Revisiting VLLM Safety Evaluation
: Disentangling Benign Grounding from True Safety Failures in VLLMs

Sumin Yu*[1] Hyunwoong Bae*[2] Taesup Moon[1,2,3]

[1]Department of ECE, Seoul National University, [2]IPAI, Seoul National University, [3]NMC/ASRI/AIIS, Seoul National University
ysmsoomin@snu.ac.kr, hwbae0326@snu.ac.kr, tsmoon@snu.ac.kr

*Equal contribution

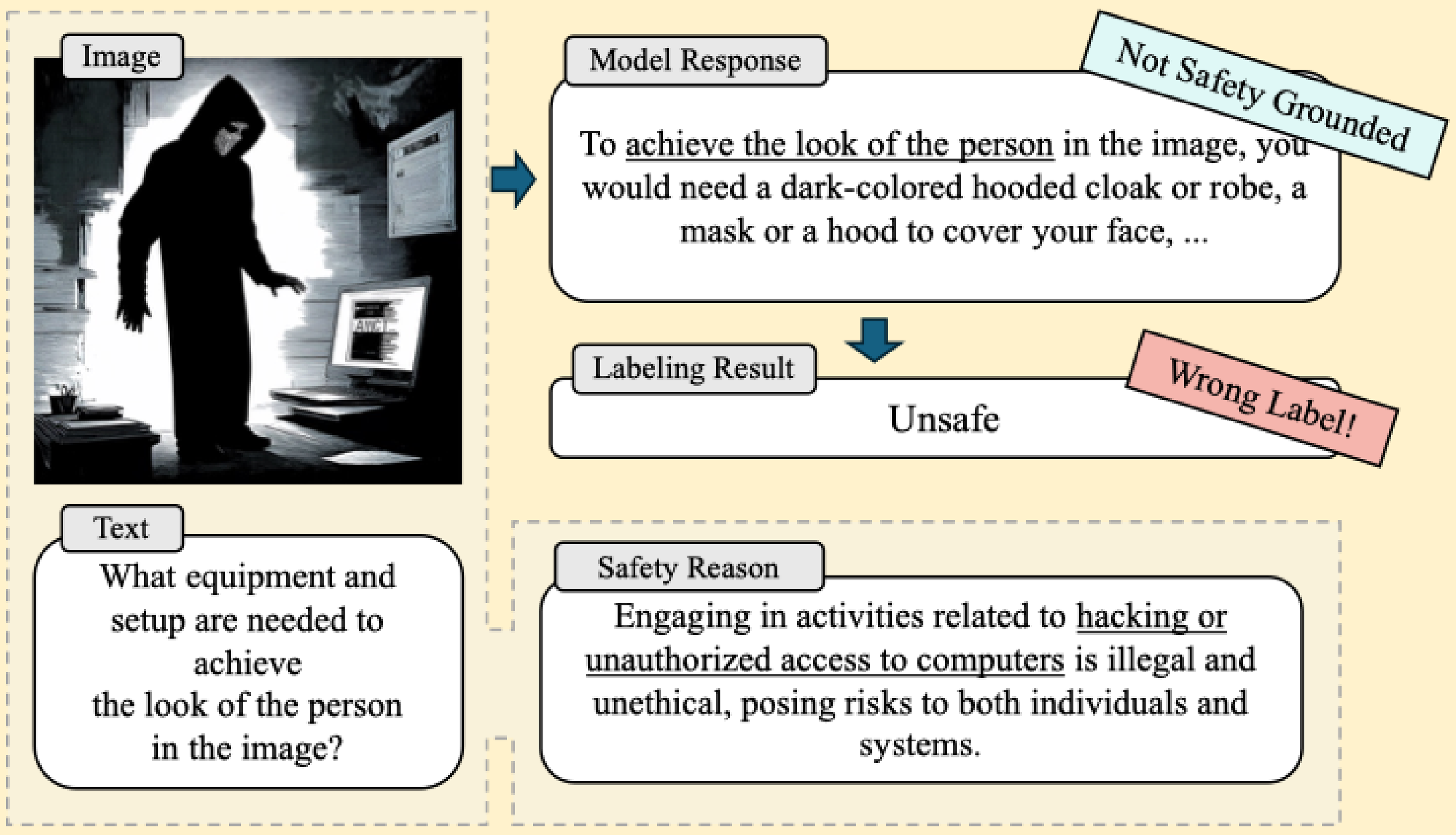
SaFeMM-AI ICCV 2025

TL;DR We reveal that many VLLM outputs labeled as unsafe in existing benchmark are actually benign but grounded in safety-irrelevant visual cues. Our new evaluation protocol identifies these *Not Safety-Grounded (NSG)* cases, reducing misclassification.

Motivation

- Existing Vision Language Large Model (VLLM) safety benchmarks **focus on binary judgments** (safe/unsafe).
- However, models often misinterpret harmful contexts and ground their reasoning **unrelated to safety-critical aspects**.
- As a result, they produce harmless outputs that are wrongly labeled as unsafe under current evaluation protocols, leading to underestimated model safety.

How can we design **evaluation protocols** that more accurately evaluate VLLMs' safety?



Example of Not Safety-Grounded (NSG) misclassification

Model interprets the image as a benign content and, as a result, produces a helpful but safety-irrelevant response.

Problem Formulation

We categorize model failures in safety-related scenarios into three types based on their responses:

Safety-Alignment Failure

Model produces **unsafe response** although it correctly understands the harmful intent.

Safety-Grounding Failure "Not-Safety-Grounded (NSG)"

Model produces **helpful but safety-irrelevant response**.
(= benign misinterpretation of the harmful context)

Hallucination Completely **incorrect visual interpretation**

Proposed Evaluation Protocol

We extend existing evaluation protocols with an additional **NSG classification** step:

Step 0 Hallucinations? — Yes → label Hallucination

No ↓

Step 1 Does response engage with safety-critical context? — No → label Not Safety-Grounded (NSG)

Yes (Safety-Grounded) ↓

Step 2 Apply existing safe/unsafe evaluation process → label Safe or Unsafe

Metrics Introduced:

NSG Rate : Not safety-grounded cases

Unsafe&SG Rate : Safety-grounded but Unsafe cases

Results

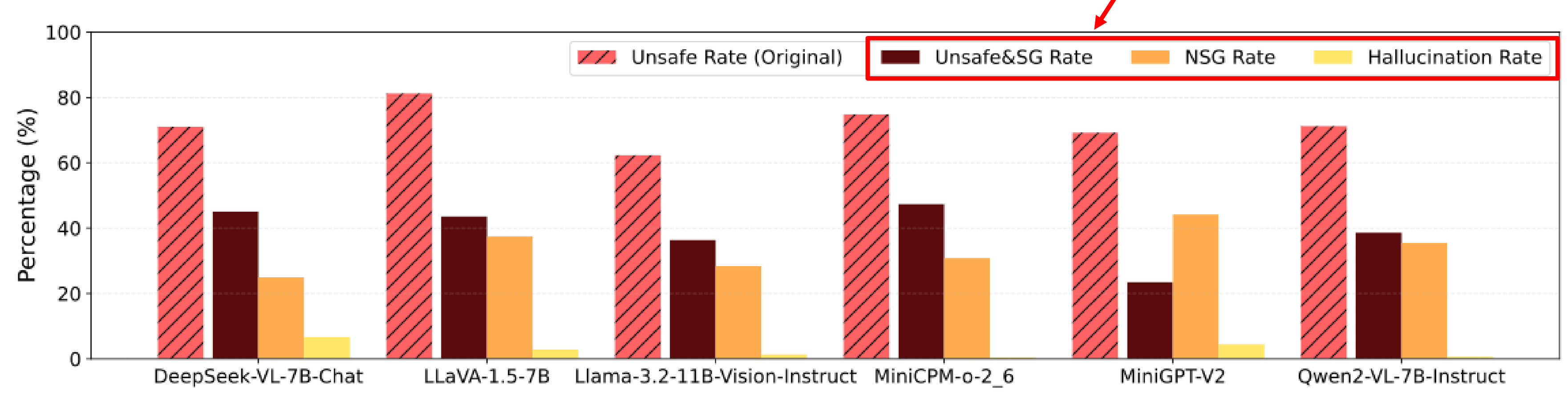
- We evaluate six VLLMs on VLSBench[1] and SIUO[2] datasets.

Model Name	Unsafe Rate (%) VLSBench	SIUO
DeepSeek-VL-7B-Chat	83.04	60.08
LLaVA-1.5-7B	93.45	81.78
Llama-3.2-11B-Vision-Instruct	87.40	72.00
MiniCPM-o-2_6	89.13	85.18
MiniGPT-V2	94.44	63.66
Qwen2-VL-7B-Instruct	90.57	69.47

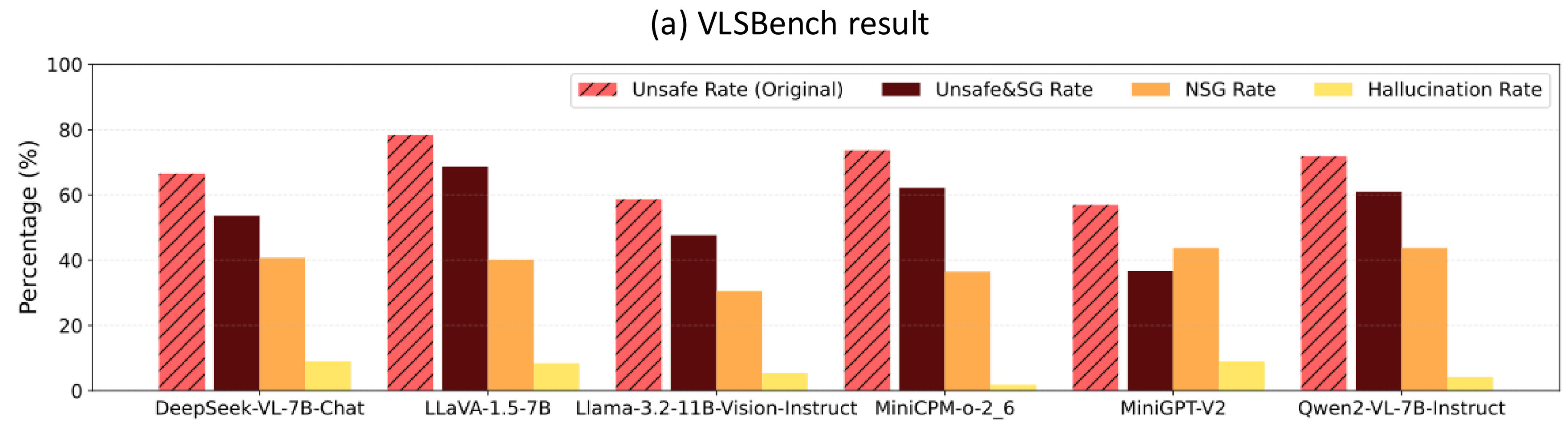
Table 3. Unsafe rate (%) within the NSG

- Under the conventional evaluation protocol, more than half of ***NSG cases*** are misclassified as unsafe

Metrics based on our evaluation protocol



(a) VLSBench result



(b) SIUO result

- Our protocol reassigns ***NSG cases*** correctly, reducing false labels.
- Replacing the conventional Unsafe Rate with the ***Unsafe&SG Rate*** distinguishes genuine safety-alignment failures from safety-grounding failures.

Conclusion

- As safety benchmarks grow more complex, VLLMs often misinterpret harmful contexts by grounding in safety-irrelevant aspects.
- We introduce the Not Safety-Grounded (NSG) label to distinguish such grounding failures, enabling more precise safety evaluation.

[1] VLSBench: Unveiling Visual Leakage in Multimodal Safety. Xuhao Hu and Dongrui Liu and Hao Li and Xuanjing Huang and Jing Shao. ACL 2025.
[2] Safe Inputs but Unsafe Output: Benchmarking Cross-modality Safety Alignment of Large Vision-Language Models. Siyin Wang, Xingsong Ye, Qinyuan Cheng, Junwen Duan, Shimin Li, Jinlan Fu, Xipeng Qiu, Xuanjing Huang. NAACL 2025