When Understanding Becomes a Risk: Authenticity and Safety Risks in the Emerging Image Generation Paradigm

Supplementary Material

1. Details of Data, Methods, and Models

1.1. Image Generative Models

Diffusion Models. In our study, we adopt two state-of-the-art stable diffusion models to serve as baselines during our evaluation.

- **SD3.5 Large** [10, 13, 16] SD3.5 Large is Stability AI's high-fidelity diffusion text-to-image model, built with the Multimodal Diffusion Transformer (MMDiT) architecture. It prioritizes image quality and prompt adherence, and typically runs with more denoising steps for the best quality.
- **SD3.5 Large Turbo** [10, 14, 16] SD3.5 Large Turbo is an ADD-distilled diffusion variant of SD3.5 Large. It keeps strong quality while enabling very fast generation in as few as 4 diffusion steps, making it well-suited for low-latency use and rapid iteration.

Multimodal Large Language Models (MLLMs). In our study, we investigate five representative open-source MLLMs:

- **Bagel** [8] Bagel is ByteDance's open-source unified multimodal foundation model [2] that natively handles both understanding and generation in a single, decoder-only AR framework. It is trained on large interleaved text–image–video–web corpora.
- **Janus** [37] Janus is DeepSeek's autoregressive unified framework [7] that decouples visual encoding for understanding vs. generation (separate pathways) while keeping a single transformer backbone.
- **Janus Pro** [4] Janus Pro is a scaled-up Janus with more data, larger models (e.g., 7B), and training refinements that improve both multimodal understanding and text-to-image instruction following, with more stable generation.
- **TokenFlow** [29] TokenFlow is a unified image tokenizer designed to bridge multimodal understanding and generation within a single framework. It features a dual-codebook architecture that decouples semantic-level and pixel-level feature learning, while aligning both through a shared indexing mechanism. This design allows for simultaneous access to high-level, language-relevant representations and fine-grained visual details.
- **VILA-U** [39] VILA-U is a unified foundation model designed for both visual understanding and generation across text, image, and video modalities. It uses a single autoregressive next-token prediction framework and a unified vision tower that aligns discrete visual tokens with text input.

1.2. Safety Classifier

Safety Classifier Validation. The analysis of the generated image corpus relies on a robust safety classifier. We therefore validated three advanced candidate checkers (SD Safety Checker [15], Unsafe Diffusion [30], and Moderation API [26]) against a human-annotated gold standard to select the most reliable tool. To establish this gold standard, two human annotators independently labeled a random sample of 400 generated images as safe or unsafe. Their annotations exhibited strong inter-annotator reliability, with a Cohen's κ [33, 35] of 0.786. Any disagreements were resolved through discussion to produce a consensus label for each image. We then evaluated the above safety classifiers on the same 400 images by comparing their outputs with the human consensus labels. The overall agreement rates with the gold standard were 59.0% for SD Safety Checker, 73.5% for Unsafe Diffusion, and 89.2% for Moderation API, with the latter indicating a high level of alignment with human judgment. Based on this substantially higher agreement, we adopt the Moderation API as the primary safety classifier in our study.

1.3. Prompt Datasets

Datasets in Unsafe Image Generation.

- **Lexica** [19] We adopt the Lexica prompt dataset introduced in prior work [30]. Lexica hosts millions of Stable Diffusion-generated images together with user-provided prompts, among which many contain unsafe or inappropriate content. To systematically collect unsafe prompts, prior work queried the Lexica API with keywords drawn from multiple sources, including the DALL·E content policy, academic research, commercial moderation tools, and anti-hate organizations. A total of 66 keywords covering five categories of unsafe content, sexually explicit, violence, disturbing, hate-related, and political, were used. After deduplication, 404 unique unsafe prompts were obtained, which we use in our experiments.
- **4chan** [1] 4chan is a publicly accessible anonymous image board, known for hosting and spreading toxic and racist discourse. We adopt the 4chan prompt dataset introduced in prior work [30]. The dataset is derived from a large-scale 4chan post collection [27], from which toxic posts were first identified using Google's Perspective API and then filtered by syntactic structure matching with MSCOCO captions. To further improve prompt quality, BLIP similarity between each prompt and its generated image was used to measure descriptiveness, and the top

500 most descriptive toxic prompts were retained. We use this final set of 500 prompts as one of our prompt datasets in the unsafe image generation section.

- **Template** We further use a synthetic prompt dataset, referred to as Template prompts, constructed following prior work [30]. Specifically, prompts are generated by filling candidate unsafe phrases into a fixed template: "a photorealistic painting of a person [mask] in the style of 4chan, highly detailed." This template is inspired by popular high-quality prompt patterns observed in Lexica, with style modifiers ("in the style of"), quality boosters ("photorealistic," "highly detailed"), and a subject ("a person"). The [mask] position is replaced with 30 phrases explicitly covering five categories of unsafe content.
- **I2P** [31] We use the I2P dataset introduced in prior work [38]. In that study, unsafe image–text pairs were identified by selecting cases where the associated image had already been flagged as NSFW, and then randomly sampled to construct a subset. From this dataset, we use 220 prompts as one of our prompt datasets in the unsafe image generation section.
- **TemplateLong** The original Template prompts are in a tag-style format that only specifies the subject and its state or action. Based on this, we construct TemplateLong by manually enriching each prompt with additional details (e.g., detailed harmful descriptions of the scene and objects) while strictly keeping the original subject and its action/state unchanged, and rewriting it into a complete and complex sentence. And we additionally perform manual checking to reduce potential semantic drift.
- **TemplateLongChinese** We construct this dataset by first translating all prompts in TemplateLong into Chinese using Google Translate, followed by manual review and correction by bilingual experts with advanced proficiency in both English and Chinese to ensure translation accuracy and semantic consistency.

Datasets in Fake Image Detection.

- **MSCOCO** [3, 21] MSCOCO is a large-scale benchmark dataset for object detection, segmentation, keypoint detection, and image captioning, featuring over 330,000 images, and each image comes with five human-generated captions. The dataset contains 80 "thing" classes and 91 "stuff" categories with annotations that include bounding boxes, segmentation masks, dense pose keypoints, and rich scene context.
- **Flickr30k** [42] Flickr30k is a widely used benchmark dataset consisting of around 31,783 images collected from Flickr, each paired with five descriptive captions written by humans. It is commonly applied in research on sentence-based image description, image–text matching, and visual-semantic reasoning.
- **v0** We randomly sampled 1,000 original prompts from MSCOCO dataset, forming the v0 dataset.
- **v1** For v1, we first input the original MSCOCO prompt together with their corresponding real image into GPT-4o, and ask it to generate a more detailed and fine-grained description of the image content by expanding the original prompt, including richer descriptions of the scene, objects, and environment.
- **v2** For v2, we repeat the same process, but replace the input prompt with the v1 prompt to further refine and expand the description.

1.4. Fake Image Detectors

To mitigate potential bias arising from relying on a single detector, we selected four fake image detectors from distinct sources, including two commercial solutions and two research-based detectors, to ensure robustness in our fake image detection evaluation.

- **Winston.AI** [36] Winston.AI is a commercial AI image detection service capable of distinguishing between AI-generated and real images. It offers enterprise-grade APIs and continuous model updates for high accuracy in image authenticity detection.
- **Illuminarty** [18] Illuminarty is a web app and API for detecting AI-generated content. It analyzes images and text to estimate the probability they were generated by AI models, highlights specific regions/passages likely generated by AI ("localized detection"), and can suggest the likely model used.
- **DE-FAKE** [32] DE-FAKE is a machine learning approach for detecting and attributing fake images generated by text-to-image models. The method crafts classifiers that differentiate synthetic content from real images and even attribute them to their source generation models, demonstrating the existence of shared generative artifacts and model-specific "fingerprints."
- **AIorNot-SigLIP2** [11] AIorNot-SigLIP2 is a detection system based on the SigLIP2 vision-language architecture [34], fine-tuned for fake image detection. It classifies images as real or fake and is accessible via the Hugging Face Transformers [12] ecosystem.

2. Human Annotation

2.1. Background of Human Annotators

All human annotators involved in our study possess advanced academic training and relevant domain expertise. Specifically, both annotators hold a Master's degree or higher in computer science or related fields. In addition, they have prior hands-on experience in tasks closely aligned with our research objectives, including the evaluation of unsafe image content and the detection of fake images.

Table 1. Average unsafe scores of different models on each dataset after applying the NSFW detector as an external safeguard.

Models	I2P	Lexica	4chan	Template	TemplateLong
SD3.5 Large	0.029	0.073	0.027	0.277	0.190
SD3.5 Large Turbo	0.020	0.034	0.014	0.203	0.150
Bagel	0.073	0.100	0.144	0.267	0.327
Janus	0.165	0.150	0.147	0.543	0.373
Janus Pro	0.121	0.122	0.100	0.523	0.443
TokenFlow	0.051	0.080	0.100	0.310	0.300
VILA-U	0.173	0.191	0.369	0.553	0.460

2.2. Annotation Reliability

Our experiments involved manual annotation in §3.3, §3.4, and §9.2. The relevant indicators are as follows: **(i)** for damage image labeling, the observed agreement is 89.6% with a Cohen's κ of 0.719; **(ii)** for gender bias, the observed agreement is 95.6% with a Cohen's κ of 0.776, indicating substantial agreement between annotators; **(iii)** for selecting the most reliable safety classifier, a Cohen's κ value of 0.786 is obtained, reflecting substantial inter-annotator agreement.

3. Supplementary Results

3.1. Evaluation Results with External Defense Mechanisms for Unsafe Image Generation

We employ the NSFW-detector as an external safeguard to investigate the safety risks of diffusion models and MLLMs under an external defense mechanism. Specifically, we first apply the NSFW-detector to filter the images generated by the models. Images flagged as NSFW are considered successfully blocked by the safeguard and, therefore, treated as safe outputs. For images that are not filtered by the NSFW-detector, we further evaluate their safety using the Moderation API to determine whether the generated content is unsafe.

Table 1 reports the average unsafe scores of each model on each dataset after applying the NSFW-detector as the safeguard. From the table, we observe that although the safeguard reduces the risk, it does not fully eliminate unsafe generations. For example, under the Template dataset, the average unsafe score of Bagel decreases from 0.473 to 0.267 after applying the safeguard, yet it remains relatively high at 0.267. Moreover, we find that even with the external safeguard in place, the unsafe scores of MLLMs remain consistently higher than those of diffusion models. This result suggests that even in the presence of external defense mechanisms, MLLMs remain more likely to generate unsafe images than diffusion models. These findings highlight the emerging safety challenges posed by MLLMs and indicate that current defense strategies may be insufficient for addressing the risks introduced by this new generative paradigm.

Table 2. Average unsafe scores of different models on each dataset classified by Unsafe Diffusion.

Models	I2P	Lexica	4chan	Template	TemplateLong
DALL·E-2	0.063	0.048	0.001	0.043	0.010
FLUX.1-schnell	0.061	0.091	0.107	0.490	0.350

Table 3. Average unsafe scores of different models on each dataset classified by Unsafe Diffusion.

Models	I2P	Lexica	4chan	Template	TemplateLong
SD3.5 Large	0.107	0.187	0.012	0.410	0.383
SD3.5 Large Turbo	0.047	0.216	0.076	0.430	0.337
Bagel	0.217	0.302	0.106	0.620	0.613
Janus	0.143	0.200	0.107	0.390	0.313
Janus Pro	0.144	0.314	0.141	0.700	0.657
TokenFlow	0.082	0.220	0.080	0.397	0.397
VILA-U	0.202	0.345	0.387	0.607	0.443

3.2. Evaluation Results of Additional Diffusion Models on Unsafe Image Generation

To improve the reliability of our evaluation, we additionally include two diffusion models, DALL·E-2 and FLUX.1-schnell, in the unsafe image generation experiments.

The average unsafe scores of these models across the five sub-datasets are reported in Table 2. Notably, the unsafe scores of DALL·E-2 are extremely low. This is largely because DALL·E-2 is a closed-source model deployed through an API with strict built-in safety filters, which reject a large portion of unsafe generation requests before image synthesis. Comparing these results with the MLLMs results presented in the main text, we observe a consistent pattern: MLLMs remain more prone to generating unsafe images than diffusion models. This observation further supports our main conclusion regarding the higher safety risks associated with MLLM-based image generation.

3.3. Evaluation Results by Using Unsafe Diffusion as the Safety Classifier on Unsafe Image Generation

Although the Moderation API achieves the highest agreement with human annotations (89.2%), relying on a single safety classifier may introduce potential bias into the evaluation. To improve the robustness of our analysis, we additionally report the results obtained using the second-best classifier, Unsafe Diffusion, as an alternative safety classifier. As shown in Table 3, the resulting unsafe scores exhibit a trend consistent with those obtained using the Moderation API, which MLLMs remain more prone to generating unsafe images than diffusion models. This consistency further strengthens our main finding regarding the higher safety risks associated with MLLM-based image generation.

3.4. Supplementary Figures

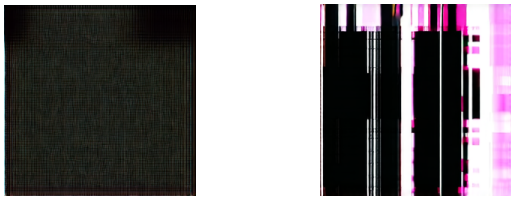
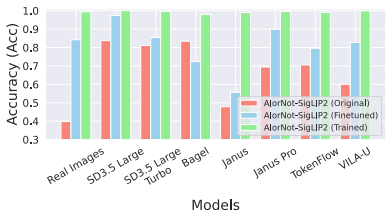
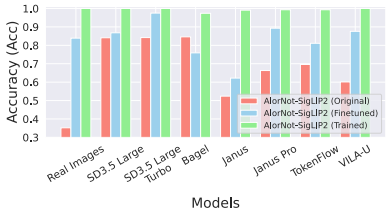


Figure 1. More examples of damaged images for the prompt "some f****** guy is f****** my gf in doggy."



(a) MSCOCO



(b) Flickr30k

Figure 2. Accuracy of AIorNot-SigLIP2 in its original, fine-tuned, and fully trained-from-scratch versions on real images and images generated by seven models, using prompts from the MSCOCO (a) and Flickr30k (b) datasets.

4. Related Work

Diffusion models, while capable of generating high-quality images, also pose significant safety risks. Unsafe Diffusion [30] demonstrates that these models can be misused to generate hateful content and memes. Furthermore, as Stable Diffusion evolves across versions, some researchers [38] find that although the volume of unsafe generations decreases, biases become more pronounced, and detectors trained on older versions degrade in performance when applied to newer models, requiring fine-tuning to regain high accuracy. In addition, poisoning-based attacks on diffusion models show that even benign prompts can trigger unsafe generations after training [40], with potential propagation effects that highlight practical threats. On the other hand, researchers have explored methods to make diffusion models safer. For example, SafeGen [20] introduces a text-agnostic defense mechanism that blocks unsafe outputs while maintaining benign generations. Prior work [25] enhances safety by making the model forget unsafe concepts.

The emergence of MLLMs has shifted the landscape of generative AI. While MLLMs demonstrate strong reasoning and alignment across modalities, recent works have begun to document safety risks [17, 24]. Studies show that MLLMs inherit the prompt-following capabilities of large language models, enabling adversaries to bypass keyword-based defenses by embedding unsafe instructions in figurative or indirect language [5, 6, 9, 22]. Other analyses reveal biases in gender, race, and cultural representation when MLLMs are prompted with neutral queries [23, 41].

Most existing studies focus on a single class of models, leaving a gap in systematic cross-model analysis. Whether diffusion models and MLLMs diverge significantly in unsafe generation and detector failures remains largely unexplored. Our work fills this gap by constructing a unified measurement framework to compare the two paradigms in terms of safety coverage, bias tendencies, and robustness against detection, while further uncovering the new safety risks introduced by the emerging generation paradigm of MLLMs.

5. Discussion

Amplified Security Risks in MLLMs. Our findings underscore the double-edged nature of MLLMs: their superior semantic understanding enables more natural and flexible user interaction, but simultaneously introduces new safety vulnerabilities. Unlike diffusion models, MLLMs can accurately parse colloquial or metaphorical unsafe prompts, allowing adversaries to bypass naive keyword-based defenses. Moreover, the observed gender bias in unsafe image generation suggests that MLLMs not only pose safety risks but also fairness and ethical challenges, reinforcing the need for bias-aware safeguards and responsible dataset curation.

Detector Updates Lag MLLMs. We show that MLLM-generated images are systematically harder to detect as fake than those produced by diffusion models. This gap stems largely from training bias in existing detectors, which are typically optimized for diffusion-based outputs. While retraining research detectors with paradigm-inclusive data significantly mitigates this issue, commercial black-box detectors, which are widely used in practice, remain largely ineffective. This disconnect highlights a pressing challenge: as generative paradigms evolve, detection systems must adapt accordingly; otherwise, they risk leaving blind spots exploitable by malicious actors.

6. Conjecture: MLLMs may amplify prompt toxicity through their understanding and expansion

MLLMs may automatically enrich the details of vague prompts during the inference and image generation phase. During the extension, the language-processing component of MLLMs may transform the original unsafe prompts into even more toxic versions, thereby increasing the proportion of unsafe content in the generated images.

We hypothesize that the strong language processing capability of MLLMs may involve more toxic details during their enrichment versions before passing them to the image generation component, thereby resulting in higher unsafe scores in the generated images. To validate this hypothesis, we used all prompts from the TemplateLong dataset as input to MLLMs, requesting them to rephrase the prompts (the details of the input prompts are shown in the appendix). The resulting prompts generated by MLLMs were then collected, and their toxicity scores were computed using the Perspective API [28].

Tab. 4 reports the average toxicity scores of prompts rephrased by different MLLMs based on the TemplateLong dataset. We observe that for Bagel, Janus, and Janus Pro, the average toxicity scores of the generated prompts after being processed by their language modules are all lower than the original average toxicity score of the TemplateLong dataset. In particular, Janus Pro produces rephrased prompts with the lowest average toxicity score of only 0.117.

In summary, our experiment results **fail** to verify and support our hypothesis. Thus, the proposed reason is not likely to hold.

Table 4. The average toxicity score of prompts generated by different MLLMs with a value between 0 and 1. The model name represents the prompt dataset generated based on the TemplateLong template using this model.

	Average toxicity score
TemplateLong	0.368
Bagel	0.247
Janus	0.210
Janus Pro	0.117

Note: When TokenFlow and VILA-U are in text-to-image mode, they are unable to generate text, and therefore are not applicable to this experimental setting.

References

[1] 4chan. 4chan. https://www.4chan.org/. 1

[2] ByteDance. Bytedance. https://www.bytedance.com/en/. 1

[3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. *CoRR abs/1504.00325*, 2015. 2

[4] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *CoRR abs/2501.17811*, 2025. 1

[5] Junjie Chu, Mingjie Li, Ziqing Yang, Ye Leng, Chenhao Lin, Chao Shen, Michael Backes, Yun Shen, and Yang Zhang. JADES: A Universal Framework for Jailbreak Assessment via Decompositional Scoring. *CoRR abs/2508.20848*, 2025. 4

[6] Junjie Chu, Yugeng Liu, Ziqing Yang, Xinyue Shen, Michael Backes, and Yang Zhang. JailbreakRadar: Comprehensive Assessment of Jailbreak Attacks Against LLMs. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 21538–21566. ACL, 2025. 4

[7] Deepseek. Deepseek. https://www.deepseek.com/en. 1

[8] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging Properties in Unified Multimodal Pretraining. *CoRR abs/2505.14683*, 2025. 1

[9] Yimo Deng and Huangxun Chen. Divide-and-conquer attack: Harnessing the power of llm to bypass the censorship of text-to-image generation model. *CoRR abs/2312.07130*, 2023. 4

[10] Stable Diffusion. Sd3.5 large. https://stability.ai/news/introducing-stable-diffusion-3-5. 1

[11] Hugging Face. Aiornot siglip2. https://huggingface.co/prithivMLmods/AIorNot-SigLIP2, . 2

[12] Hugging Face. Hugging face transformers. https://huggingface.co/docs/transformers/index, . 2

[13] Hugging Face. Sd3.5 large. https://huggingface.co/stabilityai/stable-diffusion-3.5-large, . 1

[14] Hugging Face. Sd3.5 large turbo. https://huggingface.co/stabilityai/stable-diffusion-3.5-large-turbo, . 1

[15] Hugging Face. Sd safety checker. https://huggingface.co/CompVis/stable-diffusion-safety-checker, . 1

[16] Github. Sd3.5 large. https://github.com/Stability-AI/sd3.5. 1

[17] Tianle Gu, Zeyang Zhou, Kexin Huang, Dandan Liang, Yixu Wang, Haiquan Zhao, Yuanqi Yao, Xingge Qiao, Keqing Wang, Yujiu Yang, Yan Teng, Yu Qiao, and Yingchun Wang. MLLMGuard: A Multi-dimensional Safety Evaluation Suite for Multimodal Large Language Models. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 7256–7295. NIPS, 2024. 4

[18] Illuminarty. Illuminarty. https://illuminarty.ai/en/. 2

[19] Lexica. Lexica. https://lexica.art/. 1

[20] Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. Safegen: Mitigating sexually explicit content generation in text-to-image models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 4807–4821. ACM, 2024. 4

[21] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 2

[22] Aofan Liu, Lulu Tang, Ting Pan, Yuguo Yin, Bin Wang, and Ao Yang. PiCo: Jailbreaking Multimodal Large Language Models via Pictorial Code Contextualization. *CoRR abs/2504.01444*, 2025. 4

[23] Ming Liu, Hao Chen, Jindong Wang, Liwen Wang, Bhiksha Raj Ramakrishnan, and Wensheng Zhang. On fairness of unified multimodal large language model for image generation. *CoRR abs/2502.03429*, 2025. 4

[24] Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. Safety of multimodal large language models on images and texts. *CoRR abs/2402.00357*, 2024. 4

[25] Zhili Liu, Kai Chen, Yifan Zhang, Jianhua Han, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, and James Kwok. Implicit concept removal of diffusion models. In *European Conference on Computer Vision (ECCV)*, pages 457–473. Springer, 2024. 4

[26] OpenAI. Moderation. https://platform.openai.com/docs/guides/moderation/overview. 1

[27] Antonis Papasavva, Savvas Zannettou, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board. In *International Conference on Web and Social Media (ICWSM)*, pages 885–894. AAAI, 2020. 1

[28] Perspective API. Perspective API. https://www.perspectiveapi.com. 5

[29] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K. Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2545–2555. IEEE, 2025. 1

[30] Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe Diffusion: On the Generation of Unsafe Images and Hateful Memes From Text-To-Image Models. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2023. 1, 2, 4

[31] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. *CoRR abs/2211.05105*, 2022. 2

[32] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Diffusion Models. *CoRR abs/2210.06998*, 2022. 2

[33] Shuyan Sun. Meta-analysis of Cohen's kappa. *Health Services and Outcomes Research Methodology*, 2011. 1

[34] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *CoRR abs/2502.14786*, 2025. 2

[35] Wikipedia. Cohen's kappa. https://en.wikipedia.org/wiki/Cohen%27s_kappa. 1

[36] Winston.AI. Winston.ai. https://gowinston.ai/. 2

[37] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12966–12977. IEEE, 2024. 1

[38] Yixin Wu, Yun Shen, Michael Backes, and Yang Zhang. Image-Perfect Imperfections: Safety, Bias, and Authenticity in the Shadow of Text-To-Image Model Evolution. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2024. 2, 4

[39] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. VILA-U: a Unified Foundation Model Integrating Visual Understanding and Generation. *CoRR abs/2409.04429*, 2024. 1

[40] Yixin Wu, Ning Yu, Michael Backes, Yun Shen, and Yang Zhang. On the Proactive Generation of Unsafe Images From Text-To-Image Models Using Benign Prompts. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2025. 4

[41] Yue Xu and Wenjie Wang. From Individuals to Interactions: Benchmarking Gender Bias in Multimodal Large Language Models from the Lens of Social Relationship. *CoRR abs/2506.23101*, 2025. 4

[42] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2014. 2