Structural Alignment in Worst-case Security Analysis and Multi-agent Design

Washington Garcia (UF)

Kevin Butler (UF)

Pin-Yu Chen (IBM)

Somesh Jha (UW)

Scott Clouse (AFRL/ACT3)

UF UNIVERSITY of FLORIDA

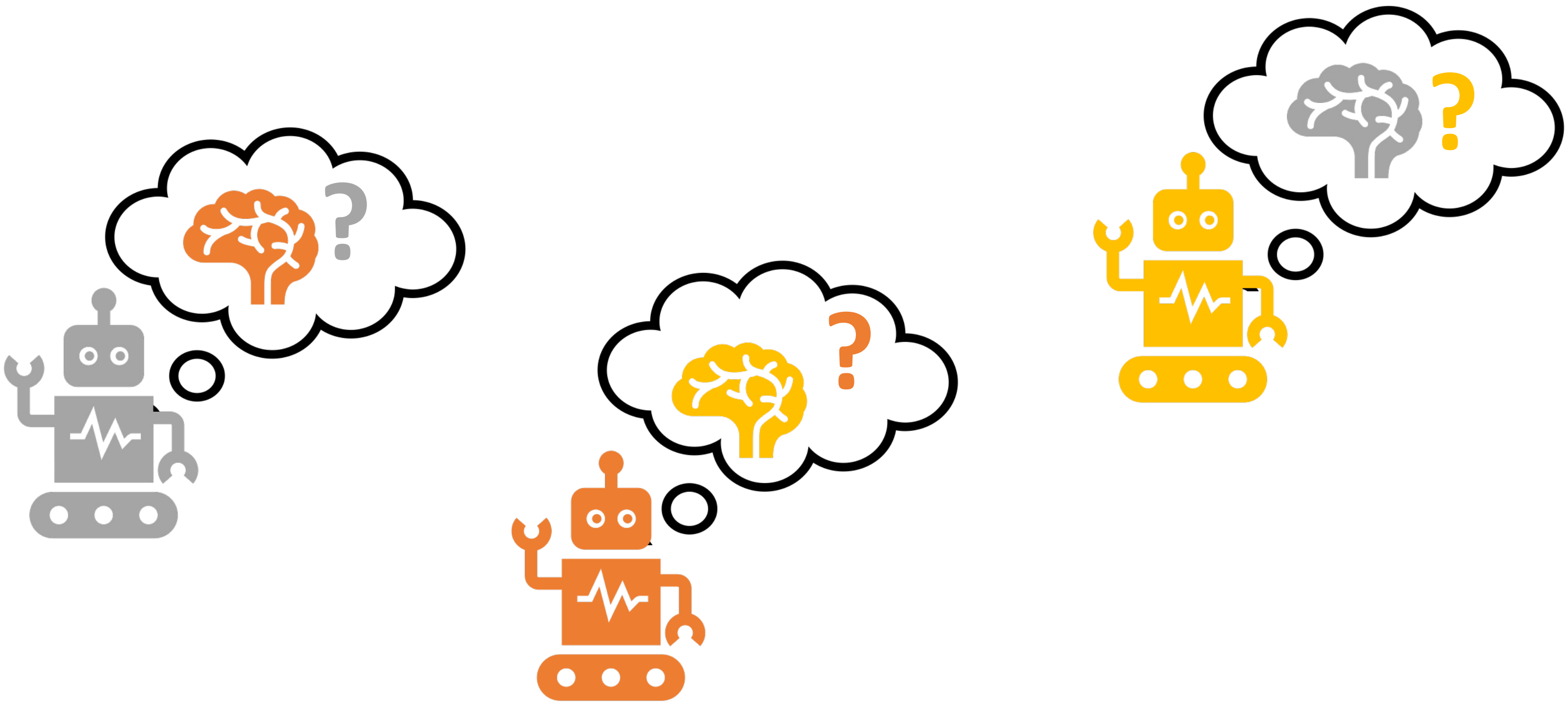
Duke UNIVERSITY

TEXAS The University of Texas at Austin

UC SANTA CRUZ

Motivation

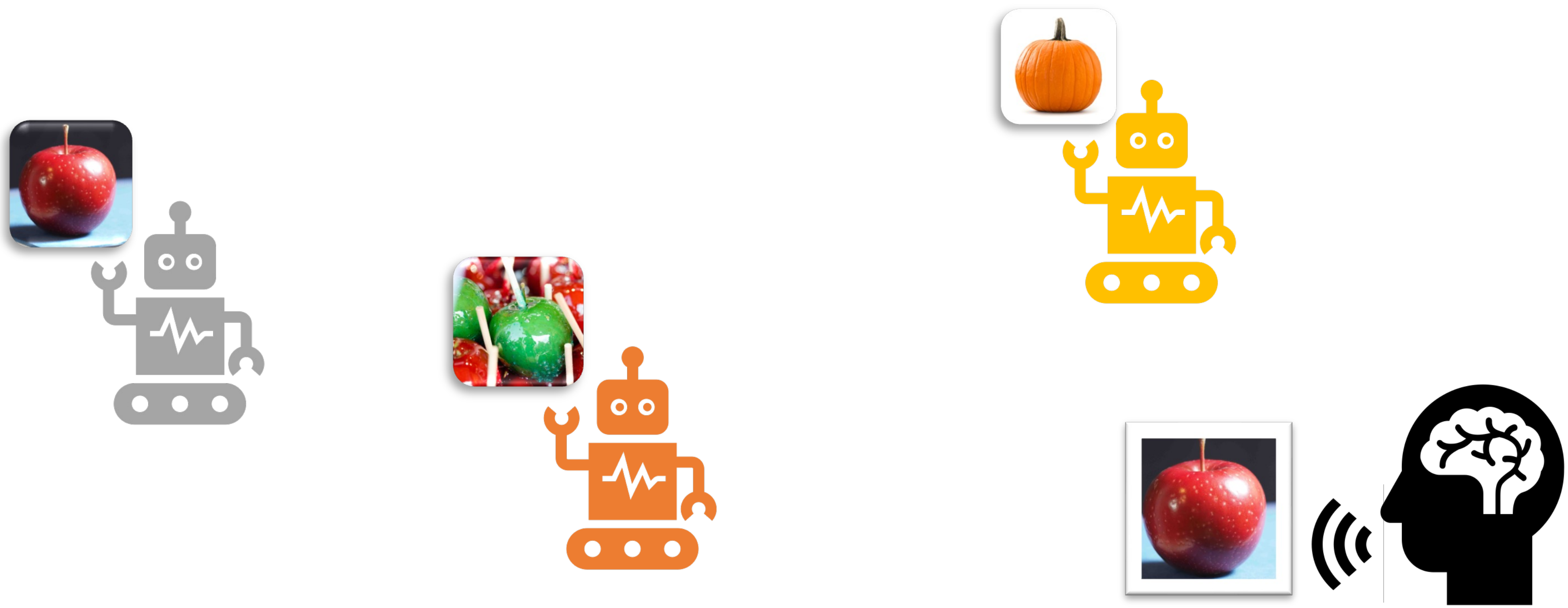
Entities in a multi-agent system must be aware of their surroundings, but also the *representational structure* of their counterparts. This can bolster situational awareness:



Motivation

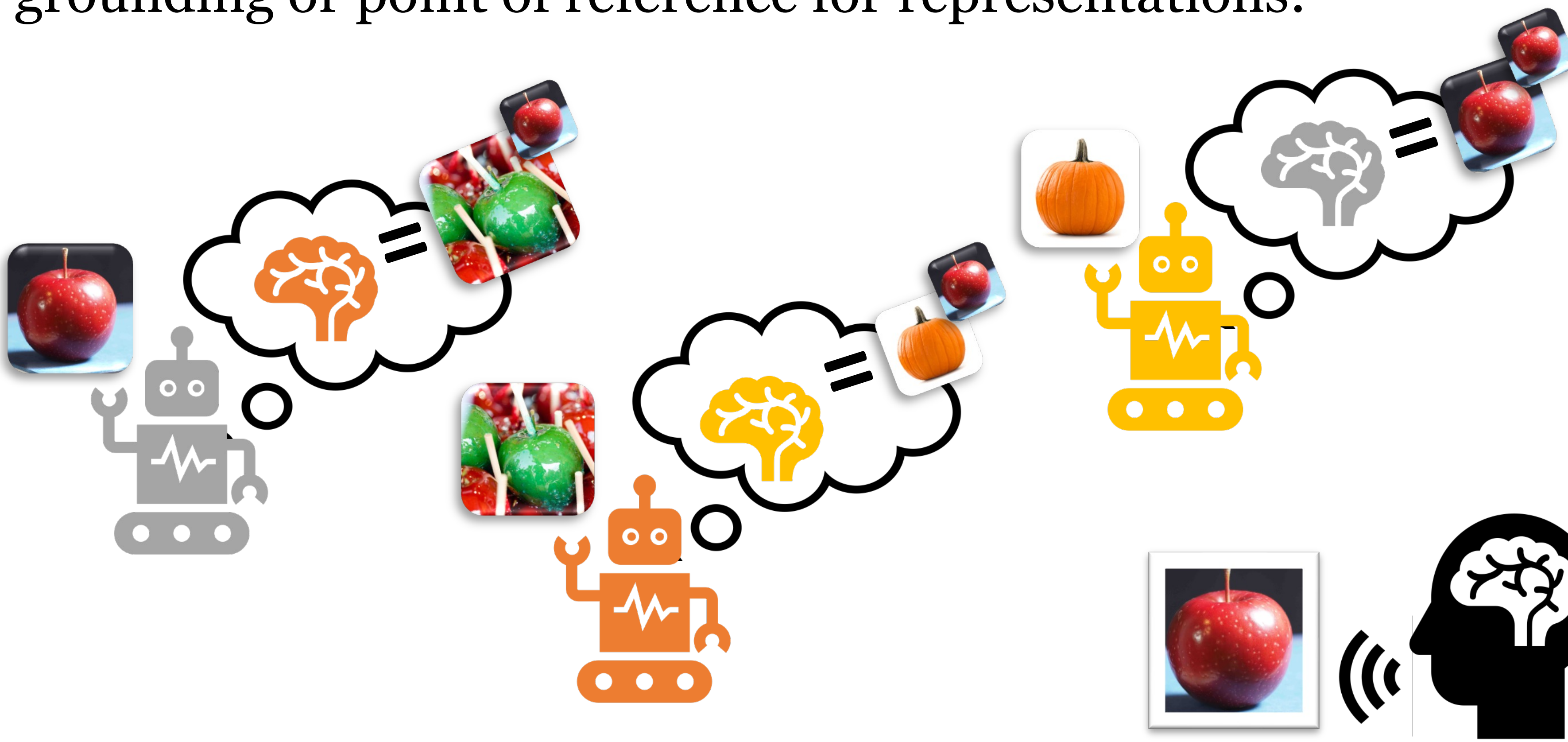
What is a good yardstick for representations?

A "human-in-the-loop" or proxy thereof can provide a grounding or point of reference for representations:



What is a good yardstick for representations?

A "human-in-the-loop" or proxy thereof can provide a grounding or point of reference for representations:



Two main questions:

1. Is agent concept resolution possible from an ***adversary's*** perspective? What's the worst-case security analysis?
2. It isn't always realistic to have a human-in-the-loop, or fine-grained data labels. What are ways around this?

A1: Expansion of our previous hard-label paper, in submission to IEEE SaTML 2023.

A2: Presented initial idea @ NAACL 2022. Introduced expansion during previous meeting. Then worked on it as part of AFRL 2022 internship.

Motivation

Two main questions:

1. Is agent concept resolution possible from an ***adversary's*** perspective? What's the worst-case security analysis?
2. It isn't always realistic to have a human-in-the-loop, or fine-grained data labels. What are ways around this?

A1: Expansion of our previous hard-label paper, in submission to IEEE SaTML 2023.

A2: We presented an initial idea @ NAACL 2022. Introduced expansion during previous meeting. Then worked on it as part of AFRL 2022 internship.

A1: Approach

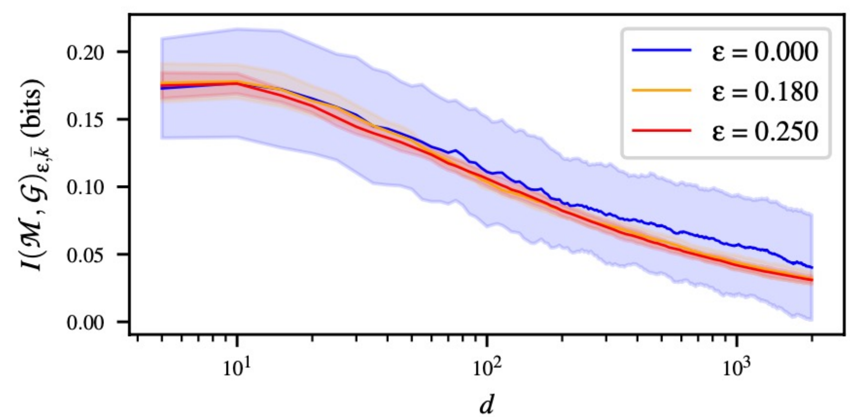
Assumptions:

1. Agent data lives on a low-dimensional manifold
2. The distribution of data points (M), true gradients (G), and ZO gradient estimates form a Markov chain:

$$\mathcal{M} \to \mathcal{G} \to \tilde{\mathcal{G}}$$

We showed previously that manifold-gradient *mutual information* can be modeled as a function of data dimension:

$$I(\mathcal{M};\mathcal{G})_{\epsilon,k} = 2\int_{\mathcal{M}^+} p(1,\mathbf{x}^+)\log\left(\frac{p(1,\mathbf{x}^+)}{p_{\mathcal{G}}(1)p_{\mathcal{M}}(\mathbf{x}^+)}\right) d\mathbf{x}^+ + 2\int_{\mathcal{M}^+} p(-1,\mathbf{x}^+)\log\left(\frac{p(-1,\mathbf{x}^+)}{p_{\mathcal{G}}(-1)p_{\mathcal{M}}(\mathbf{x}^+)}\right) d\mathbf{x}^+.$$



In practice, *does the Markov chain (MC) exist?*

First, show that the MC can be modeled by a zeroth-order (i.e., hard-label) adversary through two algorithms.

Local step neighborhood analysis (Algorithm 1):

Algorithm 1: Local Markov chain step (`MC_step`)

Input: Hard-label Gaussian process (`GP`), LIME kernel width k

Output: Sample feature coefficients $W \in \mathbb{R}^d$ and their quality score $R^2 \in \mathbb{R}$, `GP` result (`res`)

Build local neighborhood from ZO queries as Gaussian process (GP) →

initialize LIME Ridge regression trainer (`LIME`) [31]

`/* Execute GP to collect samples */`

$X, Y, \texttt{res} \leftarrow \texttt{GP}()$

Use queries to train linear model of decision boundary. →

$f_W \leftarrow \texttt{LIME}(X, \ddot{Y}, k)$

$R^2 \leftarrow f_W(X)$

return W, R^2, res

In practice, *does the Markov chain (MC) exist?*

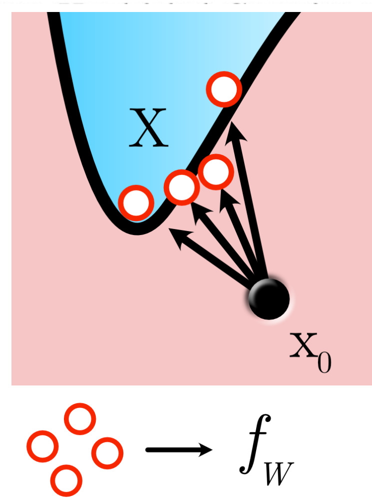
First, show that the MC can be modeled by a zeroth-order (i.e., hard-label) adversary through two algorithms.

Local step neighborhood analysis (Algorithm 1):

Algorithm 1: Local Markov chain step (`MC_step`)

Build local neighborhood from ZO queries as Gaussian process (GP) →

Use queries to train linear model of decision boundary. →



Approach

Using Algorithm 1 to model local ZO queries, we can model the whole-attack agent Markov chain through Bayesian optimization (OptiLIME).

Algorithm 2: Markov chain probing of hard-label attack

GP := (init, $\mathbf{x}_0$)
kernel width $k \leftarrow$ OptiLIME(MC_step, GP)
/* Initialize through MC_step */
Call Algorithm 1 → $W_{\texttt{init}}, R^2, \mathbf{x} \leftarrow$ MC_step(GP, k)
for $i := 1$ *to* n **do** *Hard-label attack loop*
    GP ← (approximate_gradient, $\mathbf{x}$)
    $k \leftarrow$ OptiLIME(MC_step, GP, $\mathbf{x}'$)
    /* Approximate through MC_step */
Call Algorithm 1 → $W_{\hat{\mathbf{g}}_i}, R^2, \boldsymbol{\theta} \leftarrow$ MC_step(GP, k)
    Update $\mathbf{x}$ from $\boldsymbol{\theta}$ using attack formulation
end
Get "quality" of MC models (R^2) → **return** $\mathbf{x}$, $\{W_{init}, W_{\hat{\mathbf{g}}_1}, \ldots, W_{\hat{\mathbf{g}}_n}\}$, $\{R^2_{init}, R^2_{\hat{\mathbf{g}}_1}, \ldots, R^2_{\hat{\mathbf{g}}_n}\}$

UF UNIVERSITY of FLORIDA Duke UNIVERSITY TEXAS The University of Texas at Austin UC SANTA CRUZ

Approach

Using Algorithm 1 to model local ZO queries, we can model the whole-attack agent Markov chain through Bayesian optimization (OptiLIME).

Call Algorithm 1 →

Call Algorithm 1 →

Get "quality" of MC models (R^2) →

Algorithm 2: Markov chain probing of hard-label attack

GP := (init, $\mathbf{x}_0$)

$\rightarrow f_W \rightarrow R^2_{init}$

for $i := 1$ *to* n **do** *Hard-label attack loop*

  GP ← (approximate_gradient, $\mathbf{x}$)

  $\rightarrow f_W \rightarrow R^2_{\hat{\mathbf{g}}_i}$

  Update $\mathbf{x}$ from $\boldsymbol{\theta}$ using attack formulation

end

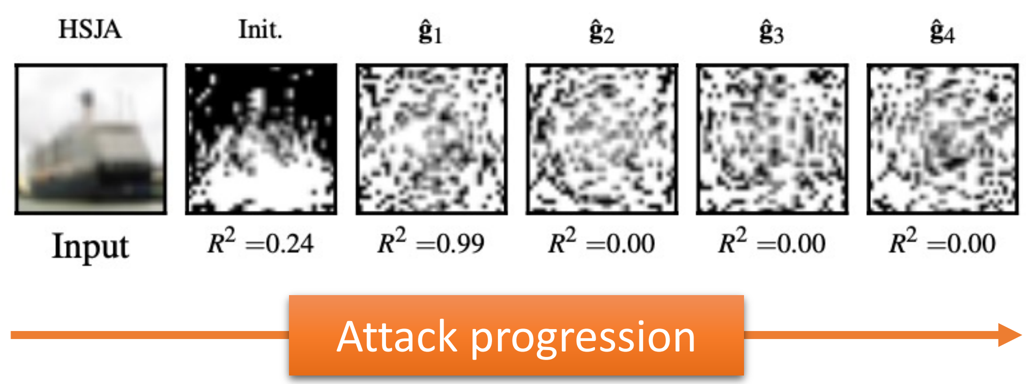
return $\mathbf{x}$, $\{W_{init}, W_{\hat{\mathbf{g}}_1}, \ldots, W_{\hat{\mathbf{g}}_n}\}$, $\{R^2_{init}, R^2_{\hat{\mathbf{g}}_1}, \ldots, R^2_{\hat{\mathbf{g}}_n}\}$

Results

Using (average) R^2 score of local models, we can answer the following:

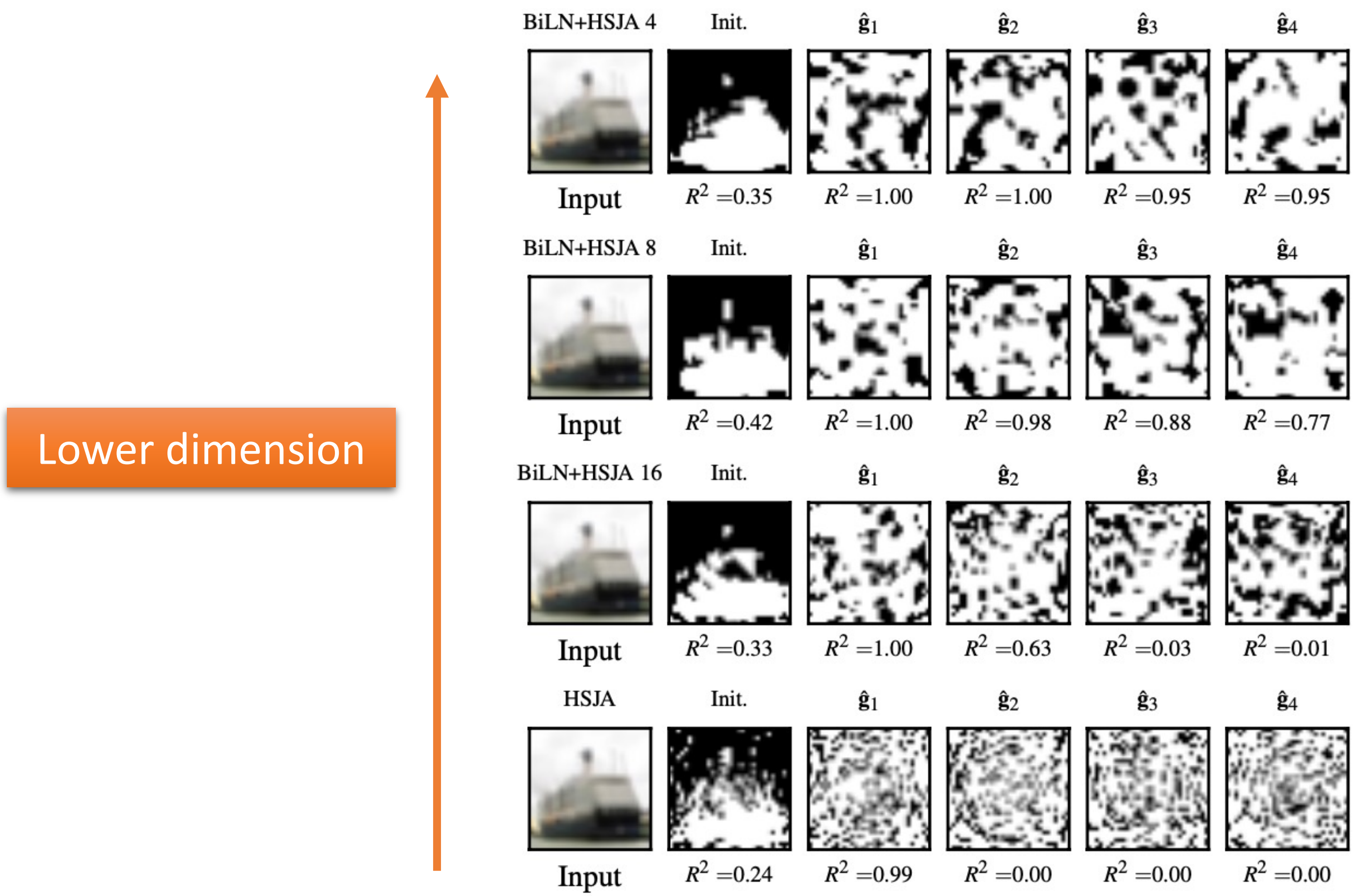
1. Are hard-label queries sufficient to model the model's semantic structure in the query neighborhood?
2. Does dimension-reduction influence our structural knowledge?

A1. Yes:

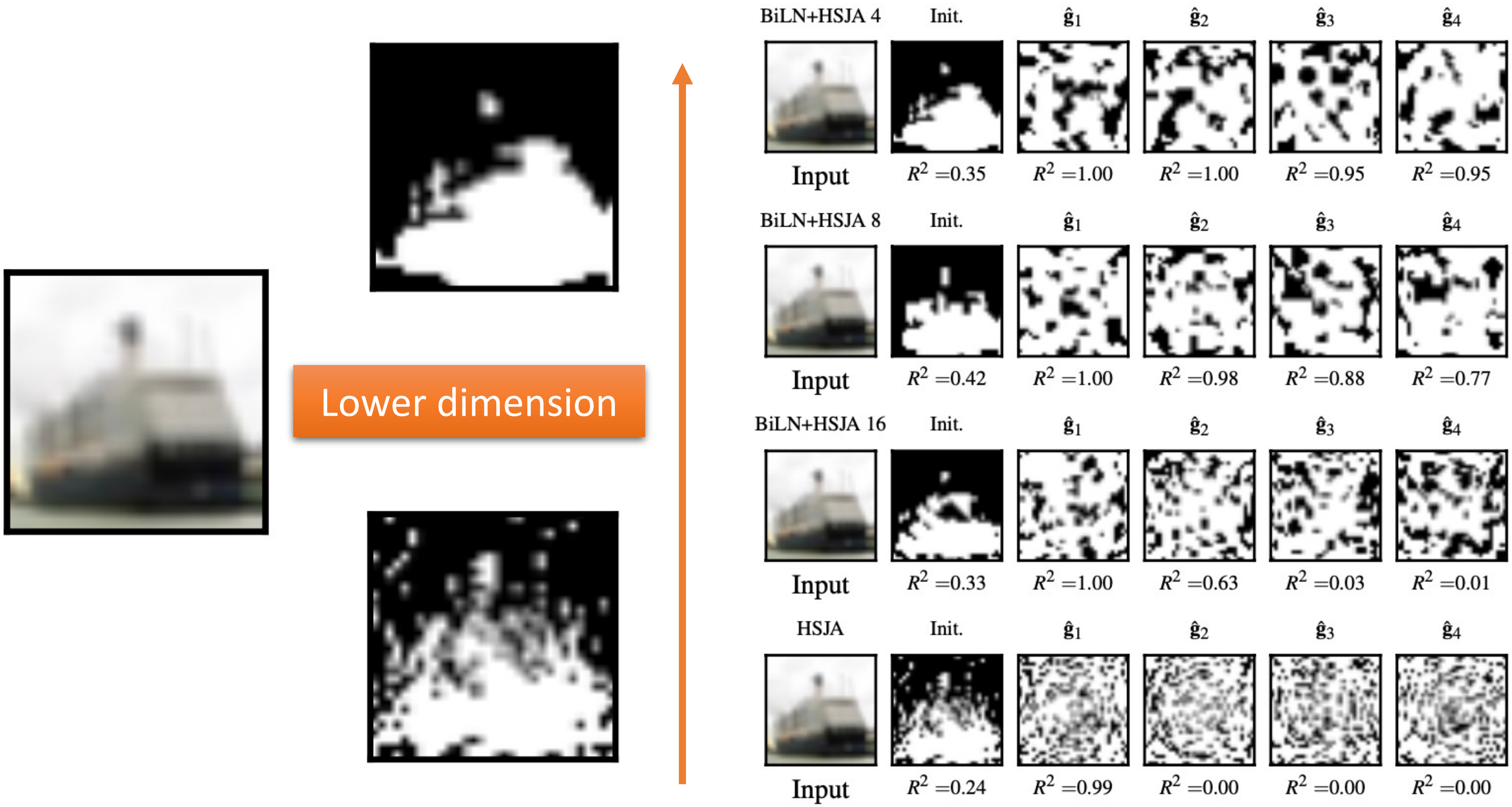


Results

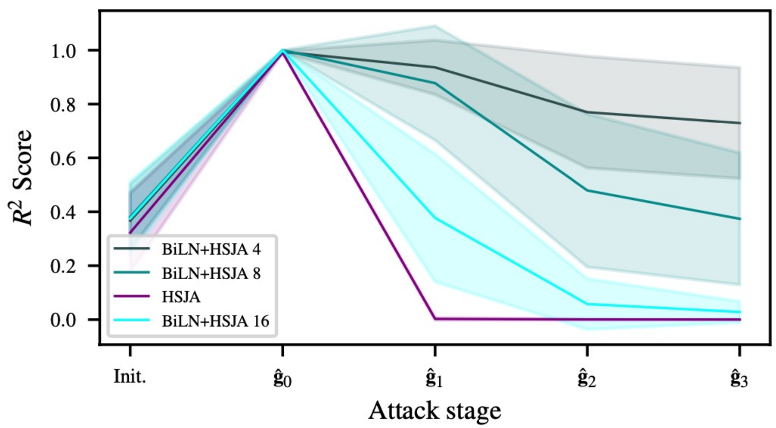
A2. Dimension-reduction leads to finer-grained structural information:



A2. Dimension-reduction leads to finer-grained structural information:



A2. Dimension-reduction leads to finer-grained structural information:



	Attack Variant	$\bar{R}^2$	FID	SR@40k (ϵ=0.031)	LPIPS
Madry CIFAR-10	HSJA	0.259	0.244	0.272	0.676±0.275
	→ BiLN 16	0.363	0.074	**0.298**	0.654±0.277
	→ BiLN 8	0.624	0.026	0.224	0.668±0.304
	→ BiLN 4	**0.779**	**0.026**	0.130	0.709±0.345
Natural CIFAR-10	HSJA	0.263	0.240	**1.000**	0.496±0.211
	→ BiLN 16	0.368	0.085	0.984	0.543±0.227
	→ BiLN 8	0.622	0.028	0.826	0.624±0.253
	→ BiLN 4	**0.759**	**0.012**	0.472	0.651±0.297

Second part of original question, what is the worst-case attack analysis?

Formulate adaptive attacks based on Algorithm 1 & 2, denoted MC and DynBiLN (cyan):

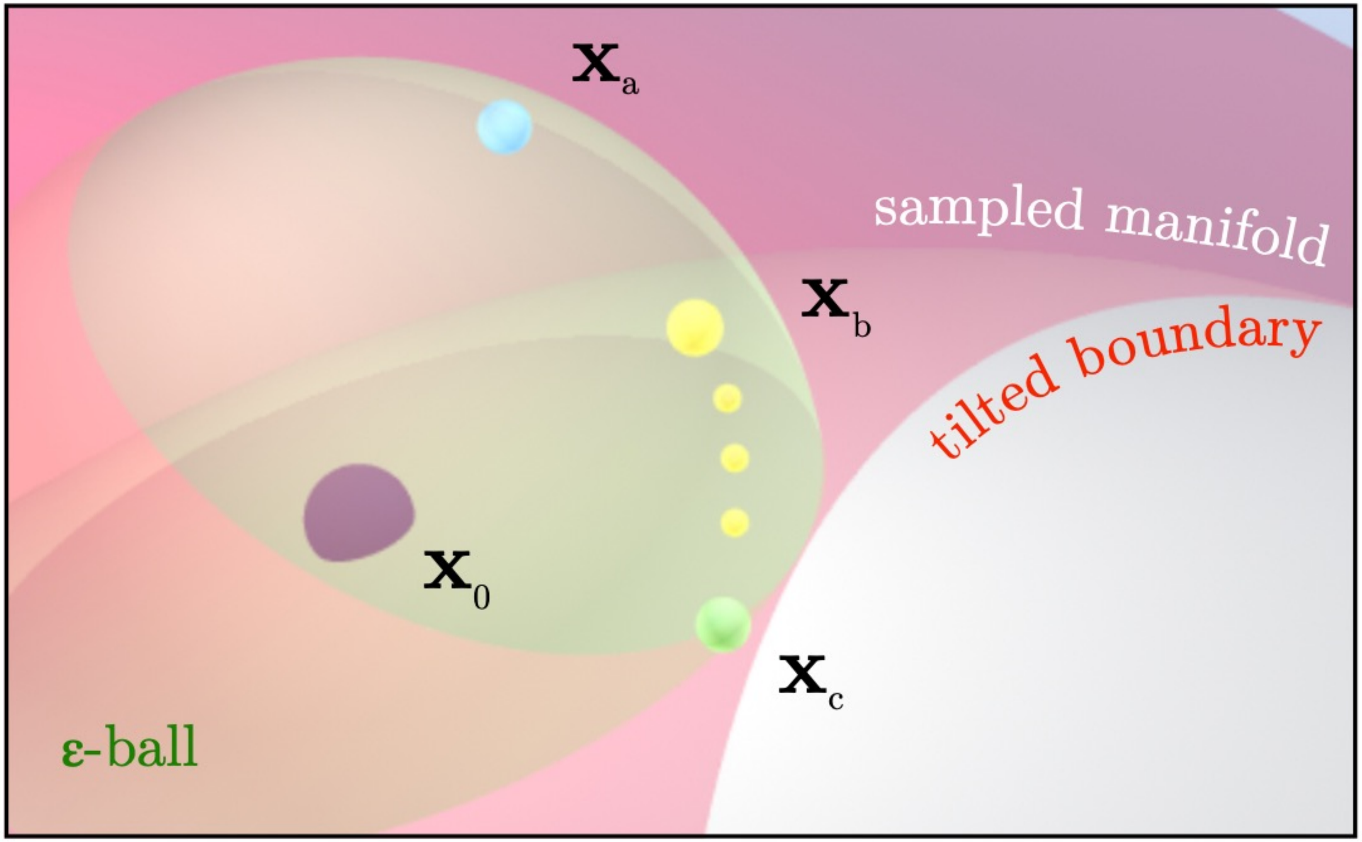
CIFAR-10

Attack Variant	FID	SR AUC (ϵ=0.031)
HSJA	0.253	0.537
→ *BiLN 4*	0.026 ↓	0.342
→ *BiLN 8*	0.023 ↓	0.574 ↑
→ *BiLN 16*	0.074 ↓	0.720 ↑
MC HSJA	0.213 ↓	0.545 ↑
→ *BiLN 4*	**0.022** ↓	0.356
→ *BiLN 8*	0.026 ↓	0.577 ↑
→ *BiLN 16*	0.068 ↓	0.705 ↑
→ *DynBiLN*	0.030 ↓	0.607 ↑
RayS	0.057	**1.000**

ImageNet

Attack Variant	FID	SR AUC (ϵ=0.031)
HSJA	1.541	0.344
→ *BiLN 16*	0.312 ↓	0.777 ↑
→ *BiLN 32*	1.085 ↓	0.771 ↑
→ *BiLN 64*	2.567	0.655 ↑
MC HSJA	1.591	0.331
→ *BiLN 16*	**0.271** ↓	0.772 ↑
→ *BiLN 32*	1.079 ↓	0.771 ↑
→ *BiLN 64*	2.287	0.615 ↑
→ *DynBiLN*	0.657 ↓	0.774 ↑
RayS	0.302	**1.000**

Geometric Interpretation



Motivation

Two main questions:

1. Is agent concept resolution possible from an ***adversary's*** perspective? What's the worst-case security analysis?

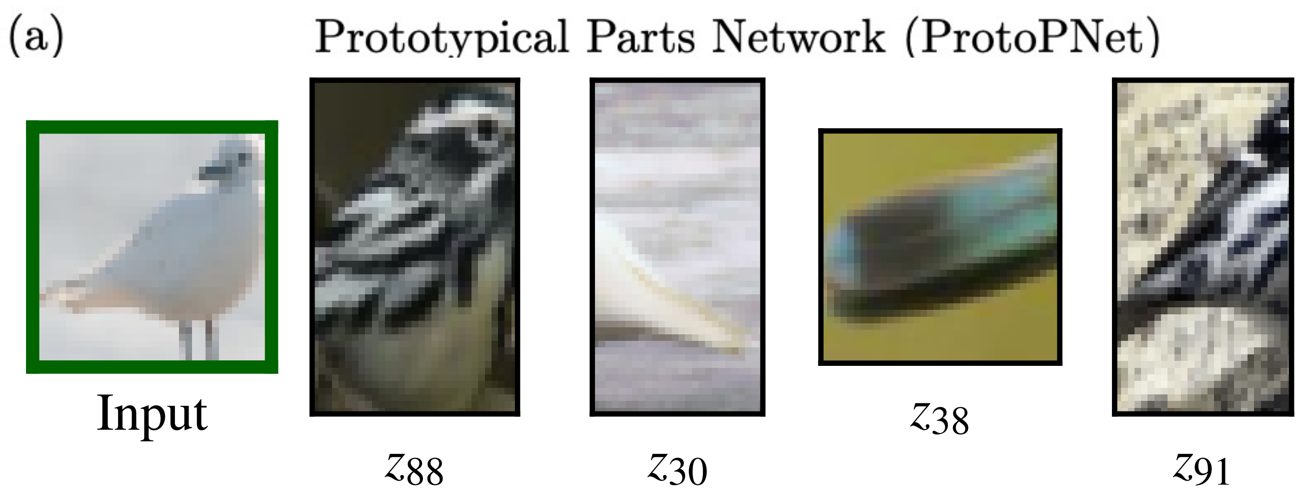
2. It isn't always realistic to have a human-in-the-loop, or fine-grained data labels. What are ways around this?

A1: Expansion of our previous hard-label paper, in submission to IEEE SaTML 2023.

A2: We presented an initial idea @ NAACL 2022. Introduced expansion during previous meeting. Then worked on it as part of AFRL 2022 internship.

A2: Approach

First, agents learn human-interpretable perceptual knowledge priors:



C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, “This Looks Like That: Deep Learning for Interpretable Image Recognition,” *arXiv:1806.10574 [cs, stat]*, 2019.

UF UNIVERSITY of FLORIDA · Duke UNIVERSITY · TEXAS The University of Texas at Austin · UC SANTA CRUZ

Multi-task learning (MTL)

- Sender solves two joint tasks:
  1. Learn to embed their top-1 activate structure ($\mathbf{z}^S$) in the message
  2. Learn to describe the target objects

- Receiver solves two joint tasks:
  1. Learn to reconstruct the sender's top-1 structure (*rec*($\mathbf{z}^S$)) from the message (*reconstruction loss*)

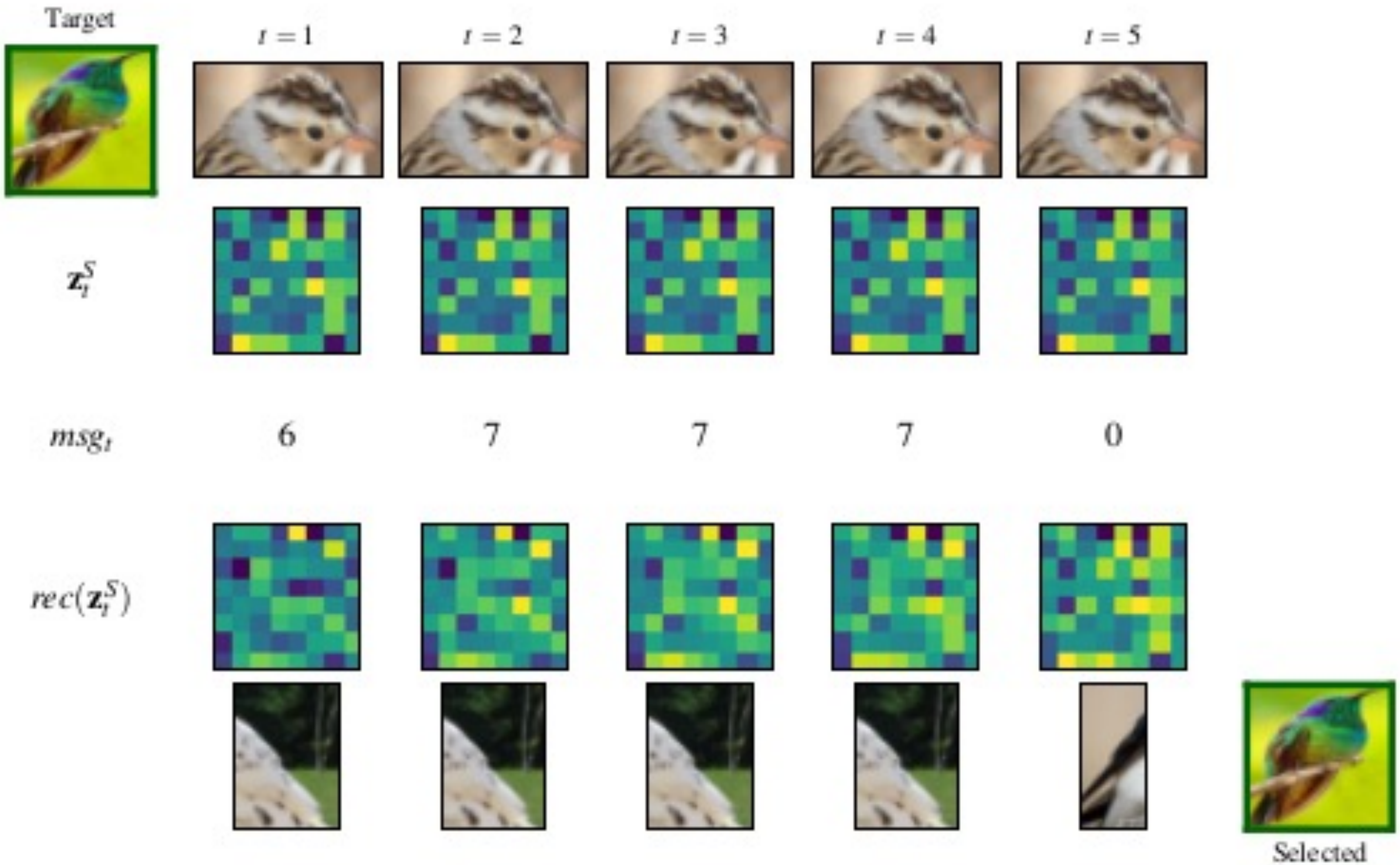
$$\mathcal{L}_{rec}(\mathbf{z}^S, rec(\mathbf{z}^S)) = \frac{1}{L}\sum_{l=1}^{L} |\mathbf{z}^S_{(l)} - rec(\mathbf{z}^S_{(l)})|$$

  2. Learn to signal the correct target object (*classification loss*)

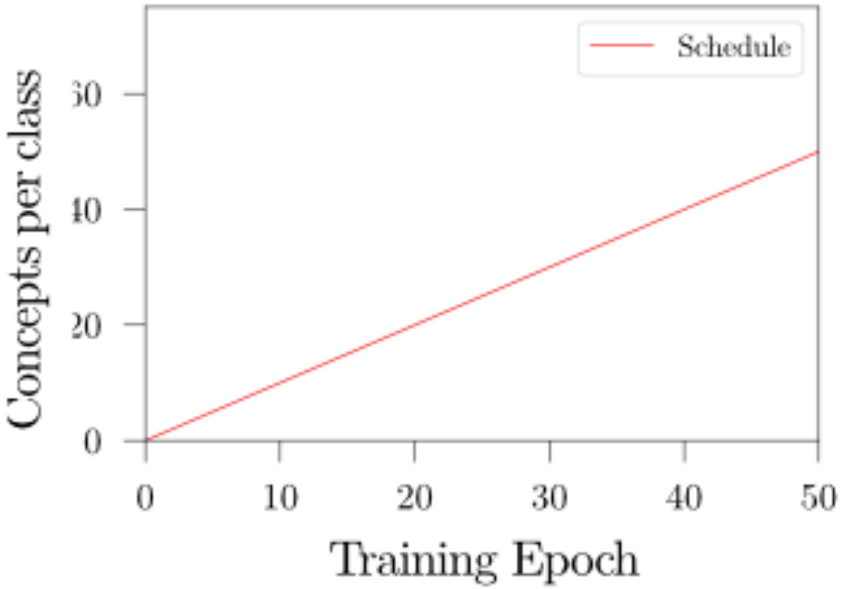
$$\mathcal{L}_{cls}(\mathbf{t}) = -\sum_{l=1}^{L} \alpha \log p(y_{(l)} = \mathbf{t} \mid msg_{(l)}),$$

$$\boxed{\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{rec}}$$

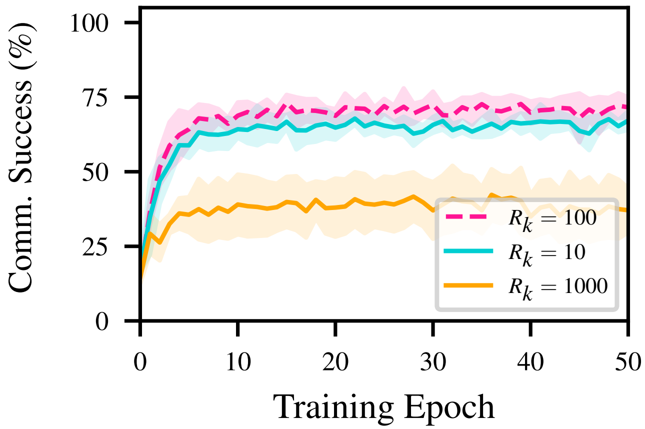
Qualitative Results

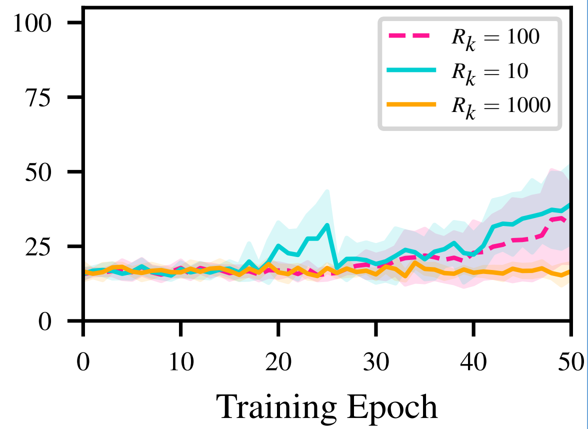


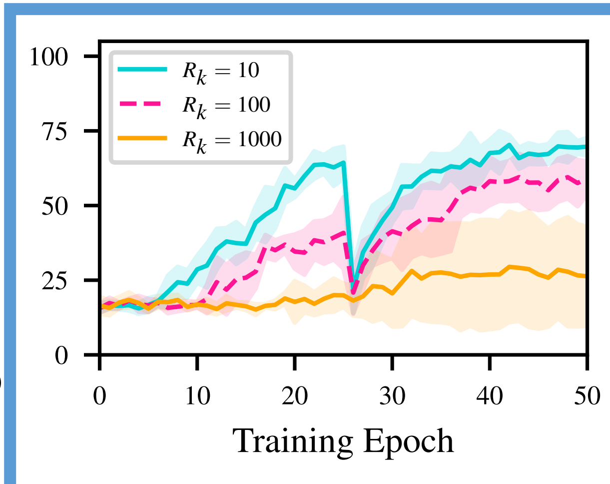
Consider a gradual expansion of the sender agent’s concept allowance, as in “Tatanka” clip from Dances with Wolves:



Concept allowance







Baseline

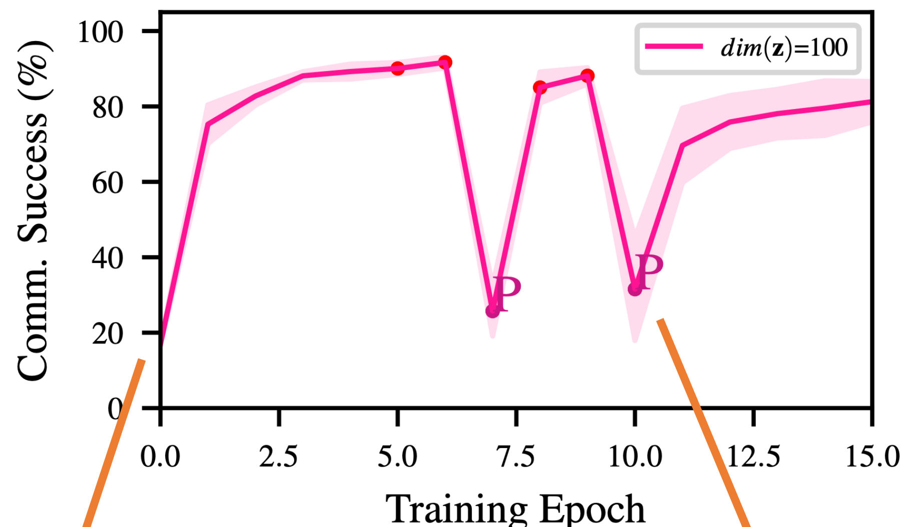
Reification (Ours)

Reification "Tatanka" Variant (Ours)

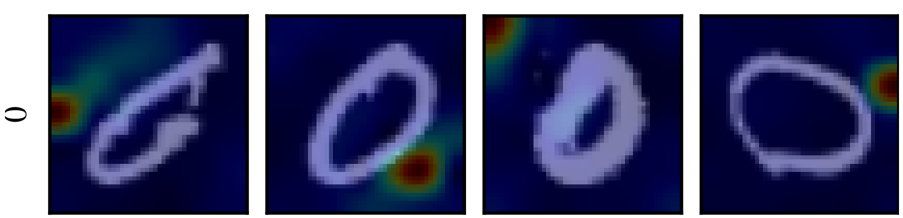
Semiotic Learning

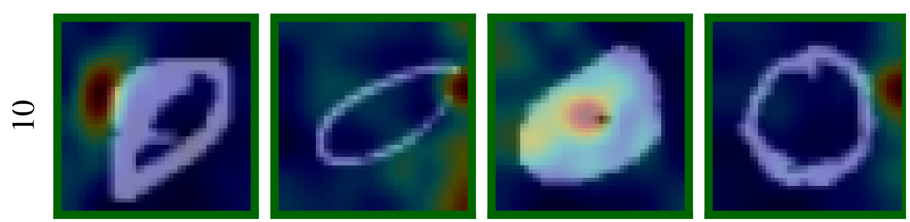
- Senders solve ~~two~~ three joint tasks:
  1. Learn to embed their top-1 activate structure in the message
  2. Learn to describe the target objects
  3. Update knowledge structure based on embedding difficulty
- Receivers solve ~~two~~ three joint tasks:
  1. Learn to reconstruct the sender's top-1 structure from the message
  2. Learn to signal the correct target object
  3. Update knowledge structure based on perceived utility of sender structure

Training instability and automatic recovery:



P = human-aligned correction





Conclusion

Adversaries can learn semantic structure in a neighborhood around a sample, and this informs geometric interpretation of generalization errors.

- Can we get the global semantic structure with few samples and queries? Implication: leakage of learned manifold
- Connection to diffusion models

Semiotic learning offers an avenue for automatic structural validation, without explicit labels!

Graduation: Dec. 2022

Joining UDRI in January

Thank you

w.garcia@ufl.edu