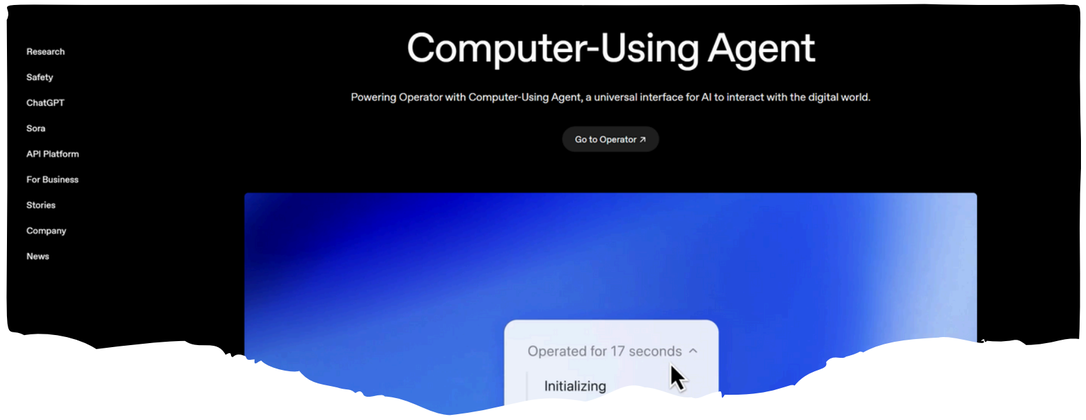
Explainable Concept Generation through Vision-Language Preference Learning for Understanding Neural Networks' Internal Representations

Aditya Taparia, Som Sagar, Ransalu Senanayake

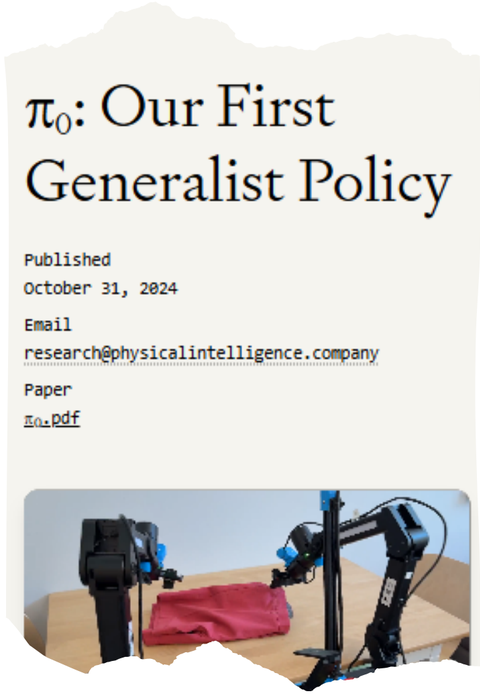
ICML
International Conference
On Machine Learning

AI Models Are Now Being Used for Decision-Making

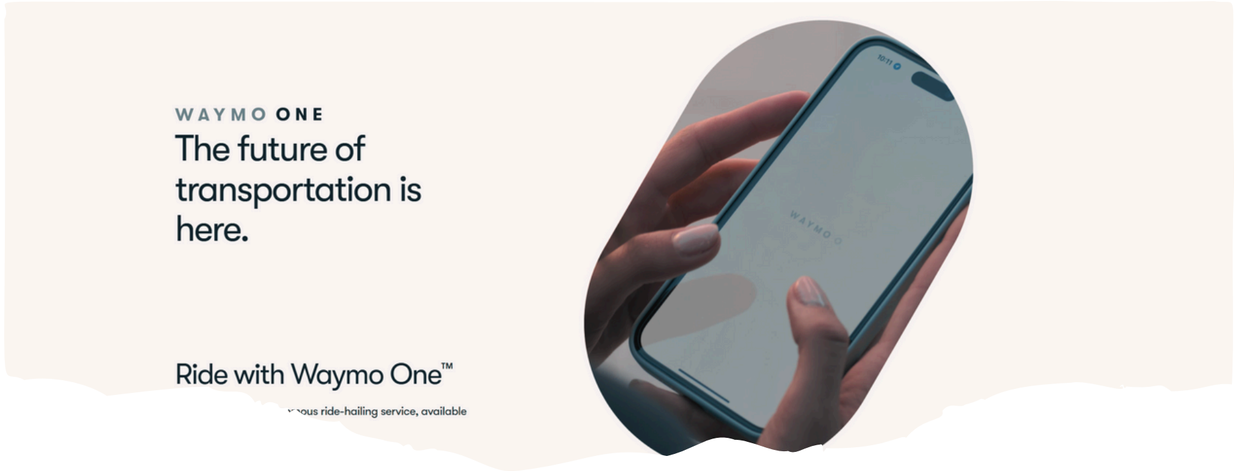


Computer-Using Agent (OpenAI)

Agentic Search (Perplexity)



Laundry Robot (Physical Intelligence)



Autonomous Driving (Waymo)

Understanding why they make a particular decision has become ever so important!

Why Do We Care?

Understanding the model helps with,

Debugging & Model Development

Helps in identifying failure cases, spurious correlations, and overfitting.

Trust & Adoption

Users and stakeholders are more likely to trust decisions they can understand.

Auditing & Regulatory Compliance

Essential for regulated domains like finance, healthcare, and law.

Traceability

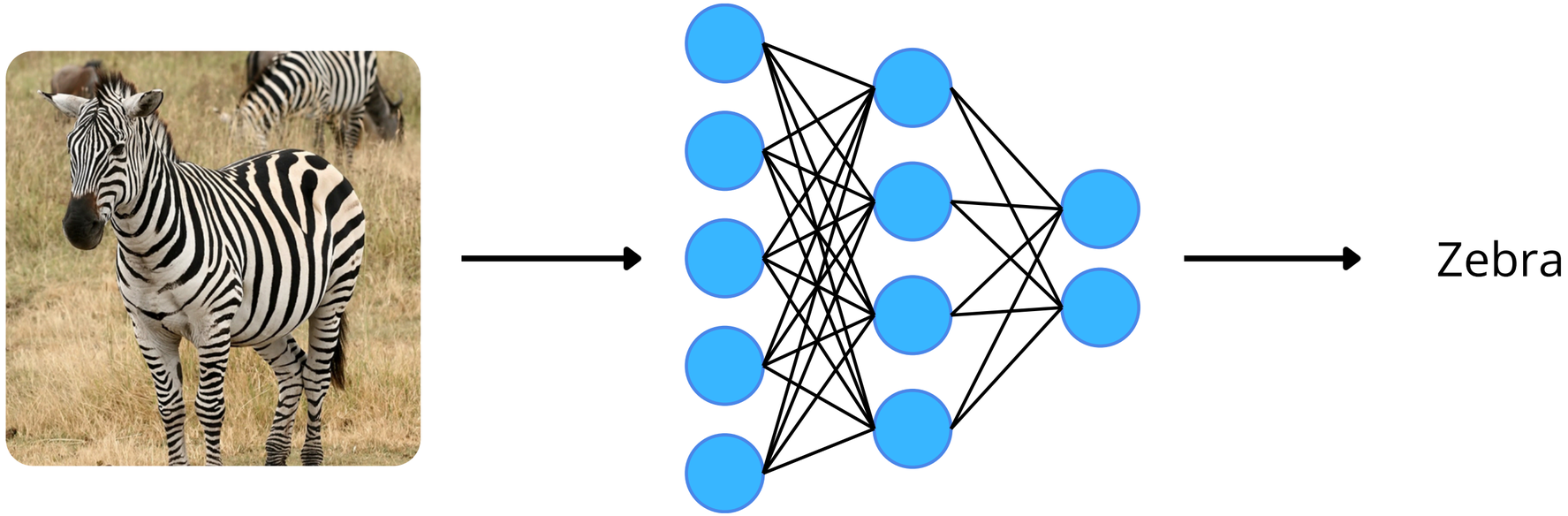
We need to trace outcomes back to causes, especially in failure scenarios.

Human-AI Collaboration

For effective interaction, humans need to understand what the AI “thinks”.

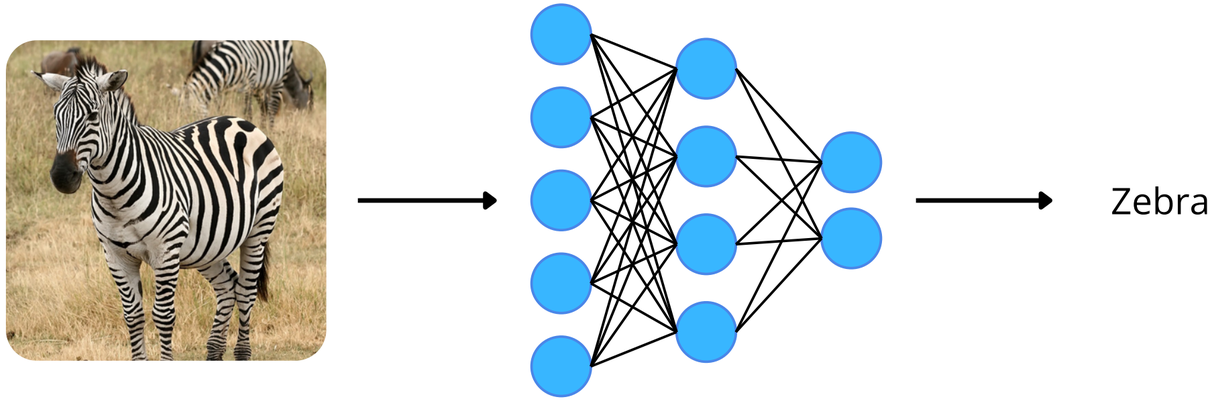
Why Is It a Difficult Problem?

Let’s consider a simple case of a image classifier.



Why Is It a Difficult Problem?

Let’s consider a simple case of a image classifier.



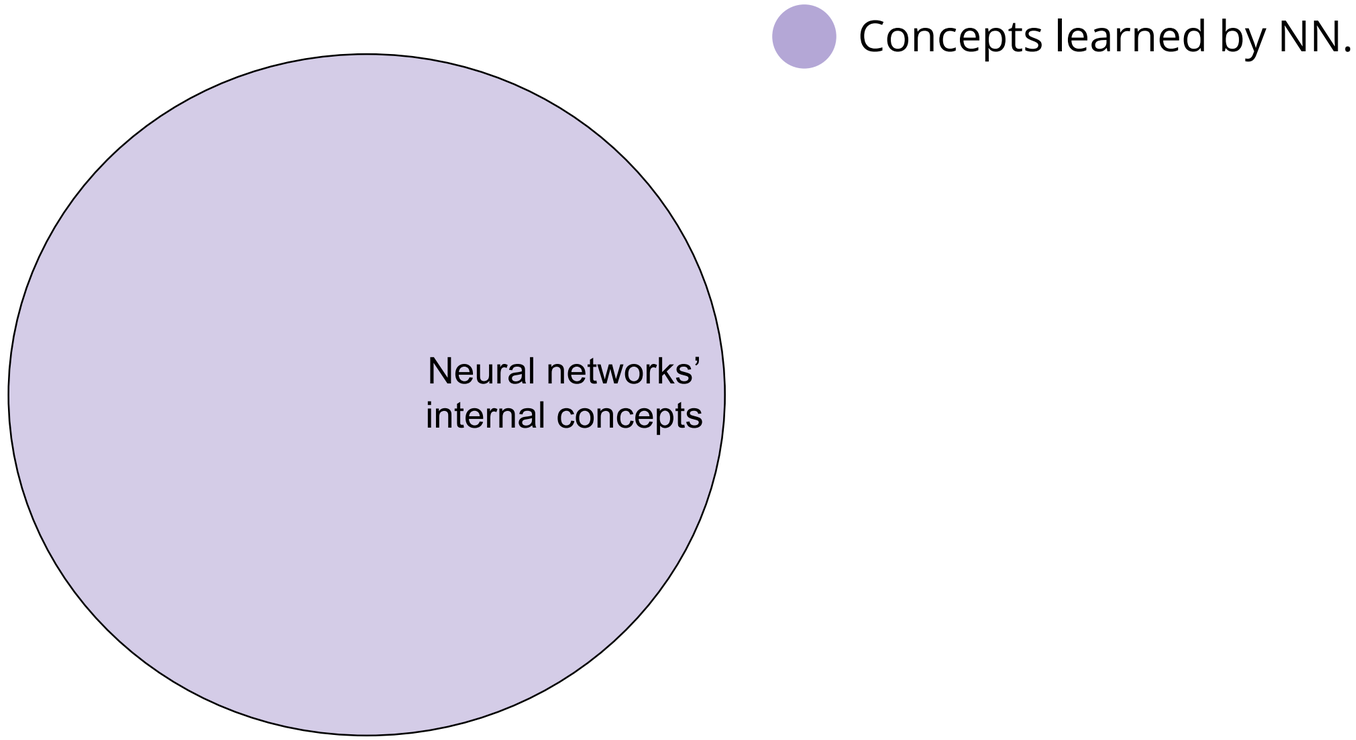
What could be the reasons?

- Four legs,
- Dry grass in background,
- Black and white stripes,
- So on...

We don’t know!!

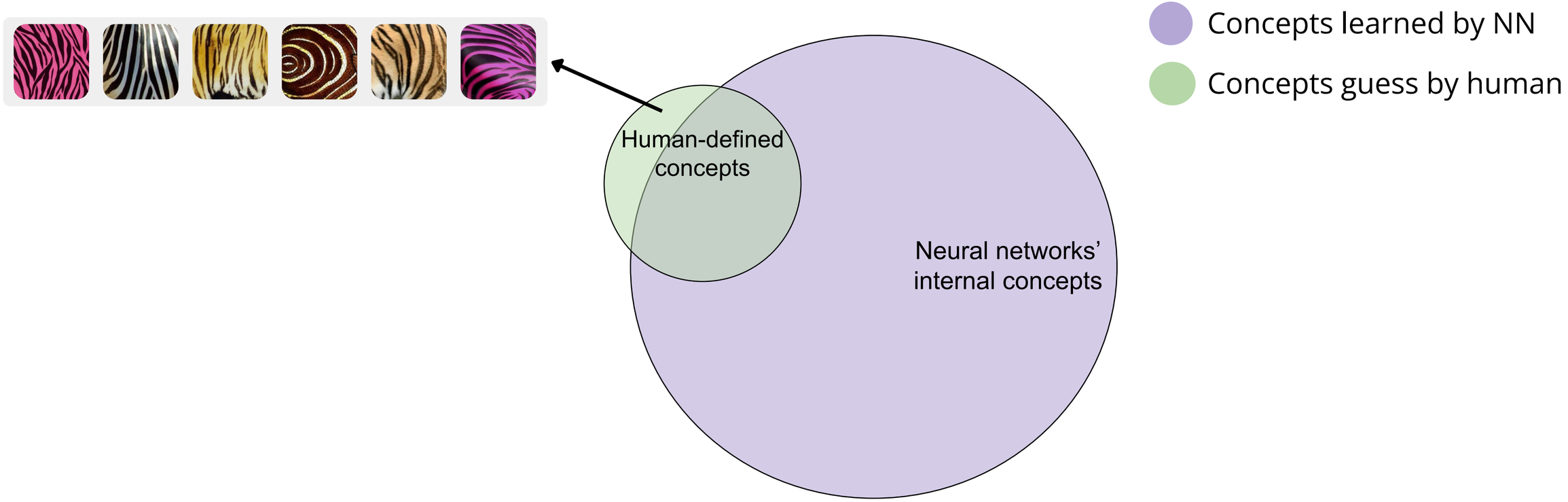
Why Is It a Difficult Problem?

There could be hundred different reasons for the NN to make that decision and identifying them with human cognition is not feasible and limits the reasoning.



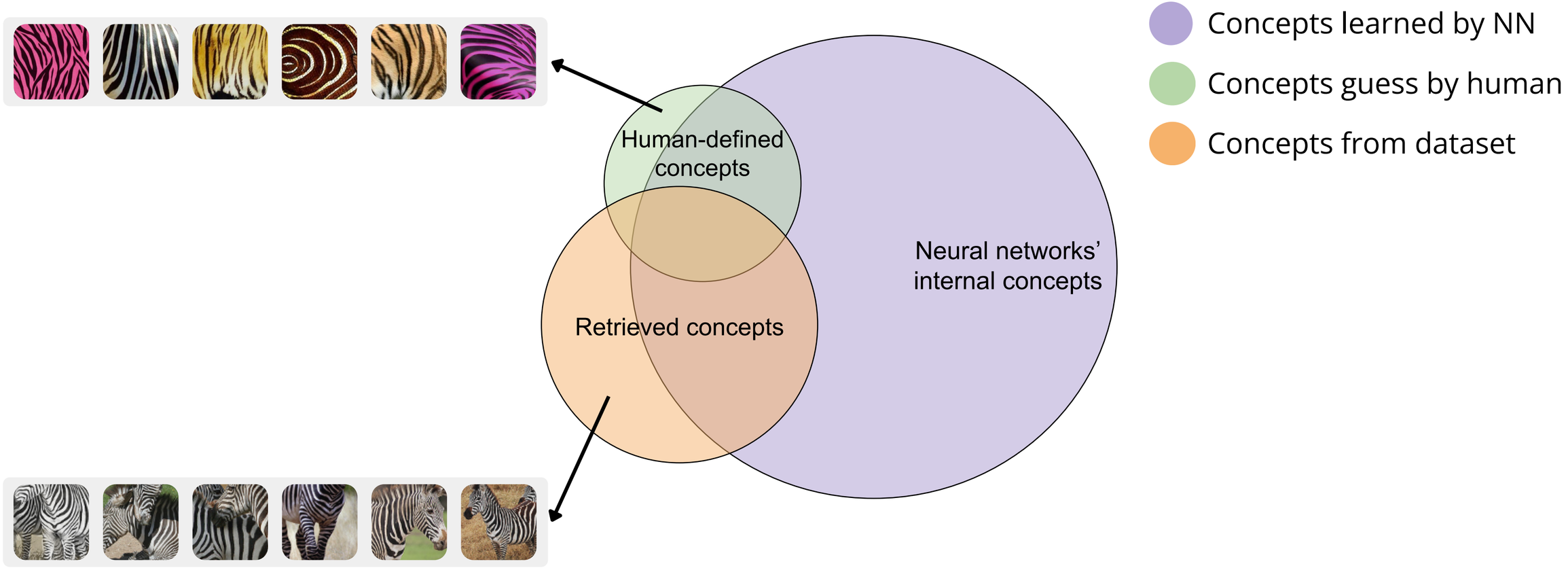
Why Is It a Difficult Problem?

There could be hundred different reasons for the NN to make that decision and identifying them with human cognition is not feasible and limits the reasoning.



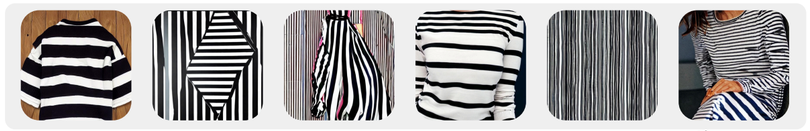
Why Is It a Difficult Problem?

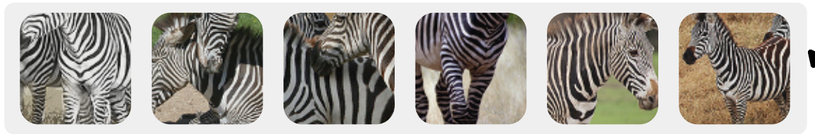
There could be hundred different reasons for the NN to make that decision and identifying them with human cognition is not feasible and limits the reasoning.

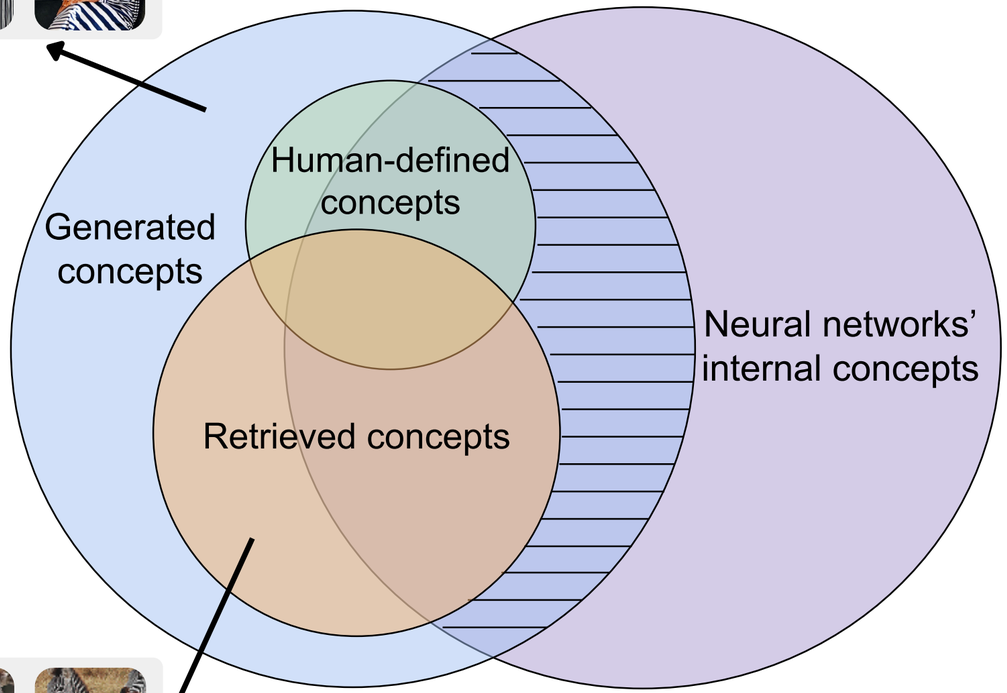


Why Is It a Difficult Problem?

There could be hundred different reasons for the NN to make that decision and identifying them with human cognition is not feasible and limits the reasoning.







Concepts learned by NN

Concepts guess by human

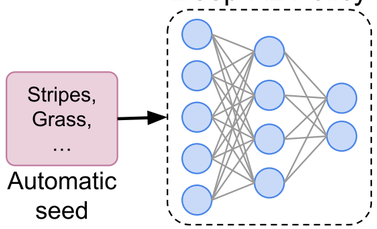
Concepts from dataset

Generated Concepts

How Do We Do It?

G_i: Group i
s_t: State at time t
r_t: Reward at time t
a_t: Action taken at time t

Deep RL Policy



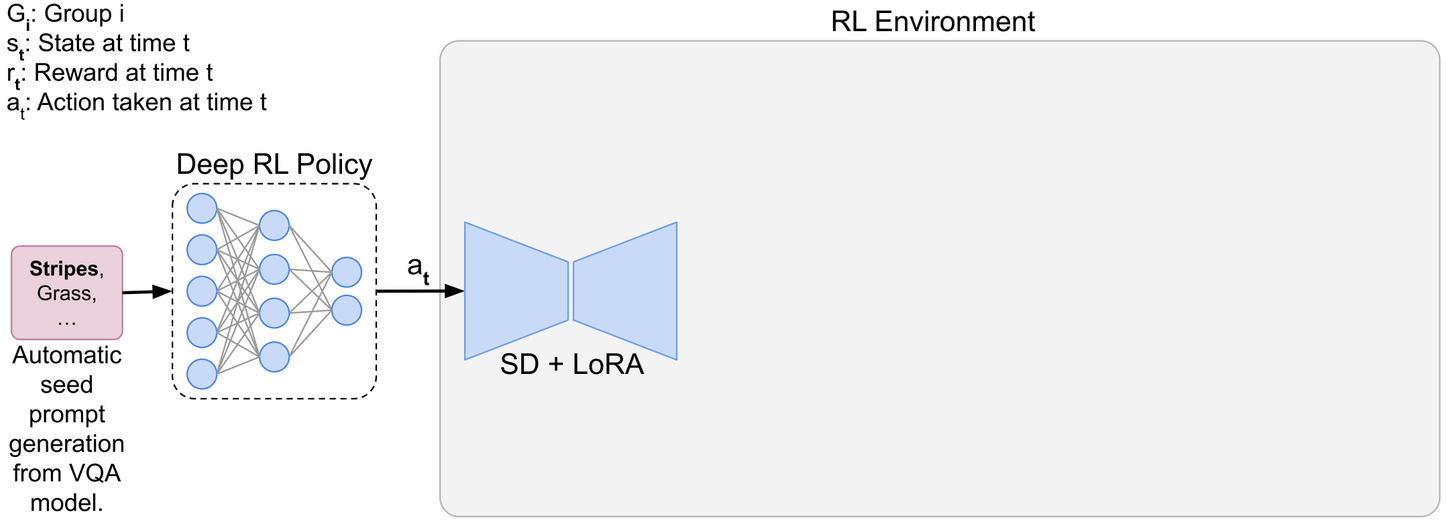
Automatic seed prompt generation from VQA model.

RL Environment

Algorithm 1 The RLPO algorithm.

Input: Set of test images $f(\cdot)$
Run pre-processing and obtain seed prompts (action space)
for each episode **do**
  for each time step t **do**
    Execute a_t by picking a seed prompt
    Generate image groups G_1 and G_2
    Evaluate TCAV scores TS_1 and TS_2
    Update SD based on the better score
    Compute reward
  end for
end for
Output: Set of concept images

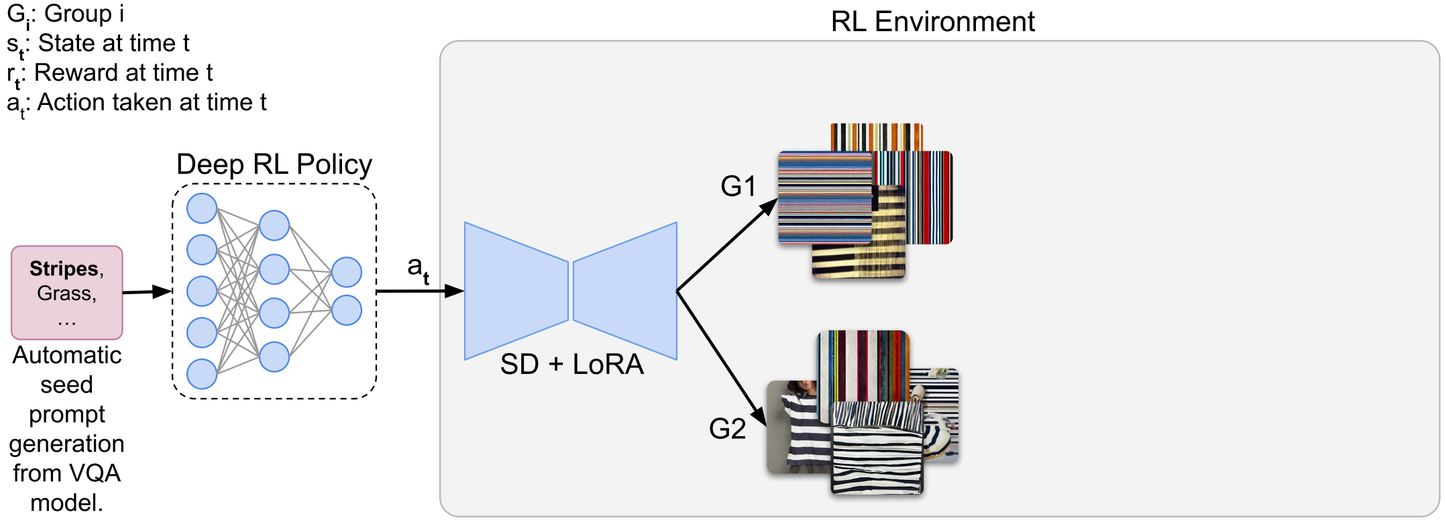
How Do We Do It?



Algorithm 1 The RLPO algorithm.

Input: Set of test images $f(\cdot)$
Run pre-processing and obtain seed prompts (action space)
for each episode **do**
  for each time step t **do**
    Execute a_t by picking a seed prompt
    Generate image groups G_1 and G_2
    Evaluate TCAV scores TS_1 and TS_2
    Update SD based on the better score
    Compute reward
  end for
end for
Output: Set of concept images

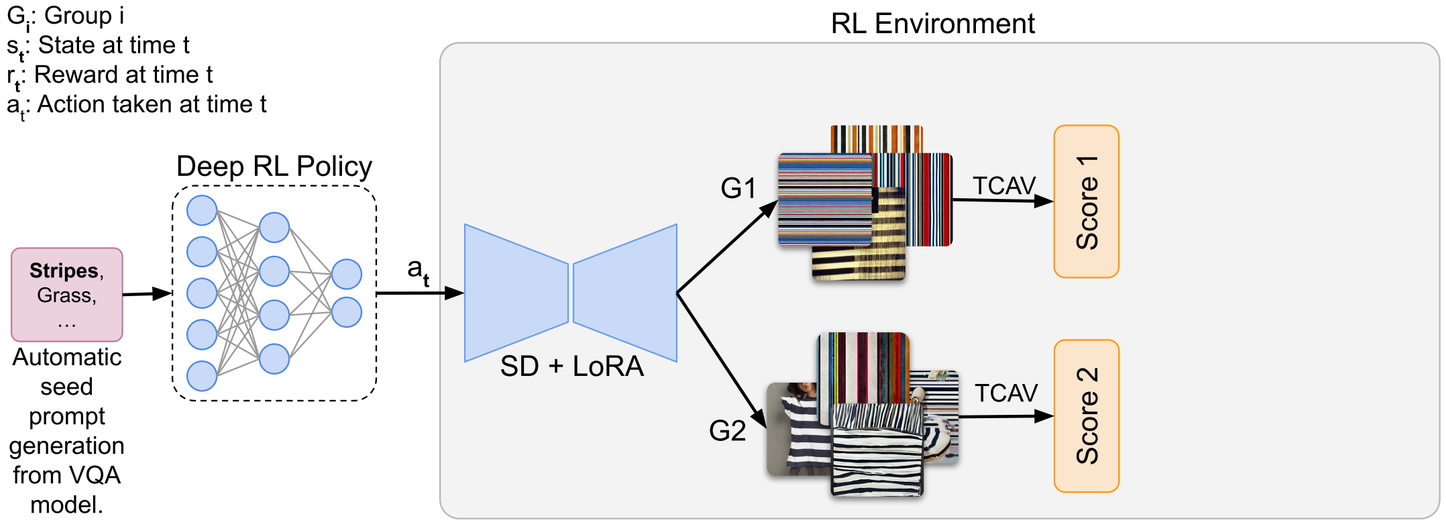
How Do We Do It?



Algorithm 1 The RLPO algorithm.

Input: Set of test images $f(\cdot)$
Run pre-processing and obtain seed prompts (action space)
for each episode **do**
  for each time step t **do**
    Execute a_t by picking a seed prompt
    Generate image groups G_1 and G_2
    Evaluate TCAV scores TS_1 and TS_2
    Update SD based on the better score
    Compute reward
  end for
end for
Output: Set of concept images

How Do We Do It?

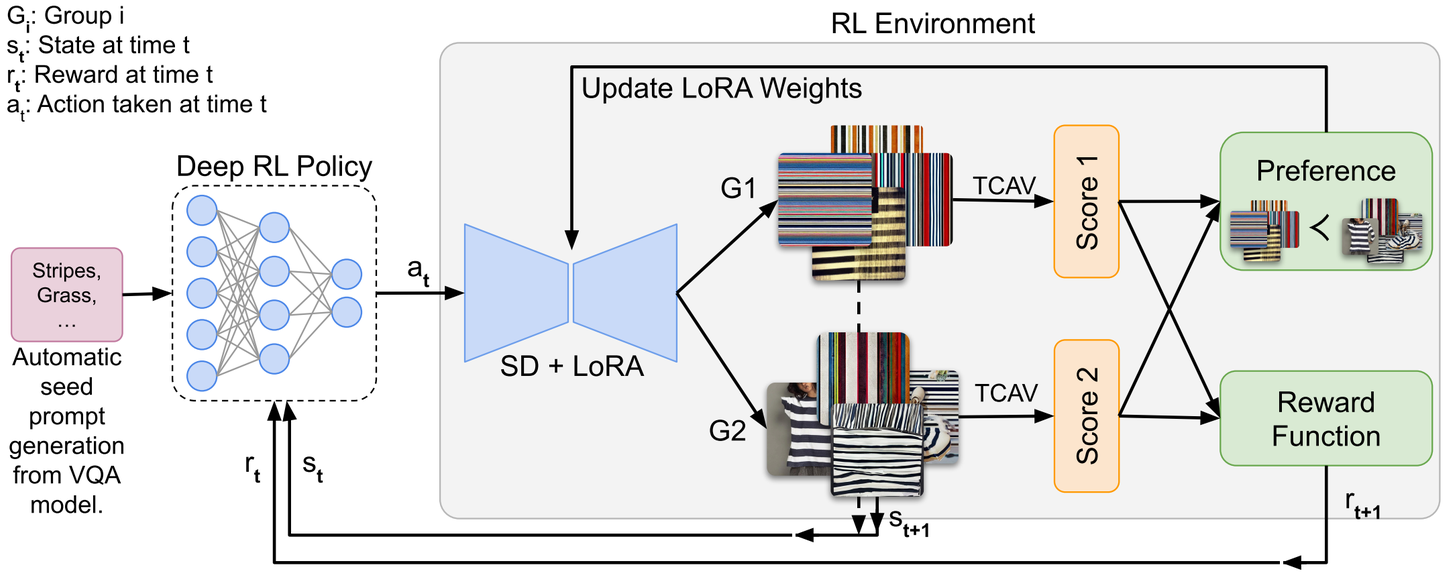


Algorithm 1 The RLPO algorithm.

Input: Set of test images $f(\cdot)$
Run pre-processing and obtain seed prompts (action space)
for each episode **do**
  for each time step t **do**
    Execute a_t by picking a seed prompt
    Generate image groups G_1 and G_2
    Evaluate TCAV scores TS_1 and TS_2
    Update SD based on the better score
    Compute reward
  end for
end for
Output: Set of concept images

Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." International conference on machine learning. PMLR, 2018.

How Do We Do It?



Algorithm 1 The RLPO algorithm.

Input: Set of test images $f(\cdot)$
Run pre-processing and obtain seed prompts (action space)
for each episode **do**
  for each time step t **do**
    Execute a_t by picking a seed prompt
    Generate image groups G_1 and G_2
    Evaluate TCAV scores TS_1 and TS_2
    Update SD based on the better score
    Compute reward
  end for
end for
Output: Set of concept images

Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." International conference on machine learning. PMLR, 2018.

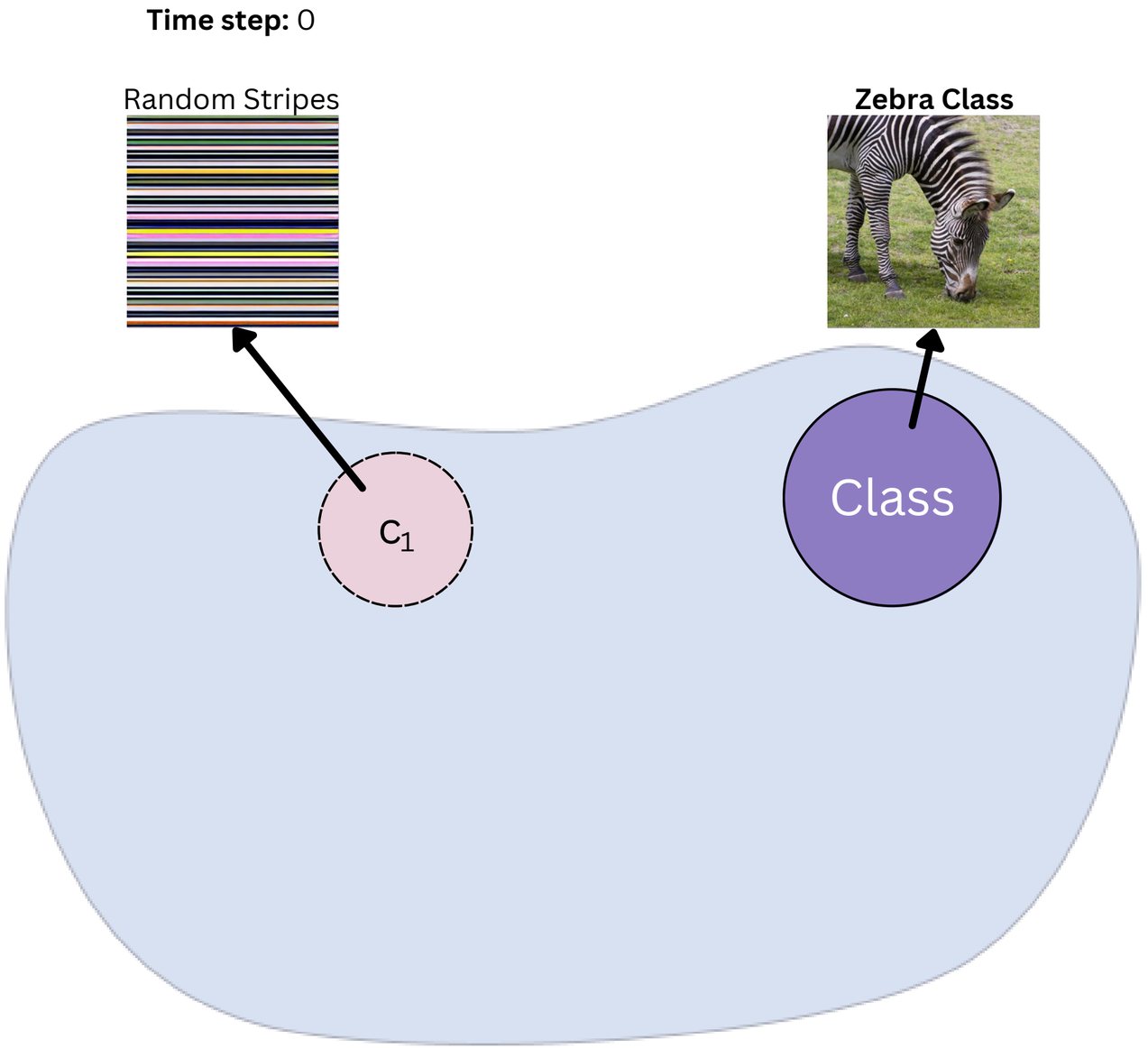
How Do We Do It?

Action Set:

- a_1: Stripes
- a_2: Mud
- a_3: Square
- a_4: Running

Selecting “Stripes” action.

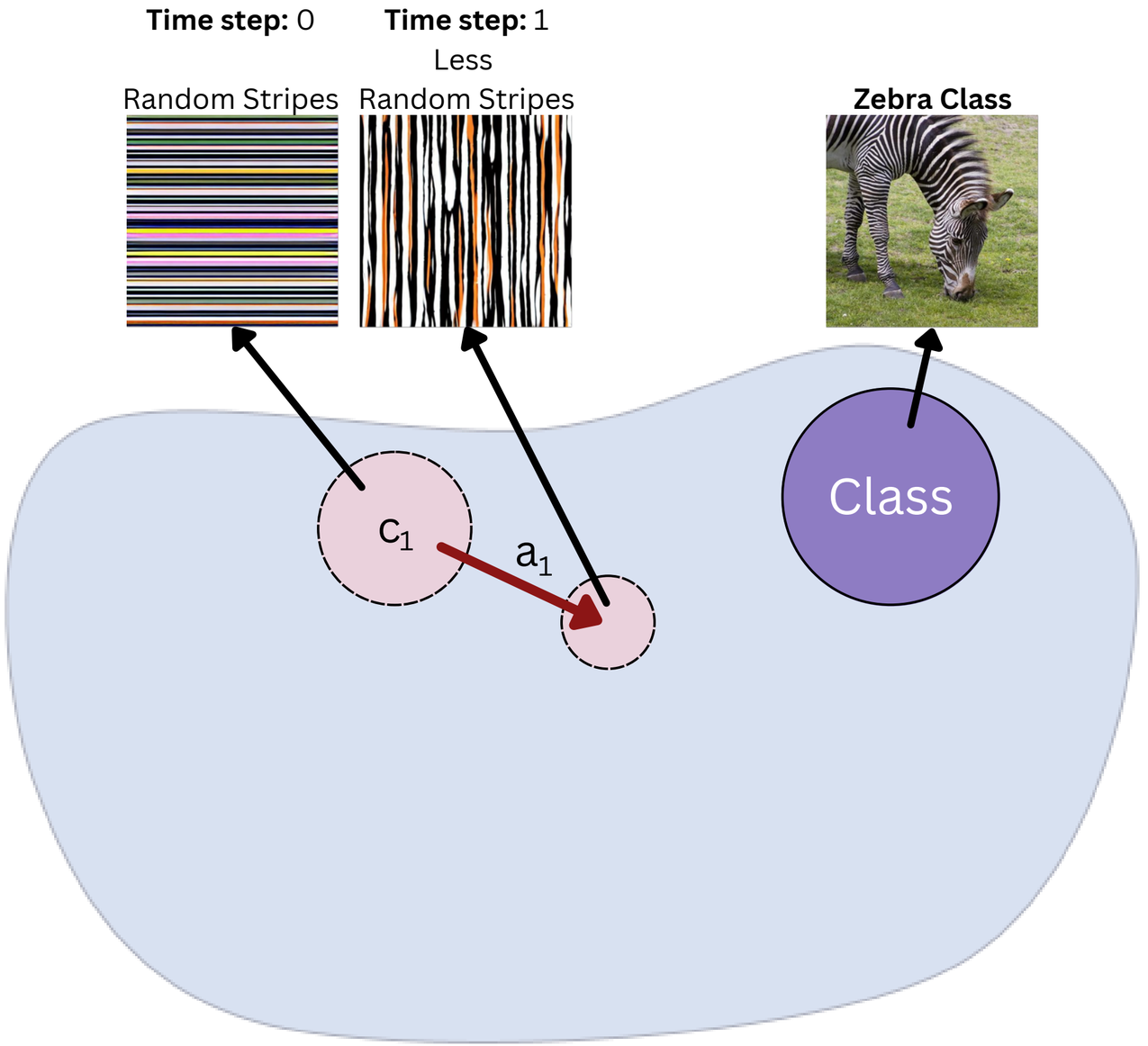
Got a high reward.



How Do We Do It?

Action Set:

- a_1: Stripes
- a_2: Mud
- a_3: Square
- a_4: Running



Selecting “Stripes” action.

Got a high reward.

How Do We Do It?

Action Set:

a_1: Stripes

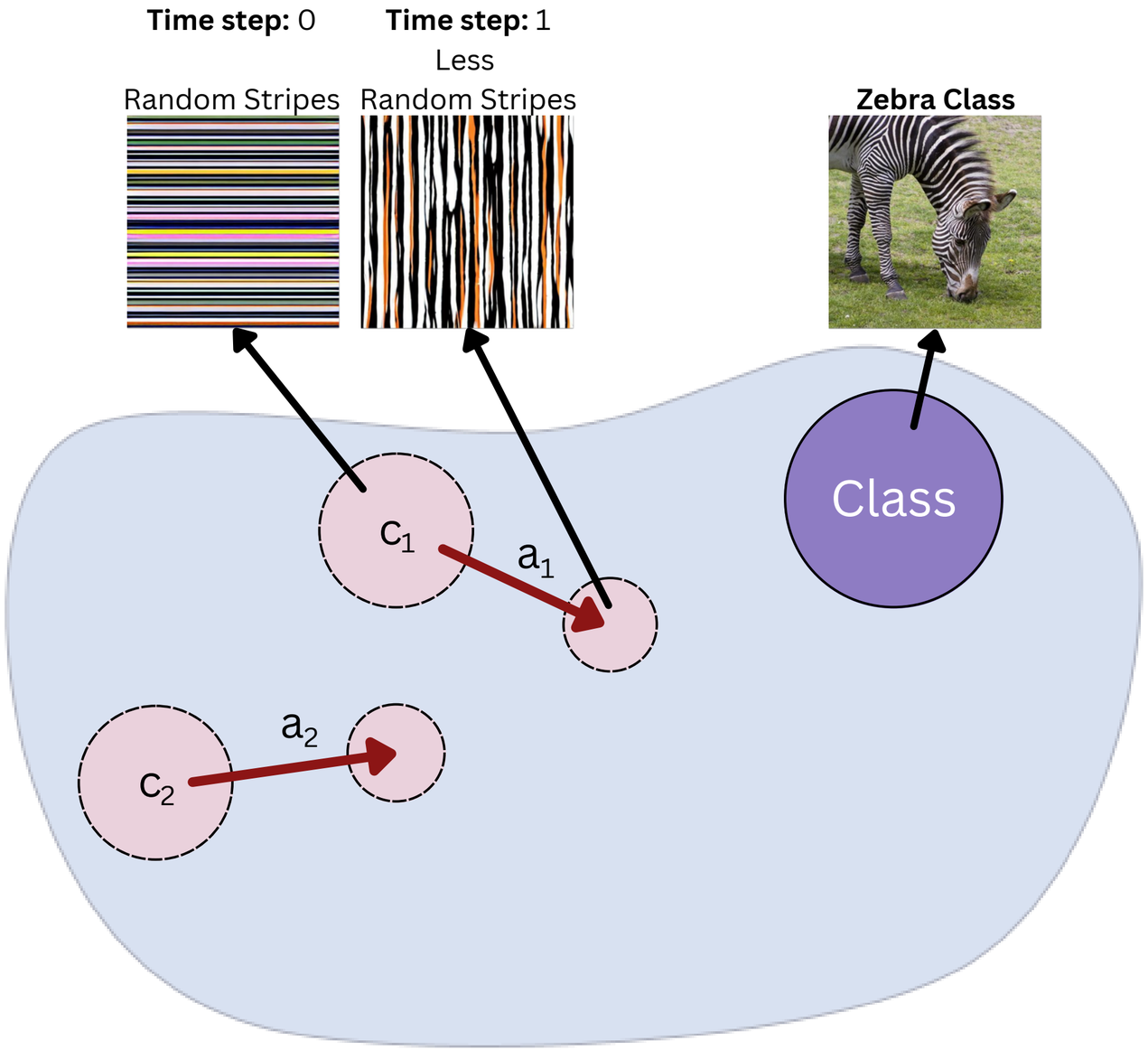
a_2: Mud

a_3: Square

a_4: Running

Selecting “Mud” action.

Okay reward.



How Do We Do It?

Action Set:

a_1: Stripes

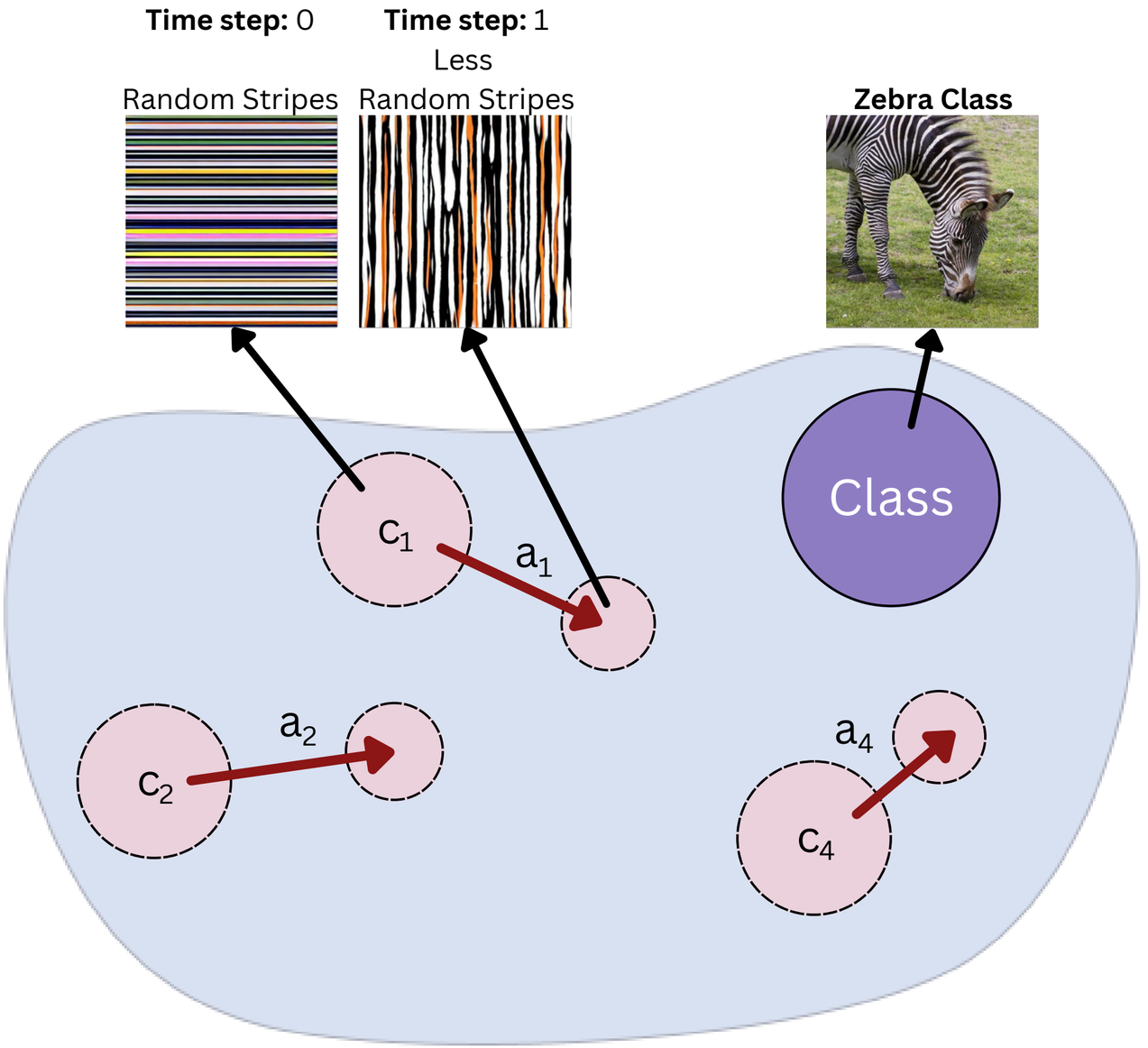
a_2: Mud

a_3: Square

a_4: Running

Selecting “Running” action.

Okay reward.



How Do We Do It?

Action Set:

a_1: Stripes

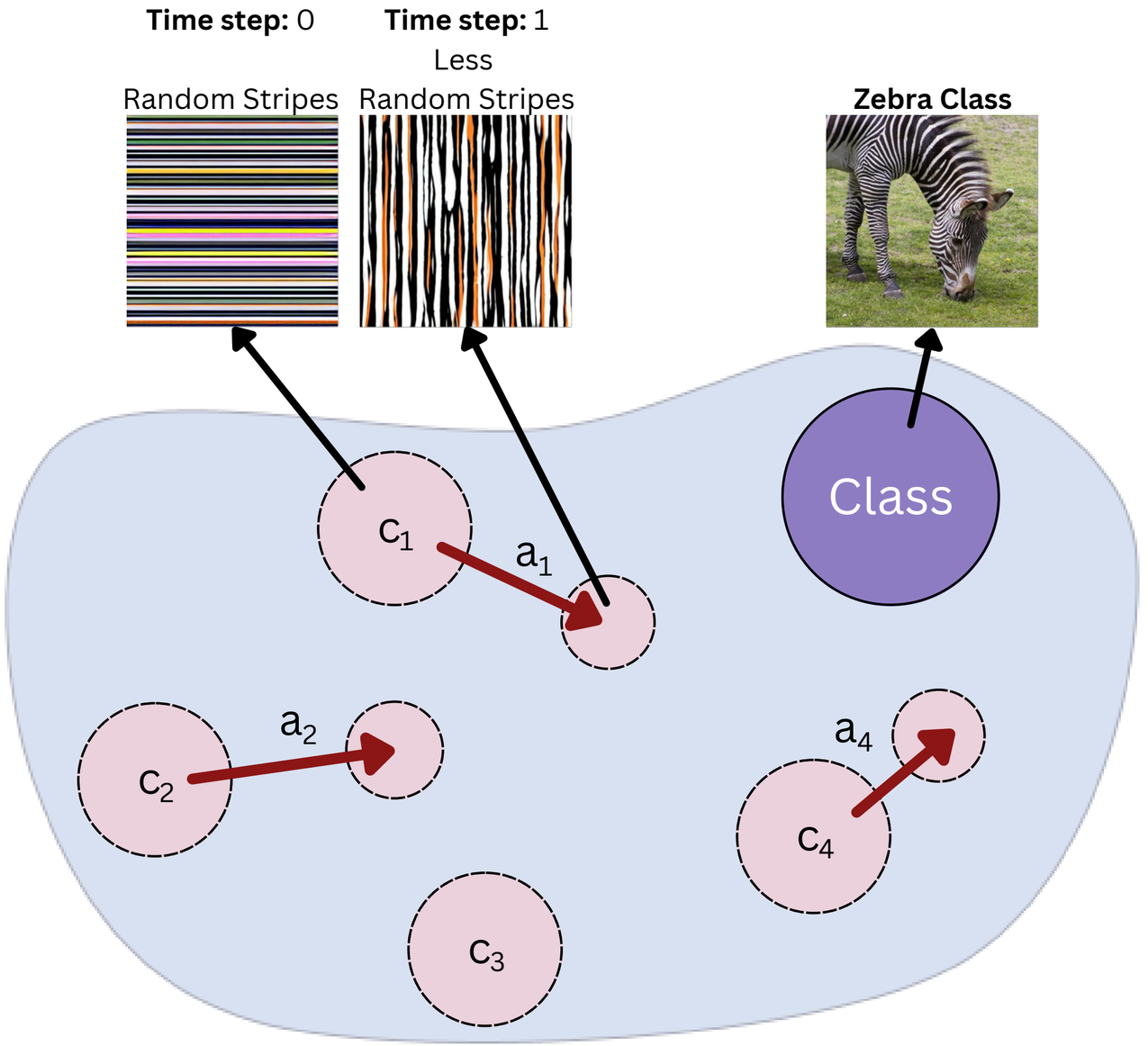
a_2: Mud

a_3: Square

a_4: Running

Selecting “Square” action.

No reward.



How Do We Do It?

Action Set:

a_1: Stripes

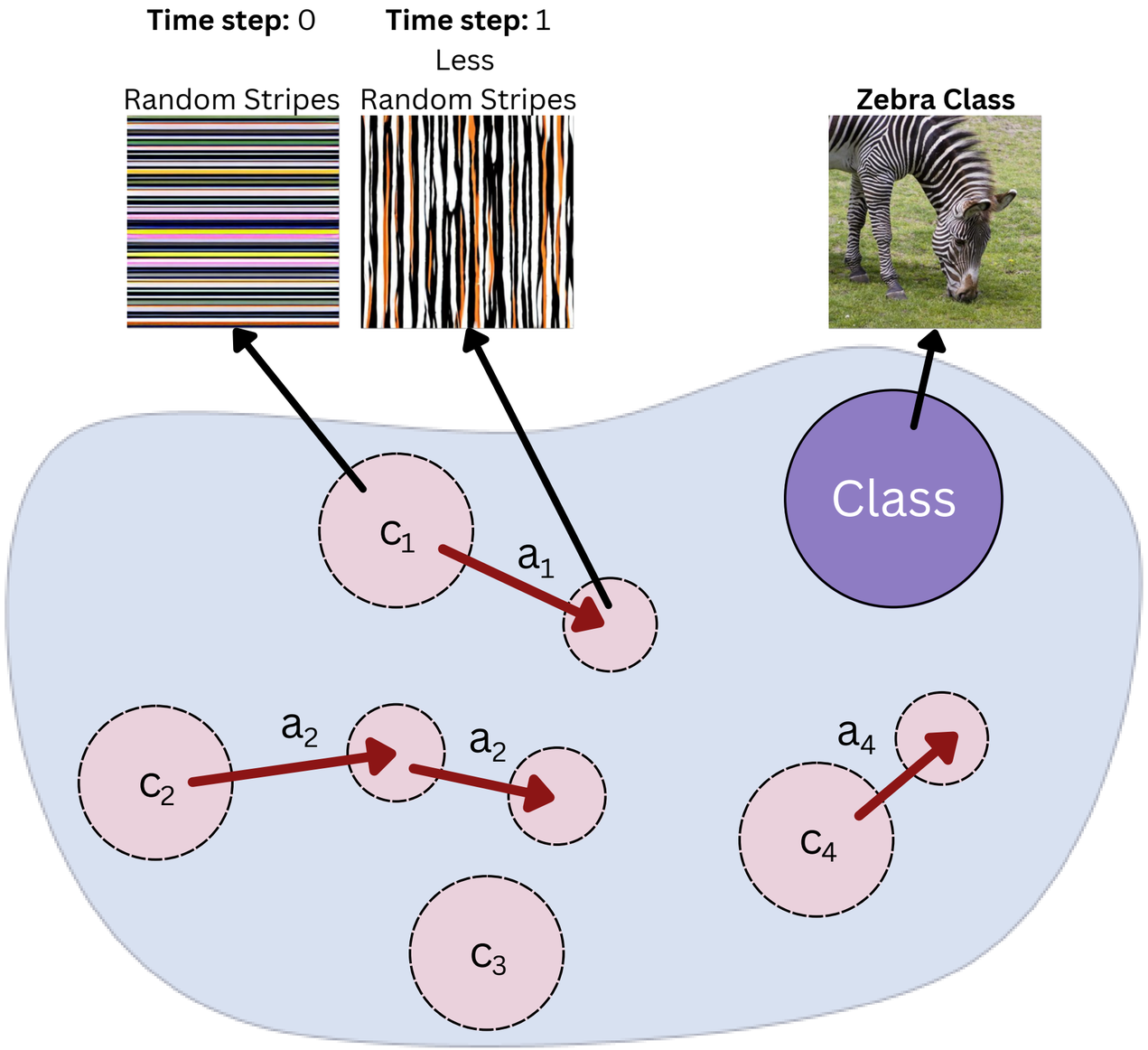
a_2: Mud

a_3: Square

a_4: Running

Selecting “Mud” action.

Okay reward.



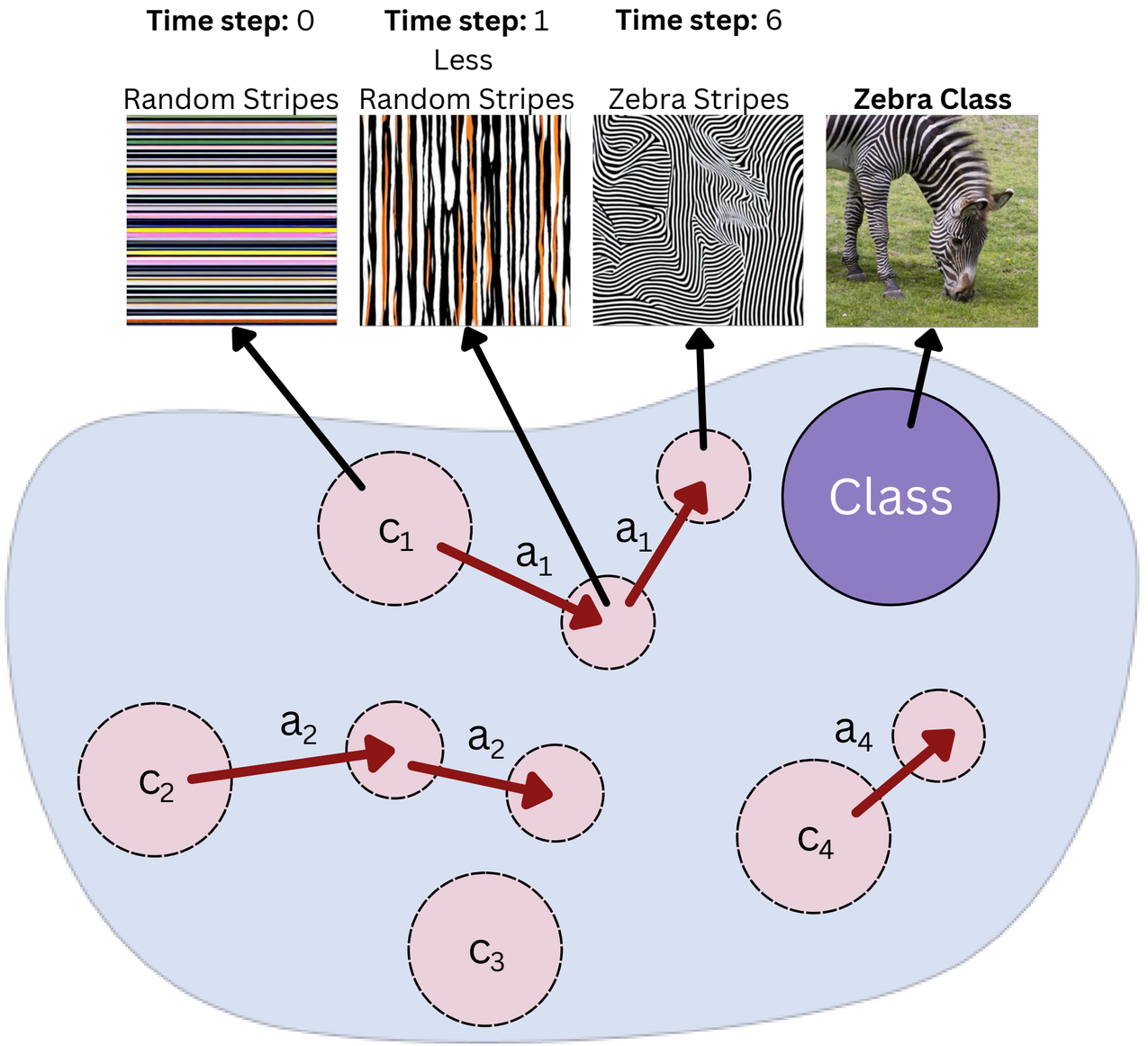
How Do We Do It?

Action Set:

- a_1: Stripes
- a_2: Mud
- a_3: Square
- a_4: Running

Selecting “Stripes” action.

High reward.



How Do We Do It?

Action Set:

a_1: Stripes

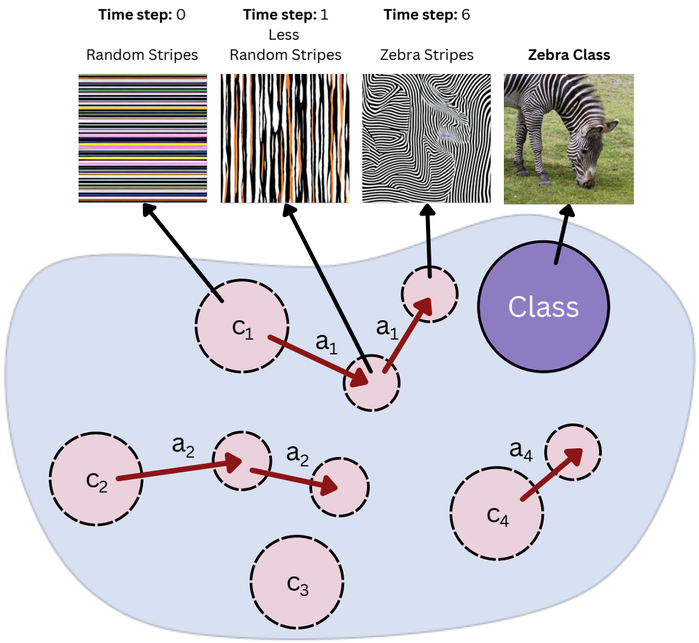
a_2: Mud

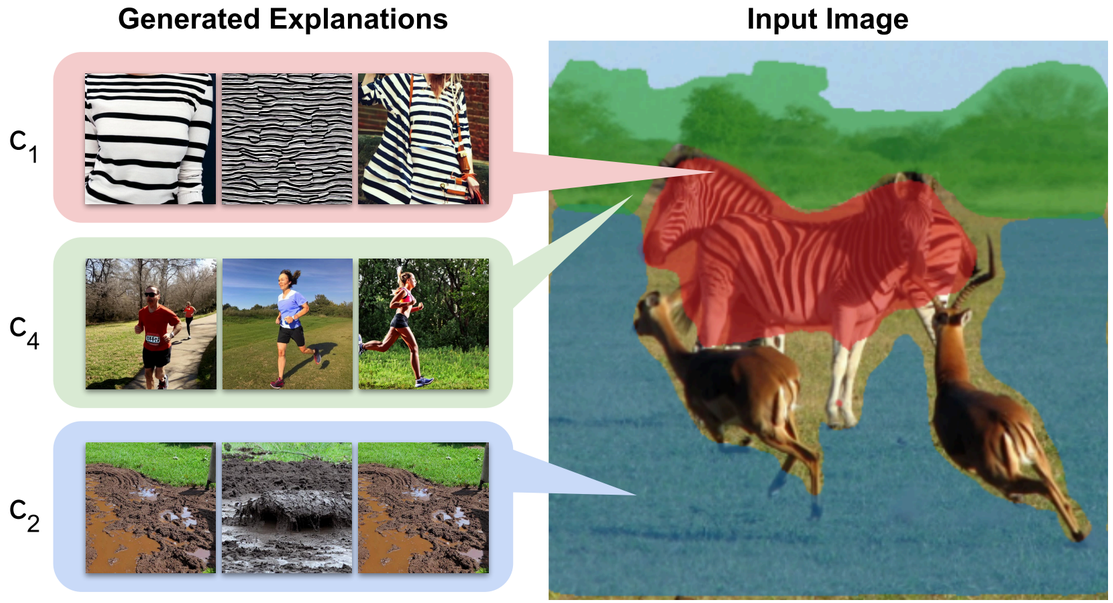
a_3: Square

a_4: Running

Selecting “Stripes” action.

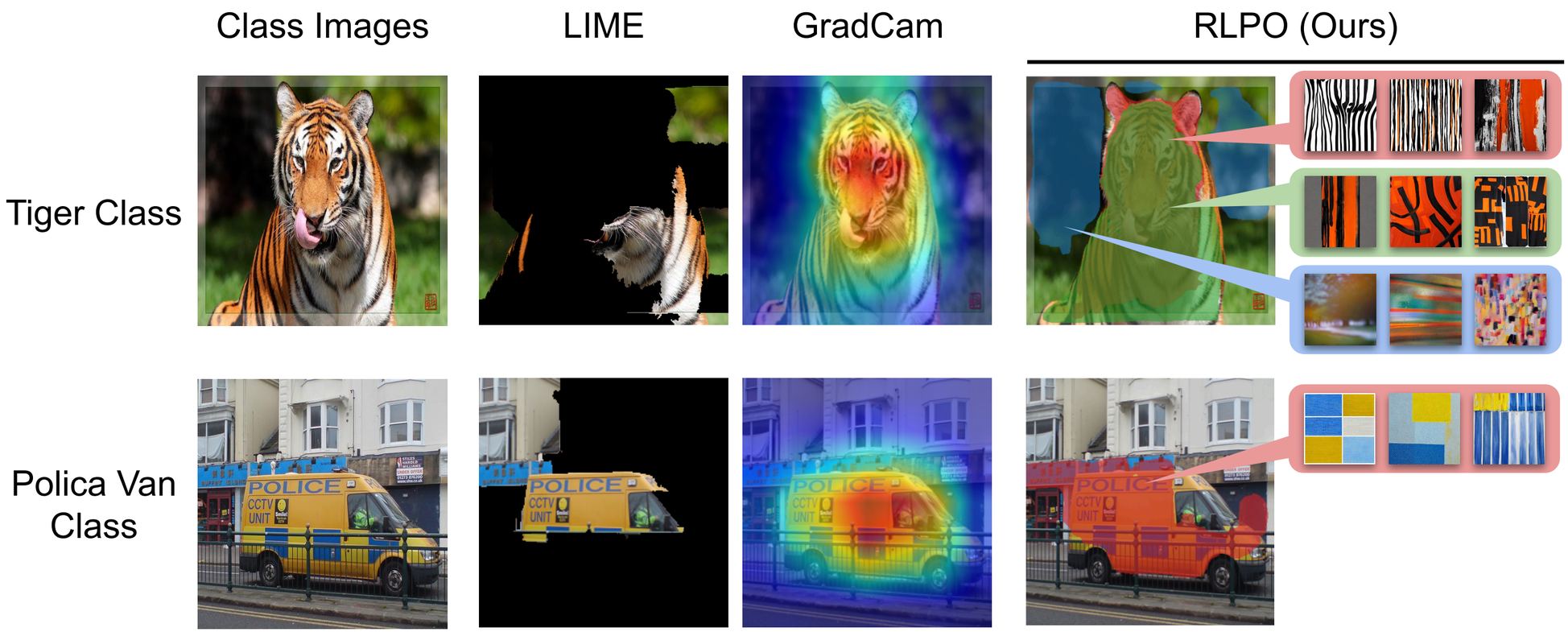
High reward.





Qualitative Results

Comparison of concepts identified by different methods. RLPO can show the correspondences between test image and different generated concepts.



Quantitative Results

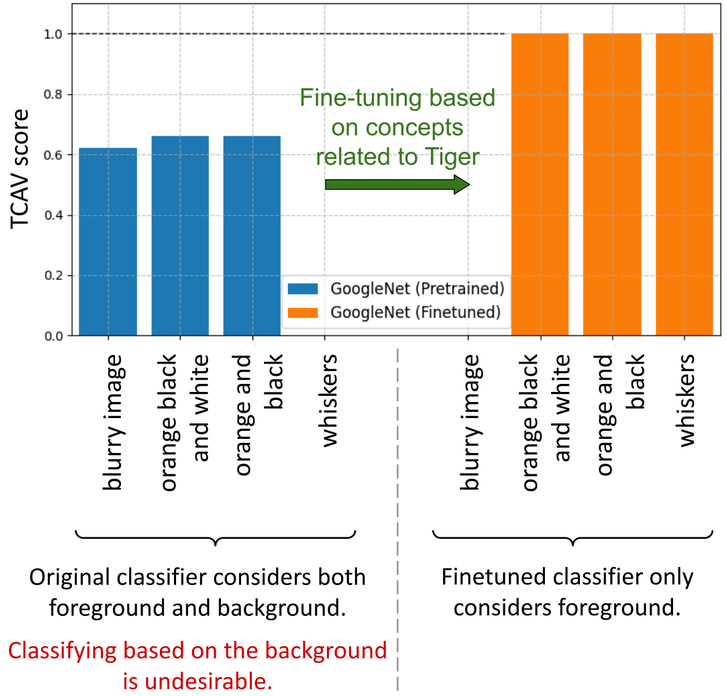
In the table below, $TS_{c,m}$ means TCAV score, CS & ED means cosine similarity and euclidean distance with CLIP embedding, and RCS & RED means cosine similarity and euclidean distance with ResNet50 embedding of class data and concepts.

Methods	Concepts	$TS_{c,m}(\uparrow)$	CS (↓)	ED (↑)	RCS (↓)	RED (↑)
EAC	C	1.0	0.76 ± 0.03	7.21 ± 0.63	0.67 ± 0.14	6.34 ± 2.16
Lens	C1	1.0	0.77 ± 0.02	7.17 ± 0.34	0.50 ± 0.18	9.70 ± 3.20
	C2	1.0	0.72 ± 0.04	8.02 ± 0.87	0.42 ± 0.10	10.90 ± 2.80
	C3	1.0	0.69 ± 0.05	8.45 ± 0.96	0.45 ± 0.05	11.03 ± 2.17
CRAFT	C1	1.0	0.76 ± 0.04	7.37 ± 0.62	0.57 ± 0.16	8.80 ± 3.20
	C2	1.0	0.72 ± 0.02	8.25 ± 0.39	0.50 ± 1.90	9.90 ± 3.40
	C3	1.0	0.73 ± 0.04	7.98 ± 0.79	0.44 ± 0.07	10.80 ± 1.90
RLPO (Ours)	C1	1.0	0.52 ± 0.04	10.48 ± 0.50	0.04 ± 0.01	16.80 ± 1.40
	C2	1.0	0.49 ± 0.02	10.65 ± 0.20	0.02 ± 0.02	17.20 ± 0.80
	C3	1.0	0.49 ± 0.02	10.74 ± 0.30	0.03 ± 0.01	17.60 ± 4.40

Use Case I: Fine-tuning Models To Focus On Relevant Concept

RLPO identified that for Tiger class, the base GoogleNet model gives equal importance to both foreground and background in the input. We then fine-tuned the model to focus more on foreground than background.

Important concepts for "Tiger" class in pre-trained GoogleNet model.

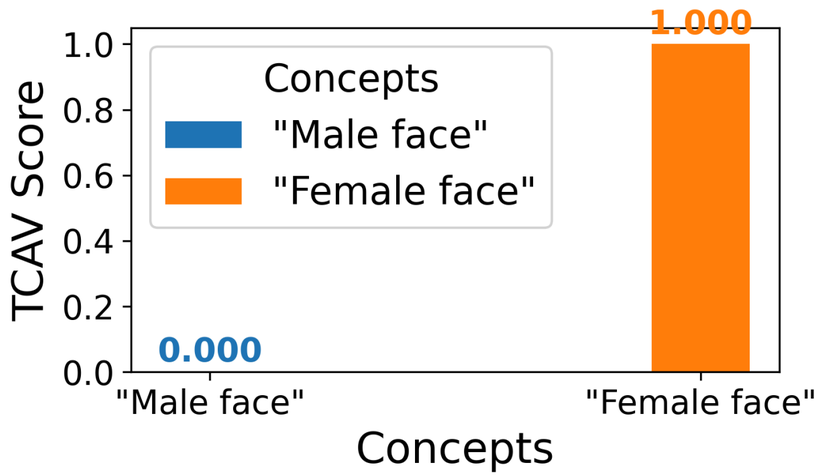


Concept shift of GoogleNet model for "Tiger" class.

Use Case II: Understanding Model Bias

Using RLPO, we generated a gender concept for a CelebA-trained blonde versus non-blonde classifier and found that,

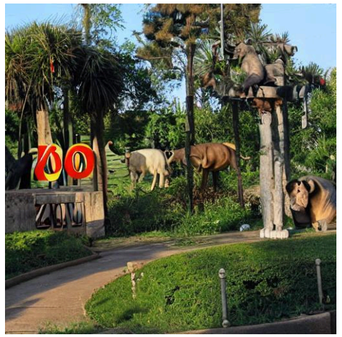
1. Concepts generated for the female face are more important than male face.
2. When we train RLPO to capture higher-level semantic concepts (such as gender), it starts combining one or more low-level concepts (such as long and blonde hair).



Other Advantages of RLPO

With RLPO, we can get explanations at **different level of abstraction**. Starting from a "zoo" seed prompt, the generated concepts evolve into tiger-like features, gaining animal shape (t=10), stripes (t=20), tiger colors (t=30), and finally a refined tiger. The classifier's prediction also shifts from "oxcart" to confidently "tiger."

Seed Prompt
"zoo"



Timestep: 0

Prediction: "Oxcart"

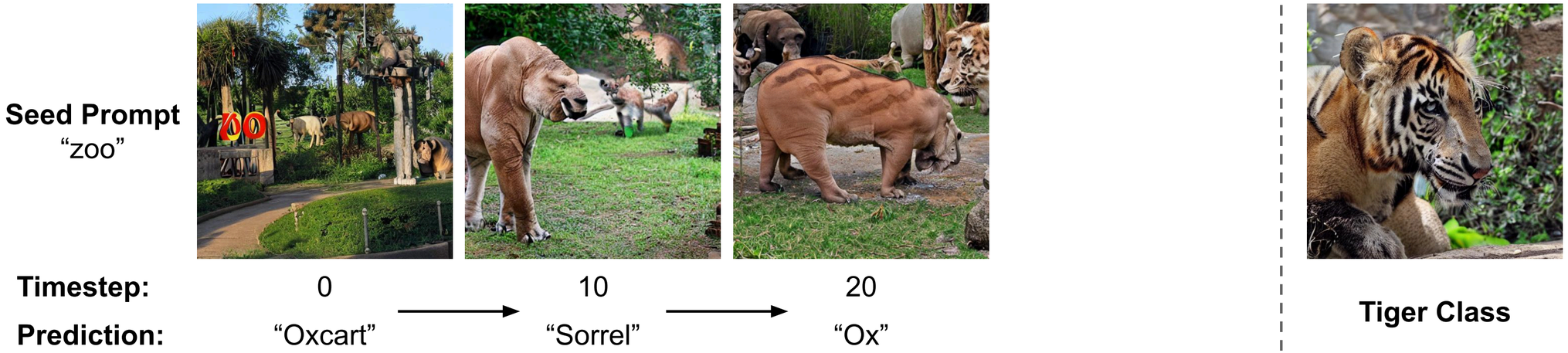
Tiger Class

Other Advantages of RLPO

With RLPO, we can get explanations at **different level of abstraction**. Starting from a "zoo" seed prompt, the generated concepts evolve into tiger-like features, gaining animal shape (t=10), stripes (t=20), tiger colors (t=30), and finally a refined tiger. The classifier's prediction also shifts from "oxcart" to confidently "tiger."

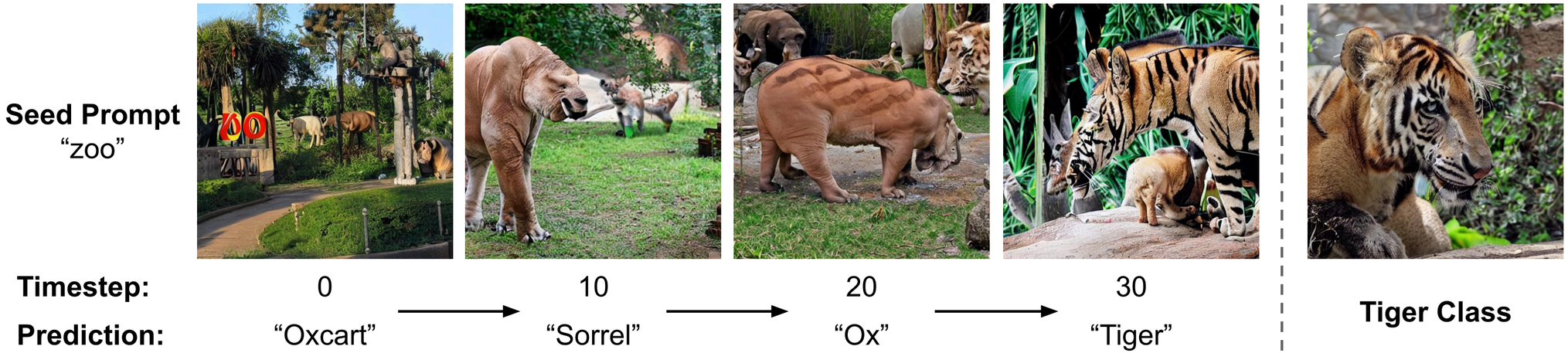
Other Advantages of RLPO

With RLPO, we can get explanations at **different level of abstraction**. Starting from a "zoo" seed prompt, the generated concepts evolve into tiger-like features, gaining animal shape (t=10), stripes (t=20), tiger colors (t=30), and finally a refined tiger. The classifier's prediction also shifts from "oxcart" to confidently "tiger."



Other Advantages of RLPO

With RLPO, we can get explanations at **different level of abstraction**. Starting from a "zoo" seed prompt, the generated concepts evolve into tiger-like features, gaining animal shape (t=10), stripes (t=20), tiger colors (t=30), and finally a refined tiger. The classifier's prediction also shifts from "oxcart" to confidently "tiger."



Beyond Images...

RLPO can also be used in identifying important concepts in sentiment analysis task. Here, instead of a diffusion model, we make use of **Mistral-7B Instruct** model to generate relevant concepts for the TextCNN sentiment classifier pre-trained on IMDB sensitivity dataset.

Positive Prompt

The customer (0.4) service team (0.3) was very helpful (0.7) and responsive when I reached out for support. They were patient and provided clear (0.3) instructions on how to address (0.4) some of the issues (0.7), which improved the situation slightly.

Generated Concepts

Customer: client, purchaser, consumer, user, shopper
Team: group, crew, unit, squad, alliance, partnership
Helpful: supportive, useful, valuable, beneficial, productive
Clear: transparent, unclouded, open, lucid, distinct
Address: speak, contact, communicate, interact, approach
Issues: problems, concerns, matters, challenges, disputes

Negative Prompt

The highly anticipated movie (0.7) turned (0.6) out to be a colossal disappointment, plagued by a weak and incoherent plot, unconvincing performances by the lead actors (0.3), lackluster (0.0) special effects (1.0), and numerous continuity errors, which collectively made it one of the worst cinematic experiences in recent memory, leaving audiences and critics (1.0) alike utterly dissatisfied and frustrated.

Generated Concepts

Effects: outcomes, consequences, impact, repercussions
Critics: reviewers, criticisms, commentators, pundits
Actors: performers, artists, thespians, players, entertain
Movie: film, motion, picture, feature, show, production
Lackluster: apathetic, bland, dull, uninspired, insipid
Turned: faced, aimed, pivoted, swiveled, rotate, reversed

Explainable Concept Generation through Vision-Language Preference Learning for Understanding Neural Networks' Internal Representations

Aditya Taparia, Som Sagar, Ransalu Senanayake
ataparia@asu.edu

Paper

Code