AI Research Nexus 2025

AI Researchers Association Annual Conference (New Zealand)

FaceMaskGPT: Building a Secure Visual Reasoning System

Xinyi Gao

Department of Computer and Information Sciences

Auckland University of Technology, 1010 New Zealand

In this talk, I will introduce our project: FaceMaskGPT. This work explores how we can build a rule-aware visual reasoning pipeline using ComfyUI and LLaMA 3.2 to automatically assess face mask compliance based on online guidance.

Our motivation stems from two key problems: First, can large vision-language models understand real-world compliance rules, such as those from WHO or CDC? And can these models apply those rules to images to decide whether someone is wearing a mask correctly?

We also address a challenge in generative models: Hallucination. These generative models generate answers for the questions that are plausible but untrue. Our goal is to build a system that learns rules automatically, applies them to image-based reasoning, and produces grounded, explainable outputs.

The core task we aim to solve is this: Given an input image, can the model judge whether the mask-wearing is correct, based on rules from the real world? To do this, we use LLaMA 3.2, a powerful multimodal model that can process both images and text. We enhance it with rules retrieved either live from the web or from a local knowledge base we built. Since hallucinations can occur when a model lacks external grounding, our solution includes both prompt-level control and rule injection, so the model always has something solid to base its answer on.

We thus discuss that how the full system works, built in ComfyUI, a visual workflow environment. We start with an image input. It passes through an NSFW Detection module. If the image contains sensitive content like violence or nudity, it is automatically replaced with a “CONTENT SENSITIVE” placeholder image. If safe enough, the image is then sent into our YOLOv11 object detection module, which identifies people and whether they appear to be wearing masks. Meanwhile, the system retrieves mask compliance rules — either from the internet (such as WHO or CDC websites), or from our local rule-based knowledge base. We hence combine the detection results and rule information into a structured prompt, which is passed to LLaMA 3.2. The model take use of both the image and the rules to reason about compliance. Finally, a post-filtering step can be applied to detect or suppress

hallucinated or unsafe output.

We input an image showing four people. YOLOv11 detects three people wearing masks. At the same time, the system retrieves a WHO guideline: “Masks must cover the nose, mouth, and chin.” We generate a prompt:“Analyze the image and answer the following questions:- Is the person wearing a mask?- If the mask is visible, is it worn correctly?” The LLaMA model answers:“The person is wearing a mask incorrectly. The mask only covers the mouth.” This answer is interpretable, safe, and grounded — because the model was provided both factual visual input and rule-based textual context.

Our early experiments show promising results: The system generates clear, interpretable answers; It helps reduce hallucinations by anchoring the model in external rules and visual detection results; The ComfyUI-based design makes every module transparent, replaceable, and extensible — so researchers can easily update components like the detection model or the rule source. This makes the system a strong candidate for trustworthy compliance analysis.

Therefore, we plan to expand the system in three main ways: We’re exploring YOLOE, a real-time object detection model that supports open-vocabulary detection. This would allow us to handle more diverse scenes with better speed and flexibility. We plan to extend the system to video stream analysis, so it can process continuous footage, such as live camera feeds in hospitals or public venues — enabling real-time monitoring. We want to generalize this pipeline to other compliance-related scenarios — for example, in sports environments like basketball to check jersey compliance, or in industrial sites to verify helmet and PPE usage.

References

Gao, X., Nguyen, M., Yan, W. (2021) Face image inpainting based on generative adversarial network. International Conference on Image and Vision Computing New Zealand.

Gao, X., Nguyen, M., Yan, W. (2022) A face image inpainting method based on autoencoder and adversarial generative networks. Pacific-Rim Symposium on Image and Video Technology.

Gao, X. (2022) A Method for Face Image Inpainting Based on Generative Adversarial Networks. Master’s Thesis, Auckland University of Technology, New Zealand.

Gao, X., Nguyen, M., Yan, W. (2023) Human face mask detection using YOLOv7+CBAM in deep learning. Handbook of Research on AI and ML for Intelligent

Machines and Systems

Gao, X., Nguyen, M., Yan, W. (2023) Enhancement of human face mask detection performance by using ensemble learning models. PSIVT.

Gao, X., Nguyen, M., Yan, W. (2023) A high-accuracy deformable model for human face mask detection. PSIVT.

Gao, X., Nguyen, M., Yan, W. (2024) HFM-YOLO: A novel lightweight and high-speed object detection model. Optimization, Machine Learning, and Fuzzy Logic: Theory, Algorithms, and Applications (Chapter 16). IGI Global.

Lee, J., Song, K. U., Yang, S., Lim, D., Kim, J., Shin, W., ... & Kim, T. H. (2025). Efficient LLaMA-3.2-vision by trimming cross-attended visual features. arXiv preprint arXiv:2504.00557.

Leu, W., Nakashima, Y., & Garcia, N. (2024). Auditing image-based NSFW classifiers for content filtering. In Proceedings of ACM Conference on Fairness, Accountability, and Transparency (pp. 1163-1173).

Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., ... & Peng, W. (2024). A survey on hallucination in large vision-language models. arXiv preprint arXiv:2402.00253.

Vallayil, M., Nand, P., Yan, W., Allende-Cid, H. (2024) Explainable AI through thematic clustering and contextual visualization: Advancing macro-level explainability in AFV systems. Australasian Conference on Information Systems.

Vallayil, M., Nand, P., Yan, W., Allende-Cid, H. (2025) CARAG: A context-aware retrieval framework for fact verification, integrating local and global perspectives of explainable AI. Applied Sciences.

Xu, G., Yan, W. (2023) Facial emotion recognition using ensemble learning. Deep Learning, Reinforcement Learning, and the Rise of Intelligent Systems, pp.146-158, Chapter 7, IGI Global.

Yan, W. (2019) Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics. Springer Nature.

Yan, W. (2023) Computational Methods for Deep Learning: Theory, Algorithms, and Implementations. Springer Nature.

Yang, B., Yan, W. (2024) Real-time billiard shot stability detection based on YOLOv8.

Deep Learning, Reinforcement Learning, and the Rise of Intelligent Systems, pp.159-172, Chapter 8, IGI Global.

Yang, G., Nguyen, M., Yan, W., Li, X. (2025) Foul detection for table tennis serves using deep learning. Electronics 2025, 14(1), 27.

Yang, G. (2025) ChatPPG: Multi-Modal Alignment of Large Language Models for Time-Series Forecasting in Table Tennis. Master's Thesis, Auckland University of Technology, New Zealand.

Zheng, A., Yan, W. (2024) Attention-based multimodal fusion model for breast cancer diagnostics. ICONIP 2024.

Zhu, W., Peng, B., Yan, W. (2024) Dual knowledge distillation on multiview pseudo labels for unsupervised person re-identification. IEEE Transactions on Multimedia.

Zhu, Y., Yan, W. (2022) Image-based storytelling using deep learning. ACM ICCCV.