Probing Large Multimodal Models (LMMs) via Semantic Information Pursuit

IARPA BENGAL Proposer's Day

Johns Hopkins University: Rama Chellappa

University of Pennsylvania: René Vidal

JOHNS HOPKINS
MATHEMATICAL INSTITUTE
for DATA SCIENCE

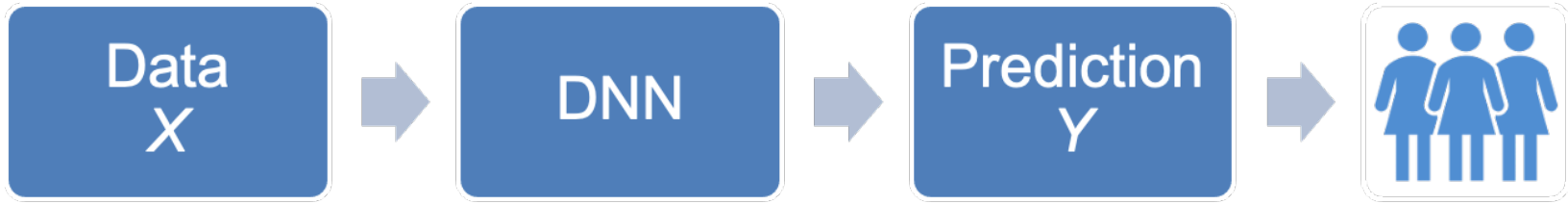
Penn
UNIVERSITY *of* PENNSYLVANIA

Recent DoD Projects

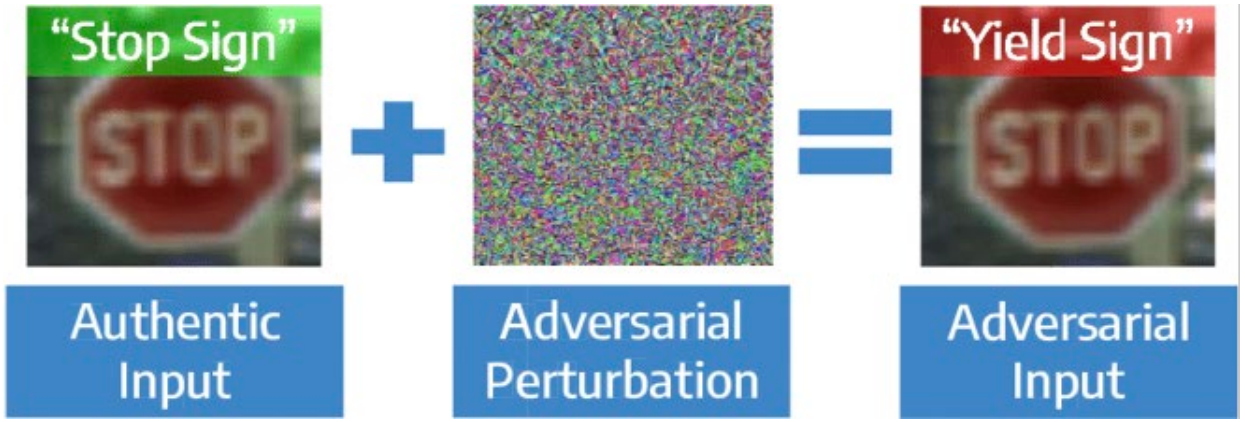
- **Rama Chellappa:**
  - IARPA: WRIVA, BRIAR, DIVA, JANUS
  - ONR MURI: Foundations of Deep Learning
  - ARO MURI: Semantic Information
  - DARPA: Guaranteeing AI Robustness Against Deception (GARD)

- **René Vidal:**
  - IARPA: WRIVA, BRIAR, DIVA
  - ONR MURI: Control and Learning Enabled Verifiable Robust AI (CLEVR-AI)
  - ARO MURI: Semantic Information
  - DARPA: Guaranteeing AI Robustness Against Deception (GARD)
  - DARPA: Reverse Engineering Deception (RED)

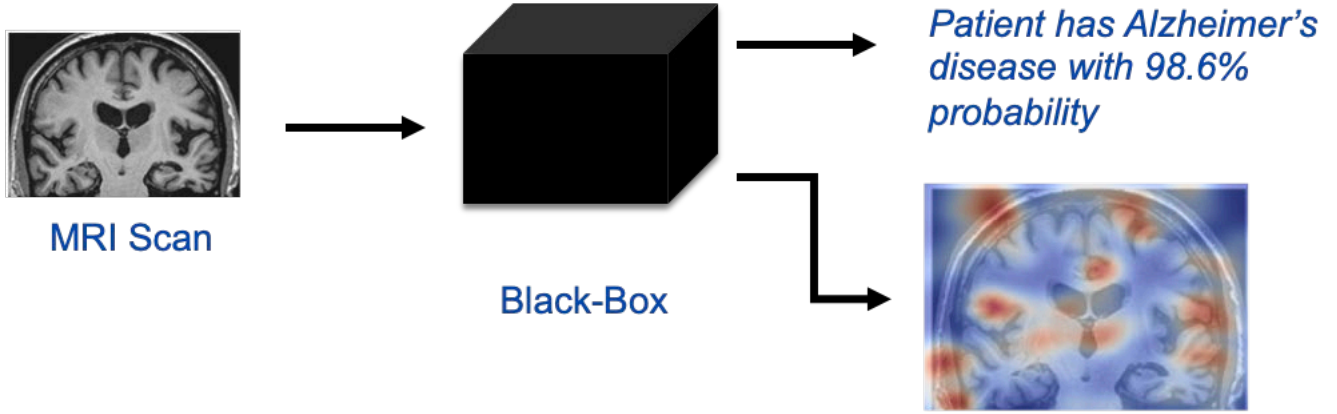
Guarantees of Performance of AI Methods



- **Robustness:**

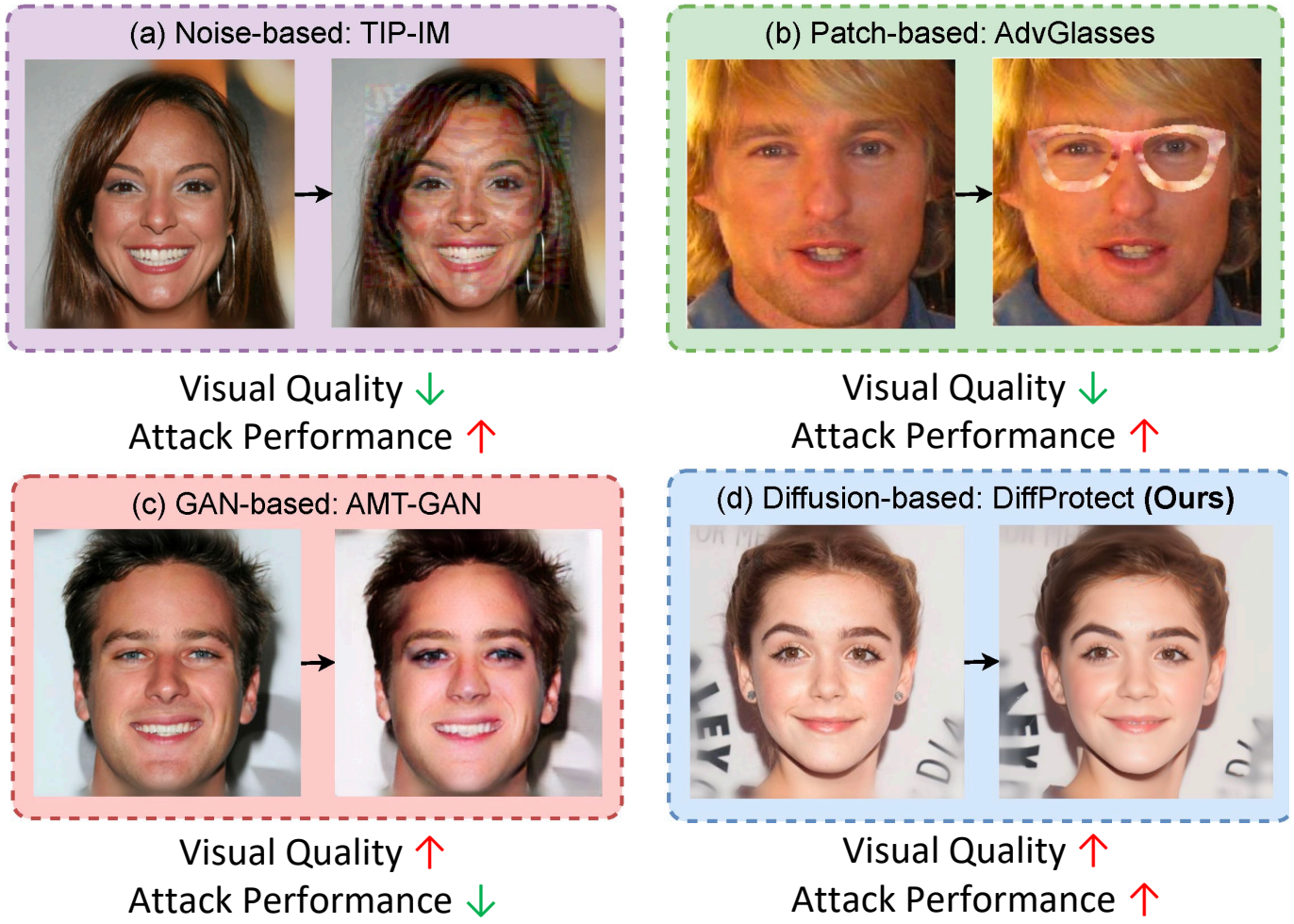


- **Explainability:**



Prior Work: Diffusion-Model Based Attacks

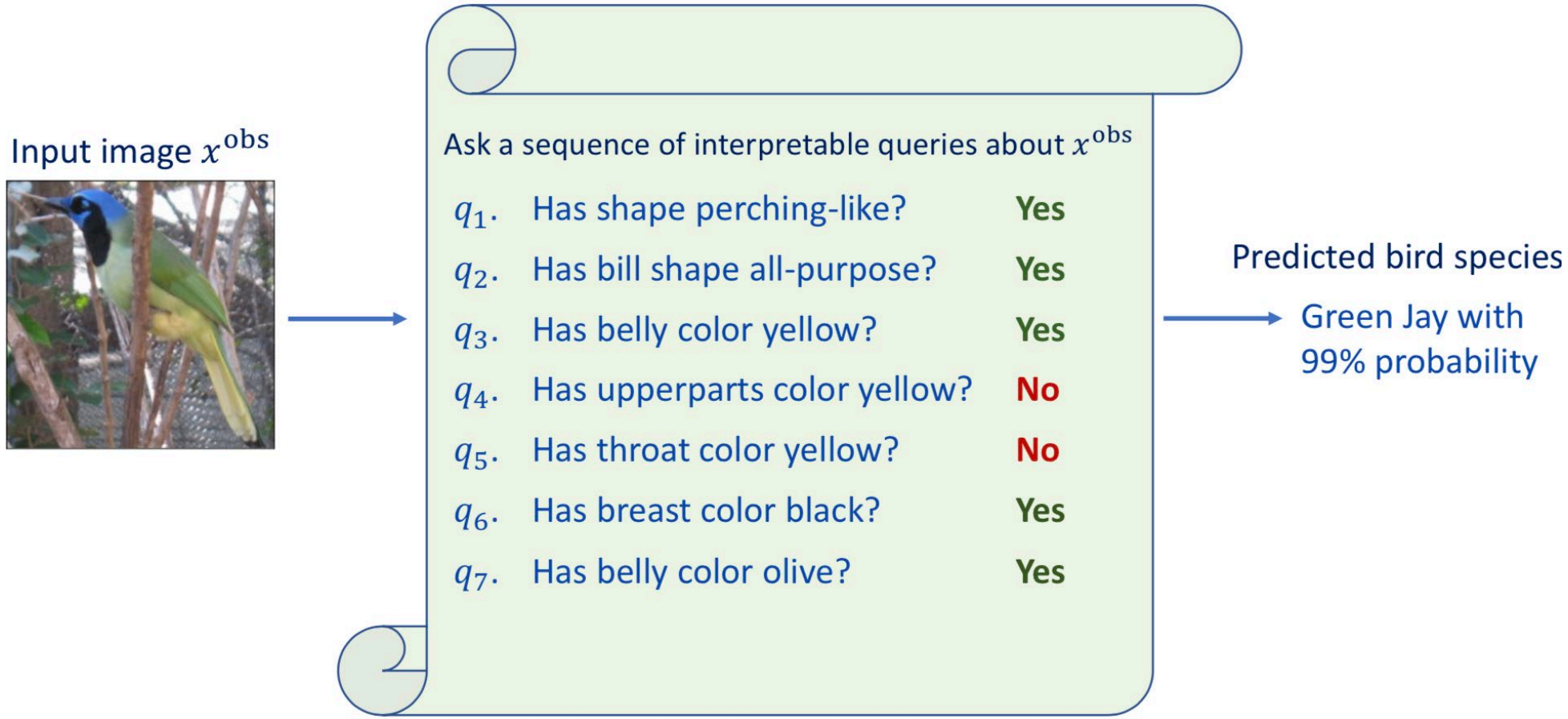
- **DiffProtect:**
  - Facial privacy protection
  - Tradeoff between attack performance and visual quality



J Liu, CP Lau, R Chellappa. DiffProtect: Generate Adversarial Examples with Diffusion Models for Facial Privacy Protection. arXiv preprint arXiv:2305.13625, 2023

Prior Work: Explainable AI by Design

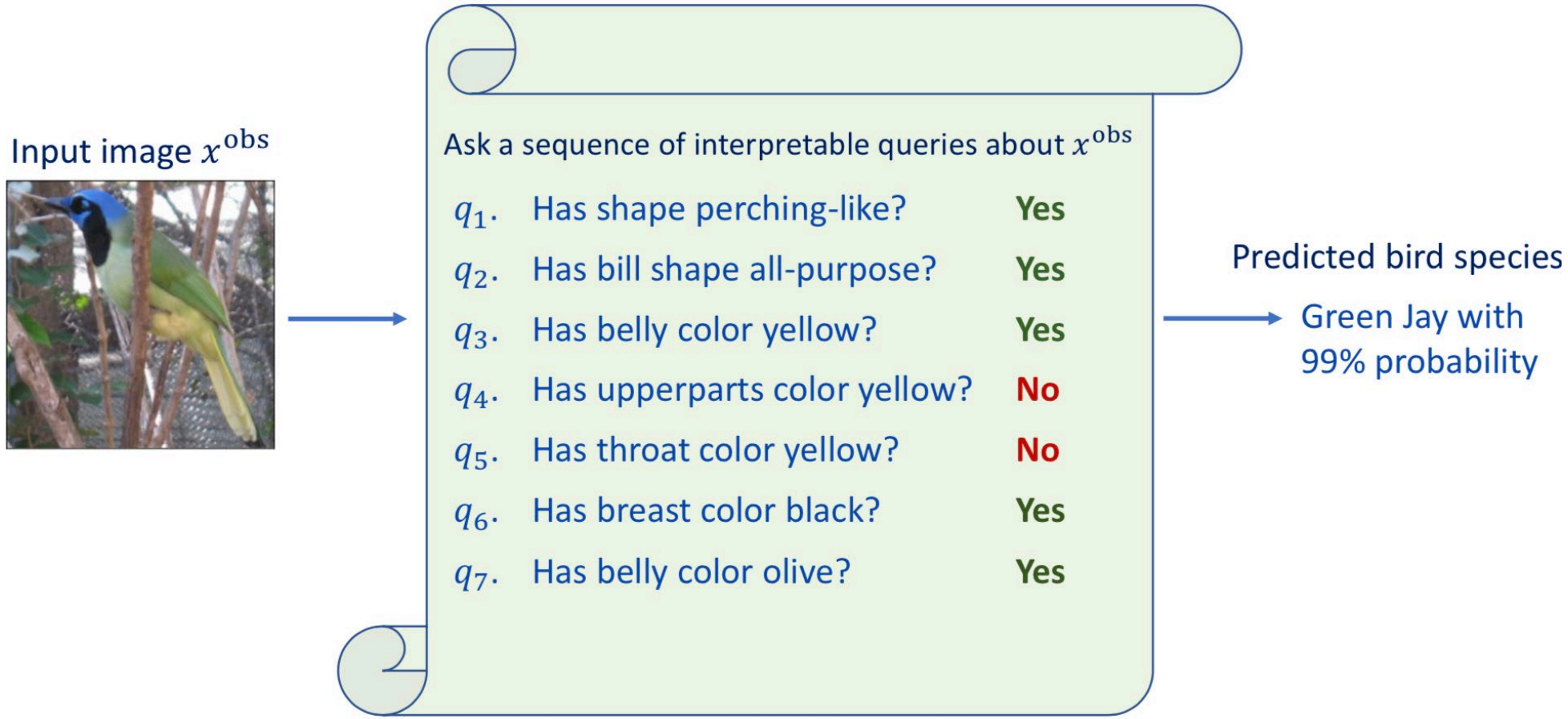
- **Semantic Information Pursuit:**
  - Map input to sequence of “questions” and “answers”
  - Make predictions based on “most informative” questions and answers



Chattopadhyay, Slocum, Haeffele, Vidal, Geman. Interpretable by Design: Learning Predictors by Composing Interpretable Queries, TPAMI 2022.
Chattopadhyay, Chan, Haeffele, Geman, Vidal. Variational Information Pursuit for Interpretable Predictions, ICLR 2023.

Prior Work: Explainable AI by Design

- **Semantic Information Pursuit:**
  - LLMs generate "imperfect questions" and LMMs generate "imperfect" answers
  - Robustness is achieved by selecting the "most informative" questions and answers



KHR Chan, A Chattopadhyay, BD Haeffele, R Vidal. Variational Information Pursuit with Large Language and Multimodal Models for Interpretable Predictions arXiv preprint arXiv:2308.12562, 2023

Proposed Work for BENGAL

- **Probing LMMs via Semantic Information Pursuit**

- Semantic information pursuit provides explanations for both “correct” and “incorrect” predictions, thus allowing us to “diagnose” the model

- When the model makes a mistake, we know why, and we can use the explanation to “correct mistakes”

- We can thus use semantic information pursuit to detect, characterize and mitigate LMM threats and vulnerabilities