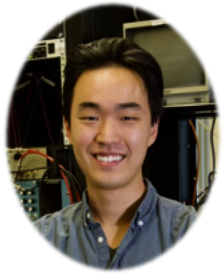
Strong and Precise Modulation of Human Percepts via Robustified ANNs



*G. Gaziv**   *M. Lee**   *Jim DiCarlo*

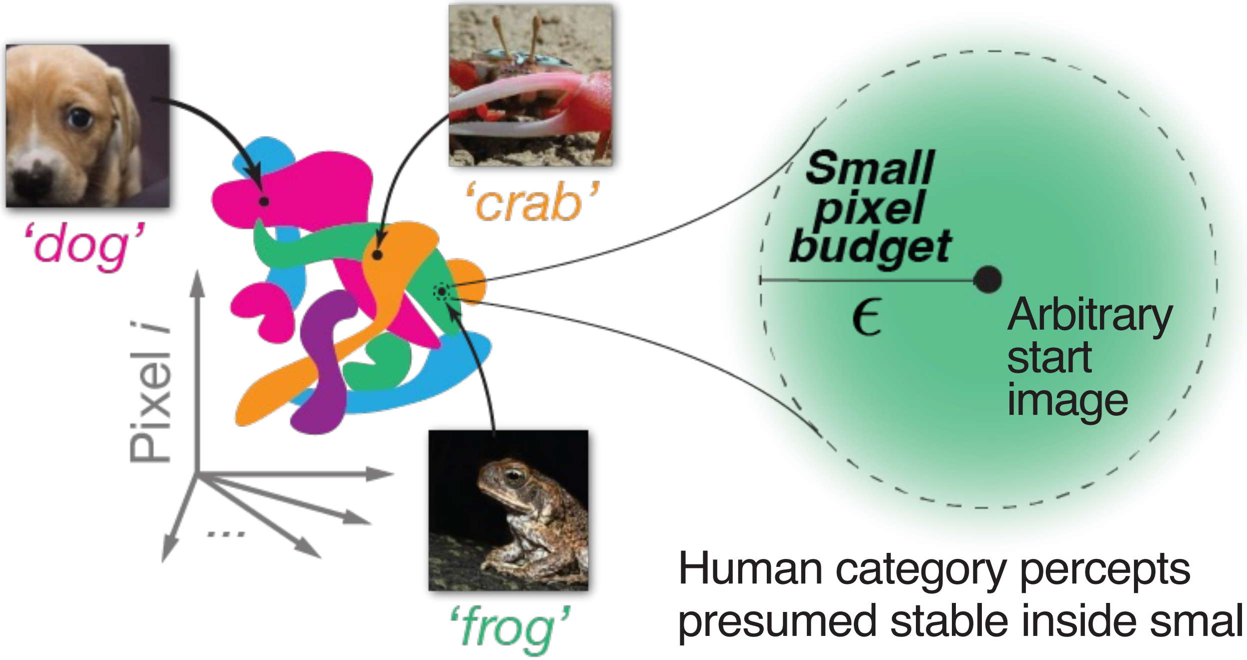
Dept of Brain & Cognitive Sciences, MIT

aka ***“Robustified ANNs Reveal Wormholes Between Human Category Percepts”***

Is perception robust as believed?

Human category percepts over pixel space

Prevailing assumption prior to this research



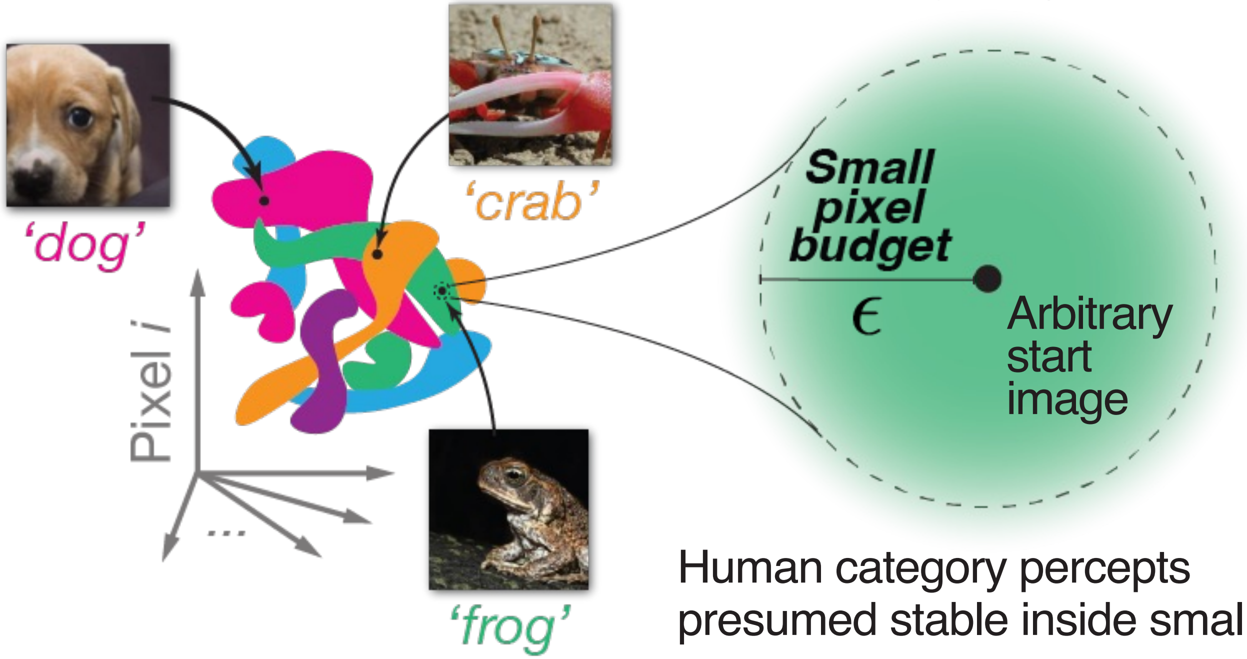
Human category percepts presumed stable inside small ϵ

Gaziv, Lee* & DiCarlo, **NeurIPS** (2023)*

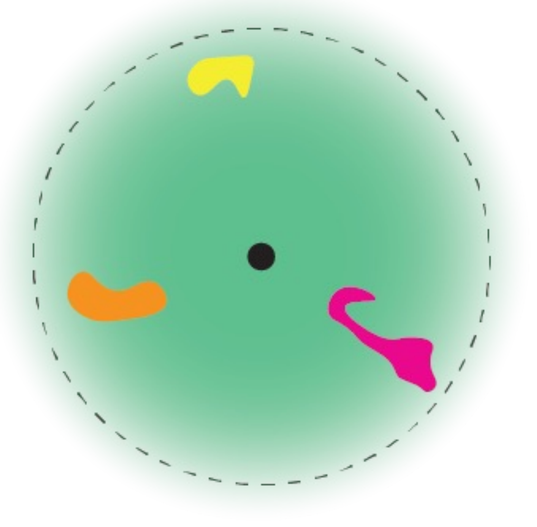
The truth of human biology near any image start point

Human category percepts over pixel space

Prevailing assumption prior to this research

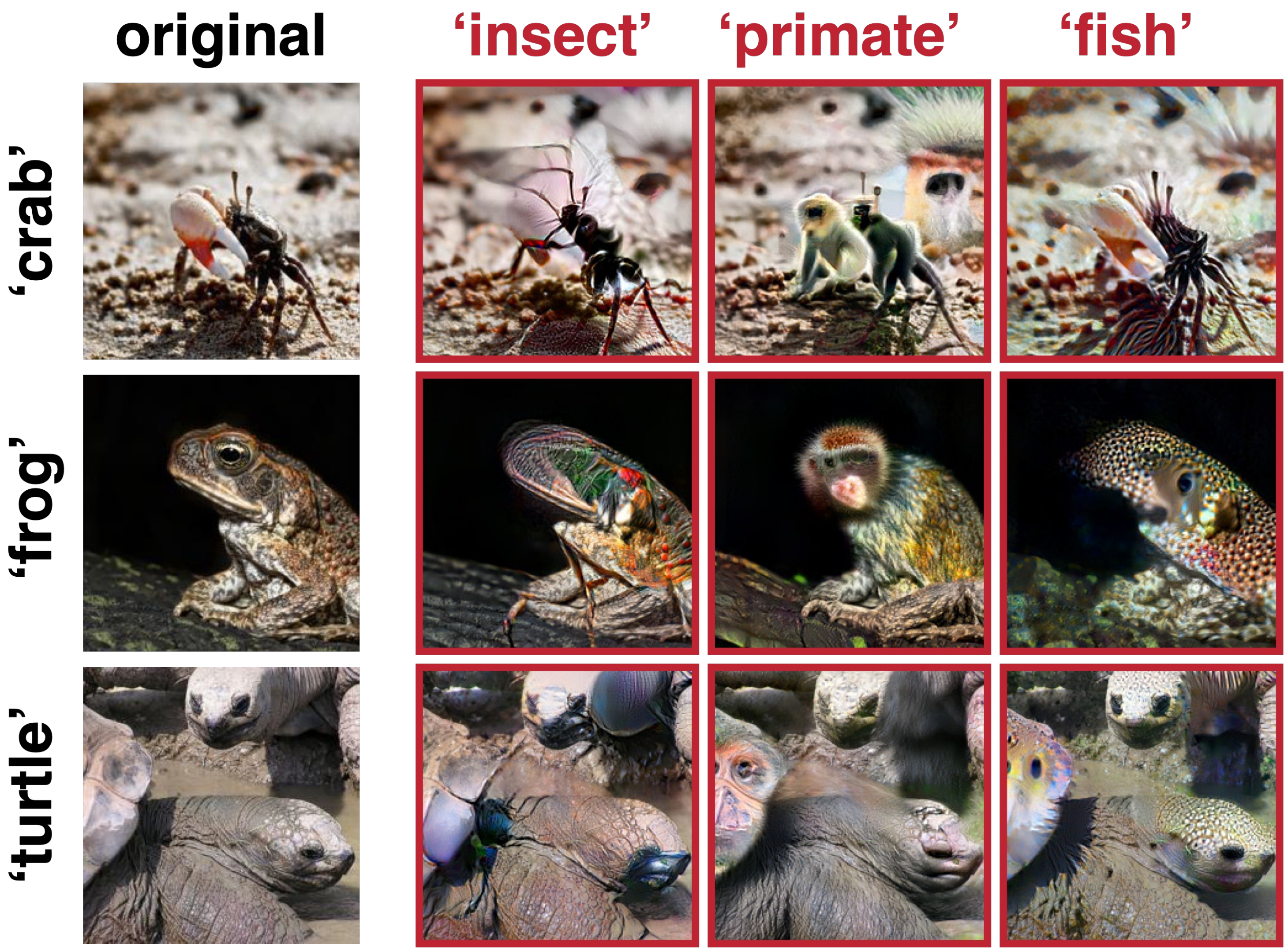


Human category percepts presumed stable inside small ϵ



Gaziv, Lee* & DiCarlo, **NeurIPS** (2023)*

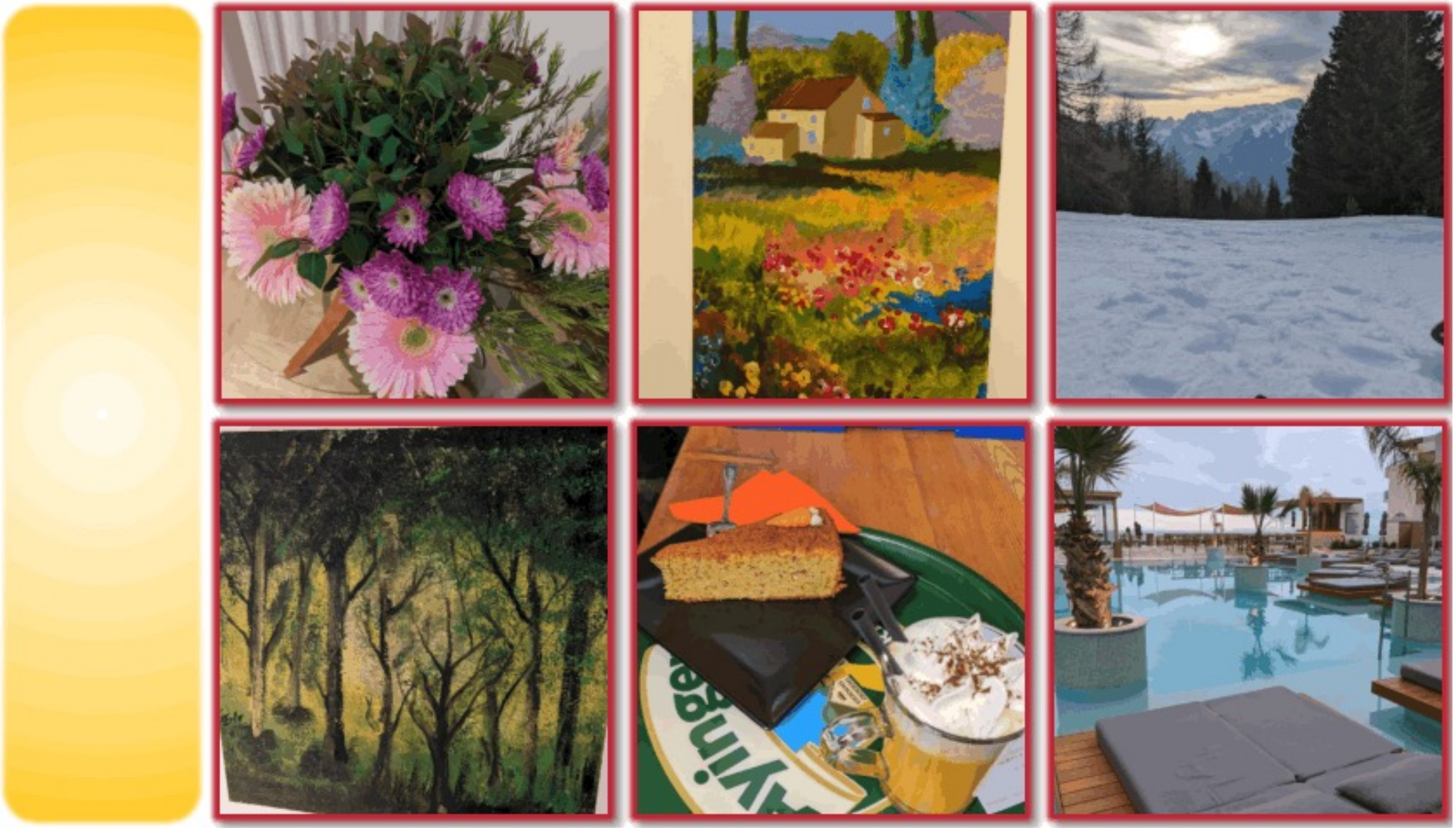
Robustified models allow for precise Targeted Modulation of human behavior in the “human-presumed-insensitive” pixel budget regime



ϵ=30

Gaziv, Lee* & DiCarlo,* ***NeurIPS*** *(2023)*

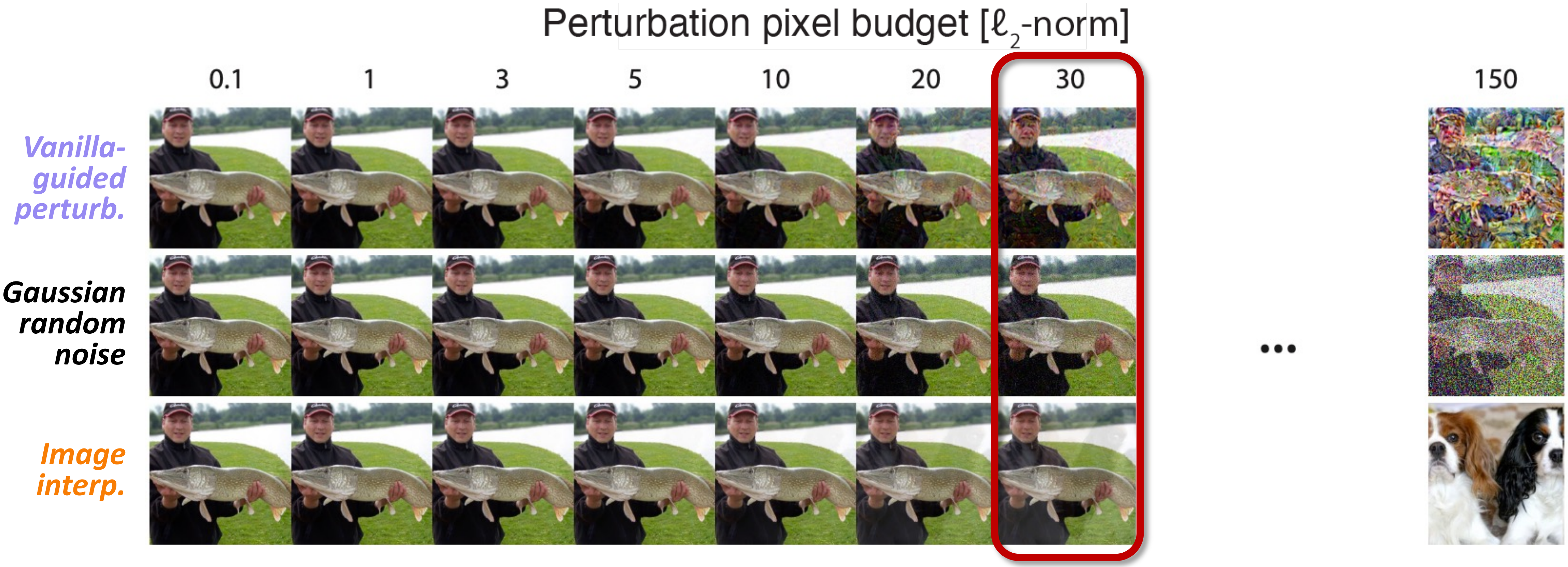
Targeted Modulation successfully modulates human reports for ***arbitrary*** start image



ϵ=30

https://github.com/ggaziv/Wormholes

Humans are (indeed) insensitive to random, vanilla-guided or interpolation-based perturbations in the “low norm” regime

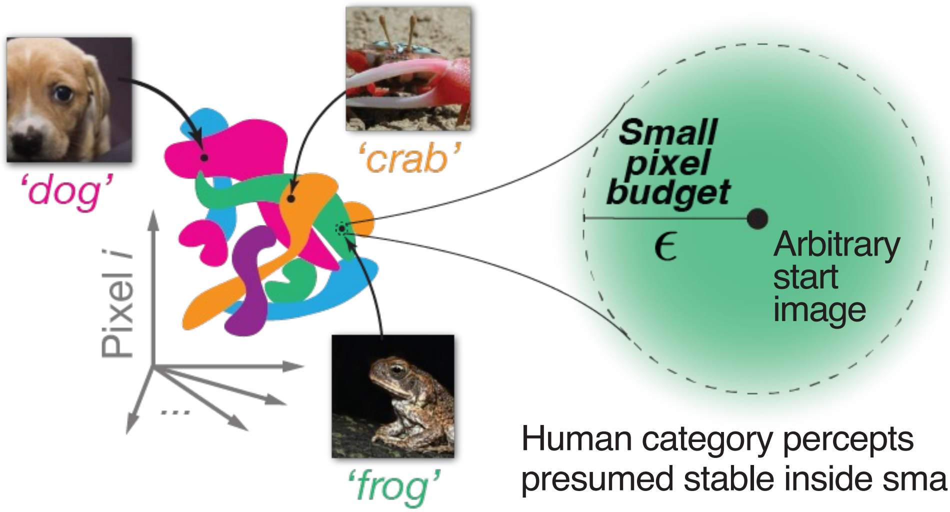


Gaziv, Lee* & DiCarlo, **NeurIPS** (2023)*

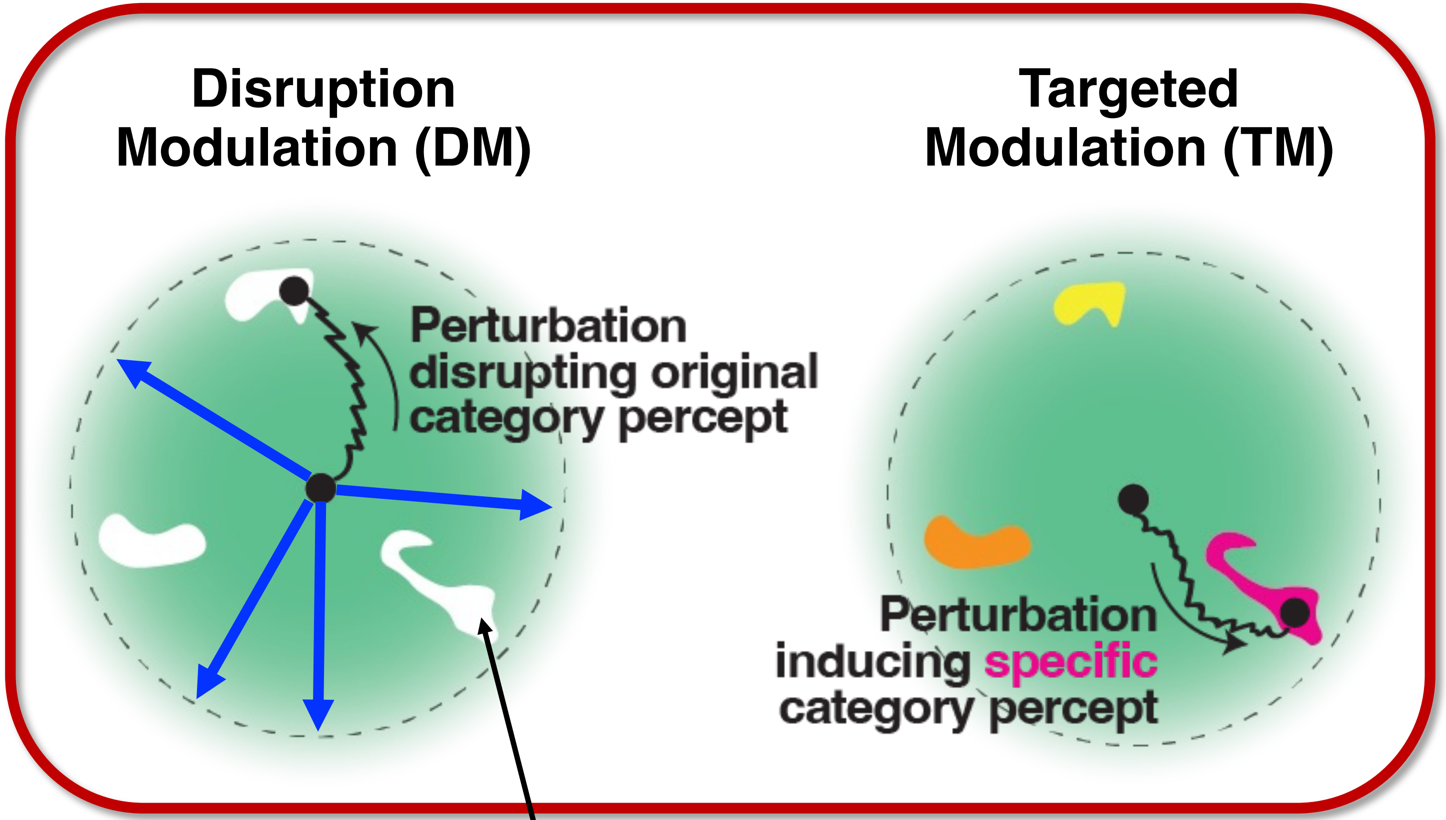
ANNs reveal “wormholes” between human category percepts

Human category percepts over pixel space

Prevailing assumption prior to this research



Human category percepts presumed stable inside small ϵ



Disruption Modulation (DM)

Perturbation disrupting original category percept

Targeted Modulation (TM)

Perturbation inducing specific category percept

Vanilla ANN vision models were unable to locate undetected “wormholes” that we now easily find

Code: https://github.com/ggaziv/Wormholes

Project Page: https://himjl.github.io/pwormholes

Gaziv, Lee* & DiCarlo, **NeurIPS** (2023)*