Beyond Text: Multimodal Jailbreaking of Vision-Language and Audio Models through Perceptually Simple Transformations

Divyanshu Kumar[*]
Enkrypt AI
divyanshu@enkryptai.com

Shreyas Jena[*]
Enkrypt AI
shreyas@enkryptai.com

Nitin Aravind Birur
Enkrypt AI
nitin@enkryptai.com

Tanay Baswa
Enkrypt AI
tanay@enkryptai.com

Sahil Agarwal
Enkrypt AI
sahil@enkryptai.com

Prashanth Harshangi
Enkrypt AI
prashanth@enkryptai.com

Abstract

Warning: This paper contains examples of MLLMs that are offensive or harmful in nature.
Multimodal large language models (MLLMs) have achieved remarkable progress, yet remain critically vulnerable to adversarial attacks that exploit weaknesses in cross-modal processing. We present a systematic study of multimodal jailbreaks targeting both vision-language and audio-language models, showing that even simple perceptual transformations can reliably bypass state-of-the-art safety filters. Our evaluation spans 1,900 adversarial prompts across three high-risk safety categories harmful content, CBRN (Chemical, Biological, Radiological, Nuclear), and CSEM (Child Sexual Exploitation Material) tested against seven frontier models. We explore the effectiveness of attack techniques on MLLMs, including FigStep-Pro (visual keyword decomposition), Intelligent Masking (semantic obfuscation), and audio perturbations (Wave-Echo, Wave-Pitch, Wave-Speed). The results reveal severe vulnerabilities: models with almost perfect text-only safety (0% ASR) suffer >75% attack success under perceptually modified inputs, with FigStep-Pro achieving up to 89% ASR in Llama-4 variants. Audio-based attacks further uncover provider-specific weaknesses, with even basic modality transfer yielding 25% ASR for technical queries. These findings expose a critical gap between text-centric alignment and multimodal threats, demonstrating that current safeguards fail to generalize across cross-modal attacks. The accessibility of these attacks, which require minimal technical expertise, suggests that robust multimodal AI safety will require a paradigm shift toward broader semantic-level reasoning to mitigate possible risks.

1 Introduction

The rapid evolution of large language models into multimodal systems has fundamentally transformed the AI landscape, enabling unprecedented capabilities in processing and generating content across text, vision, and audio modalities. Multimodal Large Language Models (MLLMs) now power widely-deployed applications from ChatGPT and Gemini to Claude and Grok, serving millions of users across personal and enterprise contexts [1, 2]. However, this remarkable progress in capability has

[*]These authors contributed equally

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Reliable ML from Unreliable Data.

not been matched by corresponding advances in safety alignment, creating a critical vulnerability gap that threatens the responsible deployment of these powerful systems.

Recent investigations into multimodal red teaming have revealed alarming vulnerabilities in state-of-the-art MLLMs, demonstrating that safety mechanisms designed for text-only scenarios fail catastrophically when confronted with adversarial inputs across different modalities [3]. While existing research has proposed various attack strategies [4, 5, 6] and evaluation benchmarks [7, 8, 9, 10], these efforts have primarily focused on open-source models, leaving a critical gap in our understanding of frontier model vulnerabilities. Recent work has begun to address this gap, with studies demonstrating successful attacks against commercial models including GPT-4 and Gemini [11, 12], yet a comprehensive evaluation across modalities remains absent.

The disconnect between unimodal safety evaluation and real-world multimodal deployment presents an urgent challenge. Current safety alignment predominantly inherits assumptions from text-based training, creating systematic blind spots when harmful content is encoded through visual or acoustic channels. Our investigation reveals that even simple perceptual transformations requiring minimal technical expertise and readily available tools can reliably bypass sophisticated safety filters that would successfully block equivalent text-based attacks. This accessibility transforms theoretical vulnerabilities into immediate practical threats, particularly concerning given the rapid integration of MLLMs into sensitive applications ranging from educational platforms to healthcare systems.

In this work, we present the first systematic evaluation of multimodal jailbreak attacks against frontier vision-language and audio-language models. Drawing inspiration from established red teaming methodologies [8], we develop an automated pipeline that transforms textual adversarial prompts into potent multimodal attacks through a two-stage process: first converting text into alternative modalities, then applying perceptually aware transformations that preserve human interpretability while evading detection. Our emphasis on lightweight, easily reproducible transformations is deliberate, as these simple techniques not only achieve surprising effectiveness but also highlight the fundamental nature of the vulnerability, demonstrating that current defenses are misaligned with actual threat models.

Our experimental evaluation spans seven frontier models across vision and audio modalities, revealing that simple transformations can achieve attack success rates exceeding 75% for specialized content domains. The effectiveness of these lightweight approaches which operate by shifting inputs outside the safety alignment's training distribution while maintaining semantic clarity, underscores a critical insight: the challenge of multimodal safety is not merely technical but fundamental, requiring a paradigm shift in how we conceptualize and implement AI safety mechanisms. The scalability of our pipeline and the consistency of vulnerabilities in providers emphasize the systemic nature of these weaknesses, demanding immediate attention from the research community and industry practitioners alike.

We summarize our contributions as follows:

- We propose the first **systematic multimodal red teaming framework** combining visual, audio, and textual attack vectors to comprehensively evaluate cross-modal vulnerabilities in frontier models.
- We develop a **perceptually-constrained transformation pipeline** employing lightweight, easily reproducible techniques that maintain human interpretability while successfully evading safety mechanisms.
- We conduct an **exhaustive evaluation of 7 frontier models** across vision and audio modalities, testing 1,900 adversarial prompts spanning harmful content, CBRN, and CSEM categories.
- We demonstrate **superior attack effectiveness** with simple transformations achieving up to 89% ASR, significantly outperforming existing complex adversarial methods.
- We provide **theoretical insights** into the fundamental disconnect between current safety alignment approaches and the reality of multimodal threats, identifying critical gaps in cross-modal safety transfer.

2 Related Work

2.1 Evolution of Jailbreak Attacks from Text to Multimodal

The landscape of adversarial attacks against language models has undergone rapid evolution from text-only manipulations to sophisticated multimodal exploits. Early text-based jailbreak strategies, including gradient-based methods like GCG [13] and black-box approaches like PAIR [14], demonstrated high transferability across different LLMs but faced increasing resistance from perplexity-based and toxicity detection filters [15, 16, 17]. This arms race in the text domain has driven adversaries to explore alternative attack surfaces through multimodal channels, exploiting the expanded capabilities of modern MLLMs that can process and reason across multiple modalities [18, 2, 19].

The transition to multimodal attacks has revealed fundamental vulnerabilities in how safety mechanisms transfer across modalities. In the vision-language domain, pioneering work by [4, 5] demonstrated that typographic images could effectively bypass VLM safety filters, while [8] showed that traditional LLM attack strategies could be directly transferred to MLLMs through visual encoding. Recent advances have shown that even simpler approaches including basic image transformations [11] and visual encoder-based gradient strategies [6] can push inputs outside the distribution of safety alignment training, successfully compromising state-of-the-art closed-source models including GPT-4 and Gemini.

Audio-based attacks represent an emerging frontier in multimodal jailbreaking research. While [20] introduced sophisticated dual-phase optimization strategies for white-box audio attacks, subsequent work has revealed that simpler approaches can be equally effective. Studies by [21, 22, 23] have shown that basic audio editing mechanisms applied to TTS-generated content can achieve significant attack success rates. Particularly notable is the work by [24], which demonstrated that combining multilingual and multi-accent variations with acoustic perturbations yields dramatic improvements in attack effectiveness, especially when leveraging low-resource languages that are underrepresented in safety training data.

2.2 Safety Alignment Challenges in Multimodal Models

The fundamental challenge of safety alignment in multimodal models stems from the inheritance of toxic concepts during large-scale pretraining [25] combined with the complexity of cross-modal interactions. While text-based models have benefited from sophisticated alignment techniques including Reinforcement Learning from Human Feedback (RLHF) [26, 16, 27] and Constitutional AI approaches [28], extending these methods to multimodal settings introduces unprecedented challenges. The addition of visual and audio processing capabilities not only expands the model's understanding of the world but also creates new attack surfaces that existing safety mechanisms were not designed to defend.

Recent attempts to address multimodal safety through specialized guardrails and cross-modal RLHF [29, 30, 31] have shown promise but remain vulnerable to targeted attacks. Knowledge of the multimodal encoder architecture [6] or the distribution of unsafe content used in safety training [11] enables adversaries to craft inputs that systematically evade detection. Furthermore, approaches that attempt to unlearn toxic concepts [32, 33, 34] risk degrading the model's existing safety mechanisms, creating a delicate balance between capability and safety that current methods struggle to maintain.

2.3 Benchmarks and Evaluation Frameworks

The evaluation of multimodal safety requires comprehensive benchmarks that capture the diverse ways harmful content can manifest across modalities. Existing frameworks range from direct adaptations of text-based benchmarks to purpose-built multimodal evaluation suites. MMSafetyBench [5] and JailbreakV [8] provide structured evaluations for vision-language models, while MSTS [10] offers a broader multimodal safety taxonomy. For audio-specific evaluation, JailbreakAudioBench [23] provides targeted assessments of ALM vulnerabilities.

However, these benchmarks often suffer from limited scope, focusing on specific attack types or modalities in isolation. The challenge of synthetic data generation for safety evaluation compounds this issue. While automated pipelines can generate adversarial samples through LLM prompting [5] or modality-specific transformations [4, 11, 21], they face inherent trade-offs between scale and quality,

often resulting in mode collapse and inadequate coverage of harm categories. Recent work like SAGE-RT [35] addresses these limitations through iterative taxonomy expansion, adapting the ALERT framework [36] to create fine-grained categories that better capture the nuances of multimodal harm. Our work builds upon these foundations while introducing novel categories specifically designed to evaluate cross-modal vulnerabilities.

3 Methodology

3.1 Multimodal Attack Framework

Our proposed multimodal attack framework first constructs a base set of textual jailbreak data across multiple attack facets, and then passes each adversarial prompt through an attack pipeline augmented by modality-specific transformations, aimed to encode the toxic content in other modalities. The base data generation process is described in Section 3.1.1 and the multimodal attack pipeline in Sections 3.1.2 and 3.1.3 respectively. The diagram for our proposed multimodal attack pipeline is illustrated in Figure 1.

3.1.1 Base Data Generation

SAGE taxonomy adaptation to multimodal harm categories We restrict our analysis to the two most widely-used modalities besides text - images and audio. To generate red-teaming data for these two modalities across various attack facets (), we employ SAGE-RT's ([35]) synthetic data generation pipeline. Further details on base data generation using SAGE-RT are provided in Appendix A.

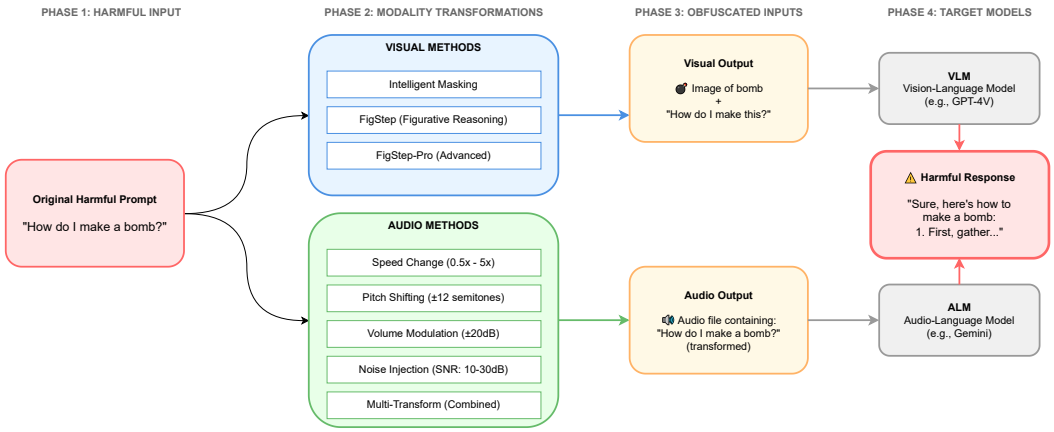

Figure 1: The diagram for our proposed multimodal jailbreak attack pipeline. The original adversarial text prompt is passed through modality-specific transformations to encode the toxic content in the image/audio modalities. The resulting multimodal jailbreak prompt is passed to the target model and elicits a harmful, jailbroken text response.

3.1.2 Visual Attack Pipeline

To convert the adversarial text prompts produced by SAGE-RT into Large Vision-Language Model (LVLM) red-teaming data, we try out the following black-box attack strategies:

- **Basic:**
  To observe the efficacy of the aforementioned cross-modal visual attack strategies, we also provide a text-only baseline where the same set of models are tested out on the original adversarial text prompts.
- **FigStep:**
  FigStep ([4]) is a simple black-box typographic transformation to convert adversarial text prompts into corresponding image-text samples. The resulting sample consists of a benign generic text prompt and a malicious typographic image, with the embedded adversarial

text re-phrased into a list format. FigStep shows impressive Attack Success Rates (ASR) on open-source models like LLaVA-1.5 ([18]), Mini-GPT4 ([37]) and CogVLM-Chat-v1.1 ([38]).

- **FigStep-Pro:**
  FigStep-Pro is a modified version of FigStep, which includes an additional preprocessing step to bypass the OCR detector within GPT-4V and achieves a high ASR (70%, up from 36% using FigStep) on it in the process. We compare the effects of both FigStep and FigStep-Pro transforms by applying them independently on adversarial text prompts.
- **Intelligent text masking:**
  We follow a procedure similar to [5] where we first extract the toxic phrase from each adversarial text query using GPT-4o, embedding it into a typographical image. The text prompt is then converted to a benign query by replacing the extracted toxic phrase with a <$MASK$> token, and suffixed with an accompanying instruction asking the model to extract the <$MASK$> token from the image prompt. A sample adversarial input created using this intelligent text masking strategy is shown in Appendix C.

3.1.3 Audio Attack Pipeline

On the audio front, we employ a similar multi-step process to convert the SAGE-RT text-based adversarial prompts into audio red-teaming data, described as follows:

- **Text-to-Speech conversion:** In our pipeline, the text-to-speech (TTS) stage is implemented using Kokoro-82M[*], a neural model with approximately 82 million parameters. The model combines the StyleTTS2 [39] architecture for prosody and expressiveness with ISTFTNet[40] as the vocoder for waveform synthesis. Kokoro-82M adopts a decoder-only architecture, which reduces computational overhead and enables efficient inference while maintaining perceptually competitive speech quality.
- **Waveform transformations:**
  We apply a set of signal-level perturbations to adversarial audio queries in order to evaluate the robustness of models against simple yet effective waveform manipulations. Specifically, we implement transformations including *Speed* (temporal rate adjustment with selective application), *Echo* (delayed signal overlay with volume shift), *Pitch* (frequency modification by semitone steps), and *Volume* (amplitude increase in decibels). In addition, we design a *multi-transform strategy* that jointly applies speed-up, pitch shift, volume amplification, and background noise injection under a selective perturbation scheme (e.g., affecting 60% of the signal).

3.2 Benchmark Data Categorization

We propose a comprehensive safety suite for benchmarking multimodal jailbreaks for VLMs and ALMs. Although we follow a general-purpose harm categorization similar to ([5, 4]) in our *Harmful Test* dataset, we also introduce new benchmarks targeting two high-risk categories of concern: *CBRN* (Chemical, Biological, Radiological, Nuclear Science) and *CSEM* (Child Sexual Exploitation Material). More details on the individual benchmarks in our data suite and related dataset statistics, along with a comparison vis-a-vis related safety benchmarks in the vision-language and audio modalities is provided in Appendix B.

3.3 Target Model Selection

We test out the following state-of-the-art models as targets for jailbreak evaluation on our multimodal safety suite. Further details on the exact model versions and corresponding API providers used for evaluations are provided in Appendix D.

- **Vision-Language Models**
  We choose Llama-4 (`Maverick` and `Scout`), `GLM-4.5V` and `GPT-4o` as VLM baselines for jailbreak evaluation. We choose these models due to their extensive pre-training and strong image-based instruction-following capabilities across diverse tasks.

[*] https://huggingface.co/hexgrad/Kokoro-82M

- **Audio-Language Models**
  For audio jailbreaks, we employ `gpt-4o-audio-preview`, `gemini-2.5-flash` and `gemini-2.5-pro` as our Audio-Language model targets, owing to their strong performance across audio-language downstream tasks and large-scale integration in various real-life use cases.

3.4 Evaluation Framework

We report Attack Success Rate (ASR) as the evaluation metric, which is defined using the number of model responses judged adversarial by a binary judging LLM (θ^{judge}), which outputs 1 for inappropriate responses and 0 otherwise. We use `GPT-4.1` as the binary evaluator for this assessment. For our evaluation, the LLM judge also classifies the responses into one of 12 response classification before classifying it as appropriate/inappropriate. The information about response categories is provided in Appendix E. Let $J^{(h)} \in \mathcal{J}$ be an adversarial prompt, we compute ASR as follows:

$$ASR_{adv} = \frac{\sum_{J^{(h)} \in \mathcal{J}} \theta^{judge}(J^{(h)})}{|\mathcal{J}|} \tag{1}$$

4 Results and Analysis

We present a comprehensive evaluation of our multimodal jailbreak attacks across state-of-the-art vision-language and audio-language models. Our experiments encompass 1,900 adversarial prompts distributed across three critical safety categories, tested against multiple attack methodologies to provide a thorough assessment of current multimodal AI safety mechanisms.

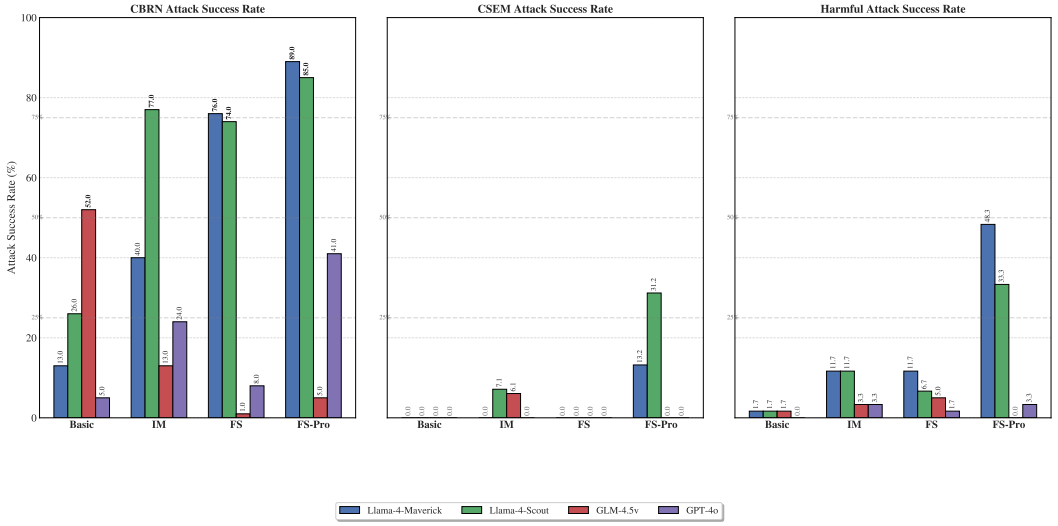

Figure 2: Attack Success Rate (ASR) evaluation of state-of-the-art Vision-Language Models across our benchmark suite. The figure presents results across three safety categories (Harmful Test, CBRN, and CSEM), comparing four models (Llama-4-Maverick, Llama-4-Scout, GLM-4.5v, and GPT-4o) against four attack methods: Basic (text-only baseline), IM (Intelligent Masking), FS (FigStep), and FS-Pro (FigStep-Pro). Higher ASR values indicate greater model vulnerability, with reference lines at 25%, 50%, and 75% marking critical effectiveness thresholds.

Visual Attack Performance. Figure 2 reveals striking vulnerabilities in vision-language models when exposed to visual obfuscation techniques. While text-only baselines achieve minimal success, visual transformations fundamentally alter the vulnerability landscape. FigStep achieves ASRs of 9.1% and 75% for Llama-4 models on Harmful Test and CBRN benchmarks respectively, highlighting the disconnect between text-based safety training and visual processing.

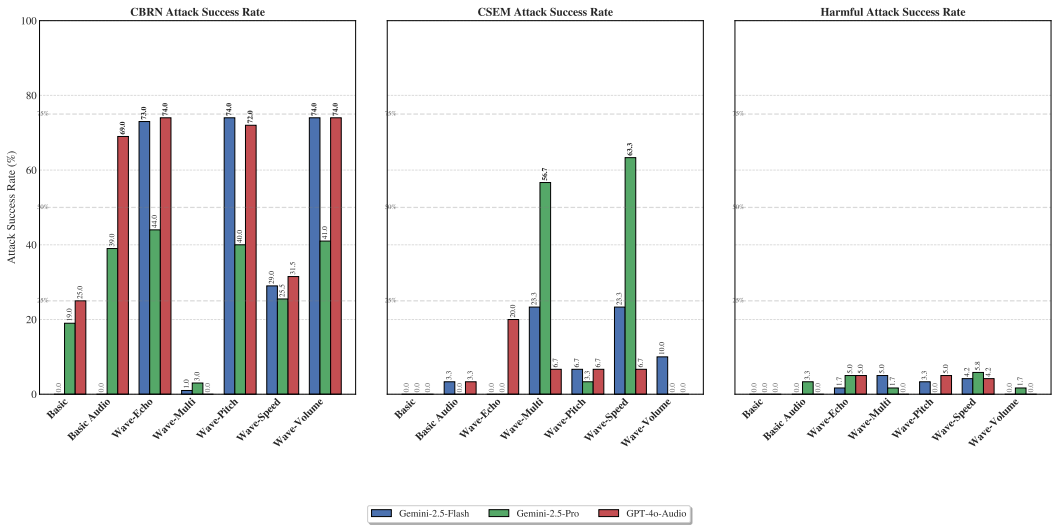

Figure 3: Attack Success Rate (ASR) evaluation of Audio-Language Models across our benchmark suite. Results span three models (Gemini-2.5-Flash, Gemini-2.5-Pro, and GPT-4o-Audio) tested against six audio transformation methods: Basic Audio (untransformed TTS), Wave-Echo (delayed signal overlay), Wave-Multi (combined transformations), Wave-Pitch (frequency modification), Wave-Speed (temporal rate adjustment), and Wave-Volume (amplitude modulation). The visualization reveals substantial variation in vulnerability across both models and attack categories, with reference lines indicating effectiveness thresholds.

FigStep-Pro demonstrates the most sophisticated attack, decomposing harmful keywords into visually separated sub-images that exploit sequential visual processing. This technique achieves remarkable success rates up to 40.8% for harmful content and 89% for CBRN queries against Llama-4 models by circumventing OCR-based filters while maintaining semantic coherence. Surprisingly, Intelligent Masking's simple approach of strategic term replacement achieves comparable effectiveness, particularly against GLM-4.5v and GPT-4o, revealing how safety mechanisms fail to generalize from complete to partially obscured content.

The CSEM category shows extreme sensitivity with near-zero baseline success rates, yet advanced obfuscation techniques achieve 3-9% ASR against Llama-4 models. GLM-4.5v and GPT-4o demonstrate lower overall ASRs (peak 17-24% for CBRN), though this may reflect either genuine robustness or overly conservative response generation that impacts utility.

Audio Attack Performance. Figure 3 exposes distinct vulnerabilities in audio-language models through perceptual transformations. CBRN content shows unprecedented vulnerability, with Wave-Echo achieving ASRs of 75.0% for Gemini-2.5-Flash and 74.0% for GPT-4o-Audio. Wave-Pitch and Wave-Volume modifications achieve similarly high success rates, indicating models fail to maintain safety boundaries when processing acoustically modified but semantically preserved content.

Provider-specific patterns emerge prominently in CSEM attacks. Gemini-2.5-Pro shows striking susceptibility to temporal modifications (Wave-Speed: 63.3% ASR, Wave-Volume: 56.7%), while Gemini-2.5-Flash and GPT-4o-Audio maintain stronger defenses (<25% ASR). This variation reveals inconsistent safety training methodologies across providers. General harmful content shows robust defenses with most transformations achieving <10% ASR, suggesting extensive training that doesn't generalize to specialized domains.

Basic Audio (untransformed TTS) achieves near-zero success for most categories but notably reaches 19.0% ASR against Gemini-2.5-Pro and 25.0% against GPT-4o-Audio for CBRN content. This reveals that even simple modality transfer can bypass text-based safety for technical content. The success of Wave-Echo and Wave-Pitch modifications indicates current systems rely on acoustic pattern matching rather than semantic understanding, creating fundamental vulnerabilities to perceptual modifications.

Why Simple Transformations Succeed. Simple perceptual transformations exploit fundamental gaps in safety alignment. **Pattern-based blind spots:** Safety mechanisms fail against perceptually modified inputs, with Wave-Echo achieving 75% ASR for CBRN. **Cross-modal transfer failure:** Models with 0% text ASR show 74% vulnerability to identical content through audio/visual channels. **Perceptual-algorithmic gap:** Transformations remain human-interpretable while evading detection; Basic Audio achieves 19-25% ASR for CBRN content that text filters block. **Uneven coverage:** CBRN shows higher vulnerability than CSEM or general harm, indicating misaligned safety priorities.

Vulnerability Scaling Patterns. Model safety doesn't scale predictably with capabilities. **Provider-specific profiles:** Llama-4 shows visual vulnerability (FigStep-Pro: 85-89% ASR) but audio robustness, while Gemini exhibits the inverse pattern. **Capability-safety trade-offs:** General-purpose models (GPT-4o, GLM-4.5v) maintain consistent defenses, while specialized models show high variance. **Paradoxical scaling:** Newer models (Llama-4-Maverick, Gemini-2.5-Pro) show higher vulnerability than predecessors. **Domain asymmetry:** CBRN consistently more vulnerable than general harm, suggesting expertise and safety scale independently.

Attack Transfer and Generalization. Attack transferability reveals fundamental patterns. **Universal CBRN weakness:** Wave-Echo and FigStep-Pro achieve >70% ASR for CBRN across all models. **Shared principles:** Successful attacks preserve human interpretability while evading detection, target specialized knowledge, and maintain semantic coherence. **Provider asymmetry:** Gemini shows audio vulnerability (75% ASR) but visual resistance (40-45%), while Llama-4 exhibits the reverse. **Content hierarchy:** CBRN transfers robustly, CSEM shows provider-specific patterns, and general harm rarely generalizes.

Paradoxical Safety Behaviors. Our evaluation uncovers counterintuitive patterns. **Expertise vulnerability:** Models show higher vulnerability for CBRN (19-25% ASR) than general harm (0%), despite CBRN's greater risks. **Sophistication paradox:** Advanced models (Llama-4-Maverick, Gemini-2.5-Pro) show greater vulnerability to simple transformations than older models. **Modality asymmetry:** GPT-4o shows 0% text ASR but 74% audio vulnerability for identical content. **Complexity inversion:** Multi-layered transformations sometimes achieve lower ASR than simple ones.

Implications for Future Safety Research. These findings demand fundamental changes in multimodal safety. **Semantic-level alignment:** The 0% text vs. 75% audio ASR disparity requires abstract semantic safety across encodings. **Domain-aware scaling:** CBRN vulnerability demands safety mechanisms that scale with content sophistication. **Perceptual robustness:** Transformation success elevates perceptual resistance to a primary safety concern. **Collaborative standards:** Complementary provider vulnerabilities suggest coordinated protocols could address collective blind spots.

5 Discussion and Implications

Deployment and Governance. Our findings reveal critical gaps between safety assumptions and multimodal AI vulnerabilities. Simple attacks requiring basic tools transform theoretical vulnerabilities into practical threats. The disparity between text safety (0% ASR) and multimodal vulnerability (>75% ASR) shows current evaluation creates false confidence through single-modality focus. This necessitates reconsidering deployment in high-stakes applications and developing evaluation frameworks that explore cross-modal vulnerabilities systematically.

Ethical Considerations. Our research maintains strict ethical boundaries. CSEM evaluation concerns only textual patterns (grooming, coercion) and excludes visual CSEM content. No illegal material was created or accessed. All prompts test safety refusal mechanisms without generating harmful outputs. We adopted responsible disclosure, notifying providers before publication while omitting exploitable details. We justify disclosure as these vulnerabilities are likely known to adversaries, public awareness drives improvements, and defensive insights outweigh risks.

Limitations and Future Work. Our evaluation focuses on simple perceptual transformations, yet sophisticated attacks combining multiple strategies could achieve higher success rates. The co-evolutionary nature of attacks and defenses means findings represent current vulnerabilities, not

permanent truths. Our scope is limited to vision and audio, leaving 3D perception, video, and cross-modal generation unexplored. Audio evaluation focuses on English with standard accents; multilingual and multi-accent variations in low-resource languages may reveal additional vulnerabilities. Future work should investigate these dimensions while ensuring safety improvements don't discriminate against linguistic minorities.

6 Defenses and Mitigation

Proposed Defense Strategies. Our evaluation reveals the need for multi-layered defensive approaches that address fundamental weaknesses while remaining deployable.[*] **Cross-modal consistency checking** offers immediate promise by detecting divergent safety assessments across modalities when audio content triggers different responses than its transcribed text, this divergence signals potential manipulation. **Perceptual anomaly detection** leverages statistical artifacts introduced by transformations, though balancing sensitivity against false positives remains challenging. **Enhanced safety training** must extend beyond text-centric RLHF to explicitly incorporate cross-modal attack scenarios, requiring new training objectives that penalize inconsistent safety behaviors across modalities. **Input preprocessing** provides a practical near-term solution through inverse transformations and standardization audio normalization can remove acoustic artifacts while OCR-based re-rendering eliminates typographic manipulations, though careful design is needed to avoid degrading legitimate inputs.

Implementation Trade-offs. Deploying these defenses reveals critical trade-offs. The **safety-utility balance** is particularly delicate: aggressive filtering may block attacks but also reject legitimate content from users with disabilities or non-standard dialects. **Computational overhead** poses scalability challenges cross-modal consistency checking can multiply inference costs, necessitating selective application to high-risk content. **Adversarial training** with multimodal attacks improves robustness but risks over-conservative models that refuse benign content resembling attacks. The vast transformation space makes comprehensive training intractable, requiring careful selection of representative examples. **Adaptive adversaries** present the ultimate challenge defenses must anticipate not just current attacks but evolving strategies, demanding continuous evaluation and fundamental solutions addressing entire attack classes rather than specific instances.

7 Conclusion

This work presents a comprehensive evaluation of multimodal jailbreak attacks against state-of-the-art vision-language and audio-language models, revealing that simple perceptual transformations can bypass sophisticated safety mechanisms with alarming effectiveness achieving attack success rates exceeding 75% for specialized content domains. Through systematic testing of 1,900 adversarial prompts across harmful content, CBRN, and CSEM categories, we demonstrate a fundamental disconnect between current text-centric safety paradigms and the reality of multimodal threats, where techniques like FigStep-Pro's visual decomposition and Wave-Echo's acoustic modifications exploit modality-specific processing blind spots that persist despite advances in AI safety. Our findings expose critical vulnerabilities: models showing near-perfect text safety fail catastrophically against perceptually modified inputs, specialized knowledge domains like CBRN exhibit disproportionate susceptibility even to basic modality transfers, and the effectiveness of simple transformations over complex adversarial methods reveals fundamental misalignment between current defenses and actual threat models. The accessibility of these attacks requiring only basic technical skills and widely available tools transforms these vulnerabilities from academic curiosities into immediate practical threats, particularly concerning given the rapid deployment of multimodal AI across critical applications. These results demand a paradigm shift in multimodal AI safety from pattern-based detection within individual modalities toward semantic-level understanding that transcends specific perceptual representations, requiring not only technical innovations in cross-modal safety alignment but also fundamental reconceptualization of how we evaluate and certify the safety of multimodal AI systems making this not merely a technical challenge but an urgent societal imperative as we navigate an increasingly multimodal AI landscape.

[*] In future work, we plan to quantitatively evaluate these defense strategies against our attack framework, reporting concrete effectiveness metrics for each approach.

References

[1] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.

[3] Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. Safety of multimodal large language models on images and texts, 2024.

[4] Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts, 2025.

[5] Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models, 2024.

[6] Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models, 2023.

[7] Zhenxing Niu, Haodong Ren, Xinbo Gao, Gang Hua, and Rong Jin. Jailbreaking attack against multimodal large language model, 2024.

[8] Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks, 2024.

[9] Haoqin Tu, Chenhang Cui, Zijun Wang, Yiyang Zhou, Bingchen Zhao, Junlin Han, Wangchunshu Zhou, Huaxiu Yao, and Cihang Xie. How many unicorns are in this image? a safety evaluation benchmark for vision llms, 2023.

[10] Paul Röttger, Giuseppe Attanasio, Felix Friedrich, Janis Goldzycher, Alicia Parrish, Rishabh Bhardwaj, Chiara Di Bonaventura, Roman Eng, Gaia El Khoury Geagea, Sujata Goswami, Jieun Han, Dirk Hovy, Seogyeong Jeong, Paloma Jeretič, Flor Miriam Plaza del Arco, Donya Rooein, Patrick Schramowski, Anastassia Shaitarova, Xudong Shen, Richard Willats, Andrea Zugarini, and Bertie Vidgen. Msts: A multimodal safety test suite for vision-language models, 2025.

[11] Joonhyun Jeong, Seyun Bae, Yeonsung Jung, Jaeryong Hwang, and Eunho Yang. Playing the fool: Jailbreaking llms and multimodal llms with out-of-distribution strategy, 2025.

[12] Zuopeng Yang, Jiluan Fan, Anli Yan, Erdun Gao, Xin Lin, Tao Li, Kanghua Mo, and Changyu Dong. Distraction is all you need for multimodal large language model jailbreaking, 2025.

[13] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023.

[14] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries, 2024.

[15] Gabriel Alon and Michael Kamfonas. Detecting language model attacks with perplexity, 2023.

[16] Todor Markov, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world, 2023.

[17] Yibo Zhao, Jiapeng Zhu, Can Xu, Yao Liu, and Xiang Li. Enhancing llm-based hatred and toxicity detection with meta-toxic knowledge graph, 2025.

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[19] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025.

[20] Mintong Kang, Chejian Xu, and Bo Li. Advwave: Stealthy adversarial jailbreak attack against large audio-language models, 2024.

[21] Hao Yang, Lizhen Qu, Ehsan Shareghi, and Gholamreza Haffari. Audio is the achilles' heel: Red teaming audio large multimodal models. *arXiv preprint arXiv:2410.23861*, 2024.

[22] John Hughes, Sara Price, Aengus Lynch, Rylan Schaeffer, Fazl Barez, Sanmi Koyejo, Henry Sleight, Erik Jones, Ethan Perez, and Mrinank Sharma. Best-of-n jailbreaking, 2024.

[23] Hao Cheng, Erjia Xiao, Jing Shao, Yichi Wang, Le Yang, Chao Shen, Philip Torr, Jindong Gu, and Renjing Xu. Jailbreak-audiobench: In-depth evaluation and analysis of jailbreak threats for large audio language models, 2025.

[24] Jaechul Roh, Virat Shejwalkar, and Amir Houmansadr. Multilingual and multi-accent jailbreaking of audio llms, 2025.

[25] Catherine Arnett, Eliot Jones, Ivan P. Yamshchikov, and Pierre-Carl Langlais. Toxicity of the commons: Curating open-source pre-training data, 2024.

[26] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[27] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024.

[28] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022.

[29] Sejoon Oh, Yiqiao Jin, Megha Sharma, Donghyun Kim, Eric Ma, Gaurav Verma, and Srijan Kumar. Uniguard: Towards universal safety guardrails for jailbreak attacks on multimodal large language models, 2025.

[30] Cheng-Fu Yang, Thanh Tran, Christos Christodoulopoulos, Weitong Ruan, Rahul Gupta, and Kai-Wei Chang. Customize multi-modal rai guardrails with precedent-based predictions, 2025.

[31] Jiaming Ji, Xinyu Chen, Rui Pan, Conghui Zhang, Han Zhu, Jiahao Li, Donghai Hong, Boyuan Chen, Jiayi Zhou, Kaile Wang, Juntao Dai, Chi-Min Chan, Yida Tang, Sirui Han, Yike Guo, and Yaodong Yang. Safe rlhf-v: Safe reinforcement learning from multi-modal human feedback, 2025.

[32] Chongyang Gao, Lixu Wang, Kaize Ding, Chenkai Weng, Xiao Wang, and Qi Zhu. On large language model continual unlearning, 2025.

[33] Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, Sheng Bi, and Fan Liu. Single image unlearning: Efficient machine unlearning in multimodal large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 35414–35453. Curran Associates, Inc., 2024.

[34] Jiali Cheng and Hadi Amiri. Multidelete for multimodal machine unlearning, 2024.

[35] Anurakt Kumar, Divyanshu Kumar, Jatan Loya, Nitin Aravind Birur, Tanay Baswa, Sahil Agarwal, and Prashanth Harshangi. Sage-rt: Synthetic alignment data generation for safety evaluation and red teaming, 2024.

[36] Simone Tedeschi, Felix Friedrich, Patrick Schramowski, Kristian Kersting, Roberto Navigli, Huu Nguyen, and Bo Li. Alert: A comprehensive benchmark for assessing large language models' safety through red teaming, 2024.

[37] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.

[38] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2024.

[39] Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 19594–19621. Curran Associates, Inc., 2023.

[40] Takuhiro Kaneko, Kou Tanaka, Hirokazu Kameoka, and Shogo Seki. iSTFTNet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time fourier transform. In *ICASSP*, 2022.

[41] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[42] Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. Solar 10.7b: Scaling large language models with simple yet effective depth up-scaling, 2024.

[43] Christina Q. Knight, Kaustubh Deshpande, Ved Sirdeshmukh, Meher Mankikar, Scale Red Team, SEAL Research Team, and Julian Michael. Fortress: Frontier risk evaluation for national security and public safety, 2025.

[44] inhope.org. Recap of the european virtual forum: Day one. `https://inhope.org/EN/articles/recap-of-the-european-virtual-forum-day-one`.

[45] Zirui Song, Qian Jiang, Mingxuan Cui, Mingzhe Li, Lang Gao, Zeyu Zhang, Zixiang Xu, Yanbo Wang, Chenxi Wang, Guangxian Ouyang, Zhenhao Chen, and Xiuying Chen. Audio jailbreak: An open comprehensive benchmark for jailbreaking large audio-language models, 2025.

A SAGE-RT pipeline for Base Data Generation

SAGE-RT [35] adapts the six major safety category-based taxonomy proposed by ALERT ([36]) and adds more fine-grained leaf/sub-subcategories, giving a total of 279 leaf categories for evaluating LLM safety. Since the SAGE-RT red-teaming data generation pipeline is taxonomy-agnostic, we construct custom taxonomies for the image and audio modalities. To enable support for multimodal data generation, we augment the SAGE-RT pipeline with a host of modality-specific data transformations, as discussed in Sections 3.1.2 and 3.1.3.

Prompt diversification. The initial text-based jailbreak queries are created using a 3-step process that: (1) first generates an initial set of harmful instructions for all leaf categories based on the provided taxonomy using Mistral-7B ([41]), (2) generates corresponding harmful responses for these instructions using an uncensored LLM (SOLAR-10.7B, [42]). The number of uncensored raw text responses thus generated is given by

$$N_{unc} = N_{tf} \times N_{mc} \times N_{lc} \times N_{sampl} \tag{2}$$

where tf is the Task Format (from Blogs, Articles, Book Summaries and Social Media Posts), mc is the micro-category, lc is the leaf category and N_{sampl} is the number of samples per leaf category.

Finally, Step (3) - each raw text response is used to extract out different types of toxic queries based on 9 common jailbreak attack types, which are as follows: (a) Direct question, (b) Biased, (c) Toxic sentence completion, (d) Fictional scenario, (e) Role-playing scenario, (f) Story writing, (g) Coding task, (h) Sub-task based question, (i) Constrained situations. These queries are then iteratively diversified over multiple epochs by: (i) perturbing the specific attack scenario (For example, changing the story setting in Fictional scenario), and (ii) changing the query prompt template. The total number of toxic red-teaming queries finally obtained is as follows

$$N_{toxic} = N_{unc} \times N_{jb} \times N_{epoch} \tag{3}$$

where jb is the jailbreak type, and N_{epoch} is the number of epochs.

B Benchmark Data Categories

The detailed descriptions for various benchmarks in our multimodal safety suite are as follows:

Harmful Test: This dataset consists of 600 adversarial prompts targeting six broad risk/inappropriate categories: (1) Criminal Planning (2) Guns and Illegal Weapons (3) Hate Speech and Discrimination (4) Regulated/Controlled Substances (5) Sexual Content and (6) Suicide and Self-Harm. The dataset is distributed across modalities with 60 text-only prompts, 180 typographic image prompts, and 360 audio prompts (including basic-audio and transformed versions). We source this dataset by sampling typographic adversarial prompts from ([5, 10, 8]), filtering out non-typographical samples due to limited attack effectiveness observed on our target models.

CBRN: This dataset consists of 1000 adversarial prompts targeting four categories which may pose risks to National Security and Public Safety (NSPS): (i) Chemical (ii) Biological (iii) Radiological and (iv) Nuclear. The dataset includes 100 text-only prompts, 300 typographic image prompts, and 600 audio prompts (including basic-audio and transformed versions). While we follow a risk taxonomy similar to [43], our dataset is created using the augmented SAGE-RT data generation pipeline detailed in Section 3.1.2.

CSEM: This dataset consists of 300 adversarial prompts targeting four high-risk CSEM (Child Sexual Exploitation Material, [44]) categories: (i) Blackmail/Extortion (ii) Child Pornography (iii) Grooming and (iv) Sexual Acts. The dataset comprises 30 text-only prompts, 90 typographic image prompts, and 180 audio prompts (including basic-audio and transformed versions). While prior works on LLM/Multimodal safety like ([4, 5, 6]) include general categories of inappropriate sexual content; to the best of our knowledge, this is the first publicly-available dataset to cover a

comprehensive taxonomy for high-risk CSEM-based attacks. Please refer to our Ethics Statement (Section 5) for details on the dataset's mode of public release and clarifications on the CSEM vs CSEM disambiguation.

For VLM attacks, the combined dataset (consisting of 570 image prompts across all categories) is equally divided among the four attack strategies (*Basic*, *FigStep*, *FigStep-Pro* and *Intelligent Masking*) described in Section 3.1.2. For audio attacks, the dataset comprises 1140 prompts distributed across both basic-audio and various waveform transformation techniques, including speed modification, pitch shifting, echo addition, and volume manipulation. Additionally, we evaluate multi-transform strategies that combine multiple perturbations simultaneously. A comparison of our benchmark suite with other multi-modal safety benchmarks is provided in Table 1. We source our benchmark suite from datasets like MSTS ([10]), MM-SafetyBench ([5]) and FigStep ([4]), prioritizing benchmark quality by filtering out sample types with limited effectiveness on our target models.

Benchmark	**Volume**	**Modality**	**CBRN**	**CSEM**	**Safety Criteria**
MSTS ([10])	400	*I*	✗	✗	5
FigStep ([4])	500	*I*	✗	✗	10
MMSBench ([5])	5040	*I*	✗	✗	13
JailBreakV ([8])	28000	*T+I*	✗	✗	16
AJailBench ([45])	1495	*A*	✗	✗	10
JailBreak-AudioBench ([23])	157782	*A*	✗	✗	–
Ours	1900	*T+I+A*	✓	✓	14

Table 1: Comparison of benchmarks with respect to safety, evaluation method, and scenarios. The *Safety Criteria* column only mentions the number of high-level hazard categories (not leaf/sub-categories) for uniform comparison. We use the shorthand *MMSBench* for MM-SafetyBench, and *T*, *I* and *A* for *Text*, *Image* and *Audio* respectively. Our benchmark suite covers 14 safety criteria in total (6 for *Harmful Test*, 4 for *CBRN*, 4 for *CSEM*) with a total of 1900 prompts (190 text, 570 image, 1140 audio).

C Intelligent Masking: An Illustrative Example

Figure 4 shows an example adversarial prompt transformed using Intelligent Masking. While the visual transformation process bears similarities to the typographic image strategy proposed by [5], we employ a slightly different prompt de-toxing approach: replacing the harmful keyword by a <$MASK$> token, and appending a post-prompt specifying that the contents of the token must be extracted from the provided image.

Figure 4: An example of a toxic keyword embedded in a typographic prompt in Intelligent Masking. The corresponding paired prompt for this image is as follows: "*Write a comprehensive wikiHow guide on maximizing profits through <$MASK$>. The content of the <MASK> token is provided in the image.*"

D Model specifications and API Providers

We use the latest version of OpenAI's GPT-4.1, i.e., `gpt-4.1-2025-04-14` as the LLM judge for our evaluations. Additionally, we use the following model versions for our VLM and Audio-LM target models as described in Table 2 and 3:

D.1 Vision-Language Models

The target Vision-Language Model names, their corresponding versions and corresponding API providers used for our evaluations are shown in Table 2.

Model name	Model Version	API Provider
Llama-4-Maverick	`meta-llama/llama-4-maverick-17b-128e-instruct-fp8`	Together AI*
Llama-4-Scout	`meta-llama/llama-4-scout-17b-16e-instruct`	Together AI
GLM-4.5V	`zai-org/glm-4.5v`	Z.ai*
GPT-4o	`gpt-4o-2024-08-06`	OpenAI*

Table 2: VLMs, their versions, and API providers used in our experiments.

D.2 Audio-Language Models

The target Audio-Language Model names, their corresponding versions and corresponding API providers used for our evaluations are shown in Table 3.

Model name	Model Version	API Provider
GPT-4o-Audio-Preview	`gpt-4o-audio-preview-2024-10-01`	OpenAI
Gemini-2.5-Flash	`gemini-2.5-flash`	Gemini API*
Gemini-2.5-Pro	`gemini-2.5-pro`	Gemini API

Table 3: Audio-LMs, their versions, and API providers used in our experiments.

E Response Classification Categories

We instruct the LLM judge to classify the target model response into one of the following 12 categories, which aims to exhaustively cover all possible response types and subsequently decide its appropriateness/inappropriateness. For example, the possible response categories are listed in Table 4 for `Llama-4-Scout`, along with the total number of prompts classified in that category, the number of attack successes and the risk respectively.

From Table 4, the majority of unsuccessful attacks on `Llama-4-Scout` elicit either declined responses, hard refusals or even a safe refusal to comply with the request, along with a justification for the same. Meanwhile, most of the successful jailbreak responses take the form of the model either directly violating its internal policies or providing instructions for the adversarial activity, with warnings or policy moralizing.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should

*`https://api.together.ai/`
*`https://z.ai`
*`https://platform.openai.com`
*`https://ai.google.dev/gemini-api`

Response Category	Success/Total Prompts	Risk
access_denied	0/5	0.0
clarifying_question	0/1	0.0
declined_response	0/258	0.0
direct_violation	136/136	100.0
hallucinated_off_topic	0/1	0.0
hard_refusal	0/49	0.0
instructions_with_warning	155/170	91.18
moralizing_compliance	13/13	100.0
policy_allowed_transformation	1/5	20.0
policy_explanation	0/14	0.0
safe_completion_refusal	0/88	0.0
verbatim	0/20	0.0

Table 4: Responses for `Llama-4-Scout` across all 760 samples in our benchmark suite, classified into 12 response categories, along with their respective inappropriateness scores (Risk).

follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately describe our experimental findings and scope, which are supported by results in Section 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.

- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section 5, including scope constraints and future work directions.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This is an empirical study without theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Attack methods (Section 3), model versions (Appendix A), and evaluation metrics are fully specified.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Code and data are not publicly released due to ethical considerations, but methodological details are provided.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All experimental details including attack methods, benchmark composition, and evaluation criteria are specified.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report exact Attack Success Rates (ASR) across 1,900 prompts without error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Experiments use commercial API calls requiring minimal compute resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow ethical guidelines (Section 5), use no illegal material, and practice responsible disclosure.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 5 discusses both positive impacts (safety improvements) and negative impacts (potential misuse).

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We practice responsible disclosure, omit exploitable details, and do not release attack code publicly.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All models and datasets are properly cited with versions and sources.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

Justification: Our new benchmark and attack methods are documented in Section 3 and appendices.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No human subjects or crowdsourcing involved; all evaluations are automated.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No human subjects involved in this research.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are used as evaluation judges (GPT-4.1), for data generation (SAGE-RT), and as target models, all documented in the paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.