

APPENDIX FOR EVOLVABLE SAFETY BENCHMARKING: A MULTI- AGENT PIPELINE FOR LVLMs

Anonymous authors

Paper under double-blind review

LLM Clarification We note that large language models were used to aid in polishing the writing of this paper, but they were not involved in the research design, experimental process, or analysis.

Ethics Statement This work on safety benchmarking for large language models (LLMs) is conducted in alignment with the ICLR Code of Ethics. Our research involves the systematic evaluation of AI models for potential harmful outputs; therefore, we have implemented strict ethical safeguards. All prompts used to elicit harmful behaviors are carefully curated from established safety benchmarks and are designed for measurement, not for generating or disseminating harmful content. The evaluated models are publicly available or used under their respective research licenses. We do not involve human subjects directly. However, we acknowledge the broader ethical implications: the benchmarks and insights generated could potentially be misused to circumvent safety mechanisms. To mitigate this, we focus on reporting aggregated metrics and trends rather than providing specific, exploitable adversarial prompts. Our goal is to contribute to the development of safer AI systems by providing rigorous, transparent, and responsible evaluation methodologies. We have no conflicts of interest to declare.

Reproducibility Statement To ensure the reproducibility of our safety benchmarking study, we provide comprehensive details and resources. Code and data will be released after the paper is accepted. For the proprietary models we evaluate (e.g., GPT-4, Claude-3), we specify the exact API versions and the exhaustive set of inference parameters (e.g., temperature, max tokens). For open-source models, we provide the specific model repository URLs and hashes. The results for all experiments are presented with the mean and standard deviation across multiple runs.

A RELATED WORK

A.1 LVLM SAFETY EVALUATION

Numerous safety evaluation benchmarks for large vision-language models (LVLMs) have been proposed. We categorize them into four types and describe them respectively.

General Safety Benchmarks These benchmarks provide broad evaluations covering multiple safety aspects such as general safety, out-of-distribution generalization, and overall trustworthiness. MM-SafetyBench (Liu et al., 2024b) is an early benchmark containing 5,040 text-image pairs across 13 scenarios. MMDT (Xu et al., 2025) builds a unified platform for comprehensive safety evaluation, covering safety, hallucination, fairness, privacy, adversarial robustness, and OOD generalization. Similarly, MultiTrust (Zhang et al., 2024c), USB (Zheng et al., 2025), and MLLMGuard (Gu et al., 2024) are unified comprehensive benchmarks that cover sufficient aspects: truthfulness, safety, robustness, fairness, privacy, etc.. Unicorn (Tu et al., 2023) is a comprehensive benchmark evaluating Out-of-Distribution (OOD) generalization and adversarial robustness. AVIBench (Zhang et al., 2024a) focuses on evaluating the robustness against adversarial visual instructions.

Jailbreak Benchmarks These datasets are specifically designed to assess the vulnerability of MLLMs to jailbreak attacks. JailBreakV-28K (Luo et al., 2024) is a widely used large-scale benchmark with 28,000 test cases, demonstrating high success rates for transferred LLM attacks. MMJ-Bench (Weng et al., 2024) is a unified pipeline for systematically evaluating jailbreak attacks and

054 defense techniques for MLLMs. Empirical studies on GPT-4o (Ying et al., 2024b) and broader land-
 055 scape analyses (Wang et al., 2024a) also contribute to understanding jailbreaking vulnerabilities.
 056

057 **Domain-Specific Benchmarks** These benchmarks evaluate safety risks in specific contexts or
 058 using real-world data formats like memes and AI-generated content. MemeSafetyBench (Lee et al.,
 059 2025) contains 50,430 instances using real meme images to evaluate VLM safety. GOAT-Bench
 060 (Lin et al., 2024) comprises over 6K memes covering fine-grained themes like implicit hate speech
 061 and cyberbullying. For generated images, ExtremeAIGC (Chandna et al., 2025) benchmarks LMM
 062 vulnerability to AI-generated extremist content. UnsafeBench (Qu et al., 2024) evaluates image
 063 safety classifiers on real-world and AI-generated images, identifying performance degradation on
 064 AI-generated content. Other fields also introduce their benchmarks. SHIELD (Shi et al., 2024)
 065 evaluates MLLM capabilities in face spoofing and forgery detection across RGB, infrared, and depth
 066 modalities. Argus Inspection (Yao et al., 2025) evaluates visual fine-grained perception and causal
 067 reasoning capabilities. VLSBench (Hu et al., 2024) resolves visual leakage issues in multimodal
 068 safety evaluation. MOSSBench (Li et al., 2024c) evaluates oversensitivity in MLLMs, measuring
 069 refusal rates for harmless queries.

070 A.2 AUTOMATED BENCHMARK (DATASET) CREATION

071
 072 Large Language Models (LLMs) or LVLMs demonstrate remarkable capabilities in data generation,
 073 enabling their use for creating or updating evaluation datasets, thereby replacing labor-intensive
 074 manual data curation processes (Liu et al., 2024a). We categorize existing methods for generating
 075 evaluation benchmarks into three types:

076 **1) Dynamic Evaluation:** Methods like DyVal (Zhu et al., 2024a), DyVal 2 (Zhu et al., 2024b),
 077 DME (Yang et al., 2024), DARG (Zhang et al., 2024d), and LatestEval (Li et al., 2024d) propose to
 078 evaluate LLMs via modifying existing benchmarks. These methods aim to enhance dataset difficulty
 079 through dynamic updates and mitigate data contamination issues. For instance, Wang *et al.* propose
 080 a self-evolving method for optimizing test prompts (Wang et al., 2024b). SDEval proposes a dy-
 081 namic safety evaluation framework for LVLMs, which can also be used for model performance
 082 evaluation (Wang et al., 2025).

083 **2) Benchmark Extension:** These methods are used to update or extend existing benchmarks.
 084 For instance, AutoBench generates question-answer pairs by retrieving topic information from
 085 a database, optimizing existing benchmarks to improve their diversity and quality (Li et al., 2024b);
 086 EvoEval explores extending existing code benchmarks to different code domains through LLM-
 087 based augmentation methods and human verification (Xia et al., 2024). Furthermore, GSM-
 088 Symbolic (Mirzadeh et al., 2024) and GSM-Infinite (Zhou et al.) are extensions based on existing
 089 datasets like GSM-8K (Cobbe et al., 2021).

090 **3) Benchmark Construction:** AutoBench utilizes LLMs to annotate image-based question-answer
 091 pairs for evaluation (Qiu et al., 2024); TaskMeAnything generates input-output pairs based on
 092 question-answer templates for building customized multimodal evaluation data (Zhang et al.,
 093 2024b); BenchAgents combines agents with human collaboration for dataset construction (Butt
 094 et al., 2024); DataGen proposes a unified framework for dataset construction and data augmenta-
 095 tion (Huang et al., 2024). AutoBench-V is the first agent-based vision-language automated evalua-
 096 tion framework (Bao et al., 2024). Additionally, LLM-as-an-Examiner (Bai et al., 2023), LLM-
 097 as-an-Interviewer (Kim et al., 2024), BenchBuilder (Li et al., 2024a), StructEval (Cao et al., 2024),
 098 TreeEval (Li et al., 2025), and DeepEval (Li et al., 2023) have also constructed various automated
 099 evaluation frameworks.

100 These methods advance automated evaluation technology to varying degrees. However, none of
 101 them establish an end-to-end LVLM safety benchmarking system. In contrast, we model and modu-
 102 larize the entire dataset construction pipeline, and are currently the only multi-agent system specifi-
 103 cally designed for LVLM safety benchmarking.

104 B DATA COLLECTION

105
 106 To construct a raw data pool for VLSafetyBench, we aggregate data from four sources to guarantee
 107 comprehensiveness and diversity: existing safety datasets, general image datasets, synthetic images,

Table 1: The two-layer safety taxonomy and sample distribution. The total number of samples in the generated benchmark is 4,000.

Category	Samples	Ratio(%)
Privacy	426	10.65
• Privacy Invasion	206	5.15
• Identity Misuse	220	5.5
Bias	419	10.475
• Discrimination	179	4.475
• Insulting & Harassing	240	6.00
Toxicity	933	23.325
• Deception	113	2.825
• Dark Themes	120	3.00
• Horror and Gore	80	2.00
• Pornograph	121	3.025
• Sexual Crimes	65	1.625
• Animal Abuse	124	3.10
• Dangerous Behavior	310	7.75
Legality	584	14.60
• Financial Fraud	203	5.075
• Terrorism	181	4.525
• Cybersecurity	200	5.00
Misinformation	491	12.275
• False Information	207	5.175
• Sensitive Information	127	3.175
• Public Misinformation	157	3.925
Health Risk	1147	28.675
• Health Hazards	235	5.875
• Environmental Harm	139	3.475
• Physical Harm	382	9.55
• Psychological Harm	391	9.775
Total	4000	100

and social media data. We follow (Zhang et al., 2025) and leverage CLIP (Radford et al., 2021) to conduct coarse filtering, ensuring that all selected images contain potentially harmful information. After filtration, the data pool comprises around 300K images, with 134K derived from existing datasets, 20K from general images, 40K generated via diffusion models, and 106K scraped from social media. Below, we detail the information of each source.

- **Existing Safety Datasets:** We select 134K image-question pairs from open-source LVLMM safety benchmarks. The involved datasets include SafeBench (Ying et al., 2024a), MLLM-Guard (Gu et al., 2024), BeaverTails-V (Ji et al., 2025), and SPA-VL (Zhang et al., 2025).
- **General Image Datasets:** We curate 20K risk images from LAION-5B (Schuhmann et al., 2022). We follow (Zhang et al., 2025) to utilize CLIP to identify harmful images. The LAION Safety Toolkit (Laion-AI, 2021) is also adopted to identify porn images.
- **Synthetic Images:** We also collect 40K synthesized images. Generative models such as Stable Diffusion 3 (Esser et al., 2024), FLUX.1 dev (Labs, 2024), and MidJourney v6 (Midjourney, 2023) are included. The generation process is guided by 20K harmful prompts generated by uncensored models. Diverse safety categories are covered.
- **Social Media Data:** We scrape 106K harmful images from the internet, including social media (Xiaohongshu¹ and X²) and news platform (Toutiao³). To ensure ethical compliance, we filter out images containing personally identifiable information (PII). In addition, we perform watermark removal on select images.

¹<https://www.xiaohongshu.com>

²<https://x.com>

³<https://www.toutiao.com>

Table 2: Safety evaluation of 35 mainstream LVLMs. In addition to ASR, we also present safety rate (SR), where $SR=1-ASR$. **Bold** indicates the best and underline indicates the second.

Model	Privacy	Bias	Toxic	Legal	Misinfo	Health	ASR	SR
Ovis2-8B	51.59	34.04	43.87	51.70	42.35	46.16	44.85	55.15
Ovis2-16B	45.86	39.72	43.68	52.13	31.12	42.37	42.73	57.27
Ovis2-34B	45.86	39.01	42.34	45.11	31.63	37.50	40.40	59.60
SAIL-VL-8B	57.32	42.55	51.70	61.49	43.88	52.49	51.70	48.30
MiniCPM-V2.6	51.59	43.97	46.16	57.66	41.84	50.25	48.23	51.77
QVQ-72B-Preview	38.56	32.62	33.61	38.65	40.00	38.24	36.35	63.65
Qwen2.5-VL-32B-Instruct	45.39	17.73	28.36	45.93	15.90	36.44	31.44	68.56
Qwen2.5-VL-72B-Instruct	30.77	6.12	26.41	29.08	16.22	20.19	22.42	77.58
Mistral-Small-3.1-24B-Instruct	31.21	10.64	22.77	31.91	15.82	19.69	21.92	78.08
Pixtral-12B-2409	57.96	39.01	52.77	58.72	44.39	46.81	50.22	49.78
InstructBLIP-Vicuna-7B	77.71	68.79	70.01	78.94	68.88	67.73	72.19	27.81
InstructBLIP-Vicuna-13B	77.78	71.43	65.48	78.46	70.27	68.35	70.12	29.88
Phi-3.5-Vision-Instruct	9.55	42.55	14.41	7.23	41.33	13.96	19.09	80.91
Kimi-VL-A3B-Instruct	57.32	48.23	46.44	52.13	42.35	52.78	49.43	50.57
InternVL3-1B-Instruct	50.00	50.00	59.78	67.27	33.33	63.25	56.27	43.73
InternVL3-9B-Instruct	45.22	46.10	45.17	53.19	36.22	51.47	46.58	53.42
InternVL3-38B-Instruct	37.58	36.17	35.55	43.62	31.63	37.94	36.84	63.16
InternVL3-78B-Instruct	33.12	30.50	29.45	34.68	24.49	32.81	30.75	69.25
Gemma-3-4B-It	35.67	14.89	20.52	34.26	17.86	17.53	22.14	77.86
Gemma-3-27B-It	36.31	9.93	16.07	29.15	16.33	17.58	19.45	80.55
DeepSeek-VL2-Tiny	62.42	73.05	60.78	60.85	67.35	60.89	63.09	36.91
DeepSeek-VL2	48.41	79.43	43.72	46.38	69.90	52.70	53.41	46.59
DeepSeek-VL2-Small	52.23	81.56	55.00	51.28	76.02	56.48	59.89	40.11
GLM-4.1V-9B-Thinking	69.43	45.39	54.10	68.30	46.94	62.59	57.50	42.50
LLaVA-V1.6-Vicuna-13B	59.87	30.50	42.47	56.60	33.16	45.00	44.17	55.83
LLaVA-V1.6-Mistral-7B	64.97	26.95	53.72	65.32	33.67	55.63	51.48	48.52
LLaVA-Next-110B	59.24	52.48	52.05	58.94	47.45	48.44	52.35	47.65
LLaMA-3.2-11B-Vision-Instruct	36.31	34.75	31.87	31.70	33.67	25.05	31.35	68.65
Grok-4	50.94	35.48	29.80	30.53	23.75	38.34	34.09	65.91
GPT-4o	<u>7.33</u>	<u>12.50</u>	<u>5.69</u>	<u>2.95</u>	6.91	<u>4.62</u>	<u>6.22</u>	<u>93.78</u>
GPT-4.1	8.05	<u>12.50</u>	7.30	4.03	6.84	5.44	7.13	92.87
Qwen-VL-Max	28.48	<u>17.39</u>	19.58	26.27	11.23	22.22	20.73	79.27
Gemini-2.5-Pro	37.50	19.82	22.41	20.79	18.47	14.43	21.41	78.59
Claude-Sonnet-4	1.32	3.55	1.88	1.55	4.15	1.89	2.22	97.78

C SAFETY TAXONOMY

We integrate the taxonomies of (Gu et al., 2024) and (Ji et al., 2025) to define a two-layer tree comprising 6 categories and 20 subcategories. The 6 categories are: *Privacy*, *Bias*, *Toxicity*, *Legality*, *Misinformation*, and *Health Risk*. The subcategories and the number of samples of each (sub)category is presented in Table 1.

D LVLm SAFETY EVALUATION

Using the benchmark generated by VLSafetyBench, we evaluate 35 mainstream LVLms in Table 2. We can observe significant safety disparity, where the safety rate ranges from 27.81% (InstructBLIP-Vicuna-7B) to 97.78% (Claude-Sonnet-4). It is observed that safety shows a general upward trend with increasing model scale within the same series, which is corroborated by the Qwen2.5-VL and InternVL3 model families.

REFERENCES

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36:78142–78167, 2023.

- 216 Han Bao, Yue Huang, Yanbo Wang, Jiayi Ye, Xiangqi Wang, Xiuying Chen, Yue Zhao, Tianyi
217 Zhou, Mohamed Elhoseiny, and Xiangliang Zhang. Autobench-v: Can large vision-language
218 models benchmark themselves? *arXiv preprint arXiv:2410.21259*, 2024.
- 219
220 Natasha Butt, Varun Chandrasekaran, Neel Joshi, Besmira Nushi, and Vidhisha Balachan-
221 dran. Benchagents: Automated benchmark creation with agent interaction. *arXiv preprint*
222 *arXiv:2410.22584*, 2024.
- 223
224 Boxi Cao, Mengjie Ren, Hongyu Lin, Xianpei Han, Feng Zhang, Junfeng Zhan, and Le Sun. Structe-
225 val: Deepen and broaden large language model assessment via structured evaluation. *arXiv*
226 *preprint arXiv:2408.03281*, 2024.
- 227
228 Bhavik Chandna, Mariam Aboujenane, and Usman Naseem. Extremeaigc: Benchmarking lmm
vulnerability to ai-generated extremist content. *arXiv preprint arXiv:2503.09964*, 2025.
- 229
230 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,
231 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John
232 Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*,
233 2021.
- 234
235 Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam
236 Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers
237 for high-resolution image synthesis. In *Forty-first international conference on machine learning*,
2024.
- 238
239 Tianle Gu, Zeyang Zhou, Kexin Huang, Liang Dandan, Yixu Wang, Haiquan Zhao, Yuanqi Yao,
240 Yujie Yang, Yan Teng, Yu Qiao, et al. Mllmguard: A multi-dimensional safety evaluation suite
241 for multimodal large language models. *Advances in Neural Information Processing Systems*, 37:
7256–7295, 2024.
- 242
243 Xuhao Hu, Dongrui Liu, Hao Li, Xuanjing Huang, and Jing Shao. Vlsbench: Unveiling visual
244 leakage in multimodal safety. *arXiv preprint arXiv:2411.19939*, 2024.
- 245
246 Yue Huang, Siyuan Wu, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Chaowei
247 Xiao, Jianfeng Gao, Lichao Sun, et al. Datagen: Unified synthetic dataset generation via large
language models. In *The Thirteenth International Conference on Learning Representations*, 2024.
- 248
249 Jiaming Ji, Xinyu Chen, Rui Pan, Han Zhu, Conghui Zhang, Jiahao Li, Donghai Hong, Boyuan
250 Chen, Jiayi Zhou, Kaile Wang, Juntao Dai, Chi-Min Chan, Sirui Han, Yike Guo, and Yaodong
251 Yang. Safe rlhf-v: Safe reinforcement learning from human feedback in multimodal large lan-
guage models, 2025.
- 252
253 Eunsu Kim, Juyoung Suk, Seungone Kim, Niklas Muennighoff, Dongkwan Kim, and Alice Oh.
254 Llm-as-an-interviewer: Beyond static testing through dynamic llm evaluation. *arXiv preprint*
255 *arXiv:2412.10424*, 2024.
- 256
257 Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024.
- 258
259 Laion-AI. Laion-safety. <https://github.com/LAION-AI/LAION-SAFETY>, 2021.
- 260
261 DongGeon Lee, Joonwon Jang, Jihae Jeong, and Hwanjo Yu. Are vision-language models safe in
the wild? a meme-based benchmark study. *arXiv preprint arXiv:2505.15389*, 2025.
- 262
263 Jiatong Li, Rui Li, and Qi Liu. Beyond static datasets: A deep interaction approach to llm evaluation.
arXiv preprint arXiv:2309.04369, 2023.
- 264
265 Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gon-
266 zalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and
267 benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024a.
- 268
269 Xiang Li, Yunshi Lan, and Chao Yang. Treeeval: Benchmark-free evaluation of large language
models through tree planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*,
volume 39, pp. 24485–24493, 2025.

- 270 Xiang Lisa Li, Evan Zheran Liu, Percy Liang, and Tatsunori Hashimoto. Autobencher: Creating
271 salient, novel, difficult datasets for language models. *arXiv e-prints*, pp. arXiv-2407, 2024b.
272
- 273 Xirui Li, Hengguang Zhou, Ruochen Wang, Tianyi Zhou, Minhao Cheng, and Cho-Jui Hsieh.
274 Mossbench: Is your multimodal language model oversensitive to safe queries? *arXiv preprint*
275 *arXiv:2406.17806*, 2024c.
- 276 Yucheng Li, Frank Guerin, and Chenghua Lin. Latesteval: Addressing data contamination in lan-
277 guage model evaluation through dynamic and time-sensitive test construction. In *Proceedings of*
278 *the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18600–18607, 2024d.
279
- 280 Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma. Goat-bench: Safety insights
281 to large multimodal models through meme-based social abuse. *arXiv preprint arXiv:2401.01523*,
282 2024.
- 283 Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi
284 Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. Best practices and lessons learned on syn-
285 thetic data for language models, 2024a.
- 286 Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A
287 benchmark for safety evaluation of multimodal large language models. In *European Conference*
288 *on Computer Vision*, pp. 386–403. Springer, 2024b.
- 290 Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. Jailbreakv-28k: A bench-
291 mark for assessing the robustness of multimodal large language models against jailbreak attacks.
292 *arXiv preprint arXiv:2404.03027*, 2024.
- 293 Midjourney. Midjourney AI, 2023. URL <https://www.midjourney.com>.
294
- 295 Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad
296 Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large
297 language models. *arXiv preprint arXiv:2410.05229*, 2024.
- 298 Ruidi Qiu, Grace Li Zhang, Rolf Drechsler, Ulf Schlichtmann, and Bing Li. Autobench: Auto-
299 matic testbench generation and evaluation using llms for hdl design. In *Proceedings of the 2024*
300 *ACM/IEEE International Symposium on Machine Learning for CAD*, pp. 1–10, 2024.
- 302 Yiting Qu, Xinyue Shen, Yixin Wu, Michael Backes, Savvas Zannettou, and Yang Zhang. Un-
303 safebench: Benchmarking image safety classifiers on real-world and ai-generated images. *arXiv*
304 *preprint arXiv:2405.03486*, 2024.
- 305 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
306 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual
307 models from natural language supervision. In *International conference on machine learning*, pp.
308 8748–8763. PmLR, 2021.
- 309 Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi
310 Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An
311 open large-scale dataset for training next generation image-text models. *Advances in neural in-*
312 *formation processing systems*, 35:25278–25294, 2022.
- 314 Yichen Shi, Yuhao Gao, Yingxin Lai, Hongyang Wang, Jun Feng, Lei He, Jun Wan, Changsheng
315 Chen, Zitong Yu, and Xiaochun Cao. Shield: An evaluation benchmark for face spoofing and
316 forgery detection with multimodal large language models. *arXiv preprint arXiv:2402.04178*,
317 2024.
- 318 Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu
319 Zhou, Huaxiu Yao, and Cihang Xie. How many unicorns are in this image? a safety evaluation
320 benchmark for vision llms. *arXiv preprint arXiv:2311.16101*, 2023.
- 322 Hanqing Wang, Yuan Tian, Mingyu Liu, Zhenhao Zhang, and Xiangyang Zhu. Sdeval: Safety
323 dynamic evaluation for multimodal large language models. *arXiv preprint arXiv:2508.06142*,
2025.

- 324 Siyuan Wang, Zhuohan Long, Zhihao Fan, and Zhongyu Wei. From llms to mllms: Exploring the
325 landscape of multimodal jailbreaking. *arXiv preprint arXiv:2406.14859*, 2024a.
- 326
- 327 Siyuan Wang, Zhuohan Long, Zhihao Fan, Zhongyu Wei, and Xuanjing Huang. Bench-
328 mark self-evolving: A multi-agent framework for dynamic llm evaluation. *arXiv preprint*
329 *arXiv:2402.11443*, 2024b.
- 330 Fenghua Weng, Yue Xu, Chengyan Fu, and Wenjie Wang. \textit {MMJ-Bench}: A comprehensive
331 study on jailbreak attacks and defenses for vision language models. *arXiv e-prints*, pp. arXiv-
332 2408, 2024.
- 333
- 334 Chunqiu Steven Xia, Yinlin Deng, and Lingming Zhang. Top leaderboard ranking= top coding profi-
335 ciency, always? evoeval: Evolving coding benchmarks via llm. *arXiv preprint arXiv:2403.19114*,
336 2024.
- 337 Chejian Xu, Jiawei Zhang, Zhaorun Chen, Chulin Xie, Mintong Kang, Yujin Potter, Zhun Wang,
338 Zhuowen Yuan, Alexander Xiong, Zidi Xiong, et al. Mmdt: Decoding the trustworthiness and
339 safety of multimodal foundation models. *arXiv preprint arXiv:2503.14827*, 2025.
- 340
- 341 Yue Yang, Shuibai Zhang, Wenqi Shao, Kaipeng Zhang, Yi Bin, Yu Wang, and Ping Luo. Dynamic
342 multimodal evaluation with flexible complexity by vision-language bootstrapping. *arXiv preprint*
343 *arXiv:2410.08695*, 2024.
- 344 Yang Yao, Lingyu Li, Jiaxin Song, Chiyu Chen, Zhenqi He, Yixu Wang, Xin Wang, Tianle Gu, Jie
345 Li, Yan Teng, et al. Argus inspection: Do multimodal large language models possess the eye of
346 panoptes? *arXiv preprint arXiv:2506.14805*, 2025.
- 347 Zonghao Ying, Aishan Liu, Siyuan Liang, Lei Huang, Jinyang Guo, Wenbo Zhou, Xianglong Liu,
348 and Dacheng Tao. Safebench: A safety evaluation framework for multimodal large language
349 models. *arXiv preprint arXiv:2410.18927*, 2024a.
- 350
- 351 Zonghao Ying, Aishan Liu, Xianglong Liu, and Dacheng Tao. Unveiling the safety of gpt-4o: An
352 empirical study using jailbreak attacks. *arXiv preprint arXiv:2406.06302*, 2024b.
- 353
- 354 Hao Zhang, Wenqi Shao, Hong Liu, Yongqiang Ma, Ping Luo, Yu Qiao, and Kaipeng Zhang.
355 Avibench: Towards evaluating the robustness of large vision-language model on adversarial
356 visual-instructions. *CoRR*, 2024a.
- 357 Jieyu Zhang, Weikai Huang, Zixian Ma, Oscar Michel, Dong He, Tanmay Gupta, Wei-Chiu Ma, Ali
358 Farhadi, Aniruddha Kembhavi, and Ranjay Krishna. Task me anything. In *Thirty-Eighth Annual*
359 *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024b.
- 360 Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huan-
361 ran Chen, Xiao Yang, Xingxing Wei, et al. Benchmarking trustworthiness of multimodal large
362 language models: A comprehensive study. *arXiv e-prints*, pp. arXiv-2406, 2024c.
- 363
- 364 Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie
365 Jin, Yu Qiao, Xuanjing Huang, et al. Spa-vl: A comprehensive safety preference alignment dataset
366 for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition*
367 *Conference*, pp. 19867–19878, 2025.
- 368 Zhehao Zhang, Jiaao Chen, and Diyi Yang. Darg: Dynamic evaluation of large language models
369 via adaptive reasoning graph. *Advances in Neural Information Processing Systems*, 37:135904–
370 135942, 2024d.
- 371
- 372 Baolin Zheng, Guanlin Chen, Hongqiong Zhong, Qingyang Teng, Yingshui Tan, Zhendong Liu,
373 Weixun Wang, Jiaheng Liu, Jian Yang, Huiyun Jing, et al. Usb: A comprehensive and
374 unified safety evaluation benchmark for multimodal large language models. *arXiv preprint*
375 *arXiv:2505.23793*, 2025.
- 376 Yang Zhou, Hongyi Liu, Zhuoming Chen, Yuandong Tian, and Beidi Chen. Gsm-infinite: How do
377 your llms behave over infinitely increasing reasoning complexity and context length? In *Forty-
second International Conference on Machine Learning*.

378 Kaijie Zhu, Jiaao Chen, Jindong Wang, Neil Zhenqiang Gong, Diyi Yang, and Xing Xie. Dyval:
379 Dynamic evaluation of large language models for reasoning tasks, 2024a.
380
381 Kaijie Zhu, Jindong Wang, Qinlin Zhao, Ruochen Xu, and Xing Xie. Dyval 2: Dynamic evaluation
382 of large language models by meta probing agents, 2024b.
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431