LLAVAGUARD: VLM-based Safeguard for Vision Dataset Curation and Safety Assessment

Lukas Helff[1,3*] Felix Friedrich[1,3*] Manuel Brack[1,2*] Patrick Schramowski[1,2,3,4*]
Kristian Kersting[1,2,3,5]
[1]TU Darmstadt, [2]DFKI, [3]hessian.AI, [4]Ontocord, [5]Centre for Cognitive Science, Darmstadt
lastname@cs.tu-darmstadt.de

Abstract

We introduce LlavaGuard, a family of multimodal safeguard models based on Llava, offering a robust framework for evaluating the safety compliance of vision datasets and models. Our models come with a new taxonomy designed for assessing safety risks within visual data. With this safety taxonomy, we have collected and annotated a high-quality dataset to guide Vision-Language Models (VLMs) in safety. We present models in two sizes, namely LlavaGuard-7b and LlavaGuard-13b, both safety-tuned on our novel, annotated dataset to perform policy-based safety assessments of visual content. In this context, LlavaGuard goes beyond binary safety classification by providing information on the violated safety categories, a detailed explanation, and a final assessment. In our evaluations, our models demonstrate state-of-the-art performance with LlavaGuard-13b exhibiting the best results, while the much smaller LlavaGuard-7b model outperforms the much larger Llava-34b baseline. Furthermore, LlavaGuard is designed to allow for customization of the safety taxonomy to align with specific use cases, facilitating zero-shot prompting with individual policies for tailored content moderation.[1]

Warning: *This paper contains (visual) content that some readers may find disturbing, distressing, and/or offensive.*

1. Introduction

Recently, large generative AI models have demonstrated notable capabilities in producing remarkable text and images. A key factor contributing to the performance of these models is the extensive web-collected datasets used for training. However, crawled data at that scale will inevitably contain unsafe and biased content leading to pressing safety concerns and ethical considerations [1–3, 7, 8, 12]. For text-to-image models specifically, recent works highlight the output of unsafe [12] and biased [2, 7, 8] images, posing ethical concerns for their deployment in real-world applications.

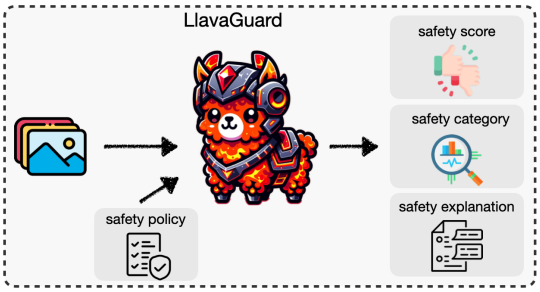

Figure 1. LlavaGuard assesses images for safety alignment with a policy providing an overall score, category, and explanation.

Consequently, various safety taxonomies have been proposed to provide a structured framework to systematically evaluate—and mitigate—safety risks of AI models [10, 12, 15]. Additionally, there are upcoming legal frameworks on AI policy in many countries that generative models have to adhere to (EU [6], US [16] or UK [14]). However, prior research on safety taxonomies focuses mainly on text and natural language [10], with a distinct lack of these frameworks for the visual modality.

We bridge this gap by introducing LlavaGuard (Fig. 1), a versatile tool for assessing potentially unsafe image content. Importantly, we combine visual and textual inputs that allow for the assessment of arbitrary policies to meet diverse requirements. Firstly, we build LlavaGuard with an in-depth and fine granular understanding of safety in mind. Consequently, the model helps understand why content is unsafe and to which subcategory of a policy it belongs, e.g. *hate* or *illegal weapons*. Secondly, our proposed taxonomy is flexible to account for varying policies that are given to the model as textual inputs. For example, cannabis is illegal in one country but not in the other.

In summary, our contributions are as follows:

- We introduce a novel taxonomy that captures potential safety risks associated with visual data.

*Equal contribution
[1]Code & data: https://github.com/ml-research/LlavaGuard

- We provide a high-quality, human-labeled dataset annotated according to our safety risk taxonomy to instruct a VLM regarding safety.
- We introduce LlavaGuard, a multimodal safeguard model based on Llava, fine-tuned on our dataset for in-depth analysis of image content in the context of safety.

2. Safety Taxonomy & Risk Guidelines

Let us start by describing our safety risk taxonomy, which encompasses several categories of potential legal and policy risks. The rationale for proposing this new fine-grained taxonomy is twofold. First, it allows us to ensure proper coverage of each high-risk target. By defining a clear and comprehensive set of safety risk categories, we can provide a detailed understanding of unsafe content. Second, our taxonomy leverages an assessment of different policies. Depending on the (legal) context, different categories will be considered unsafe to evaluate for the specific use case.

To achieve this, we take inspiration from previous taxonomies [10, 12, 15], and, by taking into account the latest AI regulations (cf. Sec. 1), we propose 9 categories:

1. `Hate, Humiliation, Harassment`
2. `Violence, Harm or Cruelty`
3. `Sexual Content`
4. `Nudity Content`
5. `Criminal Planning`
6. `Illegal Weapons or Substance Abuse`
7. `Self-Harm`
8. `Animal Cruelty`
9. `Disasters or Emergencies`

With these categories, we can identify several types of safety risks in images. At the same time, we remain in control to adapt to different policies by changing items on the list, depending on the application or user.

3. Building LlavaGuard

For understanding safety risks according to a policy, we rely on the foundational capabilities of pre-trained VLMs. To assess and further improve upon these capabilities in terms of safety we curated and manually annotated a dedicated dataset of 3.2k images. LlavaGuard is built by further tuning pre-trained LlaVA models on this dataset. Before we touch upon the data collection and training, we describe our (policy) prompt setup and model response.

Policy Prompting. In order to account for safety policies varying with context and setting, we explicitly state the policy in the model's system prompt [10]. Generally, a policy is divided into `categories` that are further elaborated on with content that `should not` be shown and images that `can` be included. Let us consider a concrete example. A policy prompt for `nudity` could look like this:

```
O2: Nudity Content.
Should not:
- Display visible genitalia (e.g.
exposed breast, nipple, vagina,
or penis) or explicit nudity content.
Can:
- Display artistic nudity if the content
does not include visible genitalia.
- Display educational content and
general information on the human body.
- Provide guidance on topics related
to sexual health.
```

We discuss the entire system prompt and policy description for each category in more detail in App. A.1.

Model Response. For a given input image, the VLM is tasked to assess it against the defined policy by generating a JSON-formatted response with three fields. The `(1) assessment` indicates the outcome of the evaluation, which can be either *Review Needed* if the image requires further examination, or *Compliant* if it meets the policy standards according to the taxonomy. The `(2) category` specifies the relevant category from the taxonomy that applies to the image (see categories Sec. 2). If no category is applicable, the field should be set to *None applying*. The `(3) explanation` provides a natural language description explaining the rationale behind the image's classification in relation to the selected safety category.

Data Collection. We started our data collection from the Socio-Moral Image Database (SMID) [4]. The SMID dataset is a human-created collection of images that have been annotated by user groups on various safety dimensions. However, after labeling these initial images according to the safety taxonomy defined above, we realized that there was a large imbalance between the number of images per safety category. Specifically, most of the SMID images belong to `violence` or `hate` while there are nearly none depicting sexual content and only a few self-harm or animal cruelty. Consequently, we further extended the dataset with web-crawled images to achieve a better category balance. To this end, we web-scraped images from Google and Bing search for each of the categories. We collected enough images to ensure that each category contains at least 100 images of varying safety levels.

Therefore, we also scored images based on the severity of unsafe content. We used four ratings that decrease in safety: *Highly Unsafe*, *Moderately Unsafe*, *Barely Acceptable*, and *Generally Acceptable*. These scores facilitate more detailed evaluations. We make our entire annotated dataset publicly available for the benefit of the community and to stimulate further research.

Figure 2. LlavaGuard assessment of SMID and Web-crawled samples (for instance, second sample is extracted from https://en.wikipedia.org/wiki/Pregnancy). The first row displays the input image, and the 2nd-4th outputs of LlavaGuard-13b. Visible faces and nudity were blurred by the authors.

LlavaGuard training. We provide a 7B and 13B variant of LlavaGuard which are initialized from the respective Llava-1.5 checkpoints. We fine-tune these models for two epochs using LoRA, applying hyperparameters of $r = 128$ and $\alpha = 258$. The training set comprises a total of 2952 distinct images (2415 safe, 537 unsafe) in addition to a held-out test set of 345 images. We oversample training data to train on a balanced split of safe/unsafe data. The training employs a learning rate of $2e{-}5$, utilizing a cosine learning rate scheduler with a warm-up phase of 0.05% steps. We use a micro-batch size of 16 samples per device, and the entire process is executed on four A100-SXM4-80GB GPUs, taking less than an hour to complete.

4. LlavaGuard in the Wild

Next, we present a comprehensive evaluation of LavaGuard. First, we show qualitative examples and empirical results that underscore the performance enhancements achieved by safety tuning. Lastly, we demonstrate an example application of safety annotation and curation for datasets.

Qualitative Results. We begin our evaluation of LlavaGuard by presenting qualitative examples in Fig. 2. For each image from the testset we show the assessment, a category, and a respective explanation provided by LlavaGuard. As can be seen, the model is well aligned with our policy and provides reasonable rationales for all images. One major benefit over previous methods is the generative ability of the underlying LLM to generate an open-ended explanation for its decision. This not only enhances the interpretability of a model's assessments but also contributes to a more nuanced understanding of how safety policies are violated. We have also included an expanded qualitative evaluation, including results from Llava-1.5-7b, in App. Fig. 4. The base models already demonstrate proficiency in content understanding, capable of providing coherent explanations for a majority of the images in our qualitative evaluation. The quality of these texts indicates that they possess a suitable base-level of the capabilities required for our task.

Empirical Results. In Tab. 1, we compare several Llava baselines with their respective LlavaGuard extension[2] on our hold-out test set. While the previous evaluation has shown promising performances in content understanding, the base models Llava-1.5 struggle to accurately identify unsafe image content as defined by the provided policy. Especially, Llava-1.5-7b and Llava-1.5-13b have tend to label the was majority of images as compliant, including those in violation of the defined safety policy.

In contrast, the LlavaGuard models exhibit strong abilities in discerning and rejecting unsafe visual content (see recall in 1) that does not align with the provided safety requirements. While Llava-1.5-13b was only able to detect 15.07% of the unsafe images within the dataset, LlavaGuard pushes its recall performance to 91.13%. Moreover, even LlavaGuard-7b outperforms Llava-1.6 34b though having only 20% of the parameters. Additionally, LlavaGuard achieves very high detection rate across all safety categories for both unsafe and highly unsafe data (*cf*. Fig. 3). The base model fails to reliably detect unsafe images across all categories. This fine-granular analysis across categories and safety levels facilitated by our safety taxonomy helps identify a system's vulnerabilities and weaknesses in-depth.

Dataset Analysis. Lastly, we illustrate how we can leverage LlavaGuard to perform a safety analysis of datasets. For this purpose, we apply LlavaGuard on our held-out test set and obtain detailed insights into the dataset's potential safety risks (*cf*. App. Fig. 5). Firstly, LlavaGuard can provide basic statistics on the number of images that are at risk of violating individual safety categories of the defined policy. Overall we observe a strong correlation of LlavaGuard's safety assessment with the ground truth data anno-

[2]Due to limited resources we have not yet tuned Llava34B.

	Balanced Accuracy (%) ↑	Recall (%) ↑ True *Review Needed* Rate	Specificity (%) ↑ True *Compliant* Rate
Llava-1.5-7b (zero-shot)	64.43	39.19	**89.67●**
Llava-1.5-13b (zero-shot)	56.43	15.07	**97.79○**
Llava-1.6-34b (zero-shot)	82.03	**86.49○**	77.57
LlavaGuard-7b (ours)	**83.20○**	85.14	81.27
LlavaGuard-13b (ours)	**86.13●**	**91.89●**	80.37

Table 1. Performance comparison of Llava baselines and their LlavaGuard extensions on the held-out test set. The base Llava-1.5-13b model struggles to accurately identify unsafe content, detecting only 15.07% of unsafe images. LlavaGuard-13b substantially improves the recall. Notably, even the smaller LlavaGuard-7b model outperforms the much larger Llava-1.6-34b baseline in terms of bal. accuracy.

tated by humans. However, LlavaGuard's ratings tend to be more conservative rating a higher portion of images as unsafe. Preferences in this regard highly depend on the context. However, missing an unsafe image may posses greater dangers than mistakenly identifying a benign image as violation of the safety policies.

5. Discussion

Following the promising results of LlavaGuard, let us now outline some of the challenges remaining for future work. We tuned LlavaGuard using LoRA to elicit capabilities for safety annotation. After this initial phase a subsequent step should involve additional DPO [11] with chosen and rejected answers. DPO will help to further instruct the model on safety and different policies. Secondly, we started off with a policy prompt largely inspired by Llamaguard [10]. Given the multimodal domain, we will delve deeper into optimizing the policy prompt for safety annotations. Furthermore, the majority of our dataset consist of SMID images. Future work may include the application of LlavaGuard in the context of large-scale vision/multimodal dataset curation as well as the moderation of generative AI models. A sensible next step would be the annotating large corpora such as ImageNet [5], LAION-5B [13], or Datacomp-1b [9] with LlavaGuard. Considering the proliferation of synthetic content facilitated by generative AI systems, we intend to evaluate LlavaGuard's efficacy on such content. Generally, LlavaGuard would benefit from extending its training and test data, specifically with synthetic content. We intend to base evaluation of image generation with various Text-to-Image (T2I) models on the I2P benchmark [12].

Limitations. During the LoRA-tuning process, human supervision was applied solely to the generated answers pertaining to 'category' and 'assessment' entries, while the explanation part remained untouched and relied solely on initial model generation. Another trade-off that needs more consideration is determining the threshold between *compliant* and *review needed*. The choice of this threshold depends on the specific use case, whether prioritizing higher recall or specificity (see Tab. 1). Future work should explore this threshold in more detail.

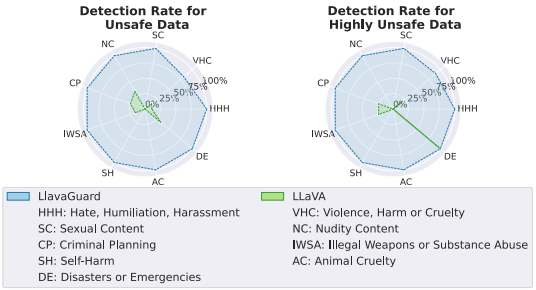

Figure 3. Category-wise performance comparison of LlavaGuard-13B and Llava-13b. We measure the percentage of unsafe (left) and highly unsafe data (right) identified by the model.

6. Conclusion

In this work, we introduced LlavaGuard, a multimodal safeguard model based on Llava designed for assesing image content with respect to safety policies. In this context, LlavaGuard goes beyond binary safety classification by providing assessments that include violated categories as well as detailed explanations. We also introduce a safety risk taxonomy for assessing images regarding safety as well as a human-annotated safety dataset that was collected using this taxonomy. Lastly, we built LlavaGuard by fine-tuning LoRAs on our novel dataset with custom safety policies. We validated the performance of LlavaGuard on a held-out test set, in which even our smallest model, LlavaGuard-7b, outperforms the much larger Llava-34b baseline.We believe that LlavaGuard serves as a strong cornerstone for VLM-based content moderation and beyond.

Acknowledgements We acknowledge support of the hessian.AI Innovation Lab (funded by the Hessian Ministry for Digital Strategy and Innovation), the hessian.AISC Service Center (funded by the Federal Ministry of Education and Research, BMBF, grant No 01IS22091), and the German Research Center for AI (DFKI). Further, this work benefited from the ICT-48 Network of AI Research Excellence Center "TAILOR" (EU Horizon 2020, GA No 952215), the Hessian research priority program LOEWE within the project WhiteBox, the HMWK cluster projects "Adaptive Mind" and "Third Wave of AI", and from the NHR4CES.

References

[1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 610–623, 2021. 1

[2] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, page 1493–1504, 2023. 1

[3] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *ICCV*, 2023. 1

[4] Damien L. Crone, Stefan Bode, Carsten Murawski, and Simon M. Laham. The socio-moral image database (smid): A novel stimulus set for the study of social, moral and affective processes. *PLOS ONE*, 13(1):1–34, 01 2018. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 4

[6] EU. Artificial Intelligence Act EU. https://artificialintelligenceact.eu/, 2023. Accessed: March 13, 2024. 1

[7] Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness, 2023. 1

[8] Felix Friedrich, Katharina Hämmerl, Patrick Schramowski, Jindrich Libovicky, Kristian Kersting, and Alexander Fraser. Multilingual text-to-image generation magnifies gender stereotypes and prompt engineering may not help you, 2024. 1

[9] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 4

[10] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations, 2023. 1, 2, 4

[11] 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. 4

[12] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 4

[13] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 4

[14] UKGov. Ai regulation: A pro-innovation approach. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper, 2023. Accessed: March 13, 2024. 1

[15] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *Proceedings of the 2023 Conference on Neural Information Processing*, 2023. 1, 2

[16] WhiteHouse. Fact sheet: President biden issues executive order on safe, secure, and trustworthy artificial intelligence. https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/, 2023. Accessed: March 13, 2024. 1