ONLP

The Role of Structure in Modern Multimodal AI Systems

Vardaan Pahuja

Department of Computer Science and Engineering

The Ohio State University

April 2026

Introducing the Speaker

2016
IIT Kharagpur
B. Tech (Hons.)

2016-17
IBM Research
Conversational AI and QA Research

2017-19
University of Montreal
M.Sc. (AI)

2019 - Present
The Ohio State University
Ph.D. in Computer Science

2022
Google Brain
Efficient Multimodal Foundation Models

2024
Microsoft Research
Computer-use Agents

Thesis Committee

Committee:

- Dr. Yu Su, Advisor
- Dr. Zhihui Zhu
- Dr. Sachin Kumar
- Dr. Xueru Zhang

What is Structure?

Structure refers to the explicit or implicit organization, relationships, or constraints inherent within data, models, or tasks.

Why Structure Matters?

- Explicit nodes, edges & hierarchies enable multi-hop reasoning

Why Structure Matters?

- Explicit nodes, edges & hierarchies enable multi-hop reasoning
- Structure acts as an inductive bias — guiding models toward solutions that reflect real-world organization, without needing to observe every possible combination

Why Structure Matters?

- Explicit nodes, edges & hierarchies enable multi-hop reasoning
- Structure acts as an inductive bias — guiding models toward solutions that reflect real-world organization, without needing to observe every possible combination
- Structured models are more sample-efficient and produce decisions that are interpretable

Structure Connects Heterogeneous Modalities: Multimodal KGs

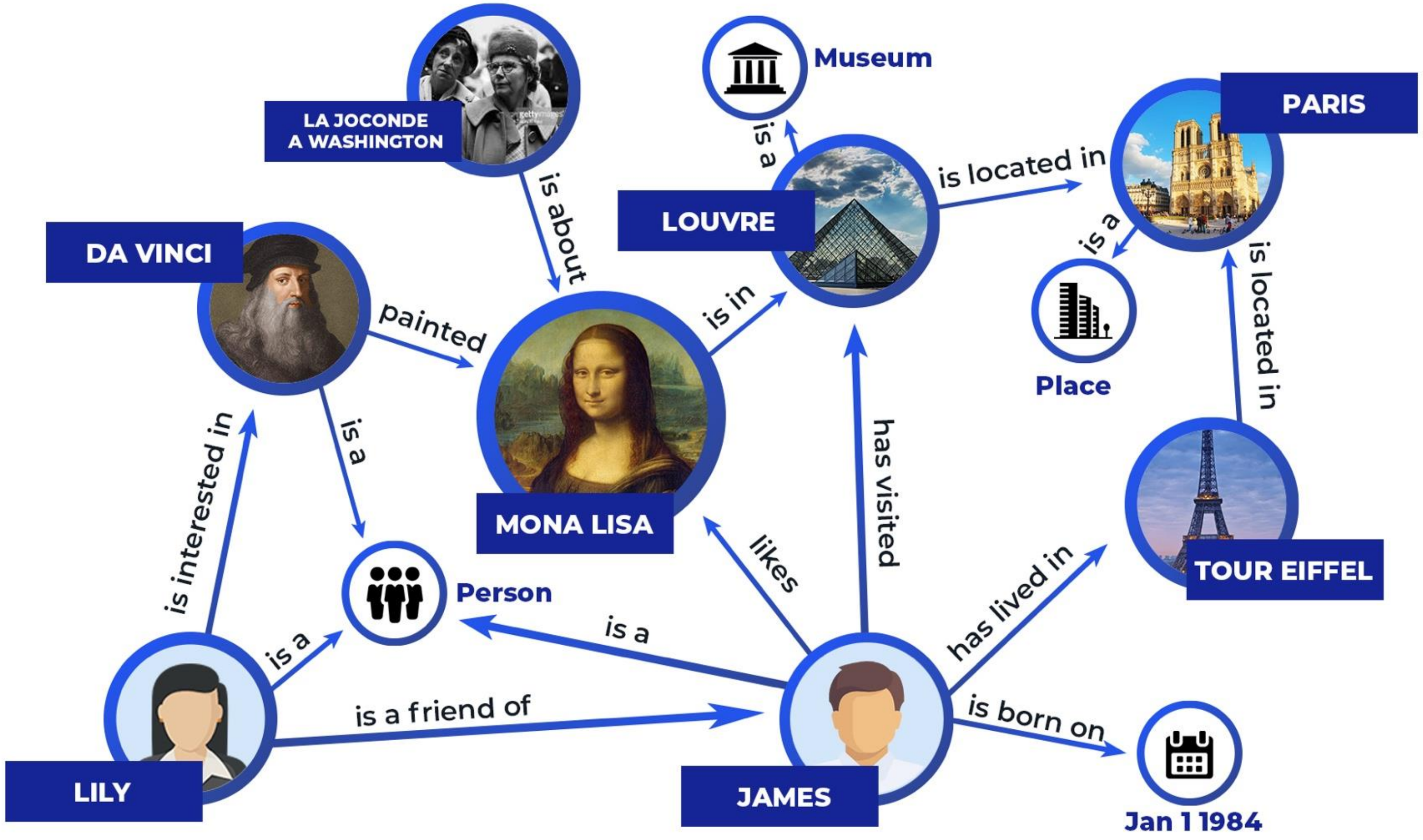
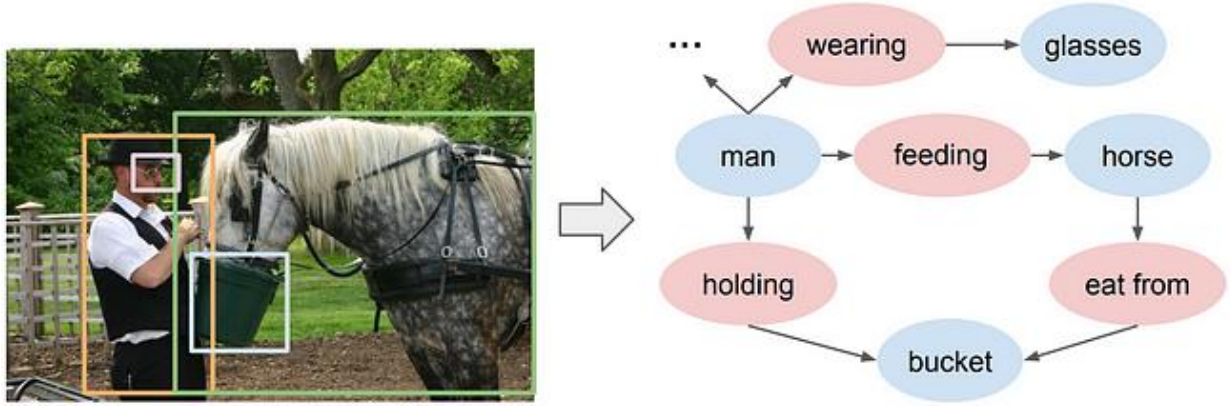


Fig. credit: Yashu Seth

Structure Connects Heterogeneous Modalities: Scene Graphs



- Scene graphs explicitly model objects and their relationships within an image or video, bridging the gap between visual and language modalities.

Fig. credit: Zhuoyi Huang and Tina Li

Structure Connects Heterogeneous Modalities: Documents

- Multimodal relationships in a document, where sections, images, captions, and page descriptions form a structured graph

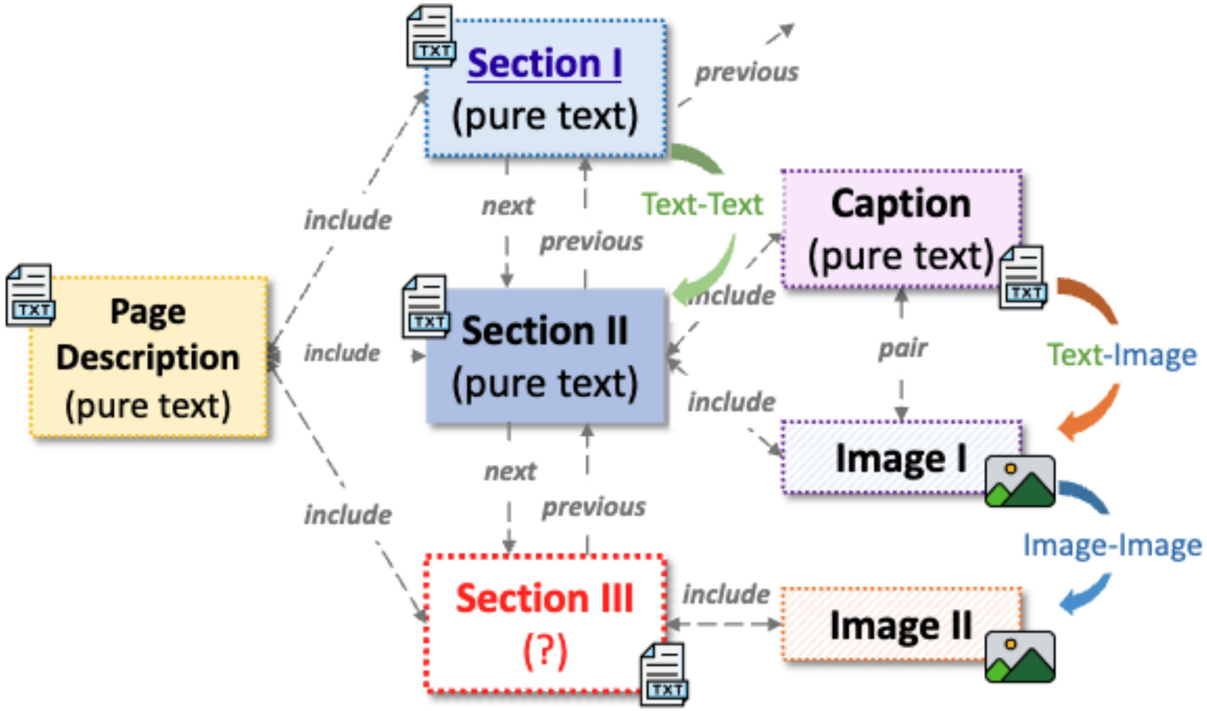


Fig. credit: Ning et al, Graph4MM: Weaving Multimodal Learning with Structural Information, ICML 2025

Structure Connects Heterogeneous Modalities: Webpages

- Elements in the DOM tree correspond to visual regions in the webpage screenshot

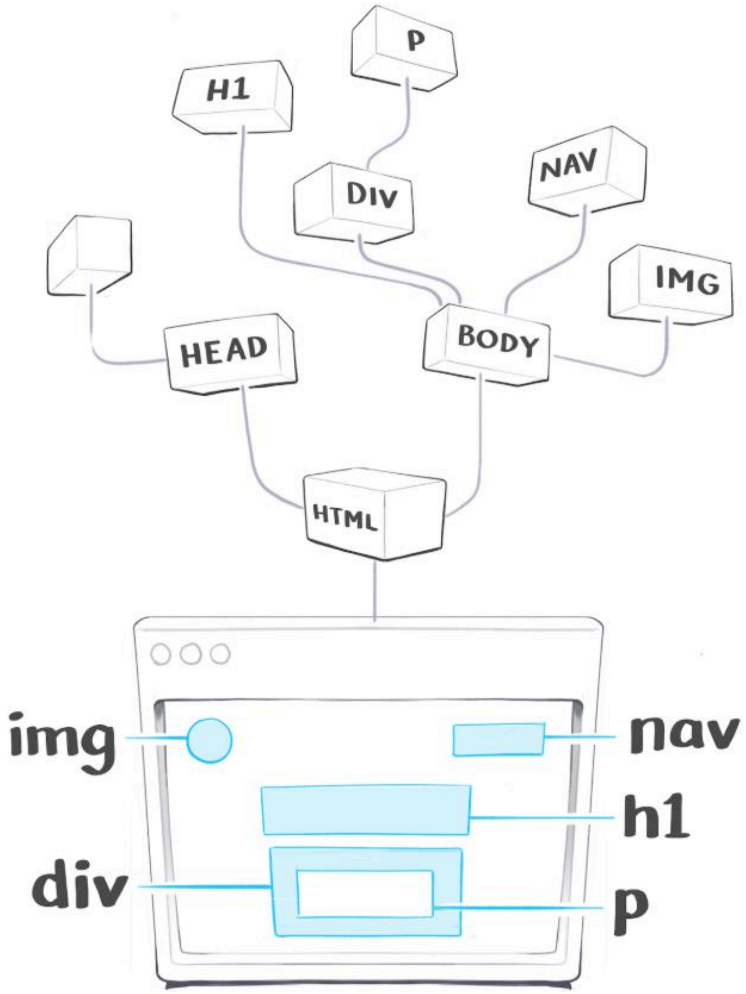
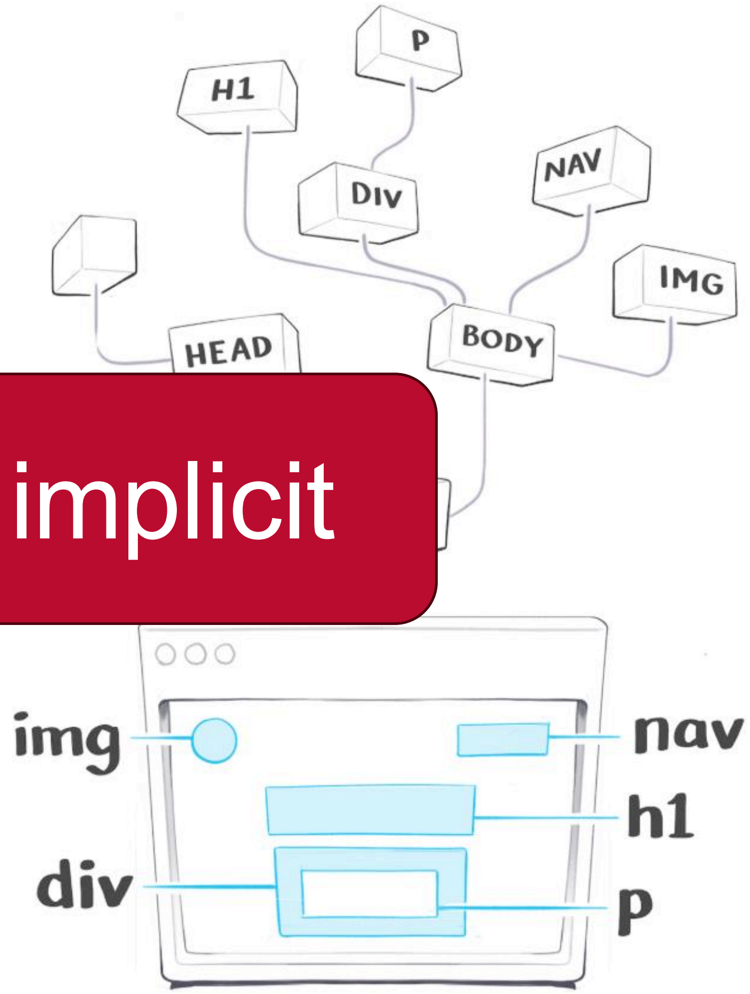


Fig credit: Hiro Nishimura

Structure Connects Heterogeneous Modalities: Webpages

- Elements in the DOM tree correspond to visual regions in the webpage screenshot



Structure may be explicit or implicit

Fig credit: Hiro Nishimura

Explicit Structure

- Multimodal KGs

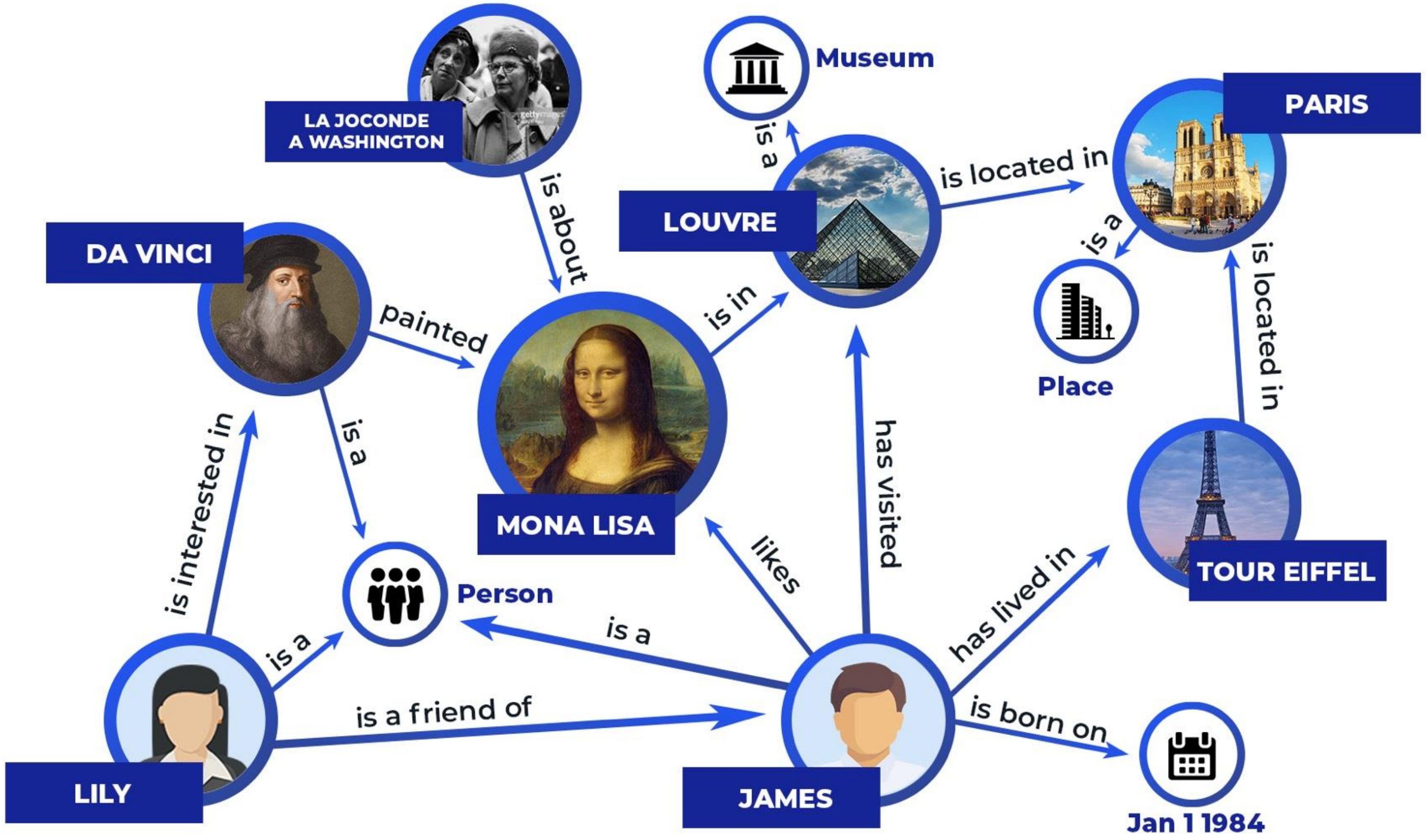
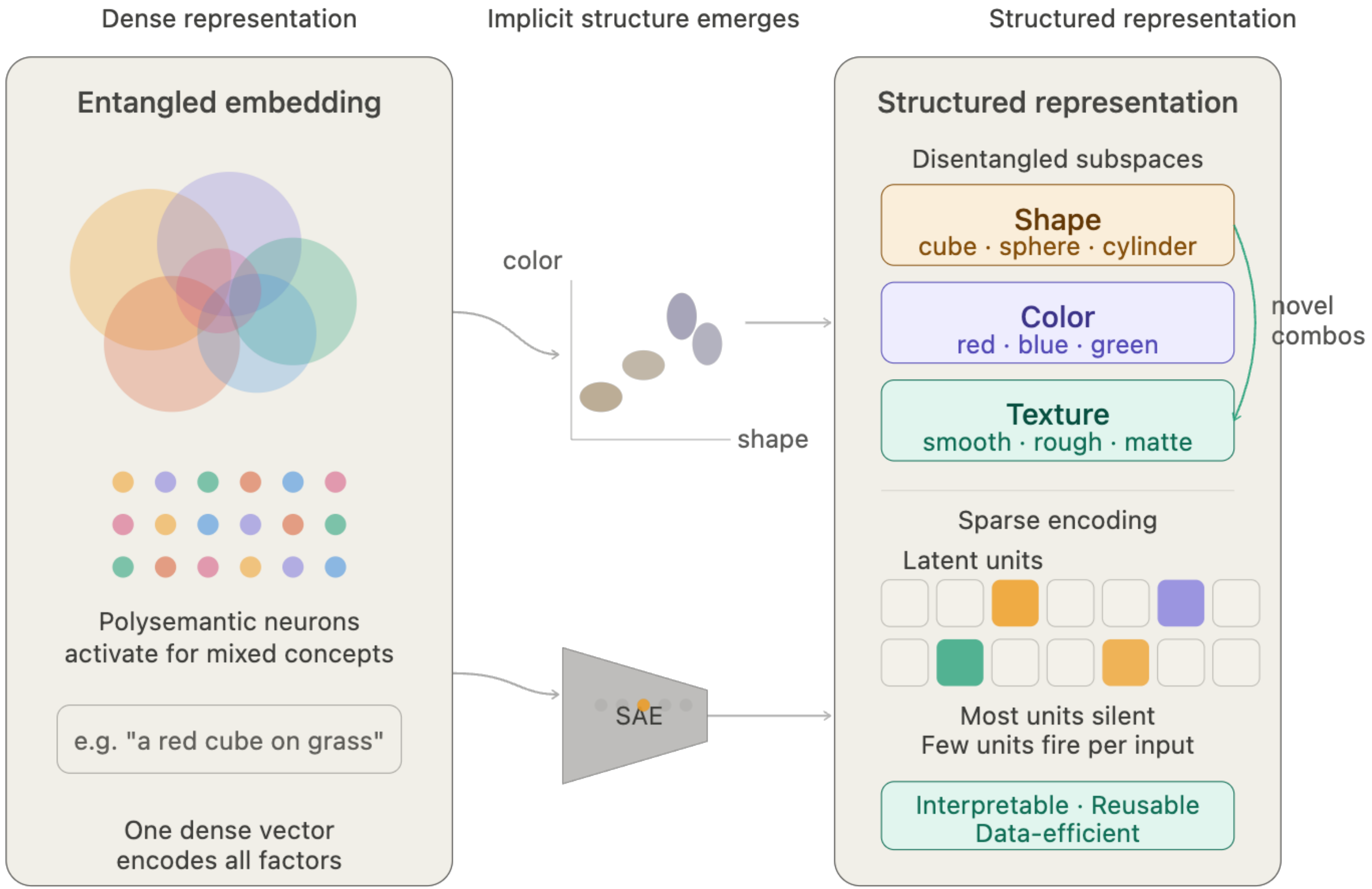


Fig. credit: Yashu Seth

Implicit Structure

- Disentangled representations



Thesis: The Role of Structure in Modern Multimodal AI Systems

CONTRIBUTION 1

Structure-aware Reasoning

Use explicit knowledge graph structure as a relational scaffold for reasoning and OOD generalization

CONTRIBUTION 2

Latent Structure Discovery

Decompose dense representations into sparse, monosemantic units aligned with interpretable biological traits

CONTRIBUTION 3

Structured Data Generation

Synthesize multimodal interaction trajectories to provide rich structured supervision for GUI agent learning

CONTRIBUTION 4

Structured Information Extraction

Impose clinical schema onto free text to transform unstructured notes into actionable structured records

Challenge 1: Structure-Aware Reasoning

- How can relevant structured graph context be leveraged efficiently to improve link prediction in knowledge graphs?
- How can heterogeneous multimodal information be integrated and reasoned over in a unified framework for robust visual recognition?

Challenge 2: Latent Structure Discovery

- How can latent structure in dense visual representations be uncovered for fine-grained visual recognition?

Challenge 3: Structured Data Generation

- How can structured multimodal trajectories be synthesized at scale to train GUI agents?

Challenge 4: Structured Information Extraction

- How can LLMs transform unstructured clinical notes into structured, schema-aligned health information?

Thesis: The Role of Structure in Modern Multimodal AI Systems

Contribution 1 — Structure-aware Reasoning

- **ACL'21** A systematic investigation of KB-text embedding alignment at scale
- **CIKM'23** A retrieve-and-read framework for knowledge graph link prediction
- **CIKM'24** Reviving the context: camera trap species classification as link prediction on multimodal knowledge graphs

Contribution 2 — Latent Structure Discovery

- **ICLR'26** Automatic image-level morphological trait annotation for organismal images

Contribution 3 — Structured Data Generation

- **ACL'25** Explorer: scaling exploration-driven web trajectory synthesis for multimodal web agents

Contribution 4 — Structured Information Extraction

- **In review** Using large language models for social determinants of health phenotyping: a systematic study

Thesis: The Role of Structure in Modern Multimodal AI Systems

CONTRIBUTION 1

Structure-aware Reasoning

Use explicit knowledge graph structure as a relational scaffold for reasoning and OOD generalization

CONTRIBUTION 2

Latent Structure Discovery

Decompose dense representations into sparse, monosemantic units aligned with interpretable biological traits

CONTRIBUTION 3

Structured Data Generation

Synthesize multimodal interaction trajectories to provide rich structured supervision for GUI agent learning

CONTRIBUTION 4

Structured Information Extraction

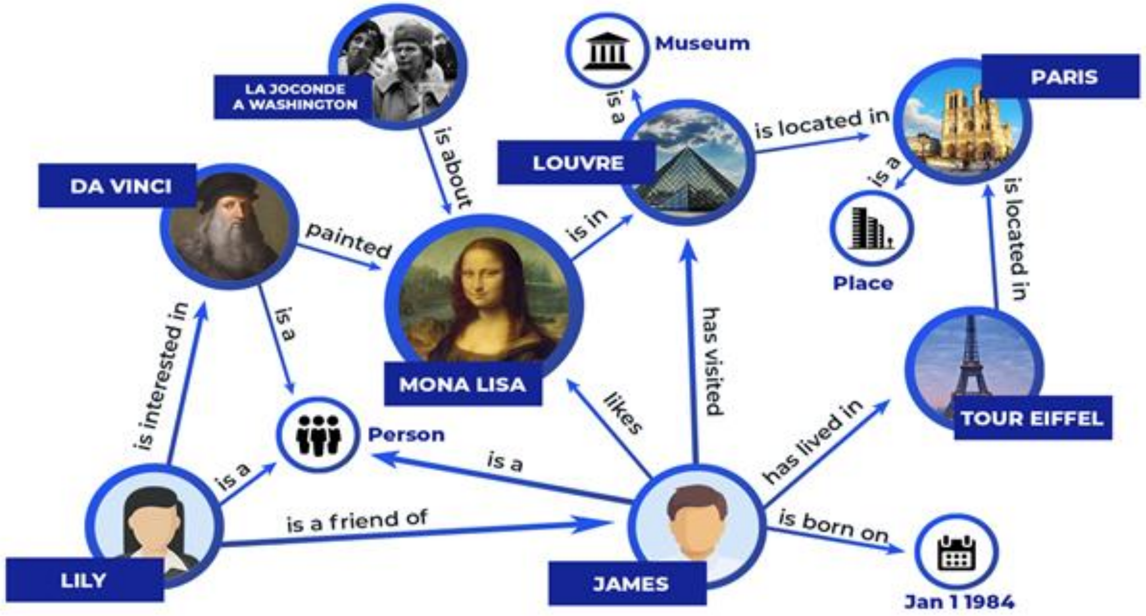
Impose clinical schema onto free text to transform unstructured notes into actionable structured records

Challenge 1: Structure-Aware Reasoning

- How can relevant structured graph context be leveraged efficiently to improve link prediction in knowledge graphs?
- How can heterogeneous multimodal information be integrated and reasoned over in a unified framework for robust visual recognition?

Knowledge Graphs

- Multi-relational graphs
- Provides structured representation for semantic relationships between real-world entities.



Pahuja, Vardaan, et al. "A retrieve-and-read framework for knowledge graph link prediction.", CIKM 2023

Knowledge Graph Completion Task

- Knowledge Graph Completion (KGC): Given an incomplete RDF triple, infer the missing entity.
- E.g. (Statue of Liberty, located in, ?)

<head> Statue of Liberty
<relation> located in
<tail> ?

Input: Structured Query

Figure credit: Hsiang-Yun Wu

Pahuja, Vardaan, et al. "A retrieve-and-read framework for knowledge graph link prediction.", CIKM 2023

Knowledge Graph Completion Task

- Knowledge Graph Completion (KGC): Given an incomplete RDF triple, infer the missing entity.
- E.g. (Statue of Liberty, located in, ?)

<head> Statue of Liberty
<relation> located in
<tail> ?

→ KGC model →

France
Italy
United States
Mexico
Canada

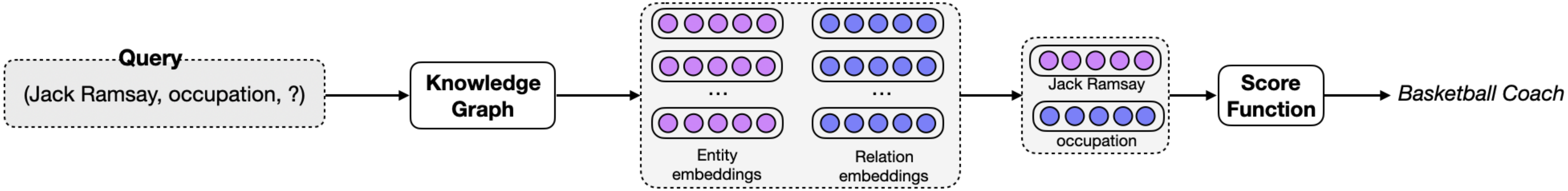
Input: Structured Query

Output: Ranked candidate entities

Pahuja, Vardaan, et al. "A retrieve-and-read framework for knowledge graph link prediction.", CIKM 2023

Motivation

- **Shallow Embedding approaches** (TransE, DistMult, RotatE, etc.)

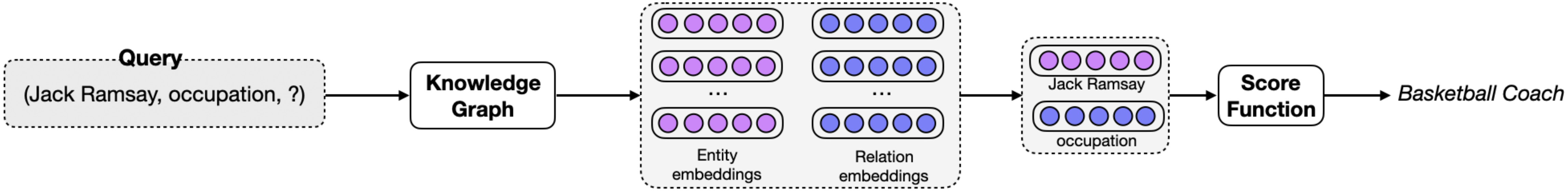


❖ Vector embeddings do not capture rich structural and semantic information of KGs

Pahuja, Vardaan, et al. "A retrieve-and-read framework for knowledge graph link prediction.", CIKM 2023

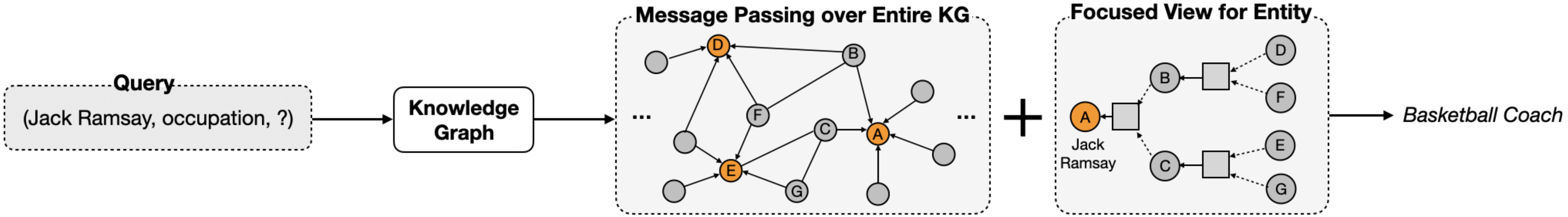
Motivation

- **Shallow Embedding approaches** (TransE, DistMult, RotatE, etc.)



❖ Vector embeddings do not capture rich structural and semantic information of KGs

- **Message-passing based GNN approaches** (R-GCN, CompGCN, etc.)



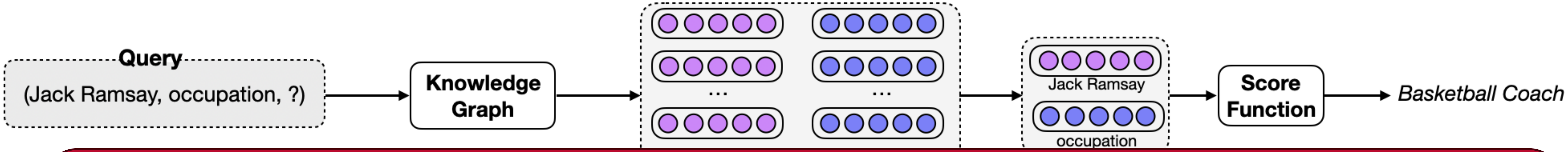
❖ Message-passing over the entire KG leads to superfluous computation.

❖ Limited scalability on large-scale KGs.

Pahuja, Vardaan, et al. "A retrieve-and-read framework for knowledge graph link prediction.", CIKM 2023

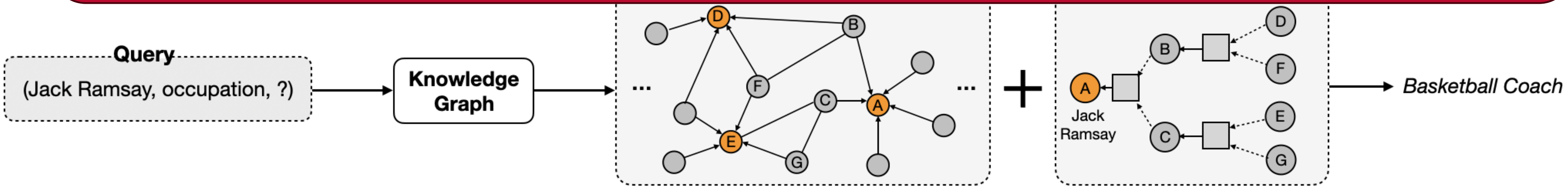
Motivation

- **Shallow Embedding approaches** (TransE, DistMult, RotatE, etc.)



Can we efficiently reason over structured neighborhood context for link prediction?

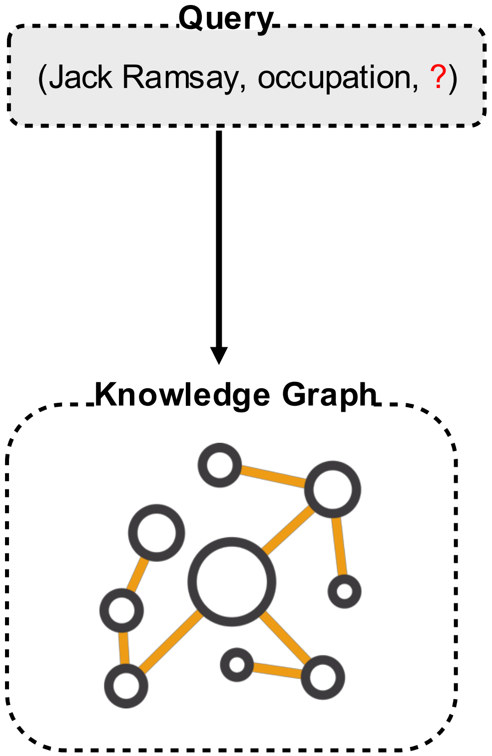
-



❖ Message-passing over the entire KG leads to superfluous computation.
❖ Limited scalability on large-scale KGs.

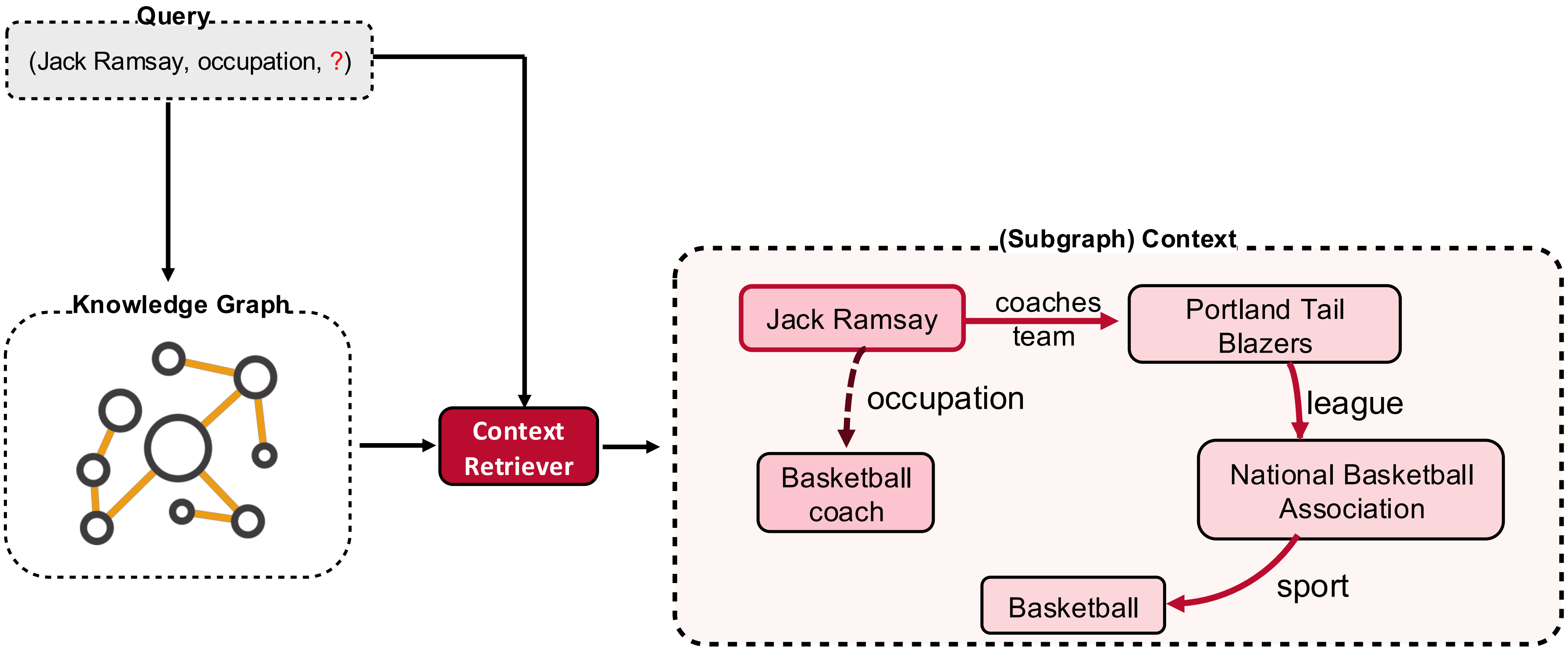
Pahuja, Vardaan, et al. "A retrieve-and-read framework for knowledge graph link prediction.", CIKM 2023

Structure-aware Link Prediction



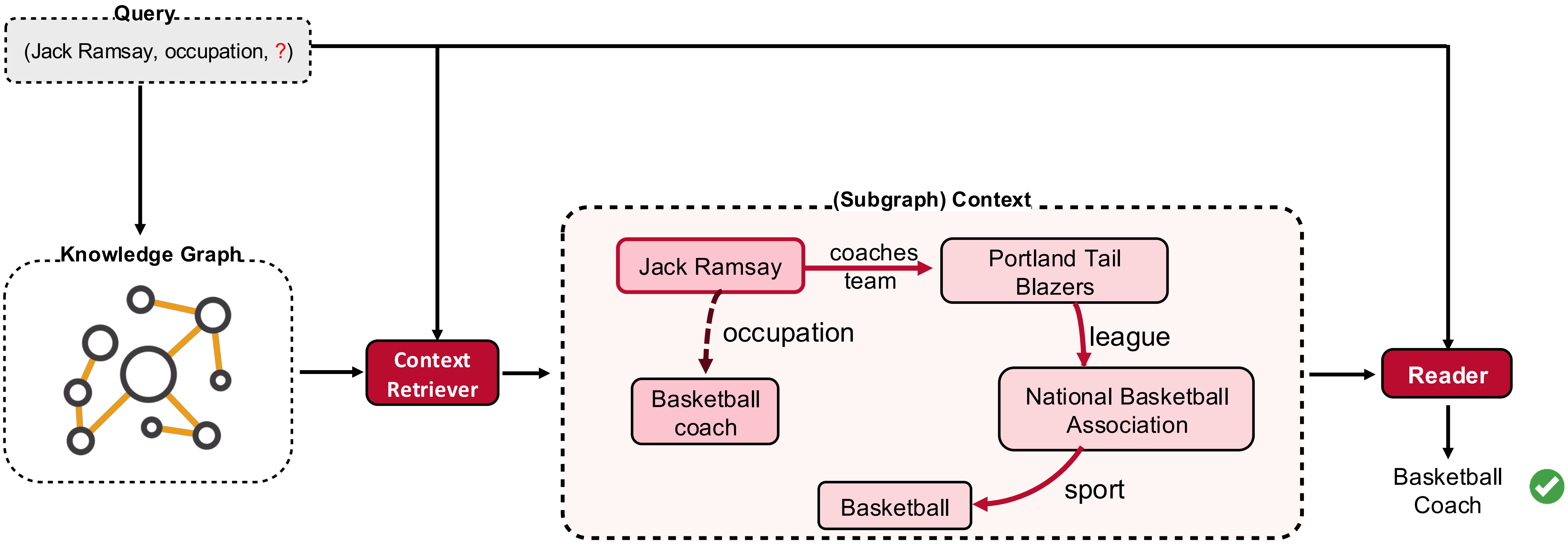
Pahuja, Vardaan, et al. "A retrieve-and-read framework for knowledge graph link prediction.", CIKM 2023

Structure-aware Link Prediction



Pahuja, Vardaan, et al. "A retrieve-and-read framework for knowledge graph link prediction.", CIKM 2023

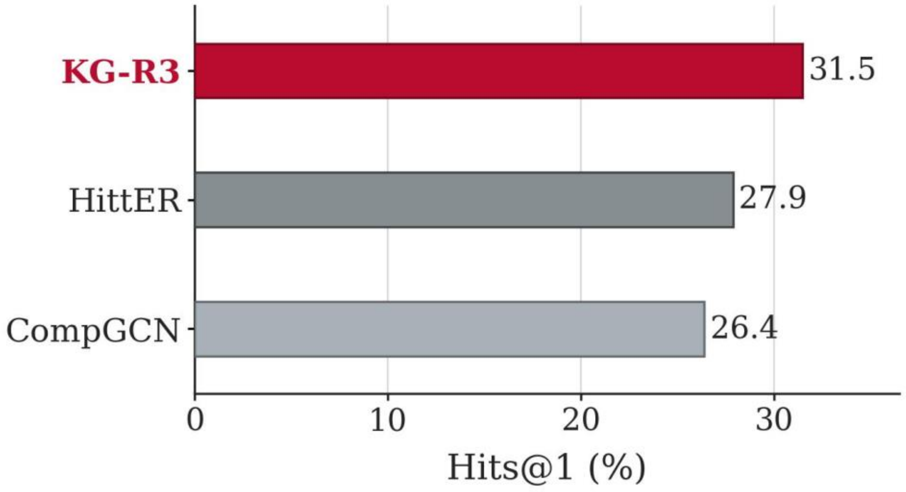
Structure-aware Link Prediction



Pahuja, Vardaan, et al. "A retrieve-and-read framework for knowledge graph link prediction.", CIKM 2023

Experiments

- 12.9% relative improvement (Hits@1) over HittER, a baseline Transformer model on FB15K-237.

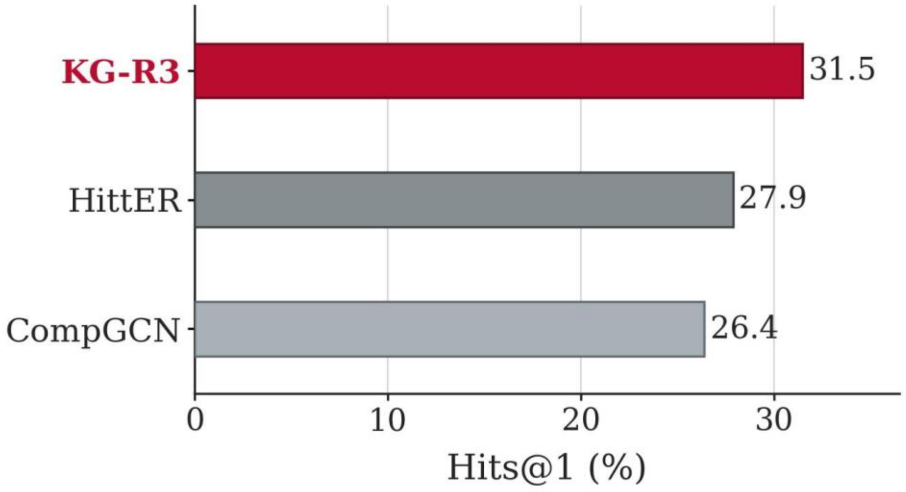


Link Prediction Performance (FB15K-237)

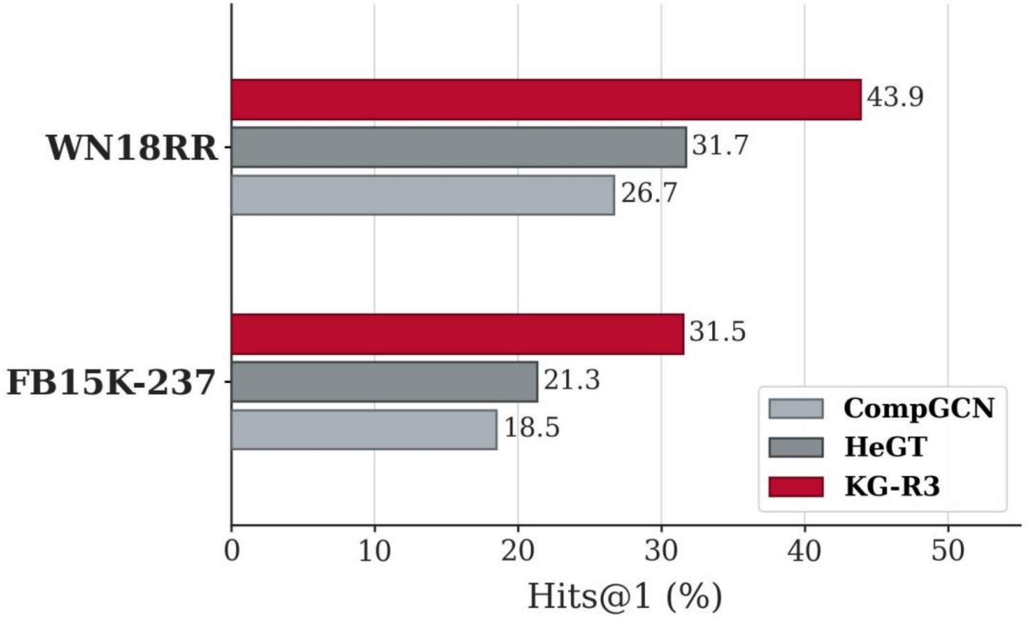
Pahuja, Vardaan, et al. "A retrieve-and-read framework for knowledge graph link prediction.", CIKM 2023

Experiments

- 12.9% relative improvement (Hits@1) over HittER, a baseline Transformer model on FB15K-237.
- Transformer-based GNN reader vastly outperforms other reader models.



Link Prediction Performance (FB15K-237)



Performance comparison (Reader module)

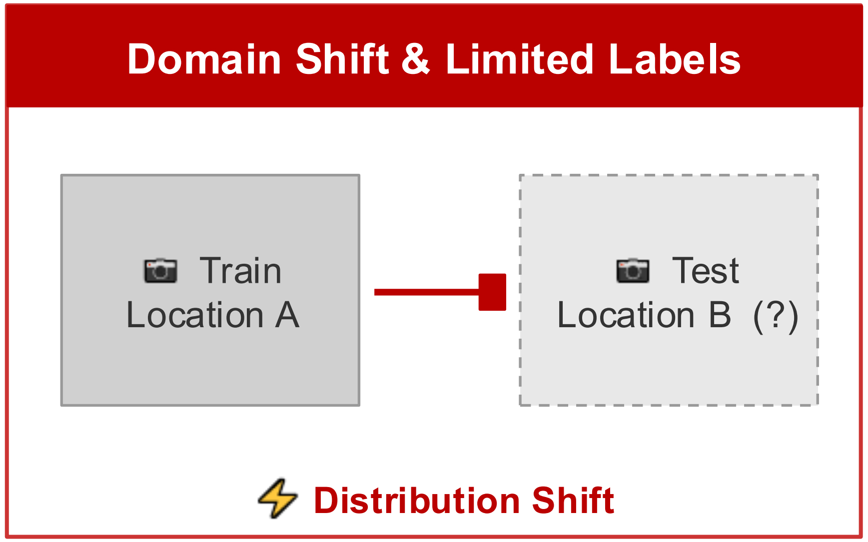
Pahuja, Vardaan, et al. "A retrieve-and-read framework for knowledge graph link prediction.", CIKM 2023

Challenge 1: Structure-Aware Reasoning

- How can relevant structured graph context be leveraged efficiently to improve link prediction in knowledge graphs?
- How can heterogeneous multimodal information be integrated and reasoned over in a unified framework for robust visual recognition?

Structure-Aware Reasoning for Visual Recognition

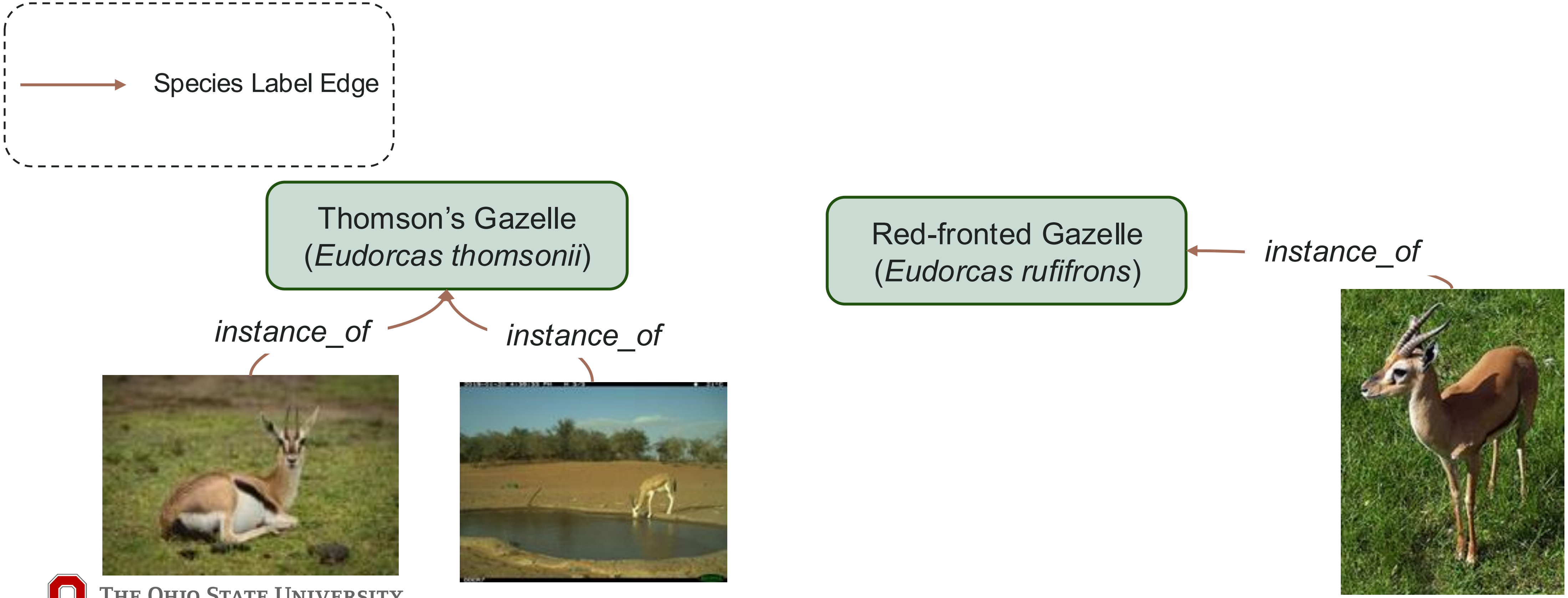
- Fine-grained species classification suffers from domain shifts and limited labeled data in real-world deployments
- Key Insight: Use invariant context (taxonomy + spatiotemporal) to overcome visual domain shifts



Pahuja, Vardaan, et al. "Reviving the Context: Camera Trap Species Classification as Link Prediction on Multimodal Knowledge Graphs.", CIKM 2024

Multimodal Data Has Structure

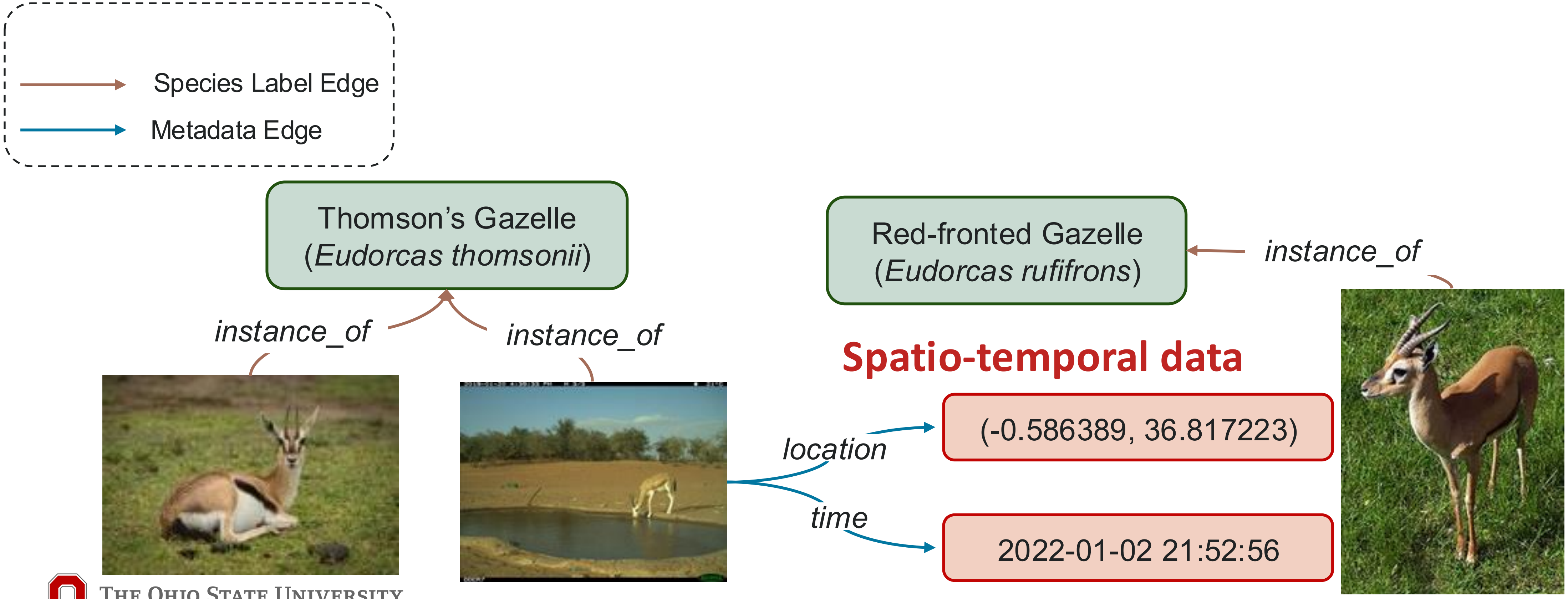
- Structured representation for semantic relationships b/w wildlife species, images, attributes and taxonomy.



Pahuja, Vardaan, et al. "Reviving the Context: Camera Trap Species Classification as Link Prediction on Multimodal Knowledge Graphs.", CIKM 2024

Multimodal Data Has Structure

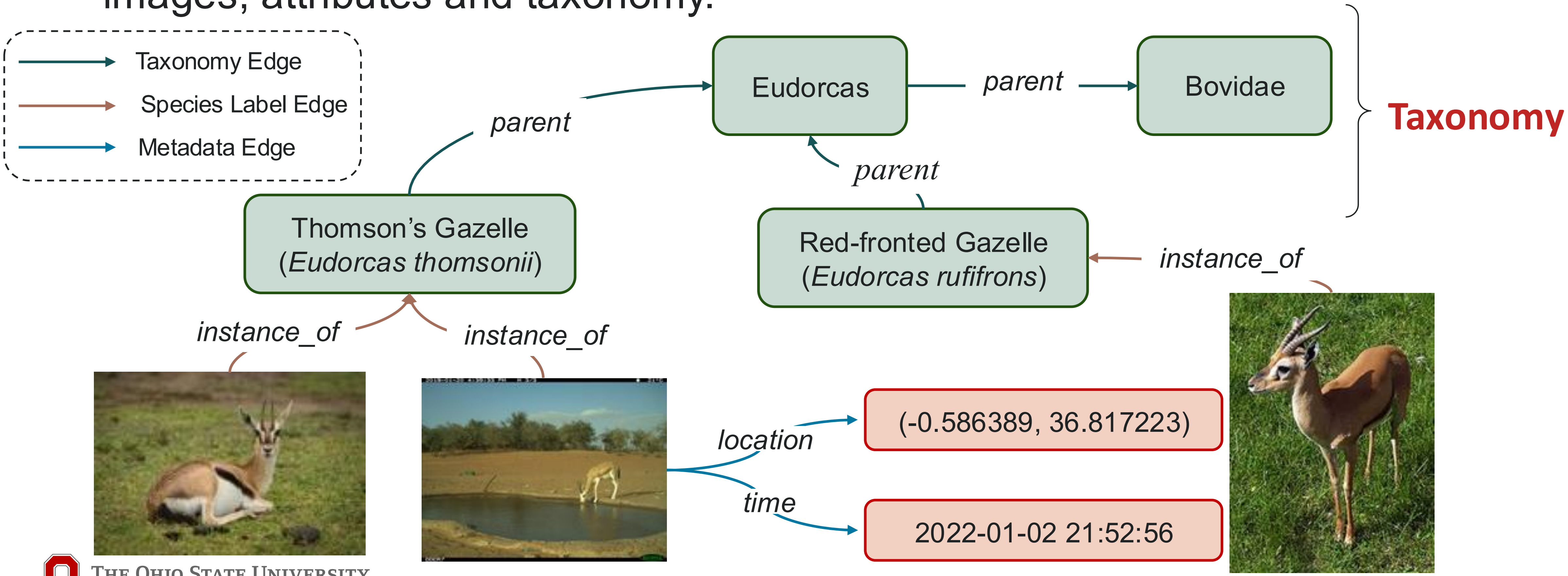
- Structured representation for semantic relationships b/w wildlife species, images, attributes and taxonomy.



Pahuja, Vardaan, et al. "Reviving the Context: Camera Trap Species Classification as Link Prediction on Multimodal Knowledge Graphs.", CIKM 2024

Multimodal Data Has Structure

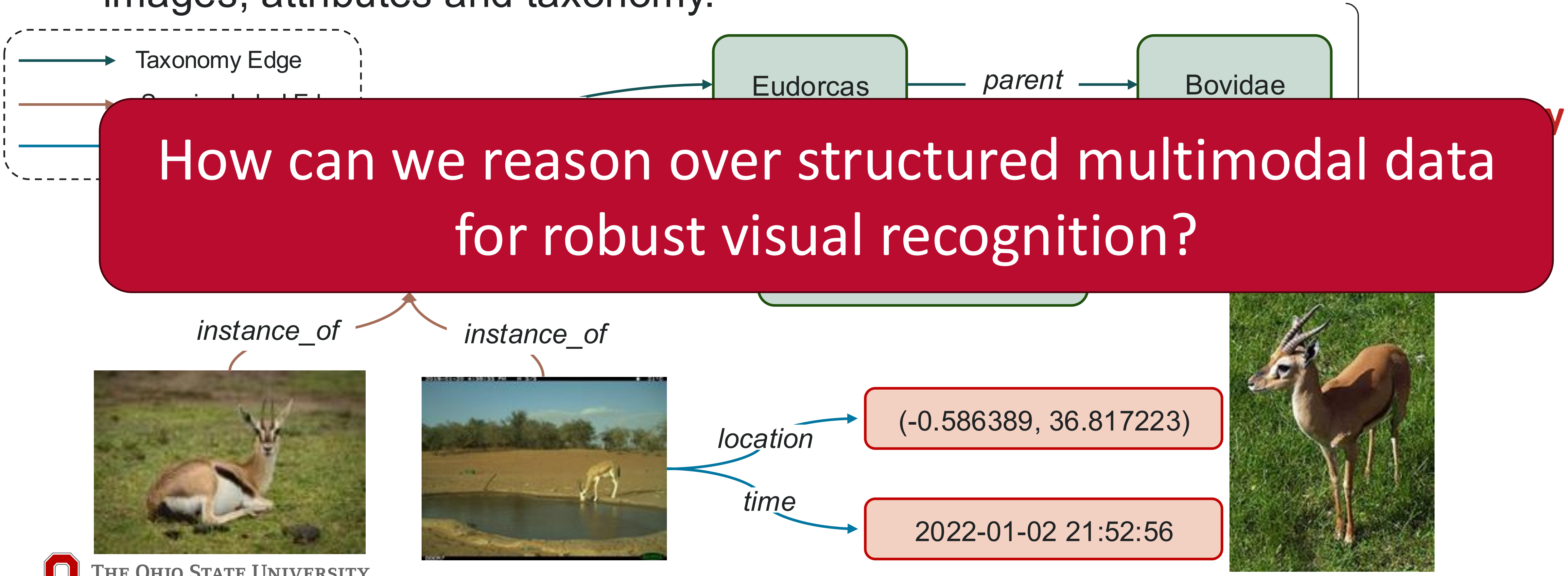
- Structured representation for semantic relationships b/w wildlife species, images, attributes and taxonomy.



Pahuja, Vardaan, et al. "Reviving the Context: Camera Trap Species Classification as Link Prediction on Multimodal Knowledge Graphs.", CIKM 2024

Multimodal Data Has Structure

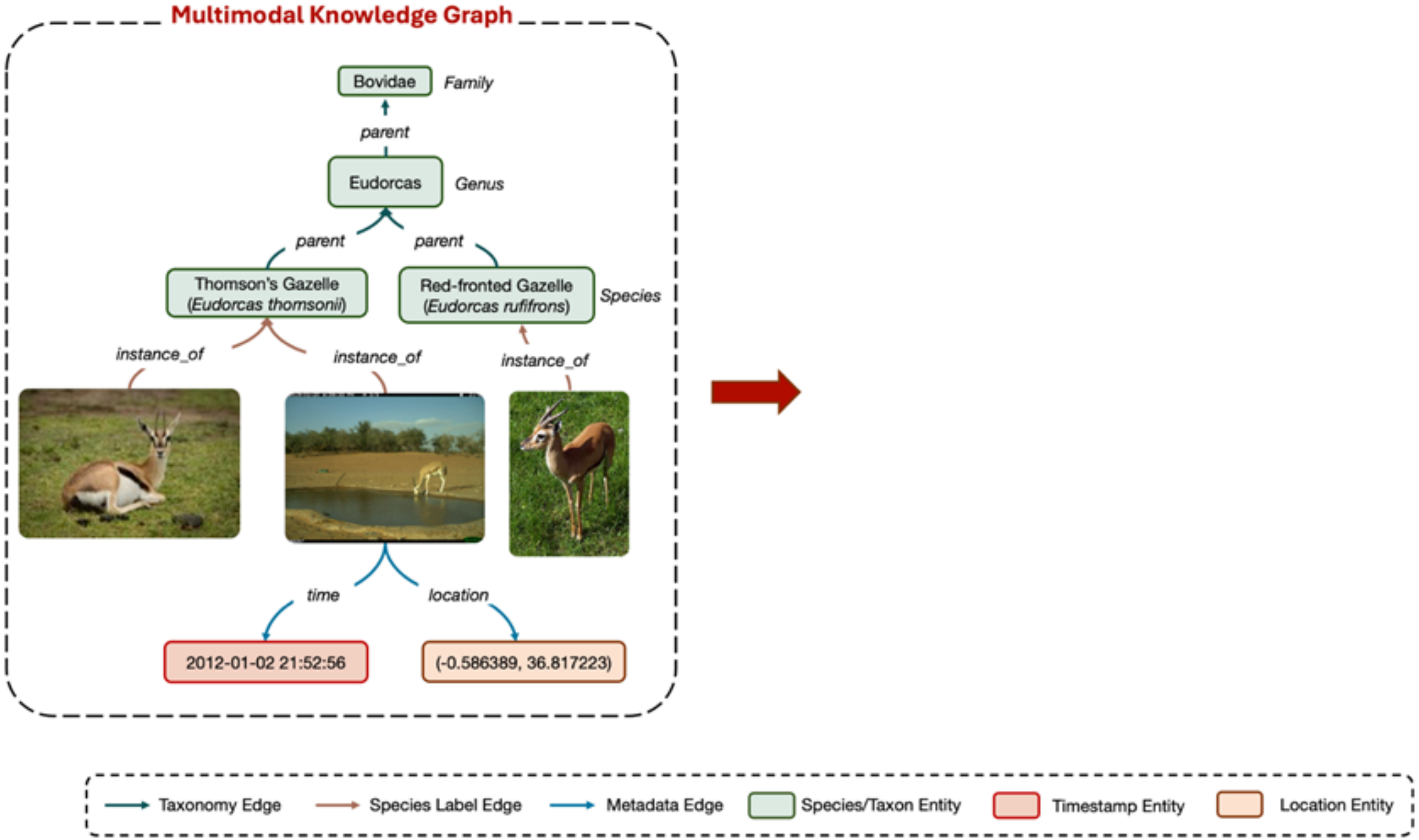
- Structured representation for semantic relationships b/w wildlife species, images, attributes and taxonomy.



How can we reason over structured multimodal data for robust visual recognition?

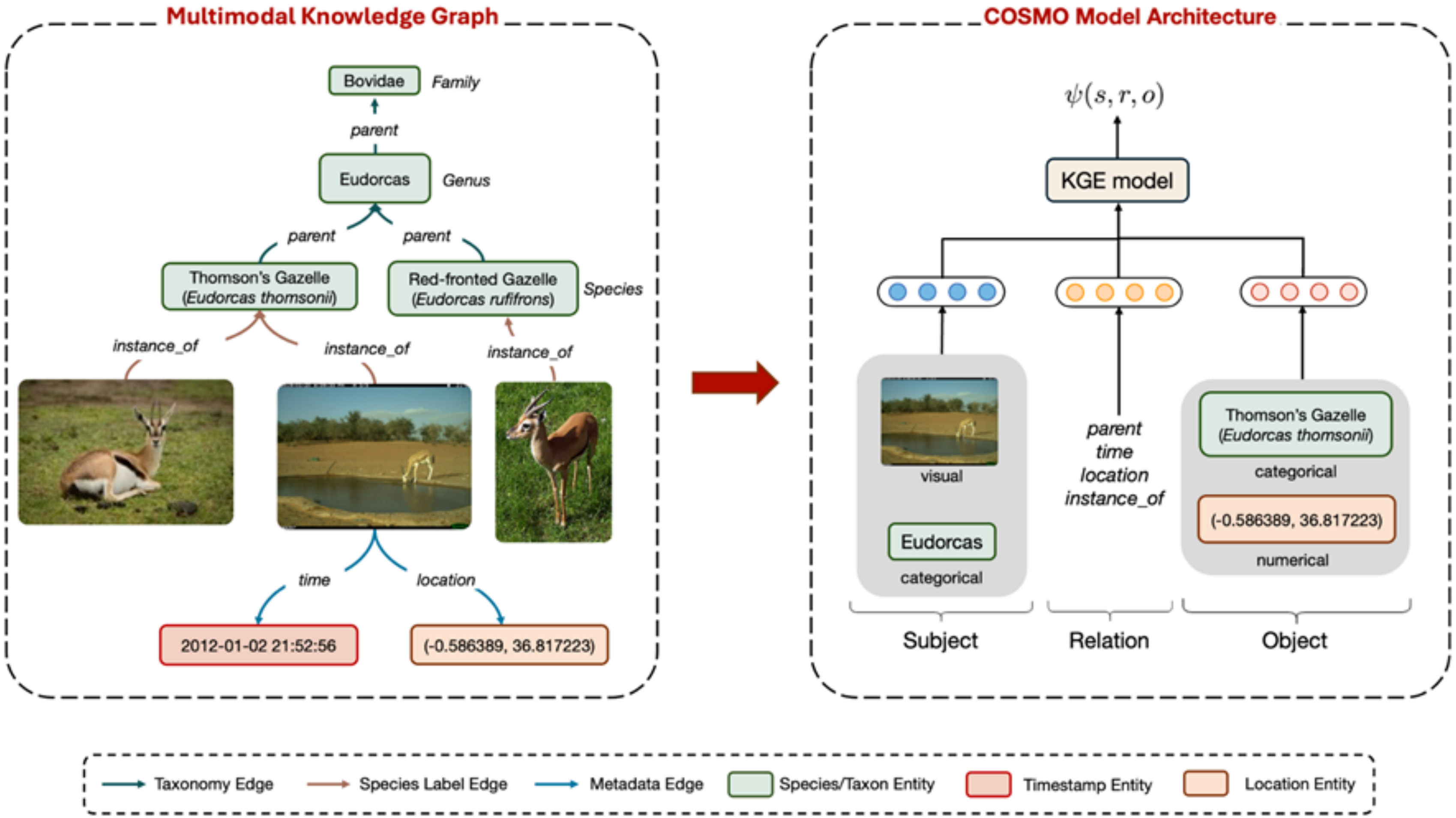
Pahuja, Vardaan, et al. "Reviving the Context: Camera Trap Species Classification as Link Prediction on Multimodal Knowledge Graphs.", CIKM 2024

How to Reason Over Structured Multimodal Data?



Pahuja, Vardaan, et al. "Reviving the Context: Camera Trap Species Classification as Link Prediction on Multimodal Knowledge Graphs.", CIKM 2024

How to Reason Over Structured Multimodal Data?



Pahuja, Vardaan, et al. "Reviving the Context: Camera Trap Species Classification as Link Prediction on Multimodal Knowledge Graphs.", CIKM 2024

COSMO: Visual Recognition using Structured Reasoning

- Task: Given a camera trap image, predict the species label

Pahuja, Vardaan, et al. "Reviving the Context: Camera Trap Species Classification as Link Prediction on Multimodal Knowledge Graphs.", CIKM 2024

COSMO: Visual Recognition using Structured Reasoning

- Task: Given a camera trap image, predict the species label
- Reformulate species classification as link prediction in the multimodal KG

<head>

<relation> instance of

<tail> ?

KGC model

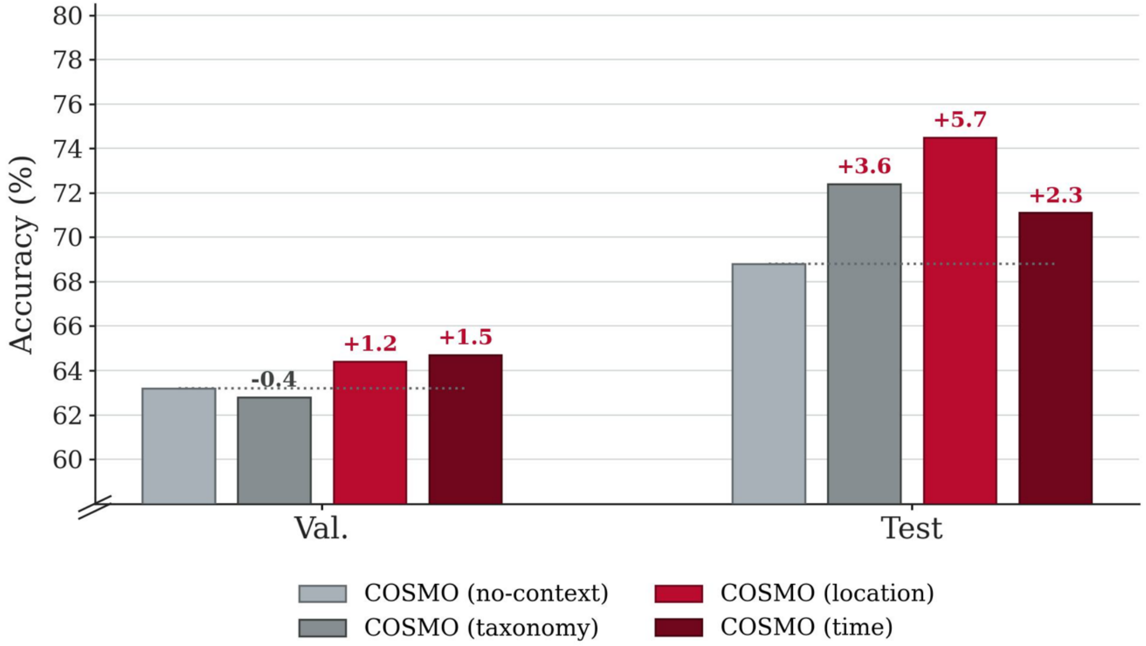
- Gazella cuvieri ✗
- Gazella marica ✗
- Eudorcas thomsonii ✓
- Eudorcas rufifrons ✗
- Eudorcas albonotata ✗

Input: Structured Query

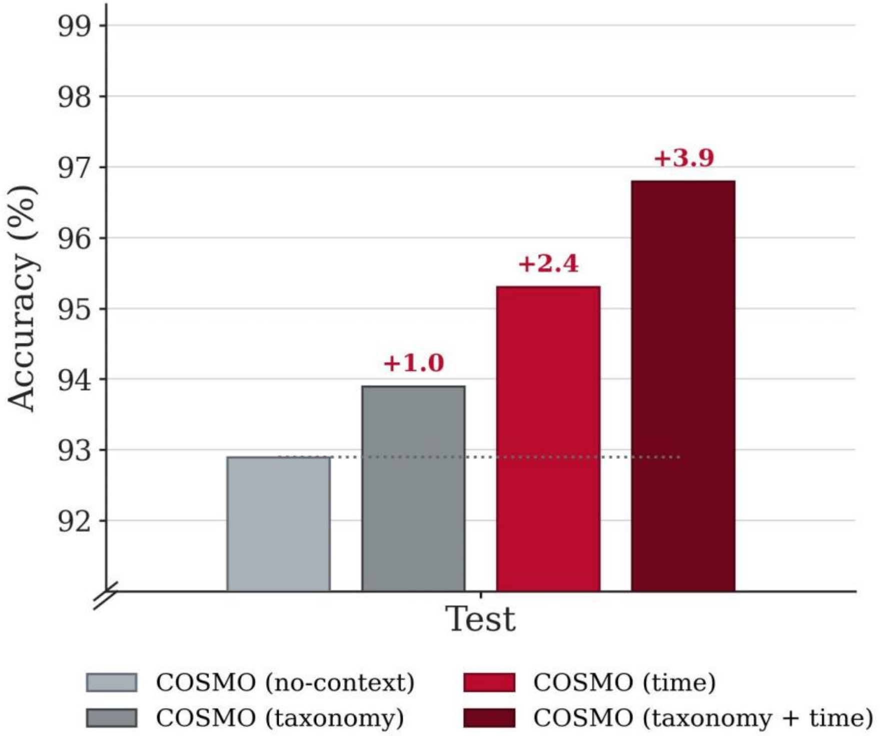
Output: Candidate species with scores

Pahuja, Vardaan, et al. "Reviving the Context: Camera Trap Species Classification as Link Prediction on Multimodal Knowledge Graphs.", CIKM 2024

Structured Multimodal Reasoning Improves Robustness



iWildCam2020-WILDS



Snapshot Mountain Zebra

Pahuja, Vardaan, et al. "Reviving the Context: Camera Trap Species Classification as Link Prediction on Multimodal Knowledge Graphs.", CIKM 2024

Thesis: The Role of Structure in Modern Multimodal AI Systems

CONTRIBUTION 1

Structure-aware Reasoning

Use explicit knowledge graph structure as a relational scaffold for reasoning and OOD generalization

CONTRIBUTION 2

Latent Structure Discovery

Decompose dense representations into sparse, monosemantic units aligned with interpretable biological traits

CONTRIBUTION 3

Structured Data Generation

Synthesize multimodal interaction trajectories to provide rich structured supervision for GUI agent learning

CONTRIBUTION 4

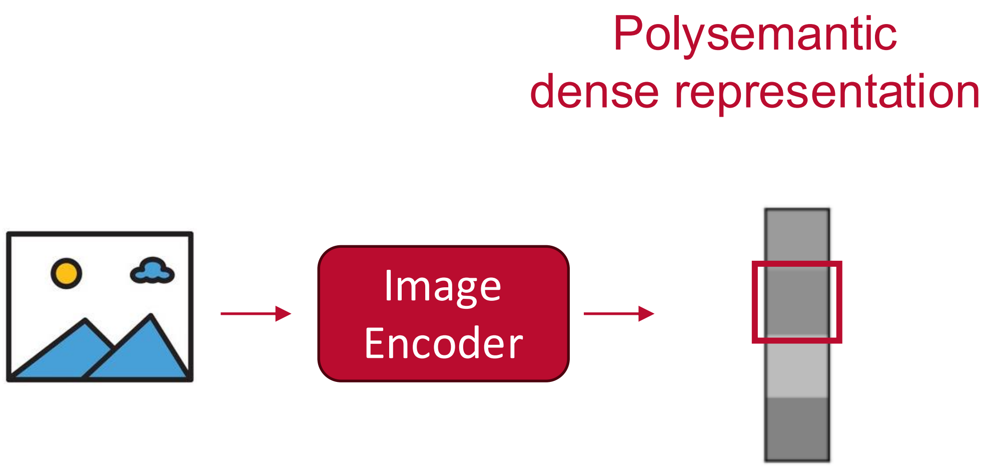
Structured Information Extraction

Impose clinical schema onto free text to transform unstructured notes into actionable structured records

Challenge 2: Latent Structure Discovery

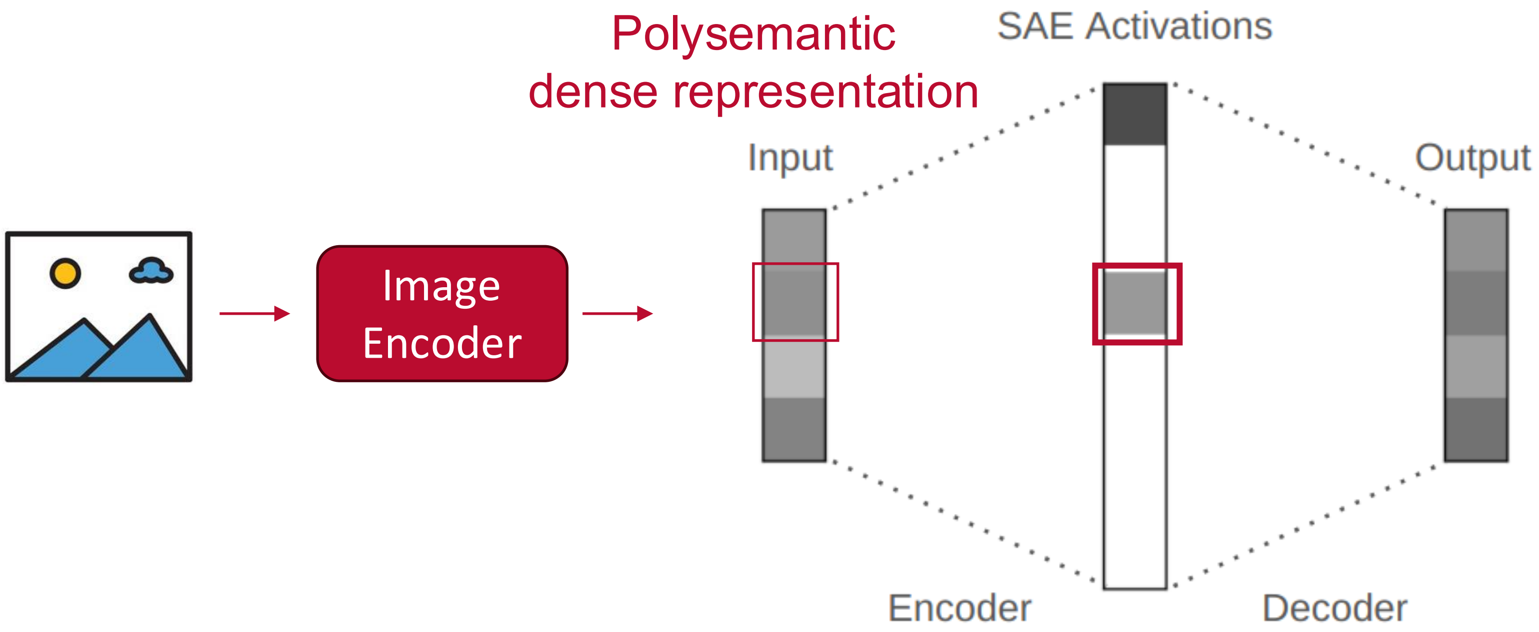
- How can latent structure in dense visual representations be uncovered for fine-grained visual recognition?

Discovering Latent Structure for Interpretable Visual Recognition



Pahuja, Vardaan, et al. "Automatic Image-Level Morphological Trait Annotation for Organismal Images.", ICLR 2026.

Discovering Latent Structure for Interpretable Visual Recognition

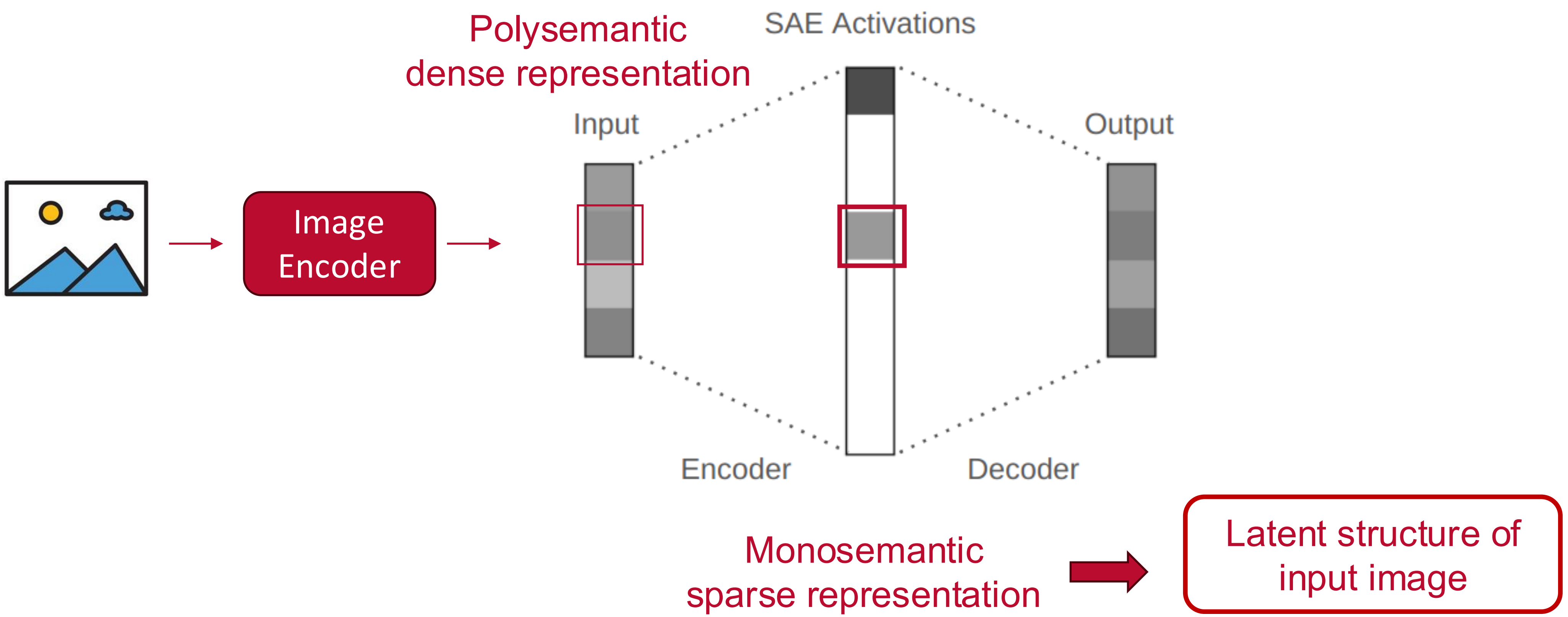


Monosemantic
sparse representation

Fig. credit: Adam Karvonen

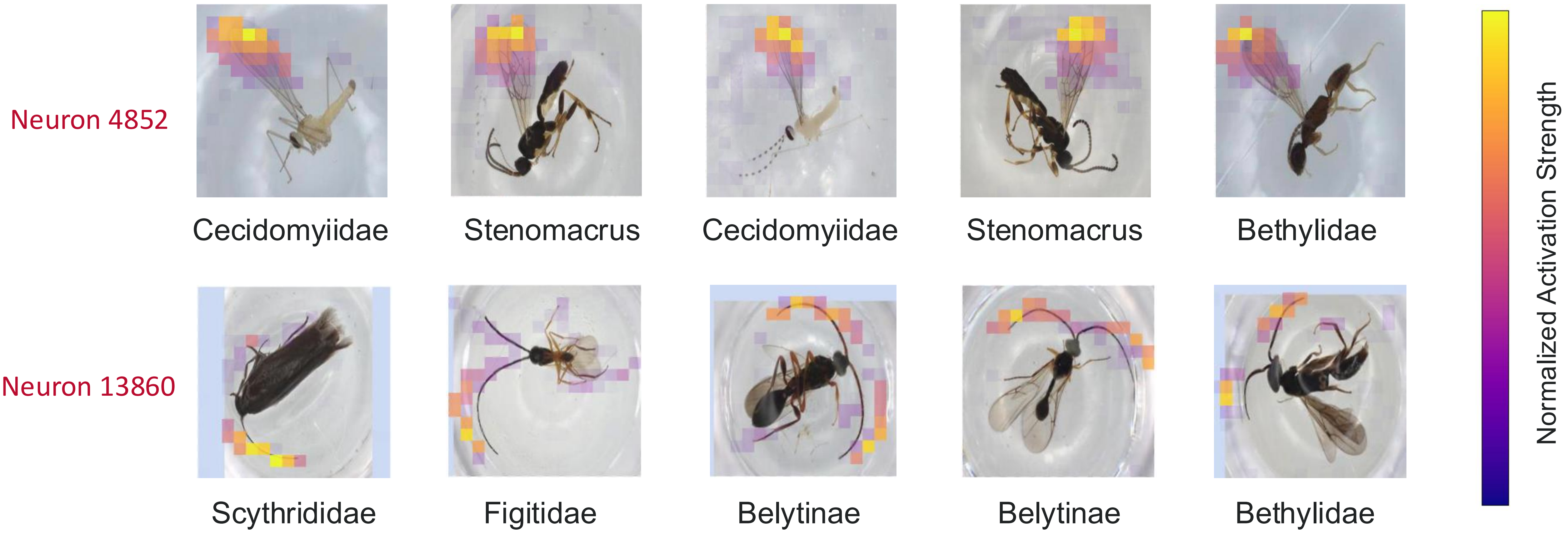
Pahuja, Vardaan, et al. "Automatic Image-Level Morphological Trait Annotation for Organismal Images.", ICLR 2026.

Discovering Latent Structure for Interpretable Visual Recognition



Pahuja, Vardaan, et al. "Automatic Image-Level Morphological Trait Annotation for Organismal Images.", ICLR 2026.

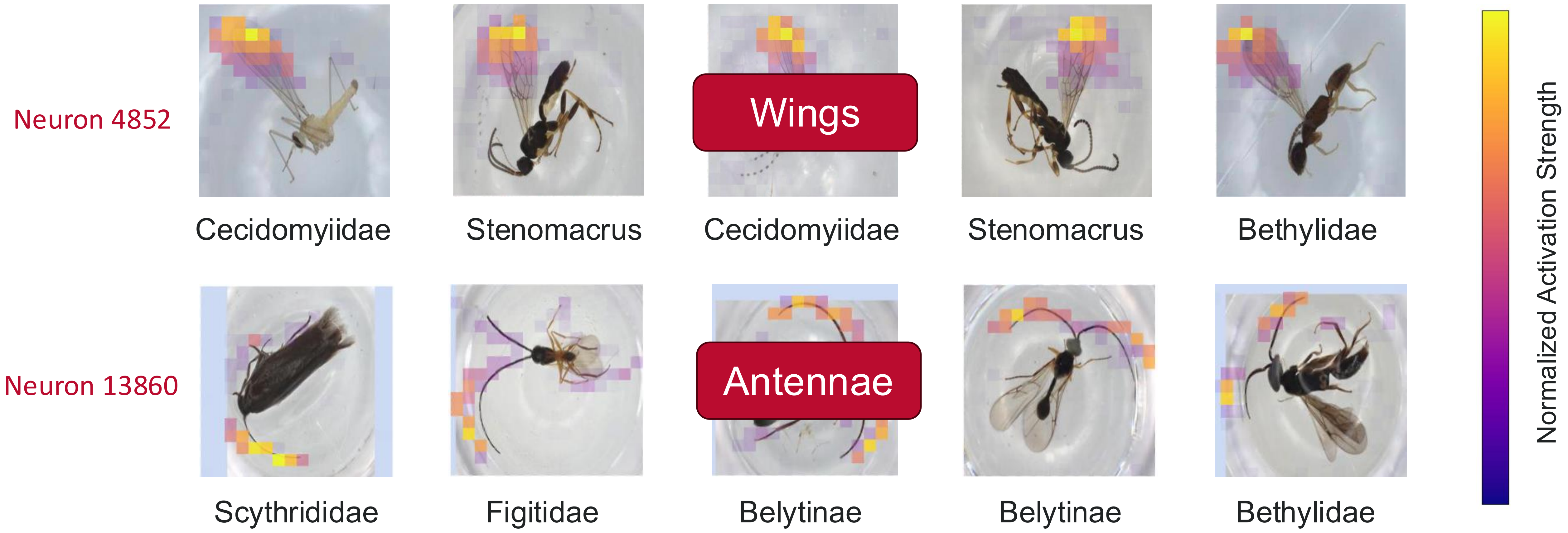
Learned SAE Latents Correspond to Biologically Plausible Traits



THE OHIO STATE UNIVERSITY

Pahuja, Vardaan, et al. "Automatic Image-Level Morphological Trait Annotation for Organismal Images.", ICLR 2026.

Learned SAE Latents Correspond to Biologically Plausible Traits

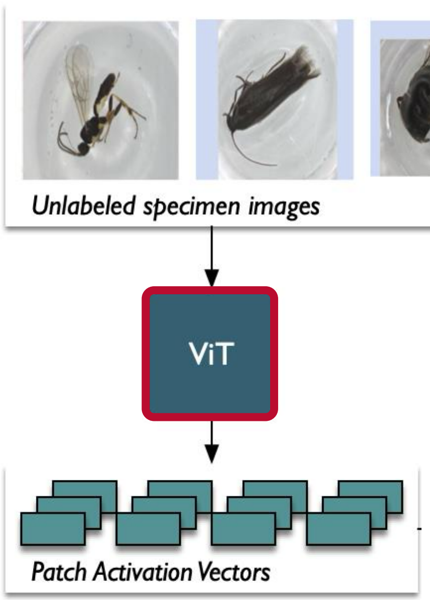


Pahuja, Vardaan, et al. "Automatic Image-Level Morphological Trait Annotation for Organismal Images.", ICLR 2026.

Discovering Latent Structure for Interpretable Visual Recognition

- SAE learns a sparse dictionary of latent units from unlabeled images
- Constraints: sparsity and non-negativity — prevent polysemantic entanglement
- High-activation units localize to spatially coherent, monosemantic regions

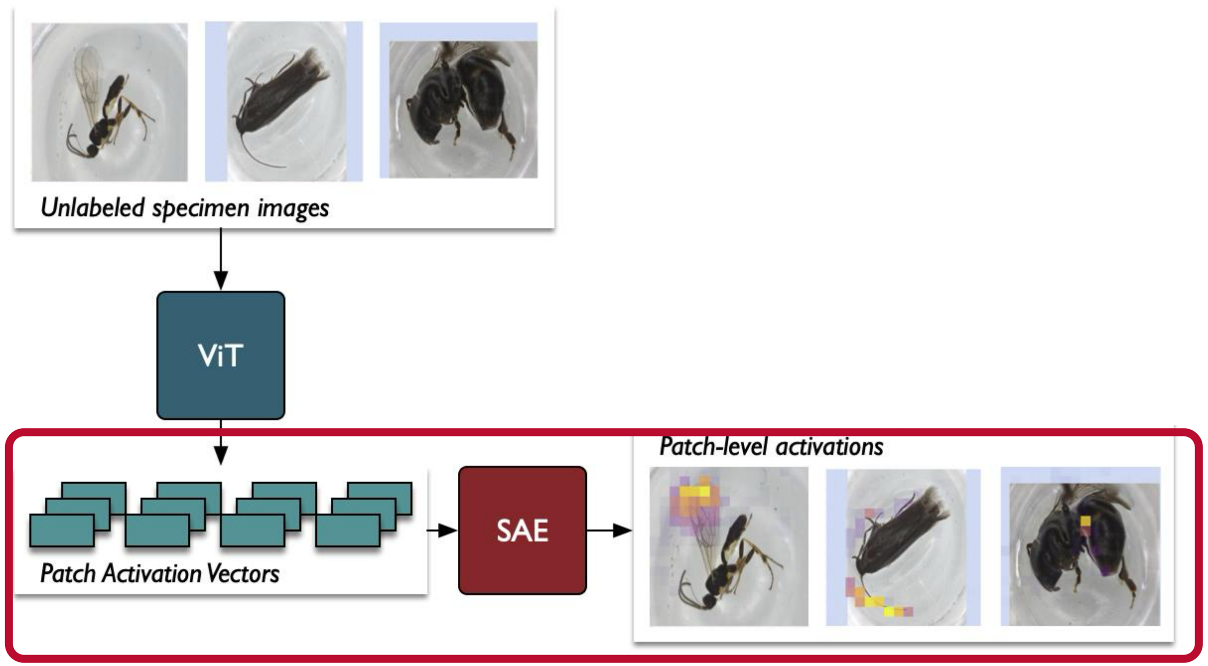
Morphological Trait Annotation Pipeline



Compute dense features using DINOv2

Pahuja, Vardaan, et al. "Automatic Image-Level Morphological Trait Annotation for Organismal Images.", ICLR 2026.

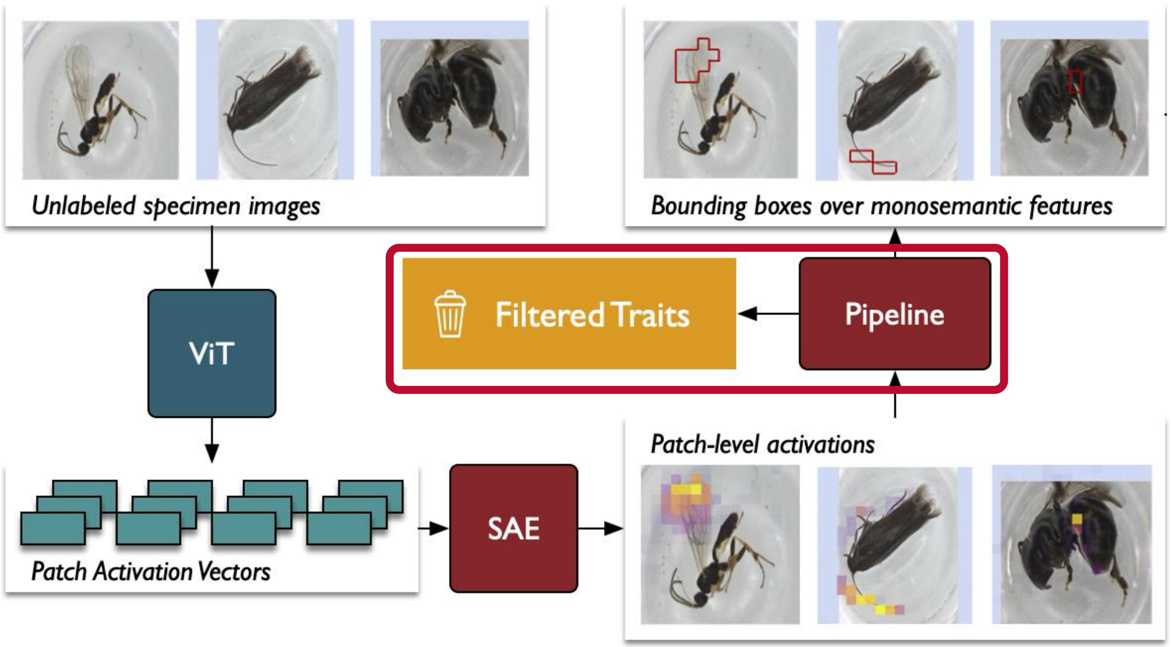
Morphological Trait Annotation Pipeline



SAE identifies
high-activation latent units

Pahuja, Vardaan, et al. "Automatic Image-Level Morphological Trait Annotation for Organismal Images.", ICLR 2026.

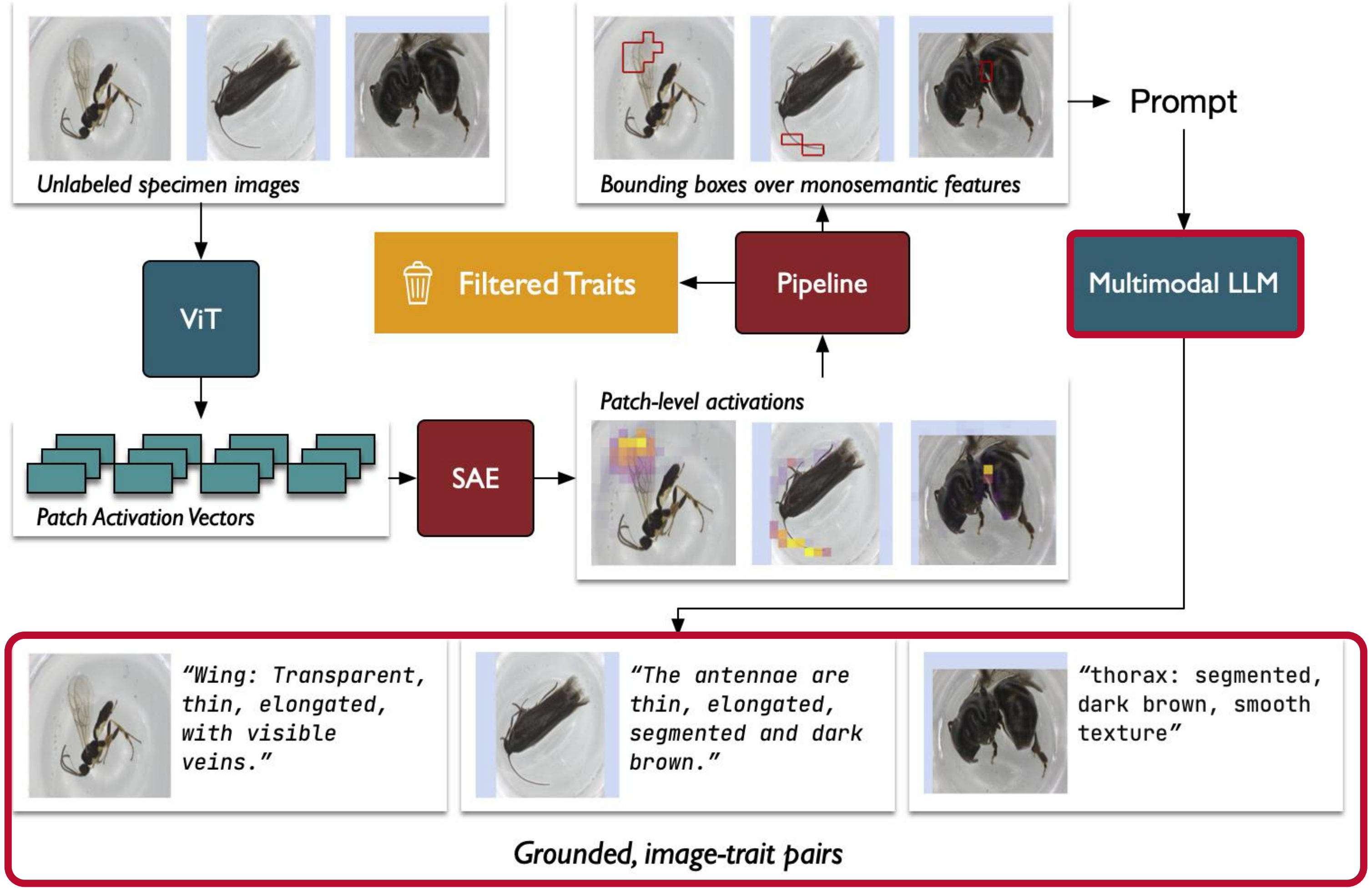
Morphological Trait Annotation Pipeline



Species-contrastive ranking: identify species distinguishing latents

Pahuja, Vardaan, et al. "Automatic Image-Level Morphological Trait Annotation for Organismal Images.", ICLR 2026.

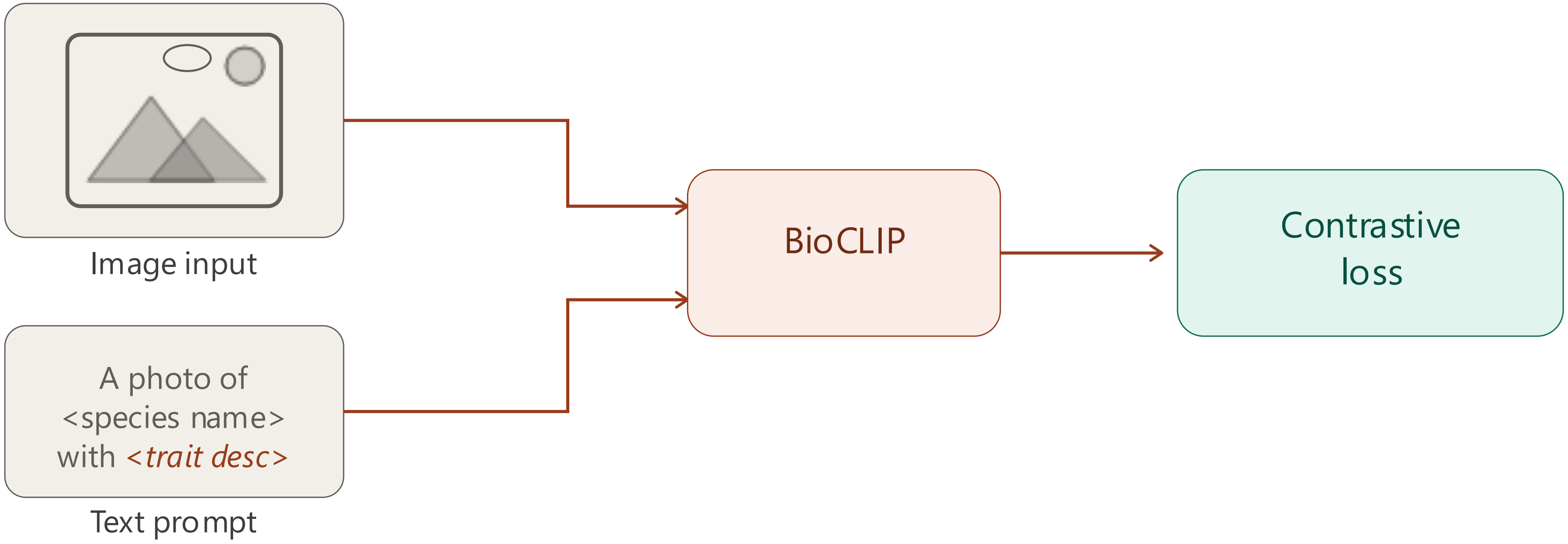
Morphological Trait Annotation Pipeline



Prompt MLLM to generate trait descriptions

Pahuja, Vardaan, et al. "Automatic Image-Level Morphological Trait Annotation for Organismal Images.", ICLR 2026.

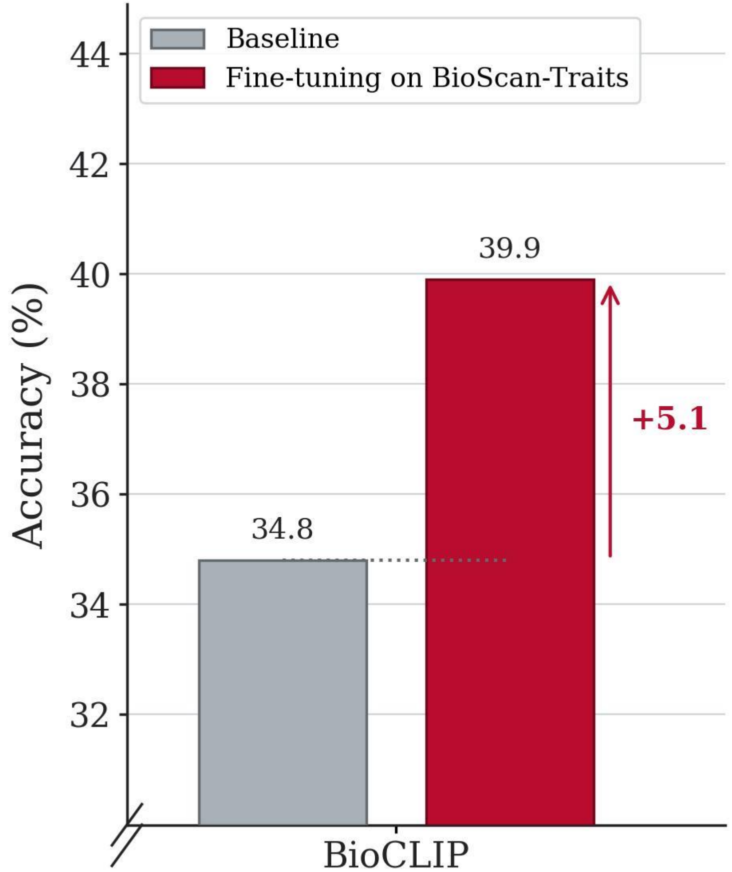
Trait-level supervision improves OOD generalization



Pahuja, Vardaan, et al. "Automatic Image-Level Morphological Trait Annotation for Organismal Images.", ICLR 2026.

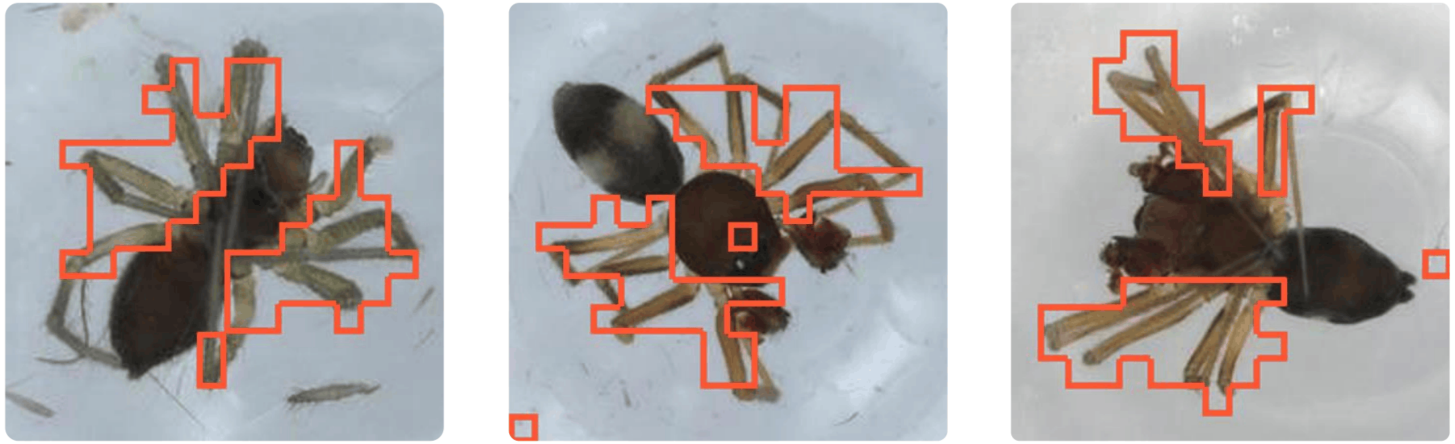
Trait-level supervision improves OOD generalization

- Finetune a biologically grounded foundation model (BioCLIP) on BioScan-Traits.
- Evaluation on Insects, an in-the-wild species classification benchmark.



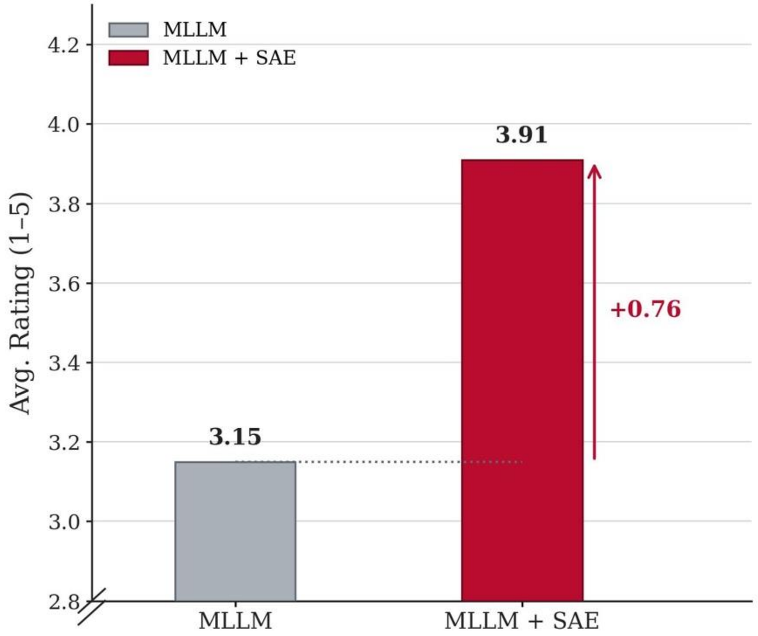
Pahuja, Vardaan, et al. "Automatic Image-Level Morphological Trait Annotation for Organismal Images.", ICLR 2026.

SAE grounding yields precise, focused trait descriptions



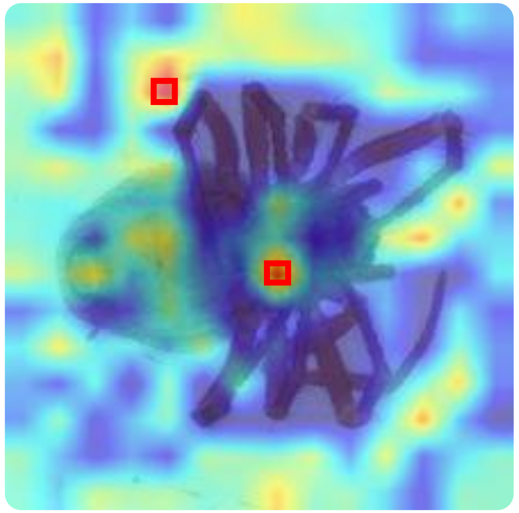
(a) (b) (c)

MLLM: Common Morphological Traits Across All Three Images: 1. Head:- *Shape*: Rounded. - *Color*: Dark brown or black. - *Texture*: Appears smooth with no distinct markings. 2. Thorax: - *Shape*: Oval and slightly elongated. - *Color*: Lighter than the head, ranging from light brown to beige. - *Texture*: Smooth, with no visible segmentation. 3. Abdomen: - *Shape*: Rounded and bulbous. - *Color*: Dark brown or black, similar to the head. - *Texture*: Smooth, with no distinct markings. 4. Legs: - *Number*: Eight legs are visible in all images. - *Shape*: Thin and segmented. - *Color*: Light brown, matching the thorax.
MLLM + SAE: **Leg**: Thin, elongated, light brown, segmented.



SAE grounding yields precise, focused trait descriptions

Grad-CAM



Trait Descriptions

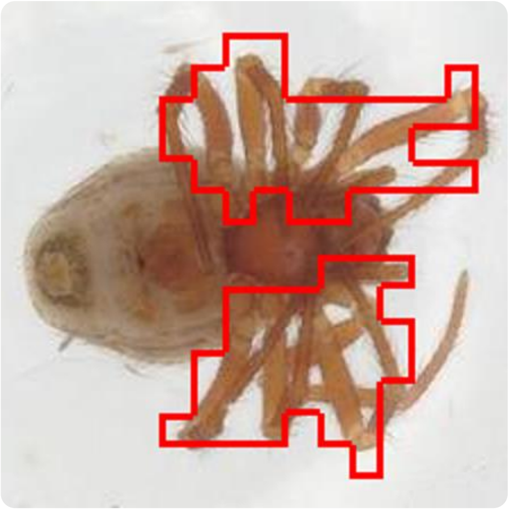
Grad-CAM: Head: Highlights the cephalothorax, emphasizing its rounded shape and the presence of sensory organs. Legs: Focuses on the anterior legs, showcasing their robust structure and segmentation. - Body: Emphasizes the abdomen, noting its segmented appearance and the presence of spinnerets at the posterior end.

❌ Interpretable

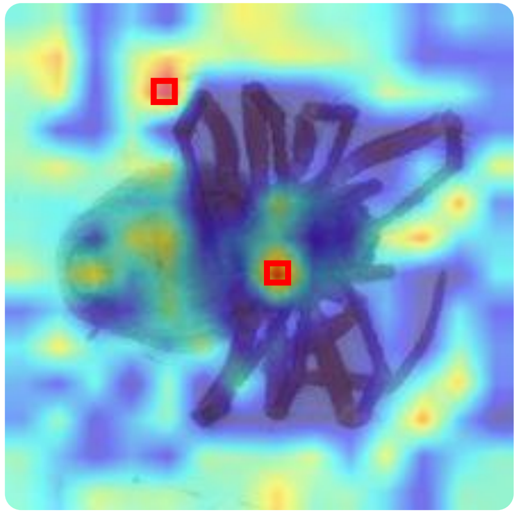
❌ Species-discriminative

SAE grounding yields precise, focused trait descriptions

BioScan-Traits



Grad-CAM



Trait Descriptions

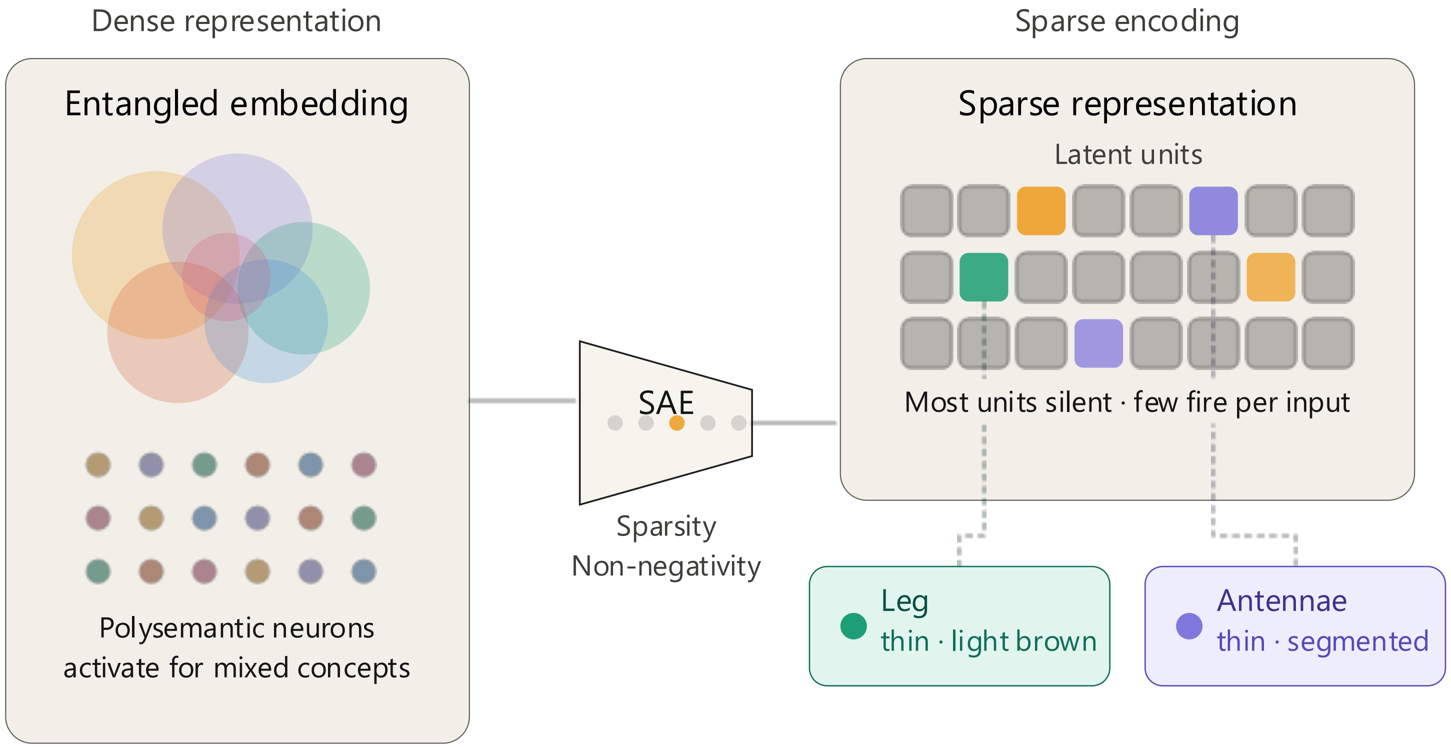
Grad-CAM: Head: Highlights the cephalothorax, emphasizing its rounded shape and the presence of sensory organs. Legs: Focuses on the anterior legs, showcasing their robust structure and segmentation. - Body: Emphasizes the abdomen, noting its segmented appearance and the presence of spinnerets at the posterior end.

Bioscan-Traits: Leg: Thin, elongated, light brown, segmented.

- ✓ Interpretable
- ✓ Species-discriminative

Key Takeaways

- SAEs yield localized, monosemantic features aligned with morphological traits
- Salient latents are verbalized into natural-language trait descriptions
- Trait-level supervision improves species classification generalization



Thesis: The Role of Structure in Modern Multimodal AI Systems

CONTRIBUTION 1

Structure-aware Reasoning

Use explicit knowledge graph structure as a relational scaffold for reasoning and OOD generalization

CONTRIBUTION 2

Latent Structure Discovery

Decompose dense representations into sparse, monosemantic units aligned with interpretable biological traits

CONTRIBUTION 3

Structured Data Generation

Synthesize multimodal interaction trajectories to provide rich structured supervision for GUI agent learning

CONTRIBUTION 4

Structured Information Extraction

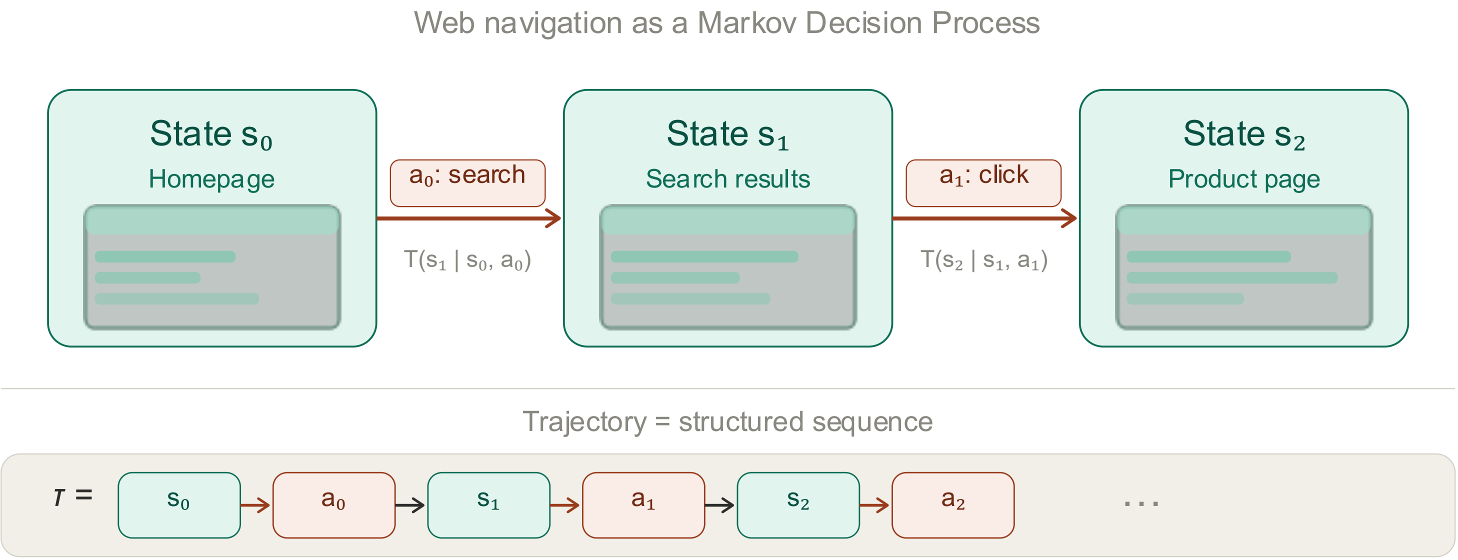
Impose clinical schema onto free text to transform unstructured notes into actionable structured records

Challenge 3: Structured Data Generation

- How can structured multimodal trajectories be synthesized at scale to train GUI agents?

Explorer: Structured Data Generation Using Exploration

- Structure can also be embedded in the data itself
- Incorporate structure in supervision used for learning



Pahuja, Vardaan, et al. "Explorer: Scaling Exploration-driven Web Trajectory Synthesis for Multimodal Web Agents.", ACL (Findings) 2025

Explorer: Structured Data Generation Using Exploration

- Leverage exploration as a key mechanism for achieving diversity in task intents.
- Explorer dynamically explores web environments to curate diverse, real-world tasks.

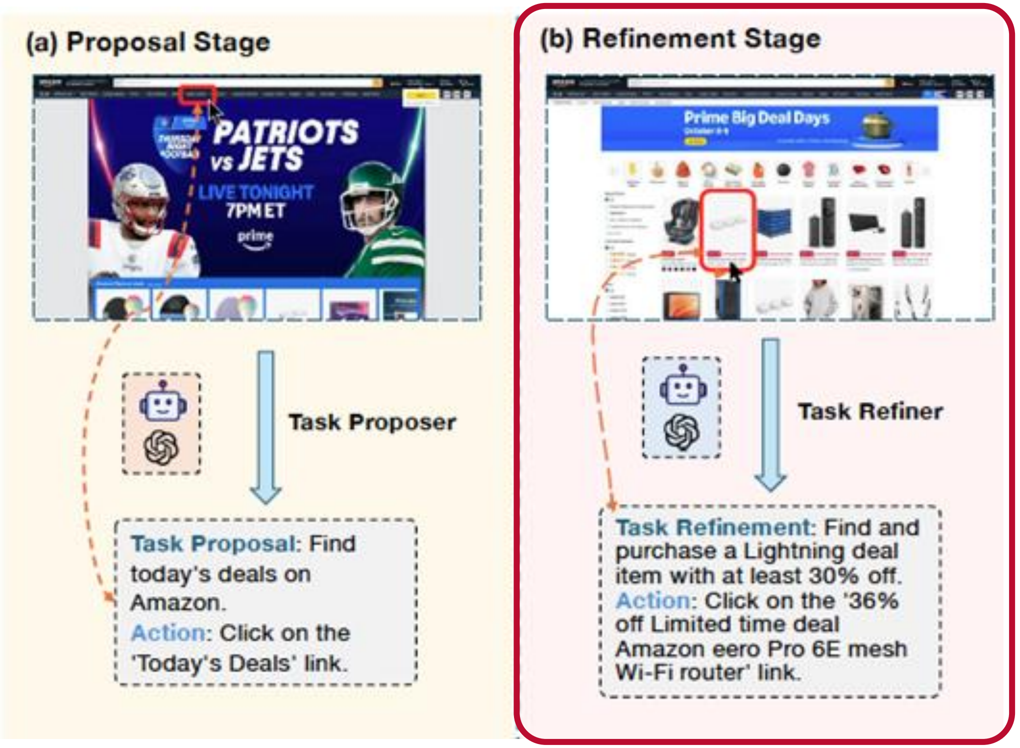
Pahuja, Vardaan, et al. "Explorer: Scaling Exploration-driven Web Trajectory Synthesis for Multimodal Web Agents.", ACL (Findings) 2025

Explorer: Structured Data Generation Using Exploration

- Propose a high level and abstract version of a real-world task

Pahuja, Vardaan, et al. "Explorer: Scaling Exploration-driven Web Trajectory Synthesis for Multimodal Web Agents.", ACL (Findings) 2025

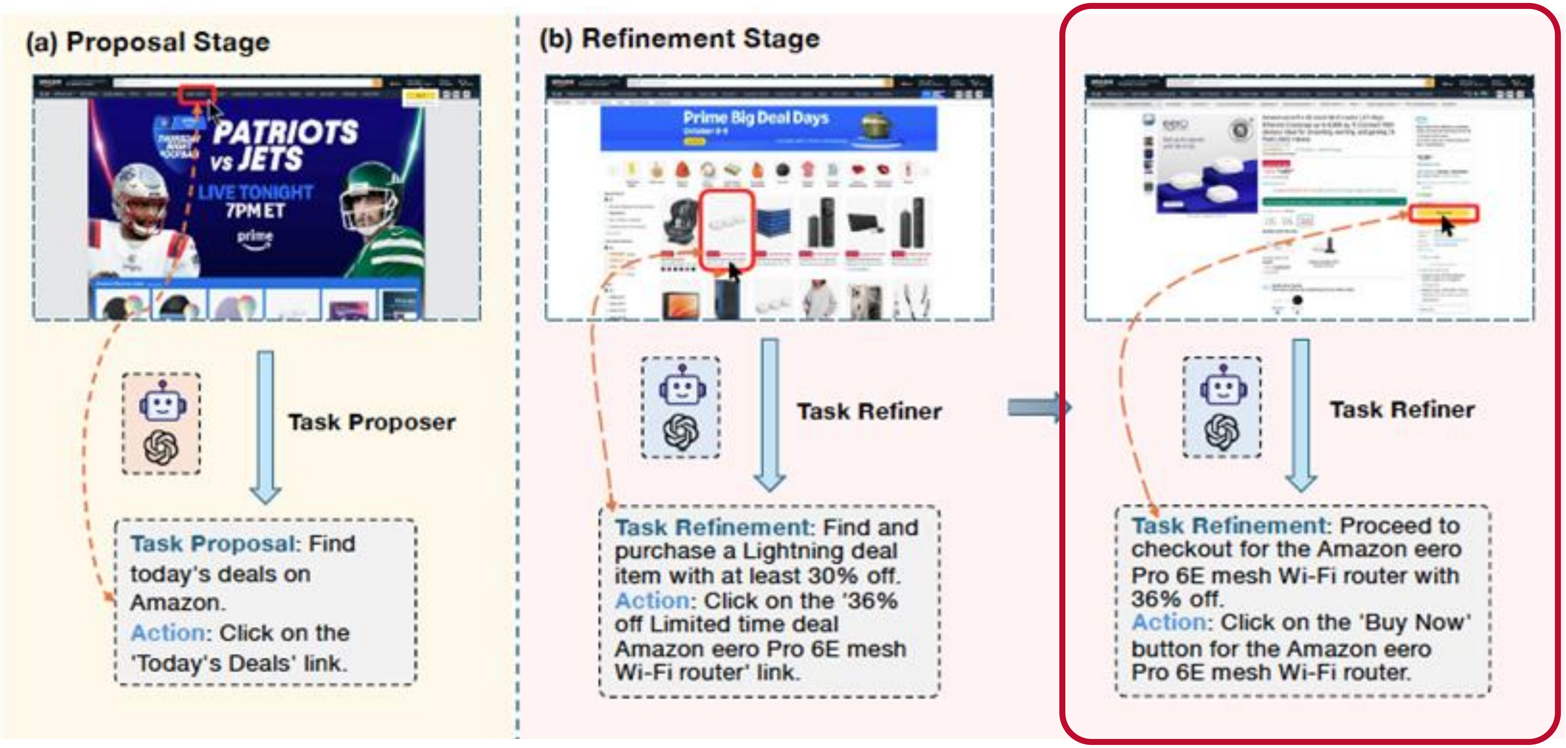
Explorer: Structured Data Generation Using Exploration



- Iteratively refine the task description via exploration

Pahuja, Vardaan, et al. "Explorer: Scaling Exploration-driven Web Trajectory Synthesis for Multimodal Web Agents.", ACL (Findings) 2025

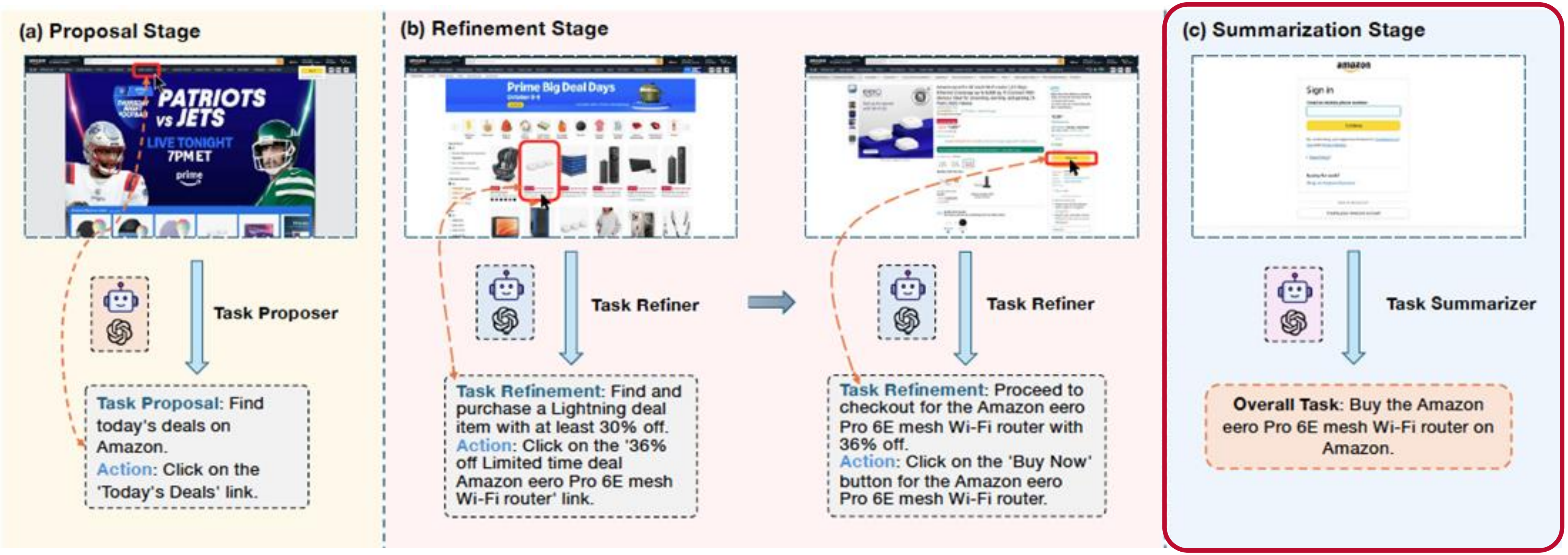
Explorer: Structured Data Generation Using Exploration



- Iteratively refine the task description via exploration
- Ensure that the updated task remains aligned with the action history

Pahuja, Vardaan, et al. "Explorer: Scaling Exploration-driven Web Trajectory Synthesis for Multimodal Web Agents.", ACL (Findings) 2025

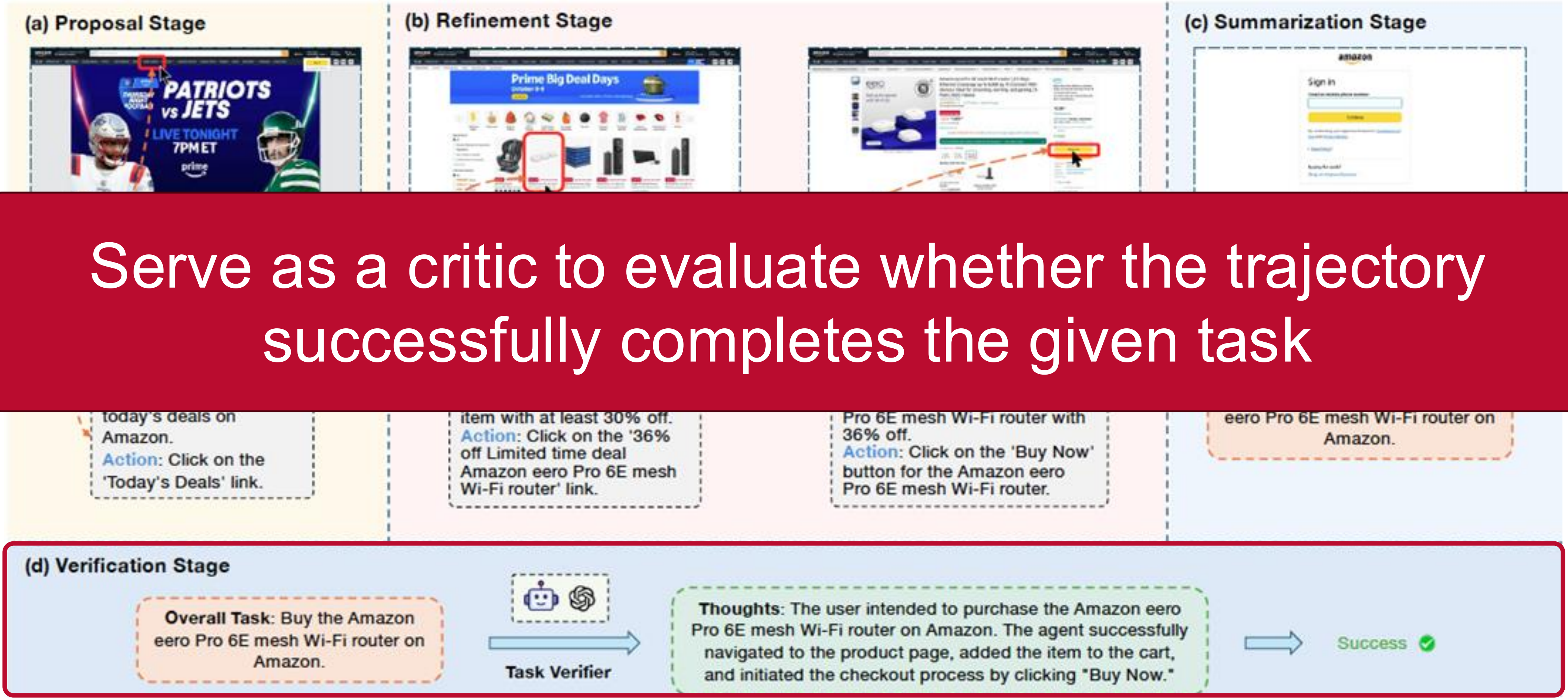
Explorer: Structured Data Generation Using Exploration



- Summarize the sequence of actions into a high-level task description

Pahuja, Vardaan, et al. "Explorer: Scaling Exploration-driven Web Trajectory Synthesis for Multimodal Web Agents.", ACL (Findings) 2025

Explorer: Structured Data Generation Using Exploration

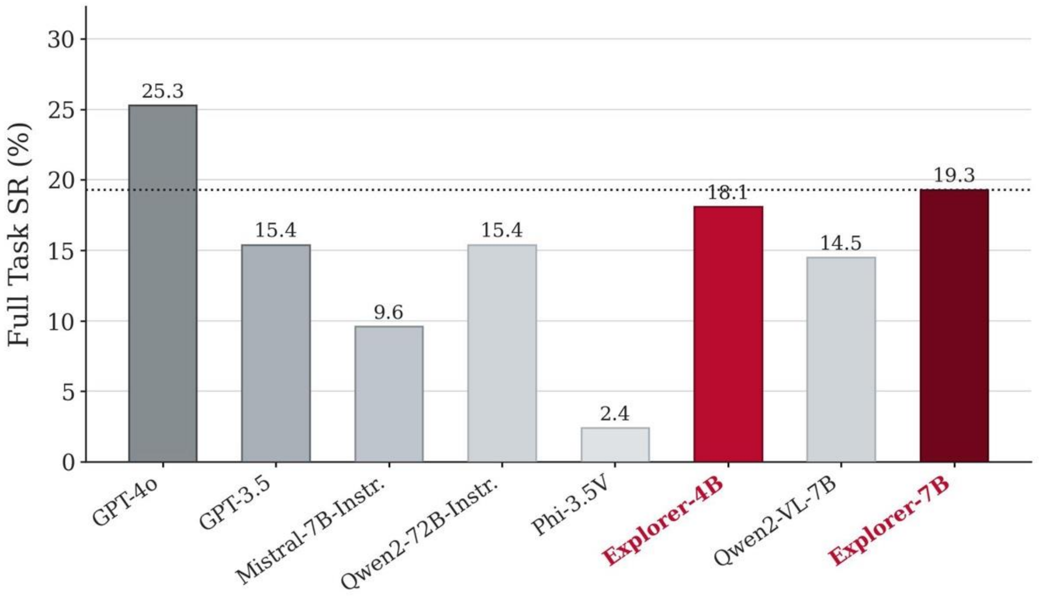


Serve as a critic to evaluate whether the trajectory successfully completes the given task

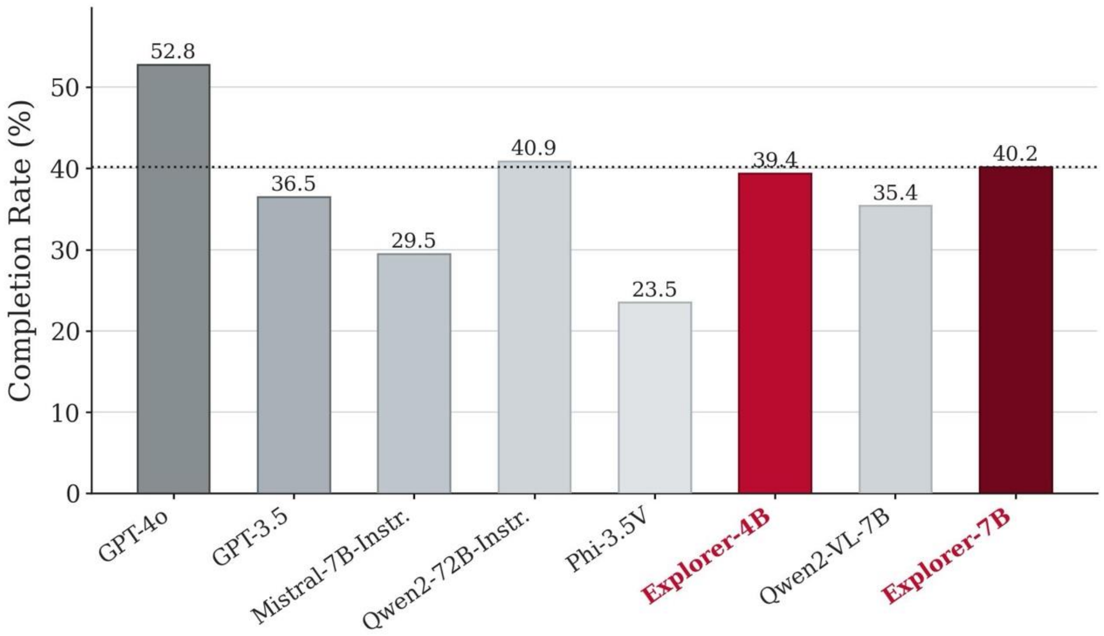
Pahuja, Vardaan, et al. "Explorer: Scaling Exploration-driven Web Trajectory Synthesis for Multimodal Web Agents.", ACL (Findings) 2025

Structured Data Training Yields Strong Web Agents

- Improvement over higher capacity LLMs
- Best M2W-Live performance for model of this size



Mind2Web-Live, Full Task Success Rate



Mind2Web-Live, Completion Rate

THE OHIO STATE UNIVERSITY

Pahuja, Vardaan, et al. "Explorer: Scaling Exploration-driven Web Trajectory Synthesis for Multimodal Web Agents.", ACL (Findings) 2025

Thesis: The Role of Structure in Modern Multimodal AI Systems

CONTRIBUTION 1

Structure-aware Reasoning

Use explicit knowledge graph structure as a relational scaffold for reasoning and OOD generalization

CONTRIBUTION 2

Latent Structure Discovery

Decompose dense representations into sparse, monosemantic units aligned with interpretable biological traits

CONTRIBUTION 3

Structured Data Generation

Synthesize multimodal interaction trajectories to provide rich structured supervision for GUI agent learning

CONTRIBUTION 4

Structured Information Extraction

Impose clinical schema onto free text to transform unstructured notes into actionable structured records

Challenge 4: Structured Information Extraction

- How can LLMs transform unstructured clinical notes into structured, schema-aligned health information?

Structured Information Extraction

Unstructured clinical note

Pt is a 47yo F presenting for DM2 follow-up. She was laid off from her part-time job 2 months ago and reports difficulty paying rent and utility bills. She currently has no reliable transportation to clinic and missed her last two appointments. She has a 9-year-old daughter at home and separated from her husband 6 months ago. Patient states she has no one to call on for help and feels increasingly isolated.

Pahuja, Vardaan, et al. "Using Large Language Models for Social Determinants of Health Phenotyping: A Systematic Study.", In Review.

Structured Information Extraction

Unstructured clinical note

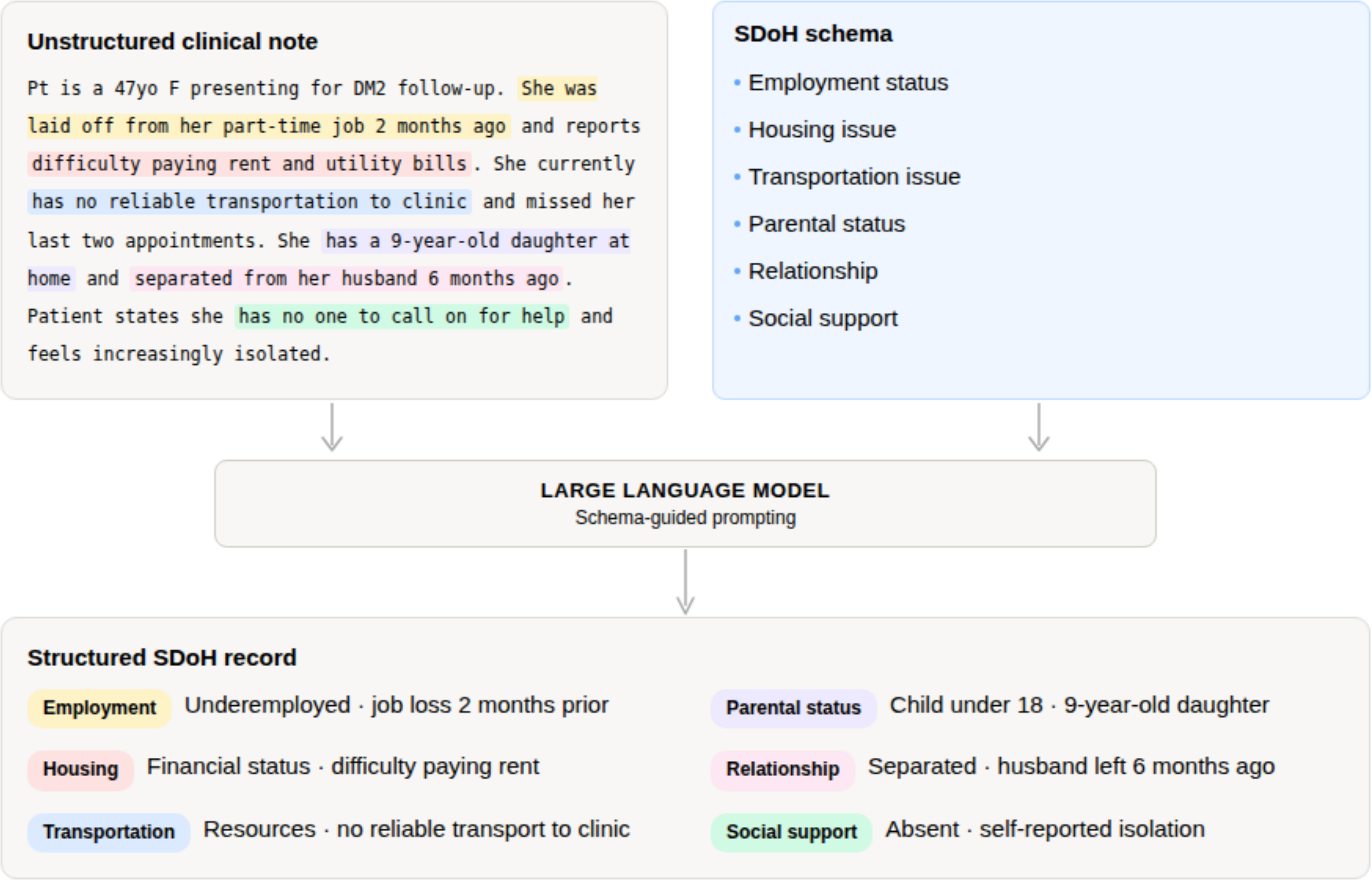
Pt is a 47yo F presenting for DM2 follow-up. She was laid off from her part-time job 2 months ago and reports difficulty paying rent and utility bills. She currently has no reliable transportation to clinic and missed her last two appointments. She has a 9-year-old daughter at home and separated from her husband 6 months ago. Patient states she has no one to call for help and feels increasingly isolated.

SDoH schema

- Employment status
- Housing issue
- Transportation issue
- Parental status
- Relationship
- Social support

Pahuja, Vardaan, et al. "Using Large Language Models for Social Determinants of Health Phenotyping: A Systematic Study.", In Review.

Structured Information Extraction



Pahuja, Vardaan, et al. "Using Large Language Models for Social Determinants of Health Phenotyping: A Systematic Study.", In Review.

Structured Information Extraction

Input data

Synthetic clinical notes

Real patient data (MIMIC-III)

SDoH categories

- Employment
- Housing
- Transportation
- Parental status
- Relationship
- Social support

→

Modeling

LLM usage paradigm

- Zero-shot prompting
- Zero-shot multi-label prompting
- Few-shot prompting
- Supervised fine-tuning
- Retrieval-augmented generation

Baseline models

- Mixtral-8x7B
- LLaMA3.3-70B
- QwQ-32B
- Phi-4-14B
- Qwen3-8B
- Qwen3-14B
- GPT-4o
- GPT-5.2
- Claude-Sonnet-4.5

→

Evaluation

Benchmarks

- Synthetic test set
- MIMIC-III

Performance metrics

- Precision
- Recall
- Macro F1 score
- Micro F1 score

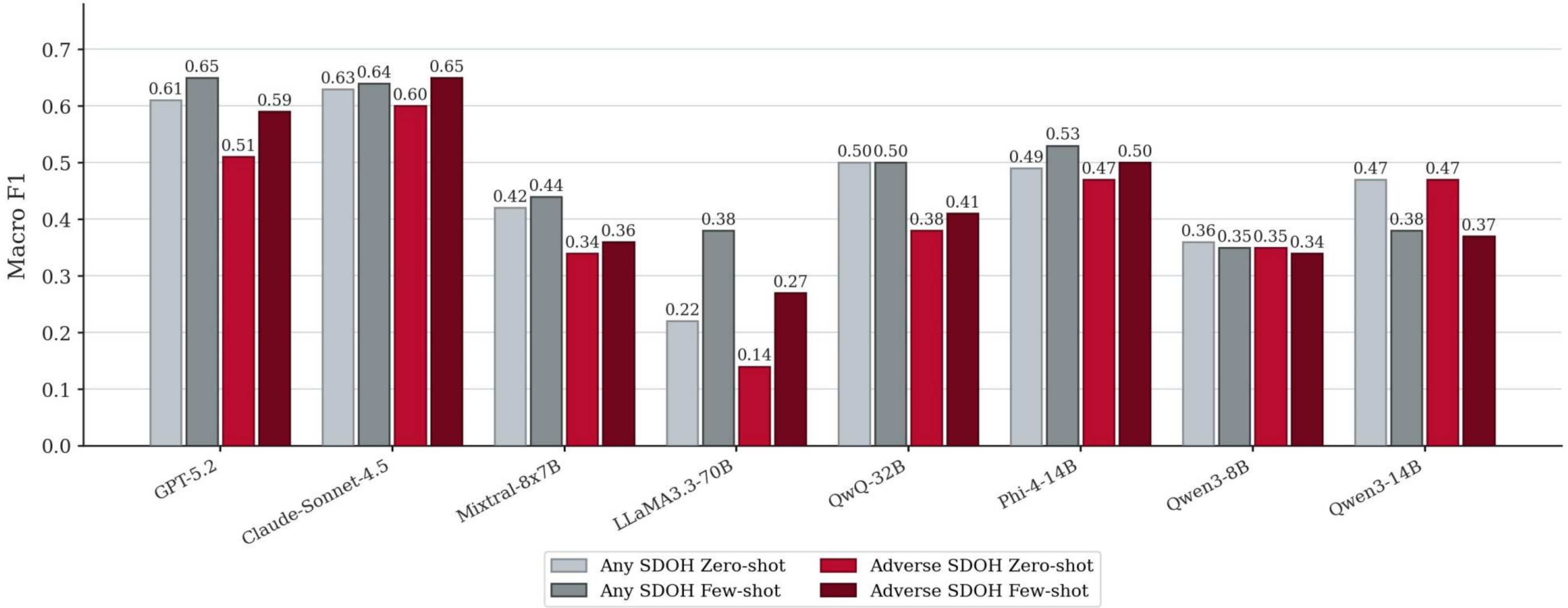
Pahuja, Vardaan, et al. "Using Large Language Models for Social Determinants of Health Phenotyping: A Systematic Study.", In Review.

Terminology

Any SDoH	Includes all social determinants, both positive and negative.	
	A patient with a stable job and good income who lives in a safe neighborhood with access to quality healthcare.	
	Positive Labels: high income, job security, safe housing	
Adverse SDoH	Negative or harmful social determinants that harm health outcomes.	
	An individual experiencing unemployment, living in substandard housing, with limited access to healthy food in a high-crime area.	
	Negative Labels: unemployed, unsafe housing, etc.	

Few-shot Classification (MIMIC-III)

- Few-shot learning improves over zero-shot for most models except the Qwen family.
- Frontier models (GPT-5.2, Claude-Sonnet-4.5) report clearer gains

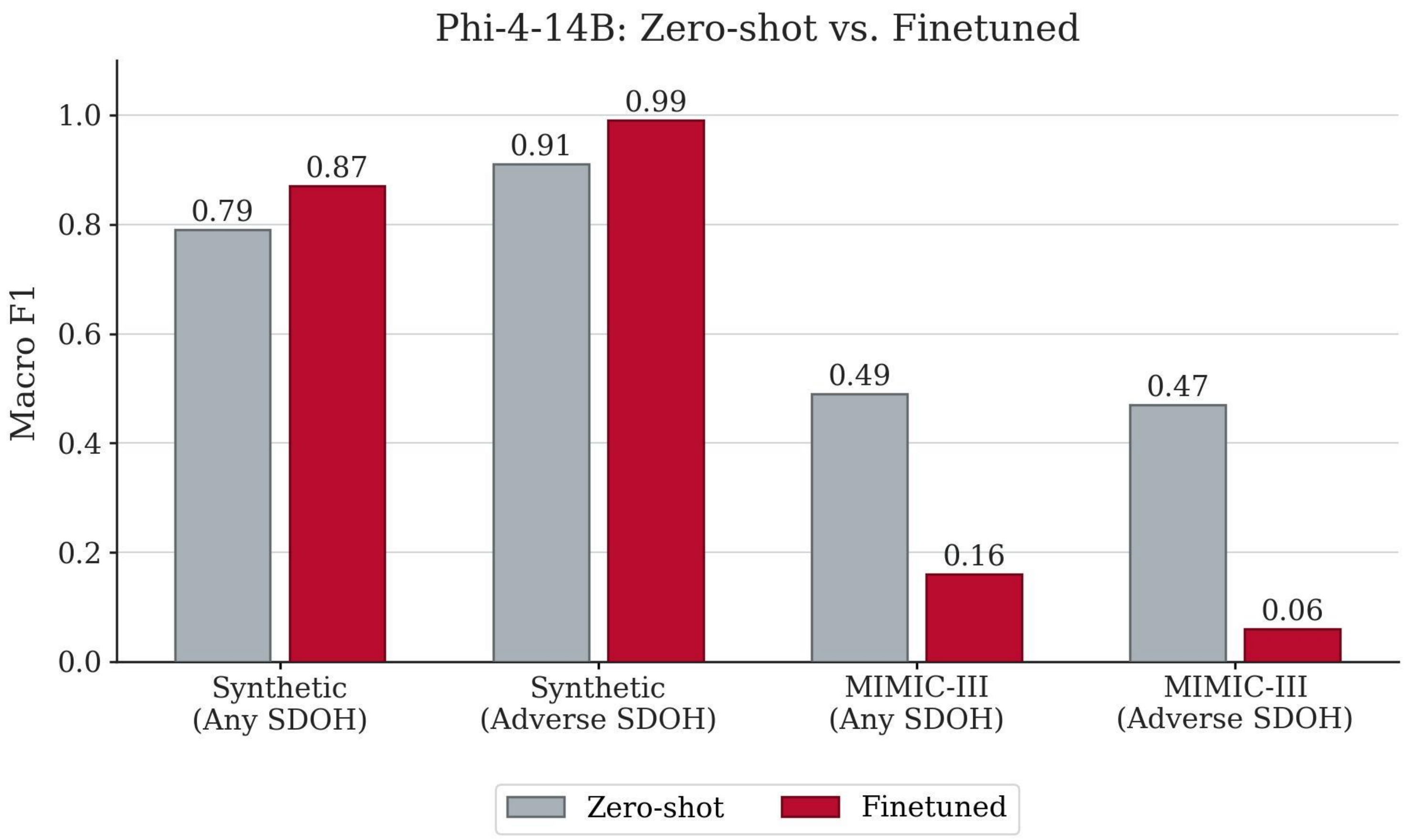


THE OHIO STATE UNIVERSITY

Pahuja, Vardaan, et al. "Using Large Language Models for Social Determinants of Health Phenotyping: A Systematic Study.", In Review.

Fine-tuning Setting

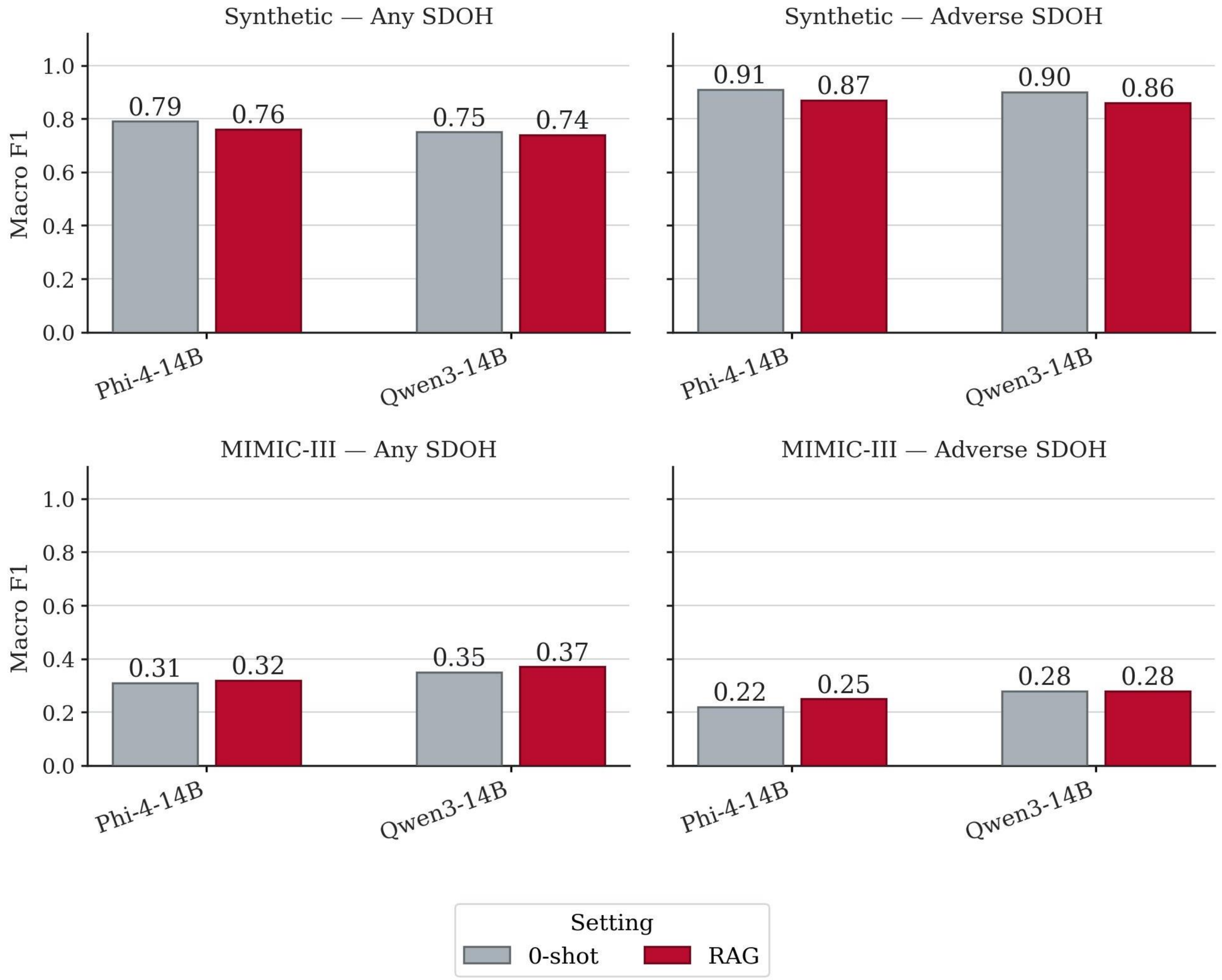
- Fine-tuning improves performance on synthetic data but fails to generalize to MIMIC-III.
- Domain mismatch and higher noise in real clinical notes drive the sharp degradation.



Pahuja, Vardaan, et al. "Using Large Language Models for Social Determinants of Health Phenotyping: A Systematic Study.", In Review.

Retrieval-augmented Generation

- RAG marginally improves on real clinical notes but not synthetic data
- Synthetic inputs carry sufficient signal - retrieval injects irrelevant context



Pahuja, Vardaan, et al. "Using Large Language Models for Social Determinants of Health Phenotyping: A Systematic Study.", In Review.

Key Takeaways

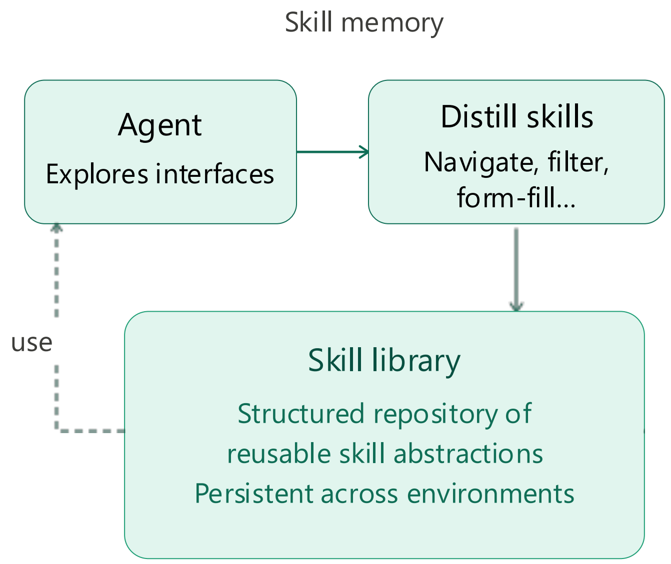
- Open-weight LLMs of moderate size match proprietary models on both benchmarks — a viable, privacy-preserving alternative for clinical NLP
- Task formulation matters: multi-label prompting and few-shot exemplars are the most robust strategies for real clinical notes
- Practical guidance for deploying LLMs for SDoH phenotyping under varying resource and privacy constraints

Pahuja, Vardaan, et al. "Using Large Language Models for Social Determinants of Health Phenotyping: A Systematic Study.", In Review.

Future Work

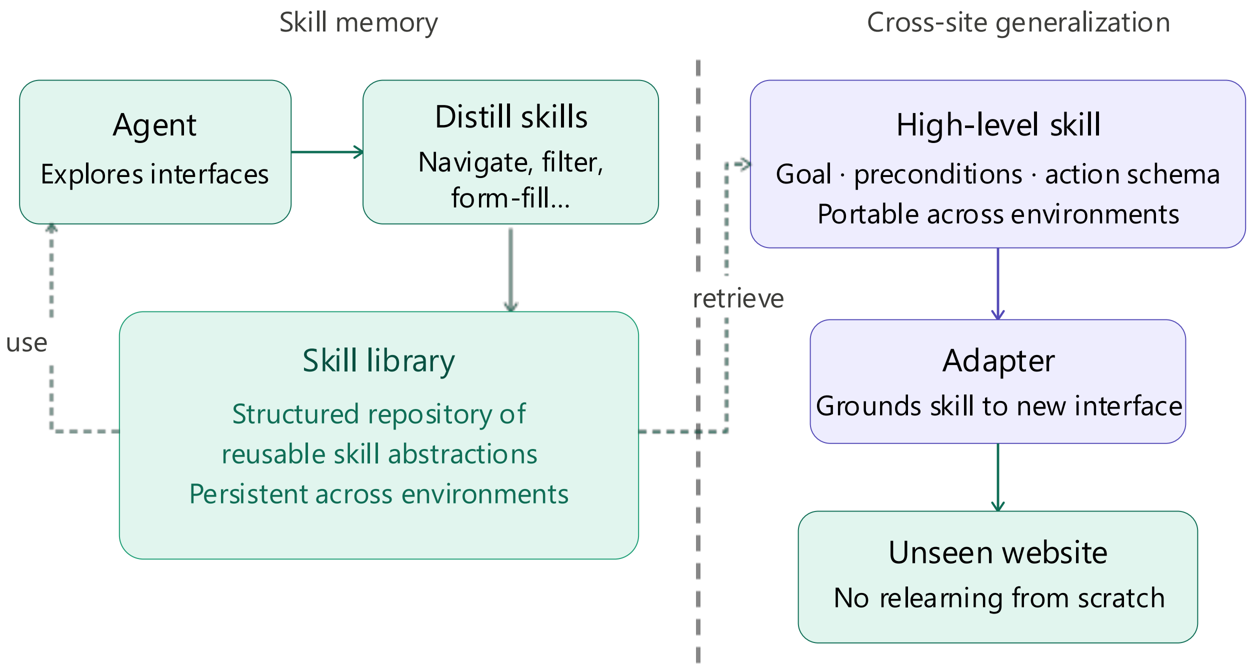
Future Prospects: Continual Learning

- Self-improving agents
- Exploration drives discovery of reusable skills across interfaces
- Skills stored abstractly in a library — portable across environments



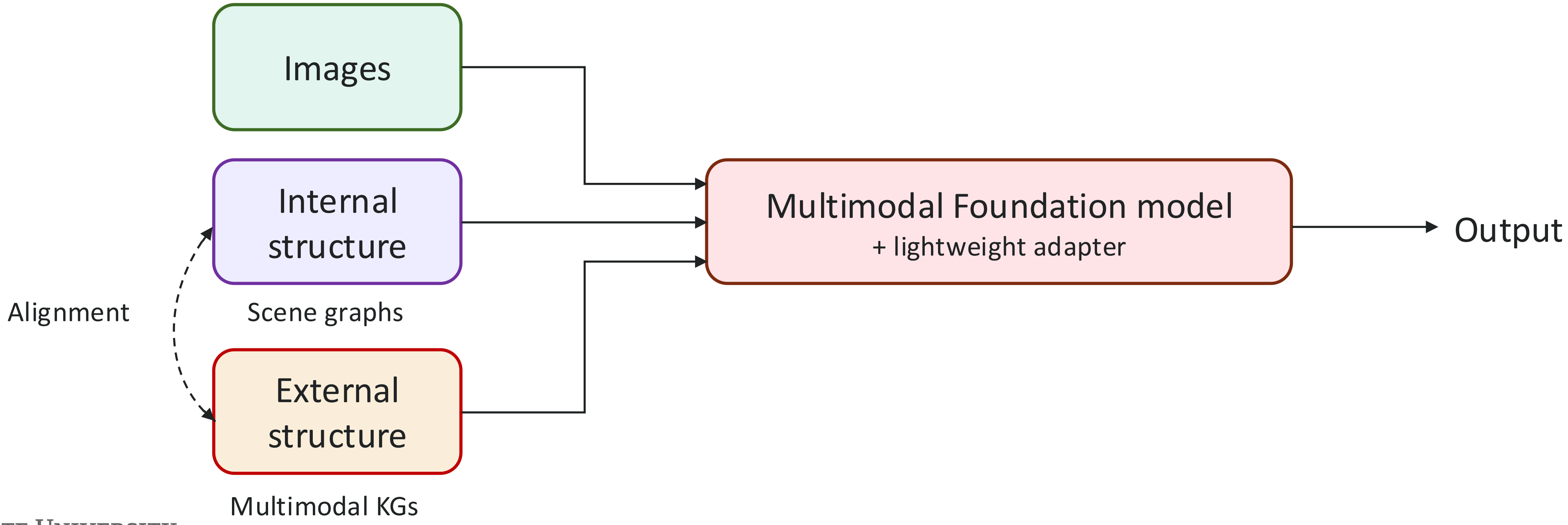
Future Prospects: Continual Learning

- Self-improving agents
- Exploration drives discovery of reusable skills across interfaces
- Skills stored abstractly in a library — portable across environments
- A lightweight adapter instantiates any skill on a new interface without relearning



Future Prospects: Structure-Aware Multimodal LMs

- Integrate perception, language, and structured knowledge within a unified framework
- Bridge perceptual structure with relational world knowledge
- Exploit structure priors while preserving general capabilities



Conclusion: Structure Remains Relevant in the LLM Era

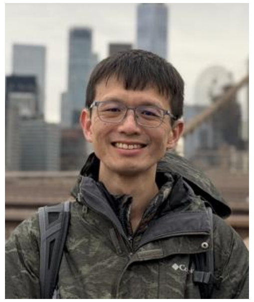
- Structure as supervision: scene graphs provide grounded SFT data, aligning CoT rationales with visual entities and relations [1]
- Structured search via Monte Carlo Tree Search enables LLM agents to plan deliberatively
- Exploration-driven trajectory synthesis has emerged as a scalable paradigm for supervised fine-tuning of GUI agents
- GraphRAG traverses relational structure to outperform flat retrieval on multi-entity and causal queries
- LLMs make knowledge bases easier to build — lowering the cost of structured knowledge acquisition

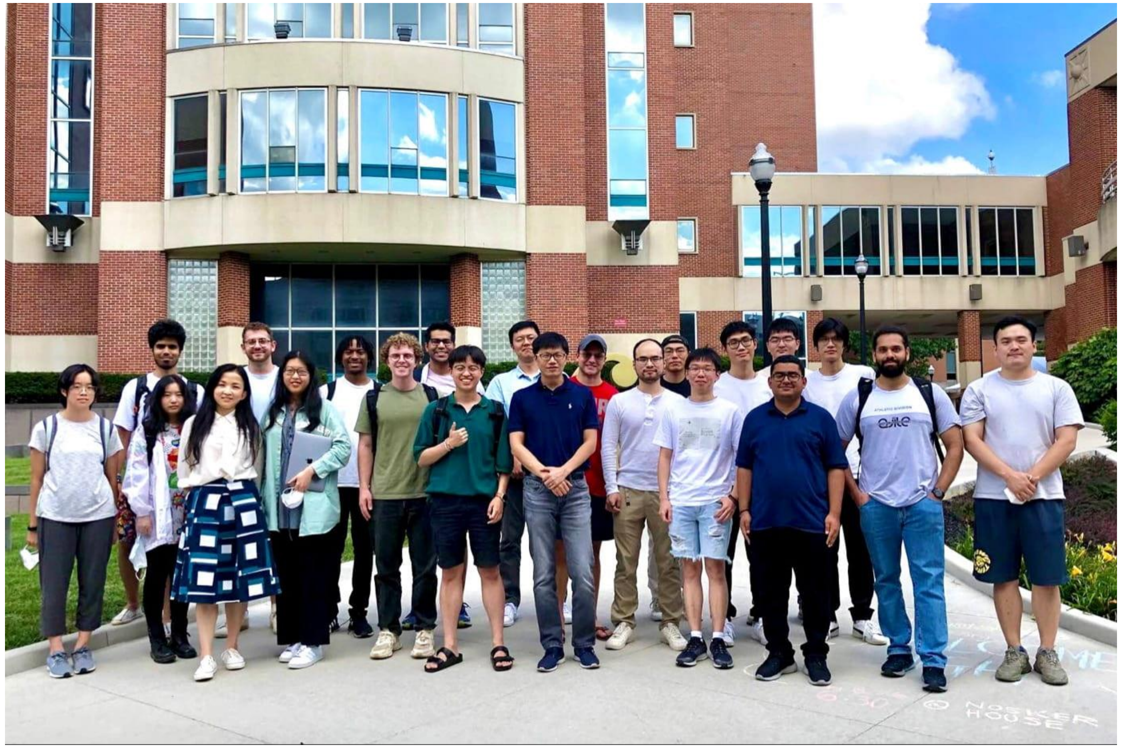
[1] SceneAlign: Aligning Multimodal Reasoning to Scene Graphs in Complex Visual Scenes, ACL 2026

Acknowledgements












Thank You!

Vardaan Pahuja
pahuja.9@osu.edu