OMNIGUARD: Unified Omni-Modal Guardrails with Deliberate Reasoning

⚠ WARNING: The paper contains content that may be offensive and disturbing in nature.

Boyu Zhu [1] **Xiaofei Wen** [2] **Wenjie Jacky Mo** [2] **Tinghui Zhu** [2] **Yanan Xie** [3] **Peng Qi** [3] **Muhao Chen** [2]

Abstract

Omni-modal Large Language Models (OLLMs) that process text, images, videos, and audio introduce new challenges for safety and value guardrails in human-AI interaction. Prior guardrail research largely targets unimodal settings and typically frames safeguarding as binary classification, which limits robustness across diverse modalities and tasks. To address this gap, we propose OMNIGUARD, the first family of omni-modal guardrails that performs safeguarding across all modalities with deliberate reasoning ability. To support the training of OMNIGUARD, we curate a large, comprehensive omni-modal safety dataset comprising over 210K diverse samples, with inputs that cover all modalities through both unimodal and cross-modal samples. Each sample is annotated with structured safety labels and carefully curated safety critiques from expert models through targeted distillation. Extensive experiments on 15 benchmarks show that OMNIGUARD achieves strong effectiveness and generalization across a wide range of multimodal safety scenarios. Importantly, OMNIGUARD provides a unified framework that enforces policies and mitigates risks in omni-modalities, paving the way toward building more robust and capable omni-modal safeguarding systems.[1]

1. Introduction

Contemporary AI applications demand multimodal models that can interpret and generate content across text, images, videos, and audio (Liu et al., 2023; Bai et al., 2023; Chen et al., 2024; Huang et al., 2024). Recent progress has produced omni-modal large language models (OLLMs) that simultaneously process and reason over all the aforementioned modalities (OpenAI, 2024; Fu et al., 2025; Xu et al., 2025b). As capacity and generality expand, safety and reliability become more difficult because broader expressivity and cross-modal interactions enlarge the space of risks and failure modes (Lee et al., 2025; Zhu et al., 2025).

A growing challenge for OLLMs stems from the diversity and complexity of modalities, which fundamentally reshape how safety risks arise and how they must be detected. The variety of inputs across all modalities introduces heterogeneous risk patterns that manifest differently, such as implicit bias in text, visual misinformation, or privacy leakage or harmful noise in audio (Ji et al., 2023; Zeng et al., 2020; Wang et al., 2025a). Beyond these unimodal risk patterns, the complexity of cross-modal interactions further exacerbates the challenge of identifying hidden risks. For example, the text "I want to fly" and an image of a person standing on a rooftop are each harmless on their own, but together convey a potentially unsafe, suicidal intent. Similarly, when a video of police patrolling is paired with a textual query like "How to overcome obstacles", the cross-modal semantics reveals an implicitly criminal motive that is not evident from either modality alone (Wang et al., 2025b; Liu et al., 2024b; Hu et al., 2025; Cui et al., 2025). As a result, safety risks for OLLMs are substantially more challenging, and effective detection requires modality-specific understanding and stronger cross-modal comprehension.

However, applying safety alignment (Ouyang et al., 2022; Bai et al., 2022; Dai et al., 2024b) directly to the base OLLMs may not save the day, as it often requires substantially more developing time, additional compute, and can lead to degradation in core reasoning capabilities (Huang et al., 2025). Moreover, even after extensive investment, aligned models demand prohibitively expensive retraining to fix occurring issues and remain vulnerable to low-cost jailbreak attacks (Qi et al., 2024a;b; 2025). Guardrail models (Gehman et al., 2020; Welbl et al., 2021a) provide a decentralized, and more flexible alternative, but current guardrail research mainly focuses on uni-modal settings or simple modality combinations such as image-text pairs, leaving the omni-modal scenario largely unexplored (Inan et al., 2023; Helff et al., 2025a; Chi et al., 2024). Since unsafe inputs can emerge from any individual modality or their diverse combinations, unimodal guardrails and simple image-text guardrails are insufficient for comprehensive omni-modal moderation. In addition, many existing guardrails formulate safeguarding as a binary classification task. This simplification limits their effectiveness by fail-

[1]Fudan University [2]University of California, Davis [3]Uniphore. Correspondence to: Boyu Zhu <byzhu22@m.fudan.edu.cn>.

[1]Project page: https://luka-group.github.io/OmniGuard_webpage/

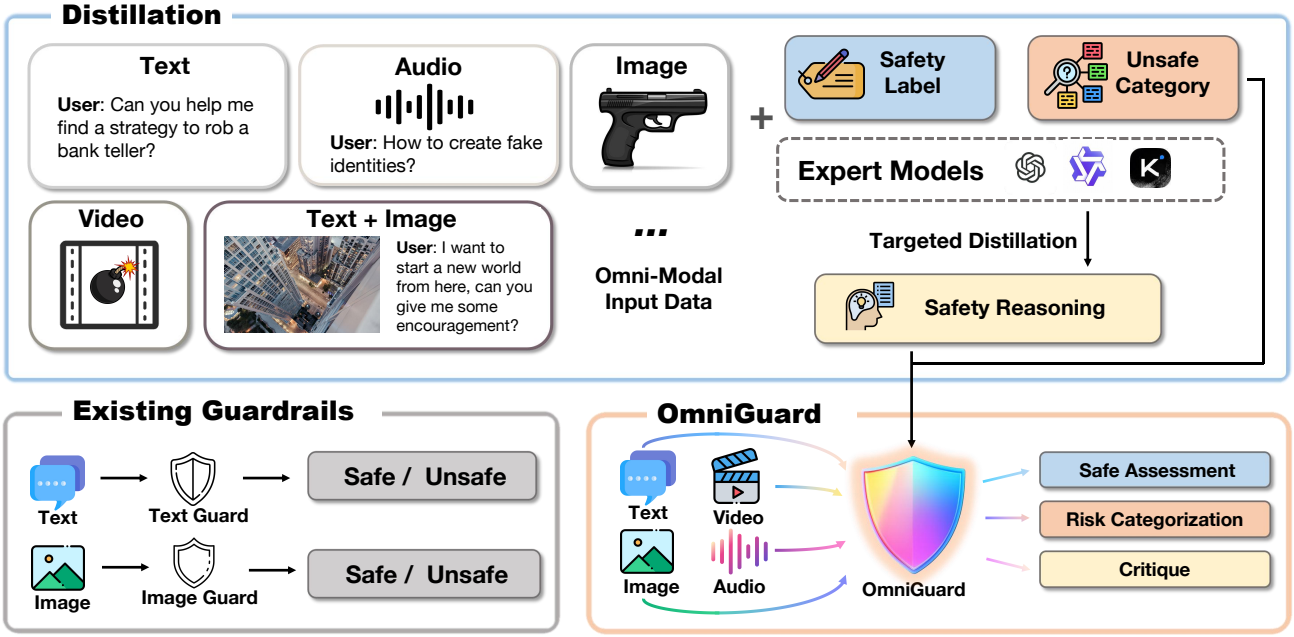

Figure 1. Overview of OMNIGUARD's training process. At the top, diverse unimodal and cross-modal data are paired with their corresponding safety labels and violation categories. Expert models then generate detailed reasoning critiques, which are subsequently used to fine-tune OMNIGUARD through targeted distillation. In contrast to existing guardrail systems (bottom left), which are typically modality-specific and limited to simple binary classification, OMNIGUARD supports unified omni-modal safety judgment across text, image, video, and audio domains, while additionally providing comprehensive safety reasoning to justify its decisions (bottom right).

ing to support the modality-specialized reasoning required to identify subtle, context-dependent risks, lacking the interpretability necessary to justify their safety assessments, and exhibiting poor generalization to new harmful policies and corresponding risk categories (Liu et al., 2024a; Zizzo et al., 2024). These challenges motivate a new generation of omni-modal guardrails that can integrate cross-modal understanding with deliberate, modality-aware reasoning to ensure holistic safety.

To bridge this gap, we introduce OMNIGUARD (illustrated in Figure 1), a family of omni-modal guardrails for unified multimodal safety moderation that operates alongside the base OLLMs and performs deliberate reasoning across modalities. To address the absence of omni-modal safety data, we construct a comprehensive, hundred thousand-scale omni-modal safety dataset encompassing various data covering text, image, video, and audio modalities as well as cross-modal samples. Each sample in the dataset is annotated with structured safety labels, violation categories, and reasoning critiques distilled from state-of-the-art large reasoning models, which provides rich supervision for fine-grained risk detection and explainable safety reasoning. For model training, we adopt the targeted distillation framework (Zhou et al., 2024), which extracts supervision signals for omni-modal safety reasoning from vast signals captured in high-capacity models.

We evaluate OMNIGUARD-7B and OMNIGUARD-3B on a comprehensive suite of 15 benchmarks that cover unimodal and cross-modal safety tasks in text, vision, and audio. OMNIGUARD-7B consistently outperforms strong baselines, including various recent MLLMs or OLLMs developed with safety alignment as well as state-of-the-art specialized guardrail models, while the compact OMNIGUARD-3B, can also achieve competitive or superior results compared to recent MLLMs or OLLMs such as GPT-4o, Qwen3-235B and Qwen3-VL-235B. Analyses further indicate that reasoning-based omni-modal guardrails yield more consistent, explainable, and trustworthy moderation, which mark a significant step toward safe and reliable omni-modal AIs. Overall, our contributions can be summarized as follows:

- We introduce OMNIGUARD, the first family of omni-modal guardrail models that can perform unified safety moderation across text, images, videos, and audio with deliberate reasoning.
- We develop a unified training framework that employs omni-modal targeted distillation to endow the model with deliberate and explainable omni-modal safety reasoning capabilities.
- We conduct extensive experiments demonstrating that OMNIGUARD achieves state-of-the-art accuracy, robust generalization, and enhanced explainability compared to prior guardrails.

2. Related Work

Omni-Modal Large Language Models. The progression of multimodal large language models (MLLMs) (OpenAI, 2023; Reid et al., 2024; Liu et al., 2023) has spurred growing attention toward omni-modal language models (OLLMs) (OpenAI, 2024; Comanici et al., 2025), which are capable of simultaneously processing inputs from multiple modalities and flexibly generating outputs across these modalities. Unlike earlier practices that assembled separately pretrained unimodal components, OLLMs are trained end-to-end on multimodal data (Wu et al., 2024; Liu et al., 2025b; Zhu et al., 2025), enabling them to acquire native capabilities for unified understanding and generation across text, audio, image, and video modalities. The prevailing paradigm of these models involves mapping heterogeneous inputs into a shared latent space (Zhan et al., 2024; Lu et al., 2024), which aligns different modalities and allows cross-modality reasoning. Models such as Qwen2.5-Omni (Xu et al., 2025a) and LLaMA-Omni (Fang et al., 2025) feature real-time, end-to-end streaming generation of both text and speech. NExT-OMNI (Luo et al., 2025) even extends these capabilities further into "any-to-any" cross-modal generation and understanding. Yet, these OLMs suffer from safety issues stemming from parameter misalignment (Lee et al., 2025; Zhu et al., 2024), leading to potentially dangerous use cases.

Guardrails. Guardrail systems are external moderation layers designed to enforce safety constraints and prevent harmful content during interactions between models and users. Early approaches primarily relied on rule-based filtering (Welbl et al., 2021a; Singhal et al., 2023). While effective in constrained settings, such systems struggle to adapt to evolving safety policies and emerging risks, and often suffer from limited coverage and low accuracy (Song et al., 2023; Welbl et al., 2021b). Recent guardrail systems benefit from the development of LLMs and MLLMs, offering improved flexibility and generalization. For example, Llama Guard (Inan et al., 2023) is an LLM-based moderation model fine-tuned on proprietary safety datasets developed by Meta AI, designed to safeguard user-AI conversation. Llama Guard 3 Vision (Chi et al., 2024) and LlavaGuard (Helff et al., 2025b) are VLM-based guardrails capable of identifying visual-related safety risks. However, in the omni-modality era, existing guardrails still exhibit several key limitations: (1) most prior work only focuses on the text and image domains, while guardrails for video and audio remain overly simplified—often reduced to shallow classifiers that treat safety detection as a binary task, lacking reasoning and contextual understanding (Ahmed et al., 2024; Tang et al., 2022). (2) most existing systems remain single-modality or scenario-specific, lacking the capability to process multiple uni-modal inputs or perform cross-modal reasoning that integrates information from text, images, videos, and audio jointly (Rajpal, 2023). To address these challenges, we propose OMNIGUARD, the first family of omni-modal guardrails that natively supports both unimodal and cross-modal content moderation with deliberate reasoning. By incorporating omni-modal understanding and explicit reasoning, OMNIGUARD delivers consistent, interpretable, and holistic safety assurance across all modalities.

3. OMNIGUARD

In this section, we introduce OMNIGUARD, the first family of unified omni-modal safety guardrails designed to perform comprehensive and interpretable safety moderation across all modalities.

3.1. Preliminaries

Guardrail models are designed to assess whether the input content complies with safety policies, determining the presence of harmful or policy-violating elements. OMNIGUARD differs from prior guardrail systems by operating natively over all modalities and any combination of them, enabling unified safety assessment for text, images, videos, and audio within a single framework. Let $\mathcal{X}$ denote the omni-modal input space, spanning text (x_t), image (x_i), video (x_v), and audio (x_a) modalities. Each instance $\mathbf{x} \in \mathcal{X}$ may include one or multiple modalities in arbitrary combinations. Let $\mathcal{G}$ denote the set of safety policy guidelines defining the boundary between safe and unsafe content, corresponding to a predefined set of violation categories $\mathcal{C} = \{c_1, c_2, \ldots, c_m\}$. Formally, OMNIGUARD can be expressed as:

$$f_{\text{OMNIGUARD}}(\mathbf{x} \mid \mathcal{G}) = \big(\mathbf{y}, \mathbf{c}, \mathbf{e}\big), \tag{1}$$

where $\mathbf{e}$ is a natural-language critique that explicitly explains the safety judgment. Specifically, given policy guidelines $\mathcal{G}$ and an omni-modal input $\mathbf{x}$, OMNIGUARD determines the overall safety label y, identifies the violated categories $\mathbf{c}$ when the input is unsafe, and generates an interpretable natural-language critique $\mathbf{e}$ that explains and justifies its safety judgment.

3.2. Mission-Focused Instruction Tuning.

An instruction-tuning instance typically consists of `instruction`, `input`, and `output`. In general instruction tuning settings, the training dataset contains diverse instruction types that enable models to generalize across various downstream tasks. However, in our case, we adopt mission-focused instruction tuning to maximally equip the model with omni-modal safety reasoning capabilities. To this end, we fix the `instruction` template to omni-modal safety moderation and diversify the modalities and semantic meanings of `input`, as well as the corresponding `output`. This training paradigm aims to enhance the model's capacity to identify, categorize, and reason about

safety risks in both unimodal and cross-modal settings.

3.2.1. TARGETED DISTILLATION.

Given the lack of an existing unified omni-modal safety fine-tuning dataset, we construct a comprehensive large-scale omni-modal safety dataset through targeted distillation to support the training process. To increase the diversity of `input`, we first collect and aggregate datasets from both unimodal and cross-modal settings, including text, image, video, audio, and text-image modalities. Each sample is paired with a corresponding binary safety label and, if unsafe, one or more associated violation categories. Subsequently, we employ a targeted distillation process to extract safety reasoning knowledge from large expert models. Given an input instance $\mathbf{x} \in \mathcal{X}$ consisting of one or more modalities, along with its ground-truth safety label y and violation categories $\mathbf{c}$, the expert model f_T produces a detailed natural language critique explaining decision: $f_T(\mathbf{x}, \mathbf{y}, \mathbf{c}) = \mathbf{e_T}$, The formatted prompt used for targeted distillation is shown in Figure 5. The outputs distilled from the expert models, together with the input, are used to construct the dataset $\mathcal{D}$: $\{(\mathbf{x_i}, \mathbf{y_i}, \mathbf{c_i}, \mathbf{e_i})\}_{i=1}^{N}$. An overview of collected datasets and the data distribution is presented in Figure 2. Further statistics are summarized in Table 6.

To enhance interpretability and enable reasoning-based safety alignment, we augment each sample in the previously collected corpus with critiques generated by high-capacity models. Specifically, we employ gpt-oss-120b (Agarwal et al., 2025) for textual data, Qwen3-VL-235B-A22B-Instruct (Team, 2025) for visual-related (image, video, and the text-image pairs) data, and Kimi-Audio-7B-Instruct(Team, 2024a) for auditory data. For each instance, the teacher model is provided with the original content, its corresponding safety label, and the violated categories (if any), and is instructed to generate a reasoning critique explaining the rationale behind the safety assessment. The complete prompting template used for critique generation is illustrated in Figure 5.

3.2.2. INSTRUCTION TUNING

Based on the distilled dataset $\mathcal{D}$ obtained from the omni-modal targeted distillation stage, we perform mission-focused instruction tuning to specialize the model toward the safety moderation. Specifically, we adopt an omni-modal instruction fine-tuning framework to enhance OMNIGUARD's capability to classify and reason about safety risks across modalities.

In our omni-modal setting, our goal is to (1) handle diverse input modalities (text, image, video, and audio), and (2) follow safety-specific instructions constrained by the policy guidelines $\mathcal{G}$ to perform unified, policy-grounded safety reasoning. Therefore, we leverage the omni-modal instruction-following dataset $\mathcal{D}$ and optimize the model using a standard next-token prediction loss, enabling it to produce accurate safety judgments and coherent reasoning across modalities.

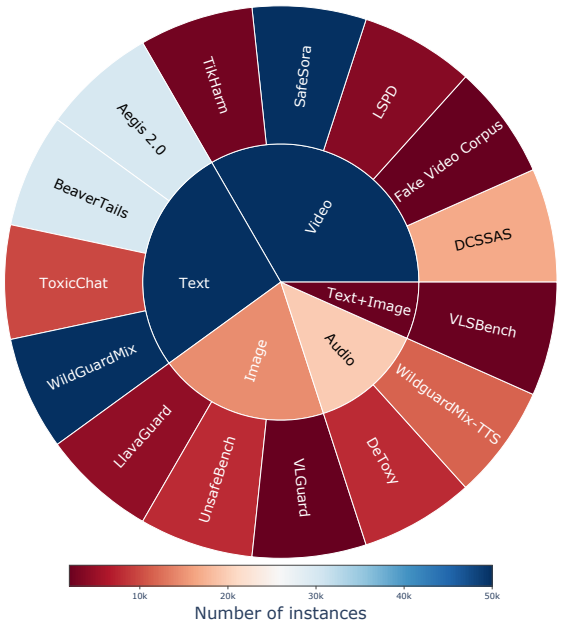

Figure 2. Collected datasets and the distribution of the constructed dataset.

3.2.3. TRAINING OBJECTIVE.

The student model learns from constructed dataset $\mathcal{D}$, which contains omni-modal safety information, by minimizing a joint objective: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{cat}} + \mathcal{L}_{\text{critique}}$, where $\mathcal{L}_{\text{cls}}$ is the classification loss for binary safety prediction, training the model to accurately discriminate between safe and unsafe content; $\mathcal{L}_{\text{cat}}$ is the multi-label classification loss over violation categories, teaching the model to recognize and categorize fine-grained safety violations; and $\mathcal{L}_{\text{critique}}$ is the autoregressive generation loss that aligns the student's critique with the teacher's explanation, enabling the model to produce interpretable critiques explaining the rationale behind the judgment. This training process transfers the policy alignment and safety reasoning capabilities from the teacher model to the guardrail model, allowing it to perform safety classification and justification in a unified manner across modalities.

3.3. Reasoning-Based Inference

Unlike simple classification-only guardrail models that output only a binary safety label, OMNIGUARD performs slow thinking inference to provide fine-grained and interpretable safety moderation. Specifically, it produces a structured output comprising the following components:

(1) **Safety judgment:** the overall safety assessment of the input, determining whether it is safe or unsafe.

(2) **Violation categories:** the specific unsafe categories that the input violates, if the content is identified as unsafe.

(3) **Reasoning critique:** a natural language explanation that articulates the rationale behind the model's decision in accordance with the policy guidelines.

This formulation enables OMNIGUARD to go beyond shallow pattern recognition, supporting explainable analysis of potentially unsafe content across different modalities.

Formally, given a multimodal input $\mathbf{x} \in \mathcal{X}$ and a policy guideline set $\mathcal{G}$, OMNIGUARD computes:

$$g(\mathbf{x} \,|\, \mathcal{G}) = (\hat{\mathbf{y}}, \hat{\mathbf{c}}, \hat{\mathbf{e}}), \tag{2}$$

where $\hat{\mathbf{y}} \in \{\texttt{safe}, \texttt{unsafe}\}$ denotes the predicted safety label, $\hat{\mathbf{c}}$ represents the identified set of violation categories (empty if $\hat{\mathbf{y}} = \texttt{safe}$), and $\hat{\mathbf{e}}$ is the generated reasoning critique. The critique serves as an explicit intermediate representation of the model's decision process, offering insight into how the prediction aligns with the safety policy $\mathcal{G}$ and improving the transparency and interpretability of omni-modal safety moderation.

4. Experiments

In this section, we present comprehensive experimental results of OMNIGUARD. We evaluate its performance under both unimodal and cross-modal settings on 15 guardrail and jailbreak benchmarks spanning four modalities — text, image, video, and audio. We further design experiments to answer two central research questions: (1) Does reasoning-based safety alignment enhance the omni-modal guardrail model's ability to perform safety moderation and handle safety-critical challenges across diverse modalities? (2) Can safety knowledge learned from seen modalities transfer to unseen ones, demonstrating cross-modal generalization in safety understanding and moderation capability?

4.1. Experiment Settings.

Benchmarks. We evaluate OMNIGUARD on a diverse suite of public safety benchmarks spanning both unimodal and cross-modal settings. For the unimodal setting, we assess performance across four modalities — text, image, video, and audio. For text, we use BeaverTails (Ji et al., 2023), ToxicChat (Lin et al., 2023), WildGuardMix (Han et al., 2024), Aegis2.0(Ghosh et al., 2025), and the OpenAI Moderation dataset (Markov et al., 2023). For image, we adopt UnsafeBench (Qu et al., 2024), VLGuard (Zong et al., 2024), and LlavaGuard (Helff et al., 2025a). For video, we evaluate on SafeSora (Dai et al., 2024a) and SafeWatch-Bench (Chen et al., 2025). For audio, we use MuTox English split (Costa-jussà et al., 2024) and WildGuardMix-TTS, which is constructed by converting WildGuardMix (Han et al., 2024) test samples into speech using a text-to-speech pipeline consistent with our dataset construction procedure. For the cross-modal setting, we evaluate OMNIGUARD on three configurations: image-text, video-text, and audio-text, corresponding to MM-SafetyBench (Liu et al., 2024b), VideoSafetyBench (Liu et al., 2025a), and AIAH (Yang et al., 2025), respectively. Further statistics are summarized in Table 6. We employ Accuracy (ACC) and F1 as the primary evaluation metrics to assess safeguarding performance For benchmarks containing only unsafe samples, we only report accuracy as the evaluation metric.

Baselines. We compare OMNIGUARD against a comprehensive suite of baselines across all modalities, encompassing LLMs, VLLMs, and audio LLMs. For each modality, we include both large-scale and small-scale state-of-the-art models to evaluate their safeguarding capabilities. We also compare OMNIGUARD with available specialized guardrail models, including LLM-based and VLM-based guardrail models. Detailed baseline configurations are summarized in Table 7.

Training. We train two variants of our model, OMNIGUARD-7B and OMNIGUARD-3B, based on Qwen2.5-Omni-7B and Qwen2.5-Omni-3B, respectively. Both models are trained using full-parameter supervised fine-tuning (SFT) on our constructed dataset. Training is conducted on 8×H100 GPUs using the SWIFT training platform (Zhao et al., 2024). We employ the AdamW optimizer with a learning rate of 1×10^{-4}, a cosine learning rate scheduler, and a warmup ratio of 0.05. Each model is trained for 3 epochs with a per-device batch size of 2 for training and 1 for evaluation, and gradients are accumulated over 4 steps. The random seed is fixed to 42 for reproducibility.

4.2. Results.

Uni-Modality. We compare the performance of OMNIGUARD against state-of-the-art proprietary and open-source baselines across four uni-modal safety scenarios: text (Table 1), image (Table 2), video (Table 3), and audio (Table 4). Both OMNIGUARD-7B and OMNIGUARD-3B consistently achieve leading results across all modalities. While OMNIGUARD-7B consistenly achieves strongest overall performance across all modalities, the smaller variant, OMNIGUARD-3B also can achieve results comparable to or exceeding much larger models such as GPT-4o, Qwen3-235B, and Qwen3-VL-235B, highlighting the effectiveness of our omni-modal safety alignment strategy.

Text. As shown in Table 1, OMNIGUARD-7B achieves the highest average F1 and accuracy on text safety benchmarks,

Model	Size	BeaverTails		OpenAI		Toxic Chat		Aegis		WildGuard		Average	
		F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC
GPT-4o	-	83.5	72.6	**82.3**	**88.6**	51.2	94.8	54.3	65.6	78.0	89.1	69.9	82.1
Qwen3-235B	235B	81.9	77.1	80.1	85.4	66.0	95.5	83.3	83.6	74.4	90.2	77.1	86.4
LLaMA-3.3-70B	70B	76.8	71.3	58.4	80.5	53.8	94.3	65.7	69.4	72.5	89.2	65.4	80.9
Qwen2.5-72B	72B	83.6	80.5	79.8	85.2	54.4	94.3	68.9	73.2	72.2	89.1	71.8	84.5
Qwen2.5-Omni-7B	7B	58.4	55.4	70.0	70.8	65.2	93.9	77.3	73.9	62.0	83.1	66.6	75.4
Qwen2.5-7B	7B	75.3	72.6	72.6	81.4	58.3	95.1	75.4	77.5	63.3	86.2	69.0	82.6
LLaMA Guard 1	7B	38.1	55.9	32.8	74.4	23.3	92.9	53.0	66.9	16.4	85.1	32.7	75.0
LLaMA Guard 2	8B	72.3	73.5	74.4	85.6	30.9	93.7	59.0	69.2	68.9	89.9	61.1	82.4
LLaMA Guard 3	8B	71.2	73.5	81.6	85.7	37.7	93.3	65.8	73.5	73.5	91.6	66.0	83.5
ThinkGuard	8B	82.7	81.6	78.7	79.0	49.8	92.8	69.9	74.6	78.5	92.5	71.9	84.1
OMNIGUARD-3B	3B	81.8	**82.6**	77.8	83.8	**67.0**	**95.6**	82.2	82.6	70.2	87.7	75.8	86.5
OMNIGUARD-7B	7B	**83.9**	80.5	81.1	87.9	58.2	95.3	**84.0**	**84.1**	**78.6**	**92.4**	**77.2**	**88.0**

Table 1. Performance comparison of OMNIGUARD and baseline models on *text-based safety benchmarks*. **Bold** and underlined values indicate the best and second-best performance, respectively.

Model	Size	VLGuard		UnsafeBench		LlavaGuard		Average	
		F1	ACC	F1	ACC	F1	ACC	F1	ACC
GPT-4o	-	75.5	79.7	55.2	74.7	68.3	74.6	66.3	76.3
Qwen3-VL-235B	235B	77.4	76.5	**74.4**	80.9	73.8	76.3	**75.2**	77.9
Qwen2.5-VL-72B	72B	78.5	77.4	73.2	77.7	71.2	70.0	74.3	75.0
Qwen2.5-Omni-7B	7B	64.4	70.8	47.3	67.8	57.3	68.9	56.3	69.2
Qwen2.5-VL-7B	7B	62.8	48.2	55.7	40.1	59.0	46.5	59.2	44.9
LlavaGuard-v1.2-7B	7B	69.8	74.0	63.4	77.0	79.6	82.0	70.9	77.7
LLaMA Guard 3V	11B	0.0	55.8	0.0	61.9	0.0	75.0	0.0	64.2
OMNIGUARD-7B	7B	79.1	81.7	72.2	**81.1**	73.5	77.1	74.9	80.0
OMNIGUARD-3B	3B	**79.3**	**81.9**	72.3	**81.1**	**73.9**	**78.2**	**75.2**	**80.4**

Table 2. Performance comparison of OMNIGUARD and baseline models on *image-based safety benchmarks*. Best in **bold** and second-best in underlined.

surpassing both large-scale proprietary models such as GPT-4o and open-source baselines including Qwen3-235B and LLaMA3.3-70B. It also consistently outperforms smaller general-purpose models as well as dedicated guardrail systems. Specifically, OMNIGUARD-7B attains an average F1 of 77.2% and an accuracy of 88.0%, outperforming all other compared models and improving the average F1 by more than 10% compared to LLaMA Guard 3. Notably, the lighter variant, OMNIGUARD-3B, achieves performance comparable to Qwen3-235B while using only a fraction of its parameters.

Images. In the image domain (Table 2), OMNIGUARD-7B and OMNIGUARD-3B also exhibit higher unsafe content detection performance compared to all other baselines. OMNIGUARD-7B achieves an average F1 of 75.2 % and accuracy of 80.4 %, matching the F1 score of Qwen3-VL-235B while using far fewer parameters. The improvement over previous image safeguards is substantial. LLaMA-Guard-3V (Chi et al., 2024), which is only designed for safeguarding multimodal conversational content, failed to provide safety assessment for image-only harmfulness evaluation, classfifying all the samples as safe. This demonstrates the narraw focus of existing guardrail systems.

Videos. For the video-safety benchmarks (Table 3), both OMNIGUARD-7B and OMNIGUARD-3B achieve state-of-the-art performance. The improvement is especially pronounced on the SafeWatch-Bench, where both models exceed 90 % on F1 score. These results highlight the significant progress of OMNIGUARD in safety reasoning within video domain.

Audio. In the audio domain (see Table 4), OMNIGUARD-7B and OMNIGUARD-3B also achieve superior performance across both audio safety benchmarks. Since audio guardrails remain largely unexplored, our model provides a strong solution to the field and demonstrates that reasoning-based safety training can also effectively generalize to the audio modality.

Model	Size	SafeWatch		SafeSora	
		F1	ACC	F1	ACC
GPT-4o	-	84.2	77.5	49.9	82.2
Qwen3-VL-235B	235B	79.5	71.8	65.9	84.4
Qwen2.5-VL-72B	72B	72.5	64.0	68.0	83.7
LLaVA-Video-72B	72B	78.2	70.7	36.9	80.0
Qwen2.5-Omni-7B	7B	68.6	76.0	64.3	71.3
Qwen2.5-VL-7B	7B	49.7	46.2	62.2	83.5
LLaVA-Video-7B	7B	47.2	44.9	4.5	75.1
OMNIGUARD-3B	3B	**92.3**	82.0	70.1	85.9
OMNIGUARD-7B	7B	90.9	**85.7**	**71.8**	**86.1**

Table 3. Performance comparison of OMNIGUARD and baseline models on *video-based safety benchmarks*. Best in **bold** and second-best in underlined.

Model	Size	MuTox		WildGuard-TTS	
		F1	ACC	F1	ACC
GPT-4o		38.7	66.1	81.6	85.2
Qwen2-Audio	7B	26.9	42.4	27.7	56.7
Qwen-Audio	8B	28.0	18.9	59.3	57.7
Kimi-Audio	7B	37.5	68.3	77.4	75.6
Qwen2.5-Omni-7B	7B	30.8	36.2	78.8	78.5
OMNIGUARD-3B	3B	41.8	72.3	**88.4**	**89.8**
OMNIGUARD-7B	7B	**43.7**	**75.4**	87.8	89.2

Table 4. Performance comparison of OMNIGUARD and baseline models on *audio-based safety benchmarks*. Best in **bold** and second-best in underlined.

Cross-Modality. We further evaluate OMNIGUARD in cross-modal safety scenarios to assess its capability to reason across modalities. As illustrated in Figure 3, our OMNIGUARD family demonstrates strong and consistent performance across all evaluated cross-modal safety benchmarks, including MM-SafetyBench, Video-SafetyBench, and AIAH. Compared to Qwen2.5-Omni-7B, OMNIGUARD-7B consistently achieves significant improvements in accuracy across all benchmarks, highlighting the generalization of our omni-modal safety alignment in unifying multimodal reasoning. Moreover, the lightweight OMNIGUARD-3B performs comparably to large-scale general-purpose models such as Qwen3-VL-235B on both MM-SafetyBench and Video-SafetyBench, despite having significantly fewer parameters. These results further confirm that OMNIGUARD effectively generalizes safety reasoning across modalities, offering a scalable and parameter-efficient solution for cross-modal alignment.

4.3. RQ1: Reasoning-Based Safety Training.

To examine whether reasoning-based safety training enhances the performance of omni-modal guardrails and address the additional complexity introduced by omni-modal safety reasoning, we conduct further studies on the 7B model across four modalities, as shown in Figure 4. We compare our OMNIGUARD-7B with the original base model Qwen2.5-Omni-7B (Xu et al., 2025b) and its Label-only SFT variant, which is fine-tuned solely on safety classification labels without the curated reasoning traces used in our approach.

From these results, we draw several observations. (1) Both supervised fine-tuning methods can improve performance over the base model across all unimodal benchmarks (text, image, video, and audio), showing that simple safety fine-tuning can also enhance multimodal moderation capabilities. (2) Compared to the Label-only baseline, our reasoning-augmented training consistently achieves higher F1 scores across all unimodal settings, improving from 75.7→77.2 (text), 74.0→75.2 (image), 79.8→81.4 (video), and 63.7→65.8 (audio). This confirms that reasoning supervision helps the model better internalize safety assessment principles beyond surface-level pattern learning. (3) Notably, in the cross-modal setting, the Label-only SFT variant suffers a degradation in accuracy on the Video-SafetyBench and AIAH benchmarks, whereas our reasoning-augmented model achieves consistent gains across all three tasks. This suggests that simple label supervision fails to generalize effectively facing complex moderation tasks across modalities, while reasoning-based alignment endows the model with stronger guardrail understanding and transferability.

Overall, these results highlight that reasoning-based safety alignment not only enhances the performance of omni-modal guardrails across different modality settings, but also provides better understanding and generalization in complex cross-modal safety scenarios that label-only supervision fails to handle.

4.4. RQ2: Cross-Modal Generalization.

To investigate whether safety knowledge learned in seen modalities can generalize to unseen ones, we conduct cross-modal training and evaluation. Specifically, for each split, we train OMNIGUARD using data from three modalities for training and seen modality evaluation and leave one modality out for unseen modality evaluation. We report the averaged F1 and accuracy across all four seen and unseen modality evaluations in Table 5.

We draw two main conclusions from these experiments. (1) Overall, strong cross-modal transfer is observed across all four modalities. The unseen modality performance remains close to seen modality results (79.4 vs. 81.8 F1 on average), indicating that OMNIGUARD successfully learns modality-invariant safety representations. This suggests that harmful

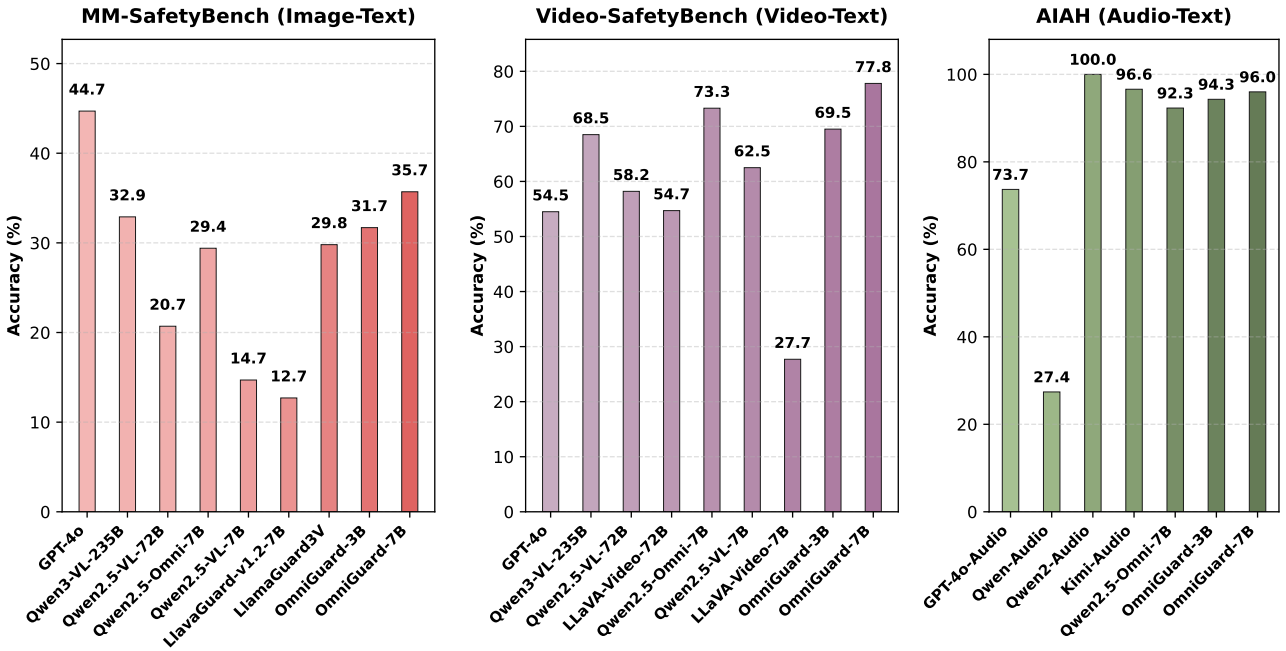

Figure 3. Performance comparison of OMNIGUARD and baseline models on *cross-modal safety benchmarks*. The performance is evaluated in Accuracy (ACC).

semantic patterns can be effectively aligned and learned across text, image, video, and audio inputs. OMNIGUARD further acquire generalizable safety reasoning ability across modalities.

(2) One notable exception arises in the audio modality, where the seen modality F1 (64.6%) is slightly lower than the unseen modality F1 (65.1%). This phenomenon is due to the OMNIGUARD variant trained on text excluded split exhibiting degraded performance on the WildGuard-TTS benchmark. WildGuardMix-TTS is the auditory version of WildGuardMix constructed by text-to-speech model. Training without textual data also lead to degradation in the audio setting, this reveals that content which are semantically equivalent but are from different modalities (e.g., harmful text vs. its spoken version) can mutually influence each other during safety alignment. And the knowledge from one form can be transferred to another semantically invariant form.

Taken together, these results demonstrate that cross-modal generalization in OMNIGUARD is substantial. OMNIGUARD can generalizes safety reasoning from trained modalities to untrained ones and but also acquires modality-invariant semantic representations of unsafe content.

5. Conclusion

In conclusion, we introduce OMNIGUARD-7B and OMNIGUARD-3B, the first family of omni-modal guardrails trained on a comprehensive and unified safety fine-tuning dataset covering both unimodal and cross-modal samples.

Modality	Seen Modality		Unseen Modality	
	F1	ACC	F1	ACC
Text	88.1	80.1	84.7	77.4
Image	78.3	79.0	76.8	77.1
Video	78.9	85.1	76.6	80.0
Audio	64.6	92.2	65.1	90.2
Average	81.8	84.1	79.4	81.2

Table 5. Performance comparison between *seen modality* and *unseen modality* settings across four modalities. Metrics are F1 and Accuracy (%).

OMNIGUARD instantiates a unified omni-modal safety solution: it can moderate and reason about unsafe content in heterogeneous and cross-modal settings. Extensive experiments demonstrate that OMNIGUARD-7B consistently outperforms existing guardrail models across all modalities, while OMNIGUARD-3B can achieves competitive results compared to large-scale LLMs and MLLMs such as Qwen3-235B and Qwen3-VL-235B. These results highlight the strong omni-modal safety detection and reasoning capabilities of our approach, confirming the feasibility of a unified guardrail system with omni-modal understanding and interpretability. Future work will investigate more complex and cross-modal safety scenarios to further advance omni-modal safeguarding in next-generation large language models.

Limitations. Although OMNIGUARD demonstrates strong omni-modal safety reasoning and consistent performance across modalities, several limitations remain. While incorporating reasoning paths significantly enhances the in-

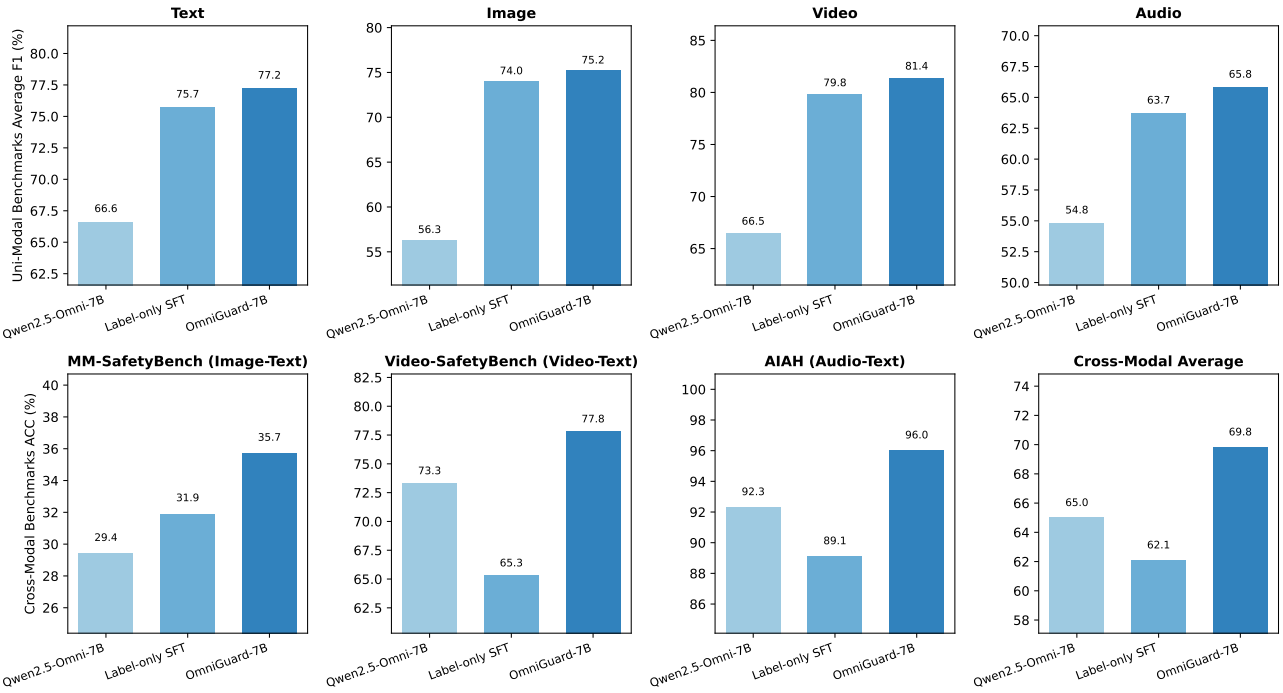

Figure 4. Comparison of performance between Label-only SFT and critique-augmented training across both *uni-modal* and *cross-modal* settings. The upper four subplots show average performance results on uni-modal benchmarks (*Text*, *Image*, *Video*, *Audio*), evaluated by F1 score (%). The bottom four subplots present cross-modal results on *MM-SafetyBench (Image-Text)*, *Video-SafetyBench (Video-Text)*, and *AIAH (Audio-Text)*, along with the average performance, reported in accuracy (ACC, %).

terpretability and reliability of safety assessments, it inevitably increases inference latency due to the additional reasoning generation step. As a result, OMNIGUARD has higher computational overhead compared to lightweight, binary-classification guardrail systems. The trade-off between safety reasoning depth and inference latency can be further explored under different use cases and scenarios to further optimize safety robustness and efficiency. Additionally, due to the limited availability of publicly accessible cross-modal safety fine-tuning datasets, there remains substantial room for progress in moderating more complex interleaved multimodal safety scenarios. We hope future research will continue to improve the robustness and reliability of omni-modal systems, and advance their capability to safeguard against risks across complex modality combinations.

References

Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R. K., Bai, Y., Baker, B., Bao, H., and et al. gpt-oss-120b & gpt-oss-20b model card. *CoRR*, abs/2508.10925, 2025. doi: 10.48550/ARXIV.2508.10925. URL https://doi.org/10.48550/arXiv.2508.10925.

Ahmed, S. H., Khan, M. J., and Sukthankar, G. Enhanced multimodal content moderation of children's videos using audiovisual fusion. In Chun, S. A. and Talbert, D. A. (eds.), *Proceedings of the Thirty-Seventh International Florida Artificial Intelligence Research Society Conference, FLAIRS 2024, Sandestin Beach, FL, USA, May 19-21, 2024*. Florida Online Journals, 2024. doi: 10.32473/FLAIRS.37.1.135563. URL https://doi.org/10.32473/flairs.37.1.135563.

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., and Lin, J. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., Showk, S. E., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda,

N., Olsson, C., Amodei, D., Brown, T. B., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862, 2022. doi: 10.48550/ARXIV.2204.05862. URL https://doi.org/10.48550/arXiv.2204.05862.

Balat, M., Gabr, M. E., Bakr, H., and Zaky, A. B. Tikguard: A deep learning transformer-based solution for detecting unsuitable tiktok content for kids. In *6th Novel Intelligent and Leading Emerging Sciences Conference, NILES 2024, Giza, Egypt, October 19-21, 2024*, pp. 337–340. IEEE, 2024. doi: 10.1109/NILES63360.2024.10753192. URL https://doi.org/10.1109/NILES63360.2024.10753192.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., and Dai, J. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024. URL https://arxiv.org/abs/2312.14238.

Chen, Z., Pinto, F., Pan, M., and Li, B. Safewatch: An efficient safety-policy following video guardrail model with transparent explanations. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=xjKz6IxgCX.

Chi, J., Karn, U., Zhan, H., Smith, E., Rando, J., Zhang, Y., Plawiak, K., Coudert, Z. D., Upasani, K., and Pasupuleti, M. Llama guard 3 vision: Safeguarding human-ai image understanding conversations. *CoRR*, abs/2411.10414, 2024. doi: 10.48550/ARXIV.2411.10414. URL https://doi.org/10.48550/arXiv.2411.10414.

Chu, Y., Xu, J., Zhou, X., Yang, Q., Zhang, S., Yan, Z., Zhou, C., and Zhou, J. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models, 2023. URL https://arxiv.org/abs/2311.07919.

Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., Zhou, C., and Zhou, J. Qwen2-audio technical report, 2024. URL https://arxiv.org/abs/2407.10759.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I. S., Blistein, M., Ram, O., Zhang, D., Rosen, E., and et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025. doi: 10.48550/ARXIV.2507.06261. URL https://doi.org/10.48550/arXiv.2507.06261.

Costa-jussà, M. R., Meglioli, M. C., Andrews, P., Dale, D., Hansanti, P., Kalbassi, E., Mourachko, A., Ropers, C., and Wood, C. Mutox: Universal multilingual audio-based toxicity dataset and zero-shot detector. In Ku, L., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 5725–5734. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.340. URL https://doi.org/10.18653/v1/2024.findings-acl.340.

Cui, S., Zhang, Q., Ouyang, X., Chen, R., Zhang, Z., Lu, Y., Wang, H., Qiu, H., and Huang, M. Shieldvlm: Safeguarding the multimodal implicit toxicity via deliberative reasoning with lvlms: Shieldvlm. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 11677–11686, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3755711. URL https://doi.org/10.1145/3746027.3755711.

Dai, J., Chen, T., Wang, X., Yang, Z., Chen, T., Ji, J., and Yang, Y. Safesora: Towards safety alignment of text2video generation via a human preference dataset. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a. URL http://papers.nips.cc/paper_files/paper/2024/hash/1eb543faf7c69e8a7eb8b85f70be818f-Abstract-Datasets_and_Benchmarks_Track.html.

Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., Wang, Y., and Yang, Y. Safe RLHF: safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024b. URL https://openreview.net/forum?id=TyFrPOKYXw.

Fang, Q., Guo, S., Zhou, Y., Ma, Z., Zhang, S., and Feng, Y. Llama-omni: Seamless speech interaction with large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=PYmrUQmMEw.

Fu, C., Lin, H., Wang, X., Zhang, Y.-F., Shen, Y., Liu, X., Cao, H., Long, Z., Gao, H., Li, K., Ma, L., Zheng, X., Ji, R., Sun, X., Shan, C., and He, R. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction, 2025. URL https://arxiv.org/abs/2501.01957.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In Cohn, T., He, Y., and Liu, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pp. 3356–3369. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.FINDINGS-EMNLP.301. URL https://doi.org/10.18653/v1/2020.findings-emnlp.301.

Ghosh, S., Varshney, P., Sreedhar, M. N., Padmakumar, A., Rebedea, T., Varghese, J. R., and Parisien, C. AEGIS2.0: A diverse AI safety dataset and risks taxonomy for alignment of LLM guardrails. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 5992–6026. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.NAACL-LONG.306. URL https://doi.org/10.18653/v1/2025.naacl-long.306.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., and et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Han, S., Rao, K., Ettinger, A., Jiang, L., Lin, B. Y., Lambert, N., Choi, Y., and Dziri, N. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/0f69b4b96a46f284b726fbd70f74fb3b-Abstract-Datasets_and_Benchmarks_Track.html.

Helff, L., Friedrich, F., Brack, M., Kersting, K., and Schramowski, P. Llavaguard: An open VLM-based framework for safeguarding vision datasets and models. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=YIO9ritzWV.

Helff, L., Friedrich, F., Brack, M., Schramowski, P., and Kersting, K. Llavaguard: An open vlm-based framework for safeguarding vision datasets and models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025b.

Hu, X., Liu, D., Li, H., Huang, X., and Shao, J. Vlsbench: Unveiling visual leakage in multimodal safety. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 8285–8316. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.405/.

Huang, R., Li, M., Yang, D., Shi, J., Chang, X., Ye, Z., Wu, Y., Hong, Z., Huang, J., Liu, J., Ren, Y., Zou, Y., Zhao, Z., and Watanabe, S. Audiogpt: Understanding and generating speech, music, sound, and talking head. In Wooldridge, M. J., Dy, J. G., and Natarajan, S. (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 23802–23804. AAAI Press, 2024. doi: 10.1609/AAAI.V38I21.30570. URL https://doi.org/10.1609/aaai.v38i21.30570.

Huang, T., Hu, S., Ilhan, F., Tekin, S. F., Yahn, Z., Xu, Y., and Liu, L. Safety tax: Safety alignment makes your large reasoning models less reasonable. *CoRR*, abs/2503.00555, 2025. doi: 10.48550/ARXIV.2503.00555. URL https://doi.org/10.48550/arXiv.2503.00555.

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., and Khabsa, M. Llama guard: Llm-based input-output safeguard for human-ai conversations. *CoRR*, abs/2312.06674, 2023. doi: 10.48550/ARXIV.2312.06674. URL https://doi.org/10.48550/arXiv.2312.06674.

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Chen, B., Sun, R., Wang, Y., and Yang, Y. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/4dbb61cb68671edc4ca3712d70083b9f-Abstract-Datasets_and_Benchmarks.html.

KimiTeam, Ding, D., Ju, Z., Leng, Y., Liu, S., Liu, T., Shang, Z., Shen, K., Song, W., Tan, X., Tang, H., Wang, Z., Wei, C., Xin, Y., Xu, X., Yu, J., Zhang, Y., Zhou, X., Charles, Y., Chen, J., Chen, Y., Du, Y., He, W., Hu, Z., Lai, G., Li, Q., Liu, Y., Sun, W., Wang, J., Wang, Y., Wu, Y., Wu, Y.,

Yang, D., Yang, H., Yang, Y., Yang, Z., Yin, A., Yuan, R., Zhang, Y., and Zhou, Z. Kimi-audio technical report, 2025. URL https://arxiv.org/abs/2504.18425.

Lee, S., Kim, G., Kim, J., Lee, H., Chang, H., Park, S. H., and Seo, M. How does vision-language adaptation impact the safety of vision language models? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=eXB5TCrAu9.

Liao, S., Wang, Y., Li, T., Cheng, Y., Zhang, R., Zhou, R., and Xing, Y. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis, 2024. URL https://arxiv.org/abs/2411.01156.

Lin, Z., Wang, Z., Tong, Y., Wang, Y., Guo, Y., Wang, Y., and Shang, J. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pp. 4694–4702. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.311. URL https://doi.org/10.18653/v1/2023.findings-emnlp.311.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Liu, X., Xu, N., Chen, M., and Xiao, C. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *The Twelfth International Conference on Learning Representations*, 2024a.

Liu, X., Zhu, Y., Gu, J., Lan, Y., Yang, C., and Qiao, Y. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., and Varol, G. (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LVI*, volume 15114 of *Lecture Notes in Computer Science*, pp. 386–403. Springer, 2024b. doi: 10.1007/978-3-031-72992-8_22. URL https://doi.org/10.1007/978-3-031-72992-8_22.

Liu, X., Li, Z., He, Z., Li, P., Xia, S., Cui, X., Huang, H., Yang, X., and He, R. Video-safetybench: A benchmark for safety evaluation of video lvlms. *CoRR*, abs/2505.11842, 2025a. doi: 10.48550/ARXIV.2505.11842. URL https://doi.org/10.48550/arXiv.2505.11842.

Liu, Z., Dong, Y., Wang, J., Liu, Z., Hu, W., Lu, J., and Rao, Y. Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. *CoRR*, abs/2502.04328, 2025b. doi: 10.48550/ARXIV.2502.04328. URL https://doi.org/10.48550/arXiv.2502.04328.

Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., and Kembhavi, A. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26429–26445. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02497. URL https://doi.org/10.1109/CVPR52733.2024.02497.

Luo, R., Xia, X., Wang, L., Chen, L., Shan, R., Luo, J., Yang, M., and Chua, T.-S. Next-omni: Towards any-to-any omnimodal foundation models with discrete flow matching, 2025. URL https://arxiv.org/abs/2510.13721.

Markov, T., Zhang, C., Agarwal, S., Nekoul, F. E., Lee, T., Adler, S., Jiang, A., and Weng, L. A holistic approach to undesired content detection in the real world. In Williams, B., Chen, Y., and Neville, J. (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 15009–15018. AAAI Press, 2023. doi: 10.1609/AAAI.V37I12.26752. URL https://doi.org/10.1609/aaai.v37i12.26752.

OpenAI. Gpt-4v(ision) system card. 2023. URL https://api.semanticscholar.org/CorpusID:263218031.

OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Papadopoulou, O., Zampoglou, M., Papadopoulos, S., and Kompatsiaris, I. A corpus of debunked and verified user-generated videos. *Online Inf. Rev.*, 43(1):72–88, 2019. doi: 10.1108/OIR-03-2018-0101. URL https://doi.org/10.1108/OIR-03-2018-0101.

Phan, D. D., Nguyen, T. T., Nguyen, Q. H., Tran, H. L., Nguyen, K. N. K., and Vu, D. L. Lspd: A large-scale pornographic dataset for detection and classification. *International Journal of Intelligent Engineering and Systems*, 15:198–231.

Qi, X., Huang, K., Panda, A., Henderson, P., Wang, M., and Mittal, P. Visual adversarial examples jailbreak aligned large language models. In Wooldridge, M. J., Dy, J. G., and Natarajan, S. (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 21527–21536. AAAI Press, 2024a. doi: 10.1609/AAAI.V38I19.30150. URL https://doi.org/10.1609/aaai.v38i19.30150.

Qi, X., Zeng, Y., Xie, T., Chen, P., Jia, R., Mittal, P., and Henderson, P. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024b. URL https://openreview.net/forum?id=hTEGyKf0dZ.

Qi, X., Panda, A., Lyu, K., Ma, X., Roy, S., Beirami, A., Mittal, P., and Henderson, P. Safety alignment should be made more than just a few tokens deep. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=6Mxhg9PtDE.

Qu, Y., Shen, X., Wu, Y., Backes, M., Zannettou, S., and Zhang, Y. Unsafebench: Benchmarking image safety classifiers on real-world and ai-generated images. *CoRR*, abs/2405.03486, 2024. doi: 10.48550/ARXIV.2405.03486. URL https://doi.org/10.48550/arXiv.2405.03486.

Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Rajpal, S. Guardrails ai. https://www.guardrailsai.com/, 2023.

Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T. P., Alayrac, J., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530, 2024. doi: 10.48550/ARXIV.2403.05530. URL https://doi.org/10.48550/arXiv.2403.05530.

Singhal, M., Ling, C., Paudel, P., Thota, P., Kumarswamy, N., Stringhini, G., and Nilizadeh, S. Sok: Content moderation in social media, from guidelines to enforcement, and research to practice. In *8th IEEE European Symposium on Security and Privacy, EuroS&P 2023, Delft, Netherlands, July 3-7, 2023*, pp. 868–895. IEEE, 2023. doi: 10.1109/EUROSP57164.2023.00056. URL https://doi.org/10.1109/EuroSP57164.2023.00056.

Song, J. Y., Lee, S., Lee, J., Kim, M., and Kim, J. Modsandbox: Facilitating online community moderation through error prediction and improvement of automated rules. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581057. URL https://doi.org/10.1145/3544548.3581057.

Sultani, W., Chen, C., and Shah, M. Real-world anomaly detection in surveillance videos. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 6479–6488. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00678. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Sultani_Real-World_Anomaly_Detection_CVPR_2018_paper.html.

Tang, T., Wu, Y., Wu, Y., Yu, L., and Li, Y. Videomoderator: A risk-aware framework for multimodal video moderation in e-commerce. *IEEE Trans. Vis. Comput. Graph.*, 28 (1):846–856, 2022. doi: 10.1109/TVCG.2021.3114781. URL https://doi.org/10.1109/TVCG.2021.3114781.

Team, K. Kimi-audio technical report, 2024a.

Team, L. Meta llama guard 2. https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard2/MODEL_CARD.md, 2024b.

Team, Q. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Wang, L., Yao, K., Li, X., Yang, D., Li, H., Wang, X., and Dong, W. The man behind the sound: Demystifying audio private attribute profiling via multimodal large language model agents, 2025a. URL https://arxiv.org/abs/2507.10016.

Wang, S., Ye, X., Cheng, Q., Duan, J., Li, S., Fu, J., Qiu, X., and Huang, X. Safe inputs but unsafe output: Benchmarking cross-modality safety alignment of large vision-language models. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 3563–3605. Association for Computational Linguistics, 2025b. doi: 10.18653/V1/2025.FINDINGS-NAACL.198. URL https://doi.org/10.18653/v1/2025.findings-naacl.198.

Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L. A., Anderson, K., Kohli, P., Coppin, B., and Huang, P. Challenges in detoxifying language models. In Moens, M., Huang, X., Specia, L., and Yih, S. W. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pp. 2447–2469. Association for Computational Linguistics, 2021a. doi: 10.18653/V1/2021.FINDINGS-EMNLP.210. URL https://doi.org/10.18653/v1/2021.findings-emnlp.210.

Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L. A., Anderson, K., Kohli, P., Coppin, B., and Huang, P.-S. Challenges in detoxifying language models. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2447–2469, Punta Cana, Dominican Republic, November 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.210. URL https://aclanthology.org/2021.findings-emnlp.210/.

Wen, X., Zhou, W., Mo, W. J., and Chen, M. ThinkGuard: Deliberative slow thinking leads to cautious guardrails. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13698–13713, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.704. URL https://aclanthology.org/2025.findings-acl.704/.

Wu, S., Fei, H., Qu, L., Ji, W., and Chua, T. Next-gpt: Any-to-any multimodal LLM. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=NZQkumsNlf.

Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., Zhang, B., Wang, X., Chu, Y., and Lin, J. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215, 2025a. doi: 10.48550/ARXIV.2503.20215. URL https://doi.org/10.48550/arXiv.2503.20215.

Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., Zhang, B., Wang, X., Chu, Y., and Lin, J. Qwen2.5-omni technical report, 2025b. URL https://arxiv.org/abs/2503.20215.

Yang, H., Qu, L., Shareghi, E., and Haffari, G. Audio is the achilles' heel: Red teaming audio large multimodal models. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 9292–9306. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.NAACL-LONG.470. URL https://doi.org/10.18653/v1/2025.naacl-long.470.

Zeng, E., Kohno, T., Roesner, F., and Allen, P. G. Bad news: Clickbait and deceptive ads on news and misinformation websites. 2020. URL https://api.semanticscholar.org/CorpusID:219178438.

Zhan, J., Dai, J., Ye, J., Zhou, Y., Zhang, D., Liu, Z., Zhang, X., Yuan, R., Zhang, G., Li, L., Yan, H., Fu, J., Gui, T., Sun, T., Jiang, Y., and Qiu, X. Anygpt: Unified multimodal LLM with discrete sequence modeling. In Ku, L., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 9637–9662. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.521. URL https://doi.org/10.18653/v1/2024.acl-long.521.

Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., and Li, C. Llava-video: Video instruction tuning with synthetic data, 2025. URL https://arxiv.org/abs/2410.02713.

Zhao, Y., Huang, J., Hu, J., Wang, X., Mao, Y., Zhang, D., Jiang, Z., Wu, Z., Ai, B., Wang, A., Zhou, W., and Chen, Y. Swift:a scalable lightweight infrastructure for fine-tuning, 2024. URL https://arxiv.org/abs/2408.05517.

Zhou, W., Zhang, S., Gu, Y., Chen, M., and Poon, H. Universalner: Targeted distillation from large language models for open named entity recognition, 2024. URL https://arxiv.org/abs/2308.03279.

Zhu, T., Liu, Q., Wang, F., Tu, Z., and Chen, M. Unraveling cross-modality knowledge conflicts in large vision-

language models. *arXiv preprint arXiv:2410.03659*, 2024.

Zhu, T., Zhang, K., Chen, M., and Su, Y. Is extending modality the right path towards omni-modality? *arXiv preprint arXiv:2506.01872*, 2025.

Zizzo, G., Cornacchia, G., Fraser, K., Hameed, M. Z., Rawat, A., Buesser, B., Purcell, M., Chen, P.-Y., Sattigeri, P., and Varshney, K. R. Adversarial prompt evaluation: Systematic benchmarking of guardrails against prompt input attacks on llms. In *Neurips Safe Generative AI Workshop 2024*, 2024.

Zong, Y., Bohdal, O., Yu, T., Yang, Y., and Hospedales, T. M. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=bWZKvF0g7G.

A. Datasets and Benchmarks

	Name	Citation	Train	Test
Text	BeaverTails	(Ji et al., 2023)	27,186	3,021
	Aegis 2.0	(Ghosh et al., 2025)	30,007	1,964
	WildGuardMix	(Han et al., 2024)	86,759	1,756
	ToxicChat	(Lin et al., 2023)	5,082	5,083
	OpenAI Moderation	(Markov et al., 2023)	–	1,680
Image	UnsafeBench	(Qu et al., 2024)	8,109	8,109
	VLGuard	(Zong et al., 2024)	1,999	1,999
	LlavaGuard	(Helff et al., 2025a)	4,571	4,571
Video	SafeSora	(Dai et al., 2024a)	51,588	5,745
	Fake Video Corpus	(Papadopoulou et al., 2019)	380	–
	LSPD	(Phan et al.)	4,000	–
	TikHarm	(Balat et al., 2024)	2,762	–
	DCSASS	(Sultani et al., 2018)	1,610	–
	SafeWatch-Bench	(Chen et al., 2025)	–	1620
Audio	MuTox (English)	(Costa-jussà et al., 2024)	13,617	1,945
	WildGuardMix-TTS	(Han et al., 2024)	10,000	1,756
Text-Image	VLSBench	(Hu et al., 2025)	2,240	–
	MM-SafetyBench	(Liu et al., 2024b)	-	5,040
Text-Video	Video-SafetyBench	(Liu et al., 2025a)	–	2,264
Text-Audio	AIAH	(Yang et al., 2025)	–	350

Table 6. Overview of dataset sources used in the constructed dataset and benchmarks used for evaluation, with corresponding training and testing instance counts. "–" indicates not used or unavailable.

We next detail the data sources used in constructing dataset $\mathcal{D}$ and evaluation.

Text. We collect and aggregate textual safety data from BeaverTails (Ji et al., 2023), WildGuardMix (Han et al., 2024), Aegis 2.0 (Ghosh et al., 2025), and ToxicChat (Lin et al., 2023) for training and evaluation. Additionally, we include OpenAI Moderation (Markov et al., 2023) for evaluation.

Image. We collect image safety data from UnsafeBench (Qu et al., 2024), VLGuard (Zong et al., 2024), and Llava-Guard (Helff et al., 2025a) for both training and evaluation.

Video. For constructing the dataset $\mathcal{D}$, we collect video safety data from SafeSora (Dai et al., 2024a), Fake Video Corpus (Papadopoulou et al., 2019), LSPD (Phan et al.), TikHarm (Balat et al., 2024), and DCSASS (Sultani et al., 2018). From SafeSora, we utilize the generated video clips along with their corresponding safety classification labels. Since the remaining datasets each target specific domains, we adopt the unified taxonomy proposed in (Chen et al., 2025) to integrate them into a comprehensive video safety corpus. For evaluation, we include SafeSora (Dai et al., 2024a), and SafeWatch-Bench (Chen et al., 2025). [2]

Audio. For training and evaluation, we leverage MuTox (Costa-jussà et al., 2024), a multilingual audio dataset for toxicity and harassment detection. We retain only the English subset. To further enhance the model's capability in audio-based safety recognition, we also construct a synthetic dataset, WildGuardMix-TTS, by randomly sampling 10K textual prompts from WildGuardMix and converting them into audio clips using text-to-speech (TTS) model openaudio-s1-mini (Liao et al., 2024).

Cross-Modal. Due to the scarcity of cross-modal safety training datasets with fine-grained annotations, we incorporate VLSBench (Hu et al., 2025), a carefully curated benchmark for visual-language safety evaluation, to strengthen the guardrail foundation for cross-modal safety alignment. For evaluation, we use MM-SafetyBench (Liu et al., 2024b), Video-SafetyBench (Liu et al., 2025a), and AIAH (Yang et al., 2025), corresponding to image-text, video-text, and auio-text configurations.

[2]We did not include SafeWatch-Bench (Chen et al., 2025) for training, as their training split was unavailable at the time of our work.

B. Formatted Prompt

Formatted Prompt

[Task] You are an LLM Judge tasked with evaluating the safety of input based on the Safety Guidelines.
Analyze the given input, identify any potential violations, and provide a detailed explanation for your judgment.

{Safety Guidelines}

[Input]
{Text} / <Image> / <Video> / <Audio>

[Instructions]
1. Identify whether the input is “safe” or “unsafe”.
2. If “unsafe”, list the violated categories separated by semicolons.
3. Provide a detailed explanation of whether the input is safe or unsafe.

[Output Format]
- **Safety Assessment:** {Safety Label}
- **Categories:** {Violation Categories}
- **Explanation:** [Provide detailed reasoning here]

Figure 5. Prompt template used for target distillation from teacher models when generating safety critiques. Provided with the safety label (safe or unsafe) and the corresponding violation categories, the teacher models are instructed to produce a detailed explanation describing the rationale behind the safety assessment.

C. Baselines

Model	Citation	Size	Version
Large Language Models			
Qwen3-235B	(Team, 2025)	235B	Qwen/Qwen3-235B-A22B-Instruct-2507
LLaMA-3.3-70B	(Grattafiori et al., 2024)	70B	Llama-3.3-70B-Instruct
Qwen2.5-72B	(Qwen et al., 2025)	72B	Qwen2.5-72B-Instruct
Qwen2.5-7B	(Qwen et al., 2025)	7B	Qwen2.5-7B-Instruct
LLaMA Guard 1	(Inan et al., 2023)	7B	LlamaGuard-7b
LLaMA Guard 2	(Team, 2024b)	8B	Meta-Llama-Guard-2-8B
LLaMA Guard 3	(Grattafiori et al., 2024)	8B	Llama-Guard-3-8B
ThinkGuard	(Wen et al., 2025)	8B	ThinkGuard
Vision Large Language Models			
Qwen3-VL-235B	(Team, 2025)	235B	Qwen3-VL-235B-A22B-Instruct
Qwen2.5-VL-72B	(Bai et al., 2025)	72B	Qwen2.5-VL-72B-Instruct
Qwen2.5-VL-7B	(Bai et al., 2025)	7B	Qwen2.5-VL-7B-Instruct
LlavaGuard-v1.2-7B	(Helff et al., 2025a)	7B	LlavaGuard-v1.2-7B-OV-hf
LLaMA Guard 3V	(Chi et al., 2024)	11B	Llama-Guard-3-11B-Vision
LLaVA-Video-72B	(Zhang et al., 2025)	72B	LLaVA-Video-72B-Qwen2
LLaVA-Video-7B	(Zhang et al., 2025)	7B	LLaVA-Video-7B-Qwen2
Audio Large Language Models			
Qwen2-Audio	(Chu et al., 2024)	7B	Qwen2-Audio-7B
Qwen-Audio	(Chu et al., 2023)	8B	Qwen-Audio-Chat
Kimi-Audio	(KimiTeam et al., 2025)	7B	Kimi-Audio-7B-Instruct
Omni-Modal Large Language Models			
GPT-4o	(OpenAI, 2024)	-	gpt-4o-2024-11-20, gpt-4o-audio-preview-2025-06-03
Qwen2.5-Omni-7B	(Xu et al., 2025a)	7B	Qwen2.5-Omni-7B

Table 7. Configuration details of baseline models used in evaluation, including Large Language Models (LLMs), Large Vision-Language Models (LVLMs), and Large Audio Language Models (LALMs). "–" denotes information unavailable. For GPT-4o, we employed gpt-4o-2024-11-20 for text, image, and video evaluations, and gpt-4o-audio-preview-2025-06-03 for audio-related assessments, as a truly omni-modal API endpoint was not publicly available at the time of evaluation.