Vipera: Towards Systematic Auditing of Generative Text-to-Image Models At Scale

Paper Demo

Yanwei Huang
huangyw@zju.edu.cn

Wesley Deng
hanwend@andrew.cmu.edu

Sijia Xiao
xiaosijia@cmu.edu

Motahhare Eslami
meslami@andrew.cmu.edu

Jason I. Hong
jasonh@cs.cmu.edu

Adam Perer
adamperer@cmu.edu

What is AI auditing?

AI auditing refers to detecting and addressing within AI by analyzing system outputs for adversarial inputs.

We focus specifically on auditing **Text-to-Image** (T2I) models, which may generate images with biases, offense, or misleading information.

Stable Diffusion

Systematic Auditing At Scale

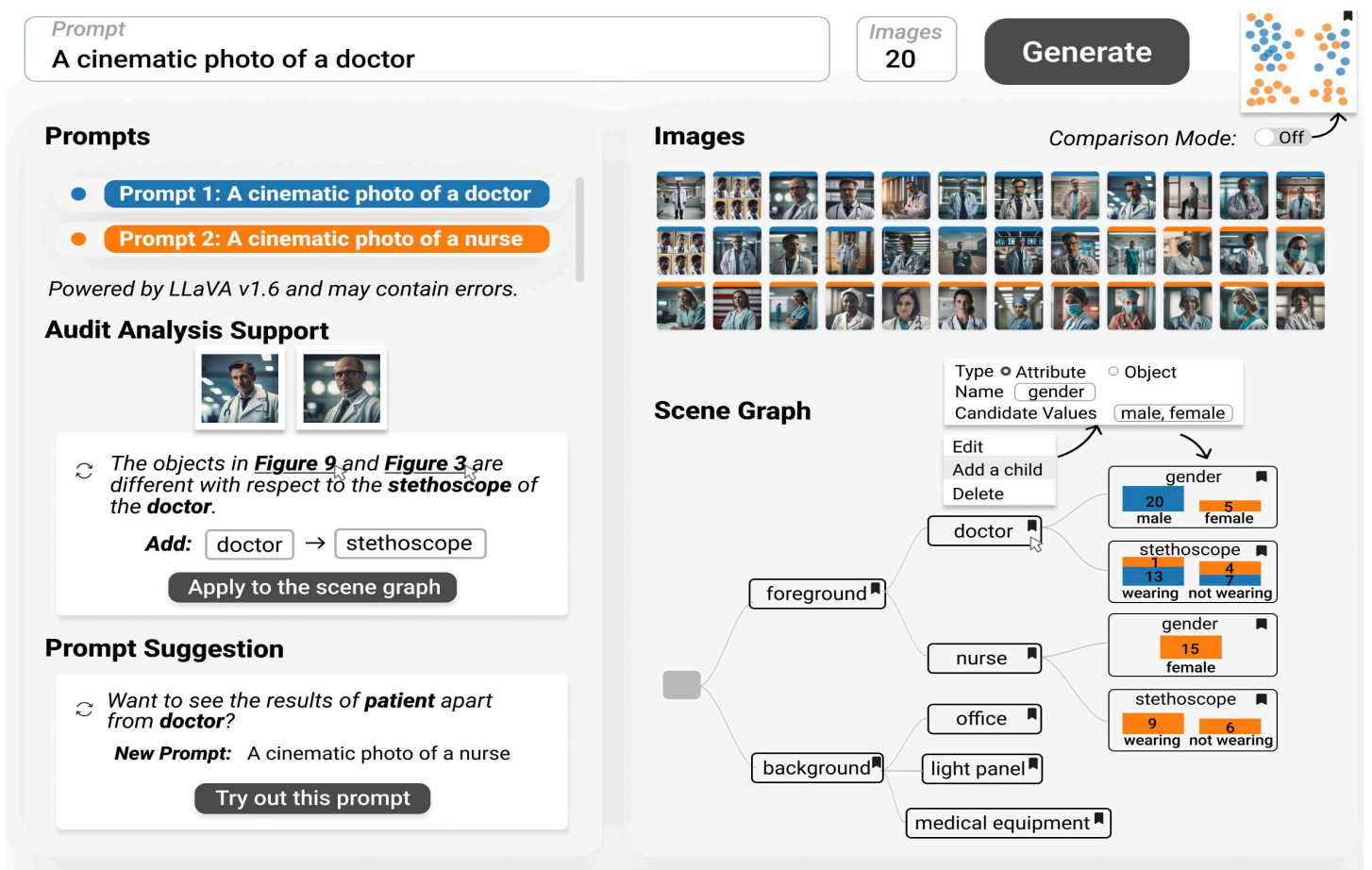
Two challenges arise in systematically exploring the large auditing space:

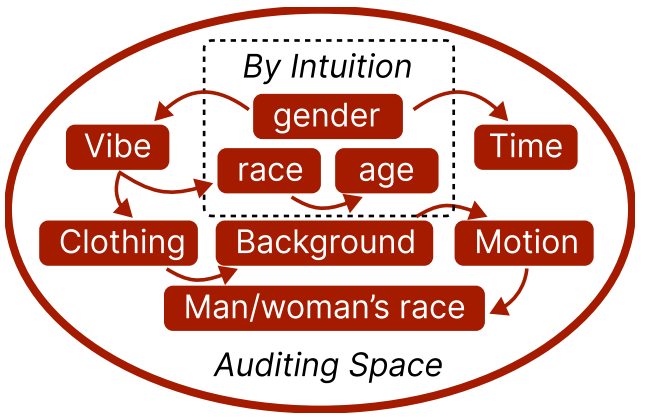
- ***Unknown Unknowns*** - How to keep auditors aware of unexplored auditing criteria?
- ***Structured Exploration*** - How to streamline auditing with formal and structured navigation?

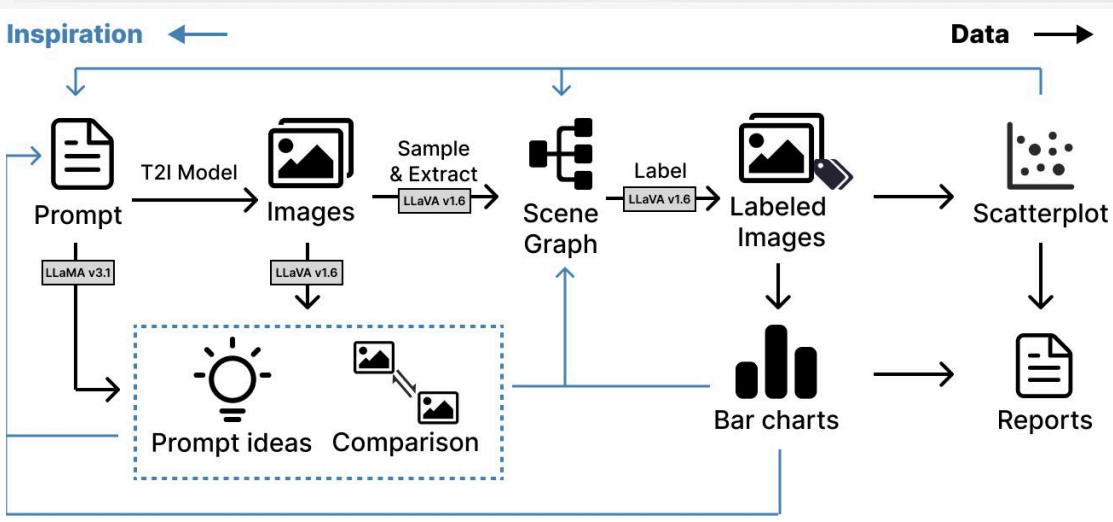
Vipera 's Design & Pipeline

***V**isual **I**ntelligence-**P**romoted **E**nd User **A**uditing*

Vipera uses **LLM-generated guidance** to keep inspiring auditors, while using **visualization guidance** to encourage structured exploration.







Findings

- **Vipera improves the breadth of auditing** by facilitating (a) the exploration of various auditing criteria, (b) the validation of user hypotheses, and (c) prompt comparisons.
- **Next, we plan to** (a) record auditing provenance for recall & storytelling, (b) enable fine-grained auditing scope customization, (c) improve intent communication in human-AI collaborative auditing.

CoALA CMU DIG Carnegie Mellon University