uncorrupted data outper-

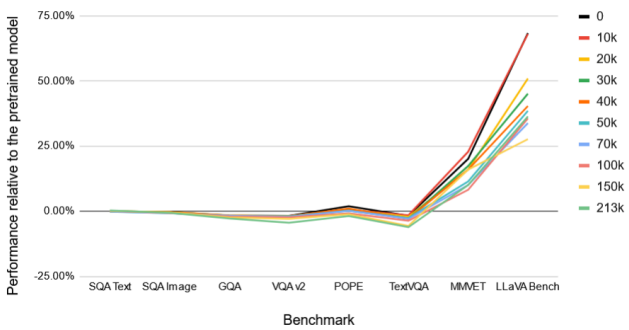

Figure 1. Graph showing the relative gain in performance by finetuning on long form data at different levels of corruption.

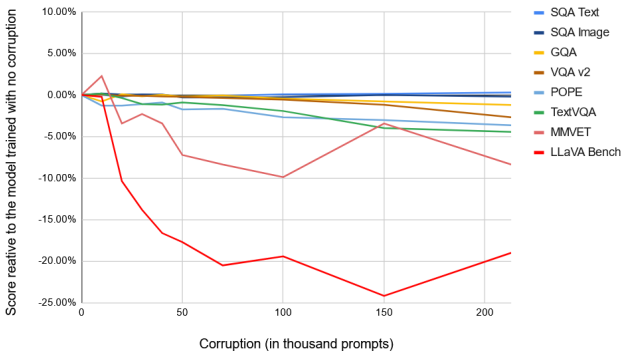

Figure 2. Graph showing the relative loss in performance when using corrupted data, compared with the model finetuned on uncorrupted data.

forms the pretrained model. The pretrained model performs as well as a model with 80 thousand samples, about one third of the total samples, being corrupted. This is most likely due to the the long form responses helping the model to better identify objects in the image.

With regards to the AI-based benchmarks, we see much more interesting results. Both the MM-Vet and the LLaVA-Bench benchmarks evaluate long form responses with the help of LLM models. Therefore this is where we expected to see the biggest effect of our data corruption. Indeed, we can see in Figure 2 that corrupting the data leads to a significant loss in performance, which scales with the data corruption. In fact, the performance of the models on the LLaVA-Bench evaluation seems to follow an decreasing exponential trend, dropping sharply at the beginning and then stabilizing at -20% accuracy when compared to the uncorrupted model. As we see in Figure 1, the performance of the uncorrupted model and the model with 10 thousand prompts corrupted is near identical, but the model with 20 thousand prompts corrupted already experienced a sharp drop in performance. This shows that the models are robust to finetuning on corrupted data, as long as that data represents less than 5% of the total finetuning data. However, anything larger than that will damage the model significantly.

If we compare this to the results on the MM-Vet evaluation, we also see a similar trend, however the results are not as bad. This highlights how not all AI-based benchmarks are equivalent. MM-Vet also shows a stronger drop in performance as the finetuning data is more corrupted (Figure 2), but does not show as strong of a decline as LLaVA-Bench. We notice that the model trained with 150 thousand corrupted samples gives particularly interesting results, performing worse on LLaVA-Bench but better on MM-Vet compare to the models trained with 100 and 213 thousand corrupted samples. We are currently conducting additional work to investigate this.

6. Human evaluations

In order to validate the degradation of the model seen in the previous benchmarks, we conducted some human evaluations, comparing the model trained on the original Detailed Caption data and the model trained on the most corrupted Detailed Caption data (213k). A simple interface with an image and a description by each of the 2 models was presented to individuals who were asked to vote for the one they preferred. This interface is shown in Appendix C.

As we can see in Table 2, both models are found to provide a comparable level of detail when describing the image. However, the model trained on the corrupted data performs a lot worse in the hallucination category, confirming that our data corruption technique is successful in significantly degrading the end models performance. We see that this increase in hallucinations directly leads to the model being less popular, validating the use of long-form benchmarks like LLaVA-Bench and the importance of improving reliability by reducing hallucinations in VLMs for real-world model usage.

Table 2. Percentage of votes per model for each category.

Category	Clean Model	Corrupted Model
Description detail	49%	51%
Hallucination accuracy	62%	38%
Overall preference	59%	41%

7. Conclusion

Our study demonstrates a clear link between training data quality and the prevalence of hallucinations in VLMs. By systematically corrupting training data and analyzing its impact, we have shown that models are robust to minor imperfections in the data, however the corruption threshold of what can significantly impair the performance of VLMs on long-form tasks remains low. Nevertheless, simpler tasks remain relatively unaffected. These findings underscore the critical importance of high-quality training data for the reliable and ethical training and deployment of VLMs in real-world applications.

The implications of our findings are significant for ensuring the reliability and responsibility of VLMs, particularly in domains requiring detailed and accurate outputs. Future work will focus on developing more sophisticated data corruption techniques, methods for detecting and mitigating hallucinations, and exploring frameworks for deploying VLMs in sensitive applications. By addressing these challenges, we can enhance the adaptability, robustness, and efficiency of foundation models, paving the way for their successful integration into various real-world scenarios.

8. Acknowledgements

We acknowledge the support from the Canada CIFAR AI Chair Program and from the Canada Excellence Research Chairs (CERC) Program. This project used compute resources provided by the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. High Performance Computing resources provided by the High Performance Research Computing (HPRC) core facility at Virginia Commonwealth University (https://hprc.vcu.edu) were also used for conducting the research reported in this work.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a Visual Language Model for Few-Shot Learning, 2022.

Ben-Kish, A., Yanuka, M., Alper, M., Giryes, R., and Averbuch-Elor, H. Mitigating Open-Vocabulary Caption Hallucinations, 2024.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3): 8, 2023.

Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., et al. PixArt-Alpha: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

Chen, W., Wang, H., Chen, J., Zhang, Y., Wang, H., Li, S., Zhou, X., and Wang, W. Y. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and 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.

Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., and 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.

Hammoud, H. A. A. K., Itani, H., Pizzati, F., Torr, P., Bibi, A., and Ghanem, B. SynthCLIP: Are We Ready for a Fully Synthetic CLIP Training? *arXiv preprint arXiv:2402.01832*, 2024.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. The Curious Case of Neural Text Degeneration, 2020.

Hudson, D. A. and Manning, C. D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Kaplan, D. Z., Gupta, K., Ramstedt, S., Roger, A., Fennell, E., Adamopoulos, G., Anthony, Q., Qi, S., Williams, A. R., Humane, P., Bhagwatkar, R., Lu, Y., and Rish, I. Robin - Visual Language Models. https://github.com/AGI-Collective/Robin/releases/tag/v1.0.0, 2023.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., and Farhadi, A. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 235–251. Springer, 2016.

Kim, J. M., Koepke, A. S., Schmid, C., and Akata, Z. Exposing and Mitigating Spurious Correlations for Cross-Modal Retrieval, 2023.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating Object Hallucination in Large Vision-Language Models, 2023.

Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., and Bai, X. Monkey: Image Resolution and Text Label Are Important Things for Large Multi-modal Models, 2024.

Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning, 2024.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. 2023.

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

Marino, K., Rastegari, M., Farhadi, A., and Mottaghi, R. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Masry, A., Long, D. X., Tan, J. Q., Joty, S., and Hoque, E. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

Mathew, M., Karatzas, D., and Jawahar, C. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., and Jawahar, C. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022.

Mishra, A., Shekhar, S., Singh, A. K., and Chakraborty, A. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pp. 947–952. IEEE, 2019.

Nous Research. A finetuning of Mistral 7B with the OpenHermes 2.5 dataset, 2023. URL https://huggingface.co/teknium/OpenHermes-2.5-Mistral-7B.

Pasupat, P. and Liang, P. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., and Saenko, K. Object Hallucination in Image Captioning, 2019.

Sidorov, O., Hu, R., Rohrbach, M., and Singh, A. TextCaps: a Dataset for Image Captioning with Reading Comprehension, 2020.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Stanisławek, T., Graliński, F., Wróblewska, A., Lipiński, D., Kaliska, A., Rosalska, P., Topolski, B., and Biecek, P. Kleister: key information extraction datasets involving long documents with complex layouts. In *International Conference on Document Analysis and Recognition*, pp. 564–579. Springer, 2021.

Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., Gao, H., Liu, J., Huang, T., and Wang, X. Emu: Generative Pretraining in Multimodality, 2024.

Svetlichnaya, S. DeepForm: Understand structured documents at scale, 2020.

Tanaka, R., Nishida, K., and Yoshida, S. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 13878–13888, 2021.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Yu, Q., Li, J., Wei, L., Pang, L., Ye, W., Qin, B., Tang, S., Tian, Q., and Zhuang, Y. HalluciDoctor: Mitigating Hallucinatory Toxicity in Visual Instruction Data, 2024.

Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*, 2023.

Zhao, Z., Wang, B., Ouyang, L., Dong, X., Wang, J., and He, C. Beyond Hallucinations: Enhancing LVLMs through Hallucination-Aware Direct Preference Optimization, 2024.

Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., and Yao, H. Analyzing and mitigating object hallucination in large vision-language models, 2024. URL https://arxiv.org/abs/2310.00754.

A. Example of the Prompt Corruption

Bellow is an example of a prompt corrupted with the method we present in Section 3. This prompt comes from the Detailed Caption dataset (Li et al., 2024) and describes the image shown in Figure 3. Major differences which lead to hallucinations have been highlighted.

Original Prompt

In this image, we see a beautiful blonde woman sitting among tall and ripe wheat ears. She is using a **laptop which rests on her lap**. Her **hand is resting on the keyboard** as she works. The woman wears a **black headband** on her head, which **matches her hair color**. The laptop screen illuminates her face as she sits in the tall grass, surrounded by the golden wheat fields.

Masked Prompt

[MASK] [MASK] [MASK] we [MASK] [MASK] [MASK] [MASK] woman sitting among tall [MASK] ripe wheat [MASK] [MASK] is [MASK] a [MASK] which [MASK] [MASK] her lap. [MASK] [MASK] is resting [MASK] [MASK] keyboard as she works. [MASK] woman wears a [MASK] headband [MASK] [MASK] head, [MASK] matches her [MASK] [MASK] [MASK] [MASK] screen illuminates her [MASK] as [MASK] sits in the [MASK] [MASK] surrounded by the [MASK] wheat [MASK]

Corrupted Prompt

Gently swaying in the breeze, we see a young woman sitting among tall stalks of ripe wheat. In her lap is a **book which she reads intently**, her eyes scanning the pages. She is **resting her feet on a keyboard** as she works. The woman wears a **white headband** on her head, which **matches her simple white dress**. The screen illuminates her face as she sits in the open field surrounded by the golden wheat stalks.

Figure 3. Image GCC_train_000583645.jpg from the Detailed Caption dataset (Li et al., 2024).

B. Training details

Tables showing training parameters needed for reproducibility studies.

Table 3. Details on the Monkey training data used, from the Monkey paper (Li et al., 2024).

Dataset	Samples
Detailed Caption (Li et al., 2024)	213k
COCO Caption (Karpathy & Fei-Fei, 2015)	82k
TextCaps (Sidorov et al., 2020)	109k
VQAV2 (Goyal et al., 2017)	100k
OKVQA (Marino et al., 2019)	18k
GQA (Hudson & Manning, 2019)	150k
ScienceQA (Lu et al., 2022)	18k
VizWiz (Gurari et al., 2018)	20k
TextVQA (Singh et al., 2019)	34k
OCRVQA (Mishra et al., 2019)	250k
AI2D (Kembhavi et al., 2016)	24k
DocVQA (Mathew et al., 2021)	118k
ChartQA (Masry et al., 2022)	84k
InfoVQA (Mathew et al., 2022)	47k
DeepForm (Svetlichnaya, 2020)	7k
KLC (Stanisławek et al., 2021)	27k
WTQ (Pasupat & Liang, 2015)	28k
TabFact (Chen et al., 2019)	91k
VisualMRC (Tanaka et al., 2021)	21k
Total	1.44m

Table 4. Hyperparameters used for model training.

Parameter	**Value**
Vision encoder learning rate	$5 \cdot 10^{-5}$
Language model learning rate	$2 \cdot 10^{-5}$
Projection learning rate	$2 \cdot 10^{-5}$
Use of fp16	True
Projection type	mlp2x_gelu
Weight decay	0
Warmup ratio	0.03
Amount of epochs	1
Batch size	128
LoRA r	128
LoRA α	256

C. Human Evaluation Example

The prompt used to generate each model's description of the image was "generate the detailed caption in English". The images used are a hand-picked selection from the Emu paper (Sun et al., 2024).

Model A is the model trained on the clean long form data from the Detailed Caption dataset and Model B is the model trained on the corrupted dataset.

Table 5. Example of the human evaluation interface

Model A	**Model B**
In this image, we see a brown dog lying on a bed with a brown blanket. The dog has a black nose and is looking directly at the camera. There is also a black and white photo of a cross in the background. The dog seems to be enjoying its time on the bed, which is located in what appears to be a bedroom.	In this image, we see a brown dog lying on a bed with a white blanket. The dog is looking at the camera with its tongue hanging out. There is also a black cat sitting on the bed next to the dog. The cat is looking at the camera with its eyes closed. The bed is made up with a white sheet and a brown blanket. There is also a brown pillow on the bed. The room is decorated with a brown rug and a brown curtain. The window is open and there is a view of the outside.

Which model payed more attention to detail?

Model A | Both | Model B

Which description did you prefer overall?

Model A | Both | Model B

Which model was more accurate with regards to hallucinations?

Model A | Both | Model B