AdaGlimpse

Active Visual Exploration with Arbitrary Glimpse Position and Scale

Adam Pardyl, ML in PL Conference 2024

ML in PL CONFERENCE 2024

group of machine gmum learning research

JAGIELLONIAN UNIVERSITY IN KRAKÓW

Warsaw University of Technology

IDEAS NCBR

NARODOWE CENTRUM NAUKI


Is AI already better than humans?

Board games?

✓ **better than the best human players**

(AlphaGo Zero, 2017)

Photorealistic art?

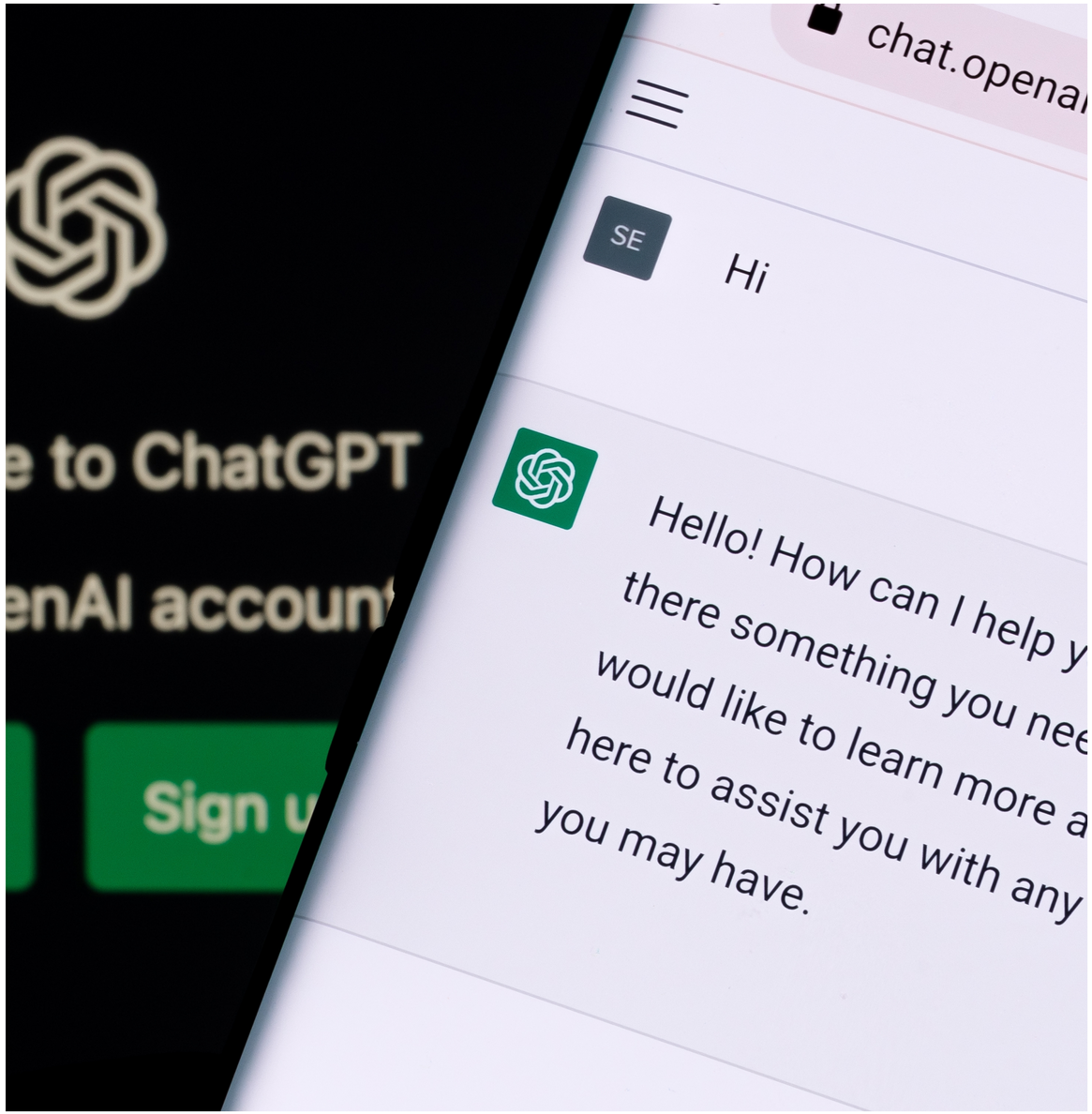
✓ better than most humans

(Stable Diffusion, 2022)

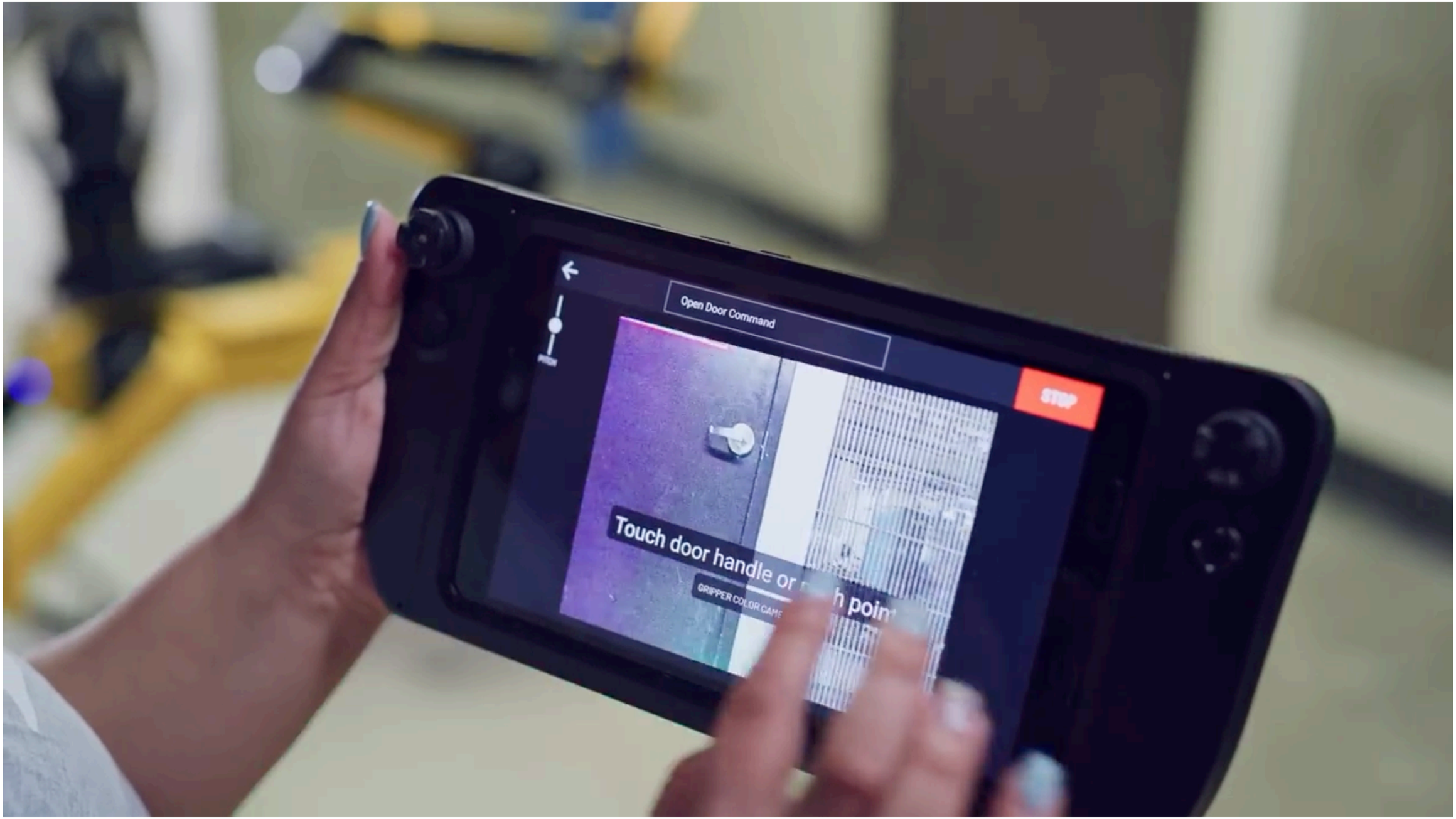
Essay writing?

✓ **better than most humans**

(GPT4, 2023)



Environment interaction?

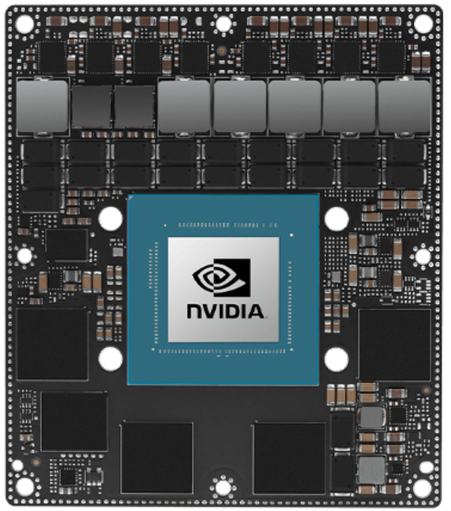

Open Door Command
STOP
Touch door handle or

Environment interaction?

…with a bit of luck and some help

(Boston Dynamics, 2024)

- No human input needed
- Unsupervised learning
- No GPU required

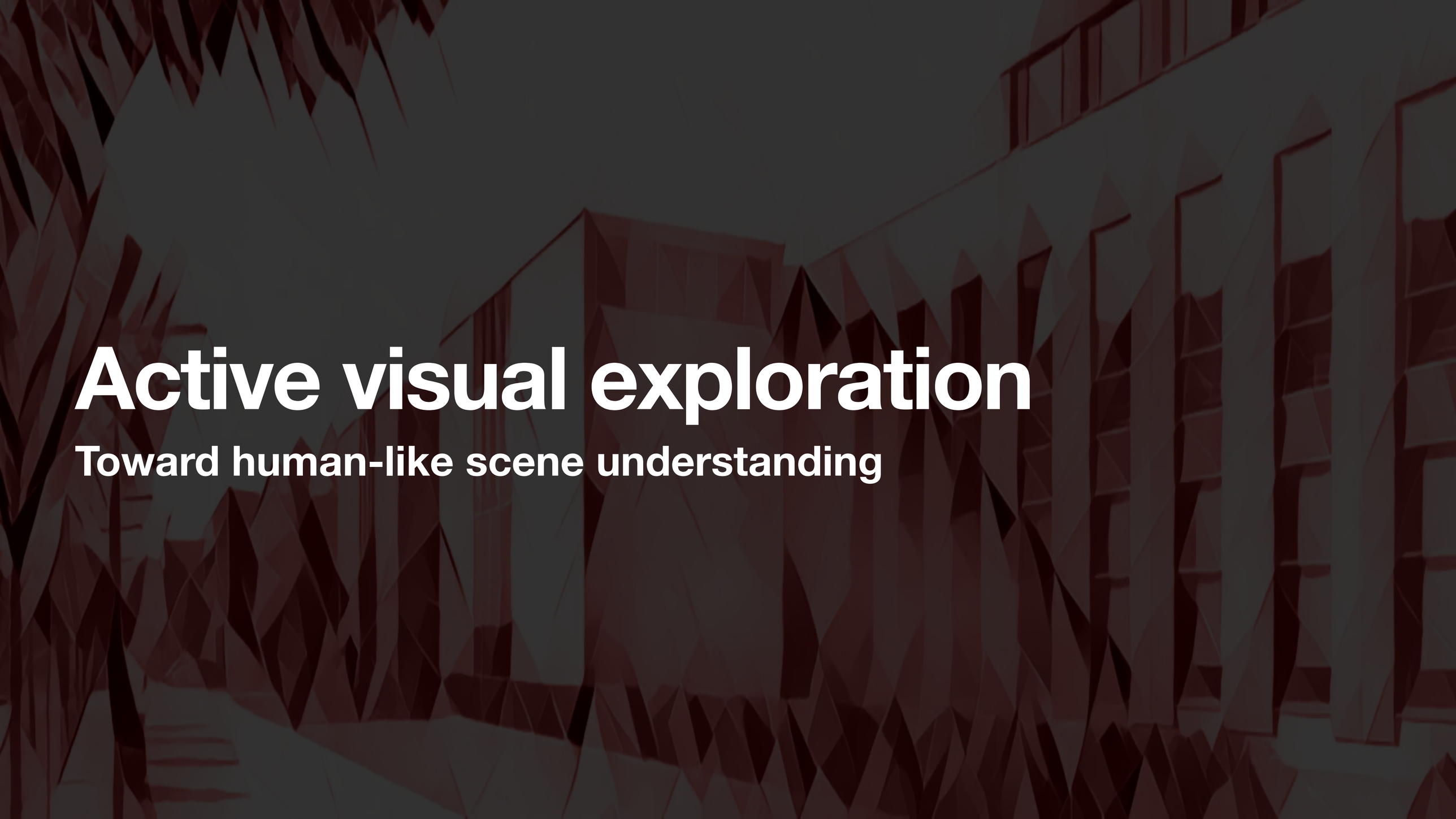



Current AI/ML solutions usually
fail in open world environments

Issues

Embodied AI

- Open vocabulary object detection
- World representation
- Spatial reasoning
- Action planning
- Sim-to-real gap
- …and many more

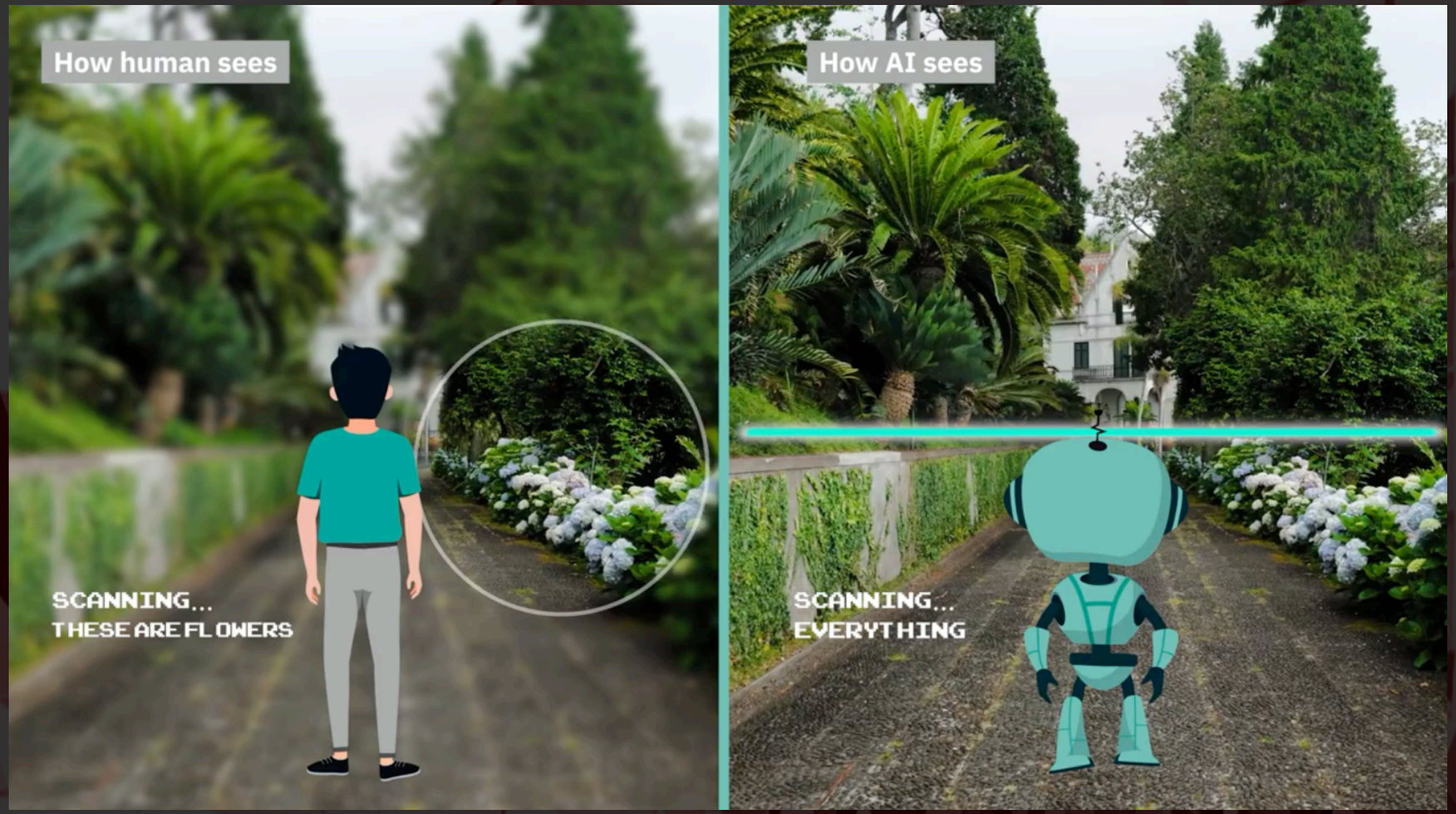


Active visual exploration

Toward human-like scene understanding

Visual exploration:

Human vs. AI



Active Visual Exploration

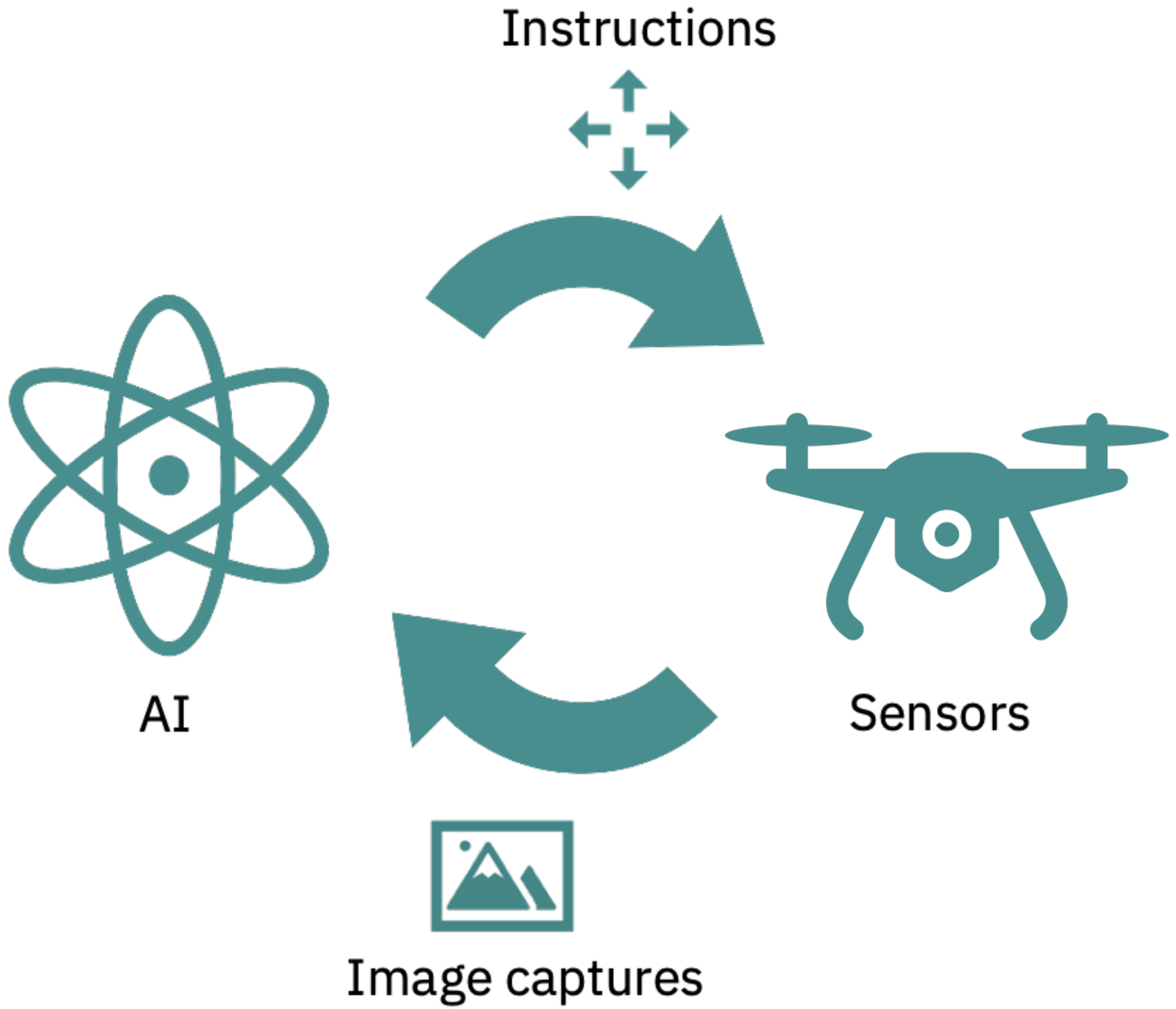
Embodied computer vision

Task:

- actively point sensors towards important objects

Goals:

- faster scene understanding
- improved power efficiency



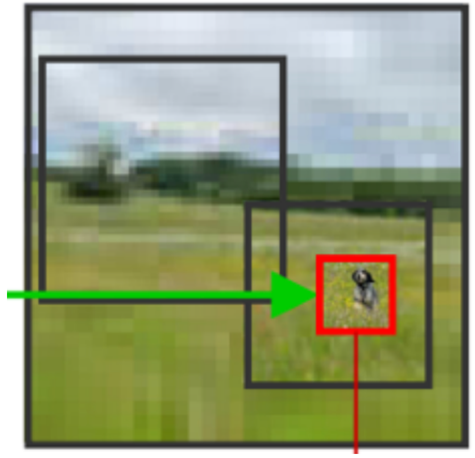
Our research progress

2D

Stationary agent
pan/tilt camera

- grid game

2.5D

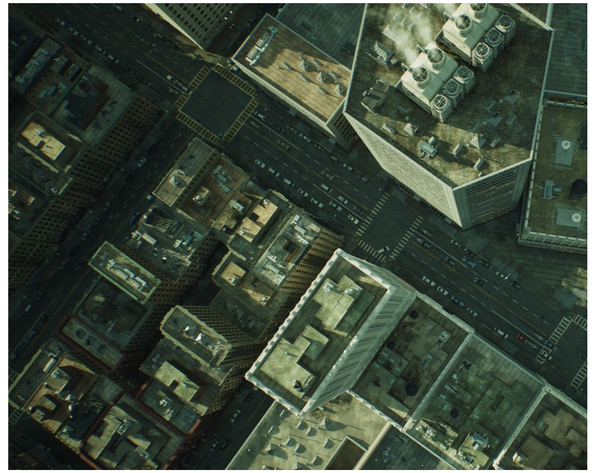


Stationary agent
pan/tilt/zoom camera

+ grid free

WIP

3D



Free movement

+ object occlusion

One year ago:

IJCAI/2023 MACAO

Active Visual Exploration Based on Attention-Map Entropy

Adam Pardyl, Grzegorz Rypeść, Grzegorz Kurzejamski,
Bartosz Zieliński and Tomasz Trzciński

IDEAS NCBR

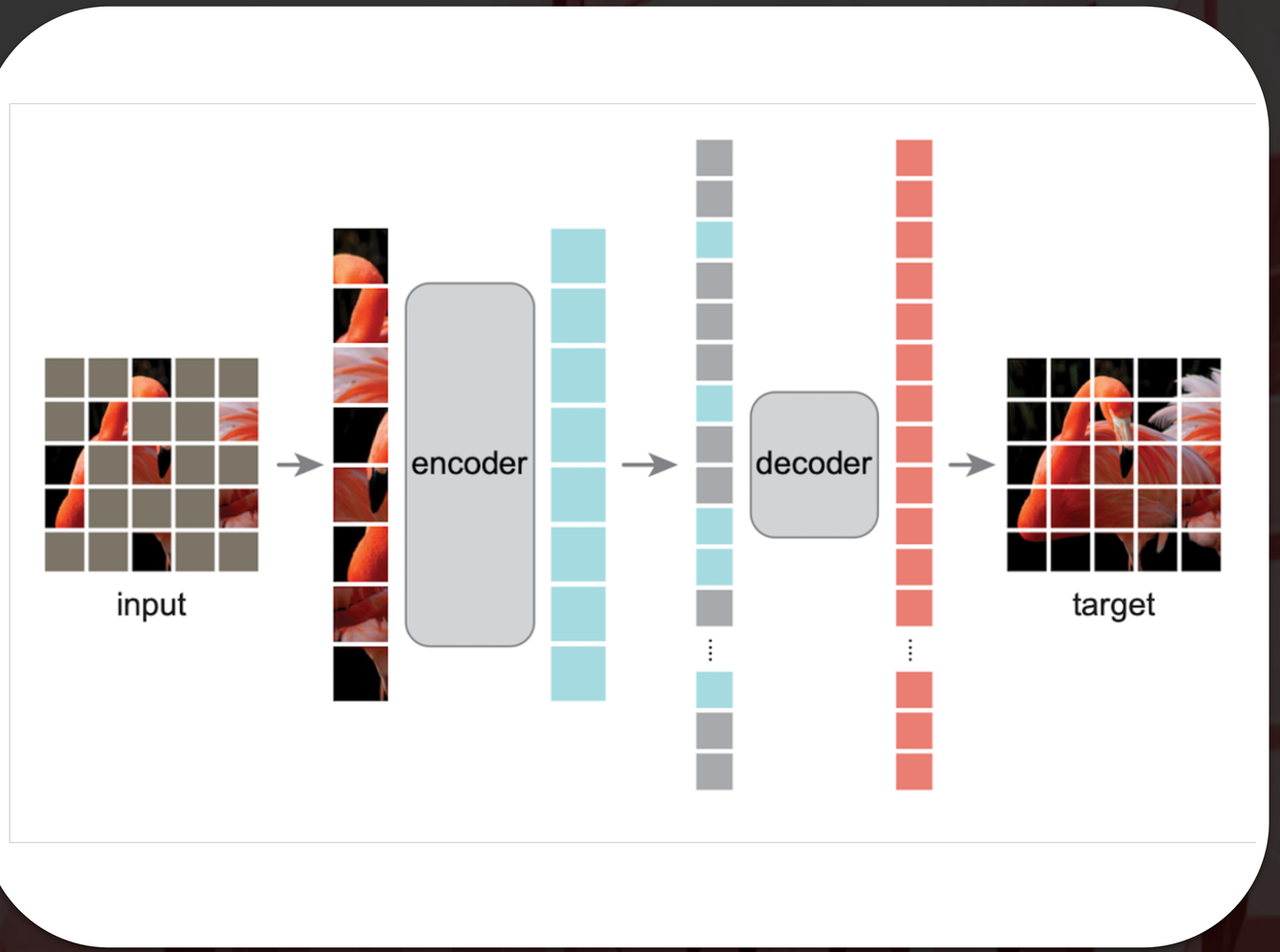
JAGIELLONIAN UNIVERSITY IN KRAKÓW

group of machine gmum learning research

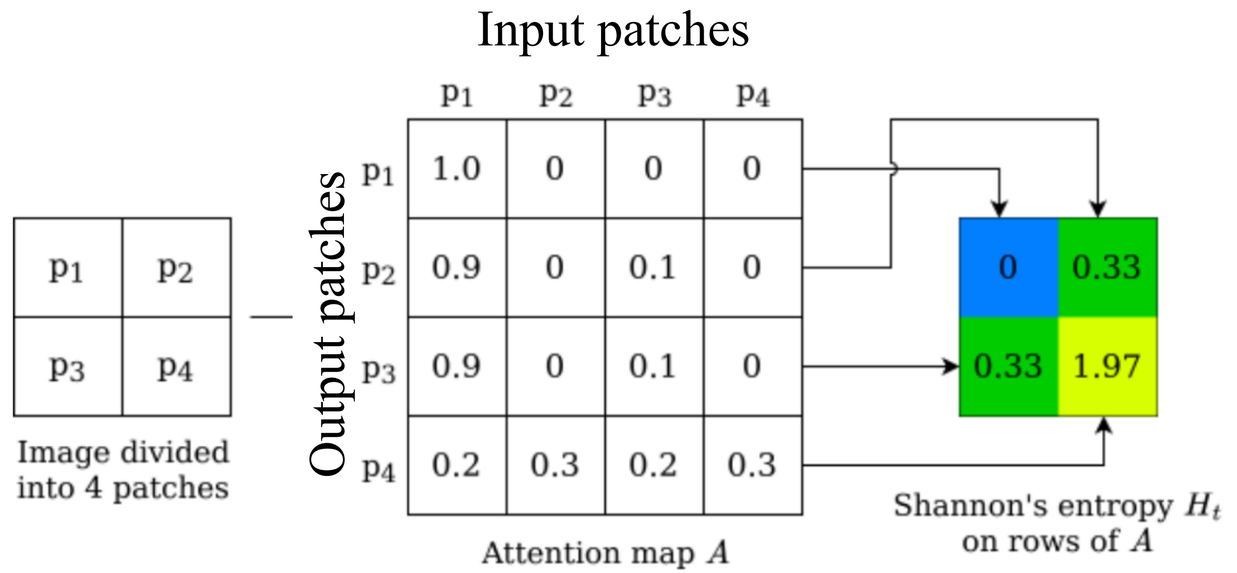
Warsaw University of Technology

Masked Autoencoder

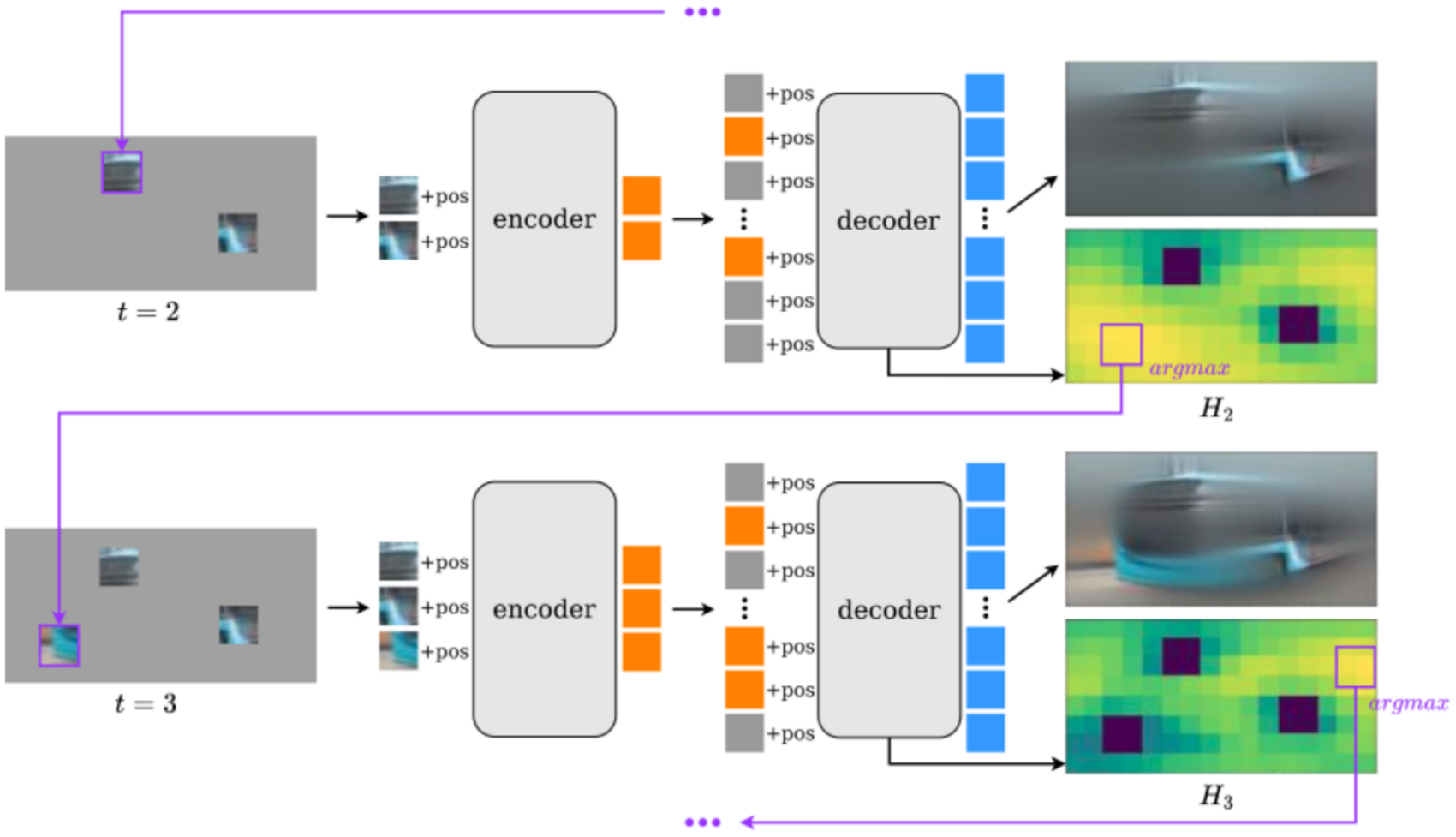
He, Kaiming, et al. "Masked autoencoders are scalable vision learners." (CVPR 2022)



Attention-map entropy



Active Visual Exploration based on Attention-Map Entropy



In practice

Example image reconstruction

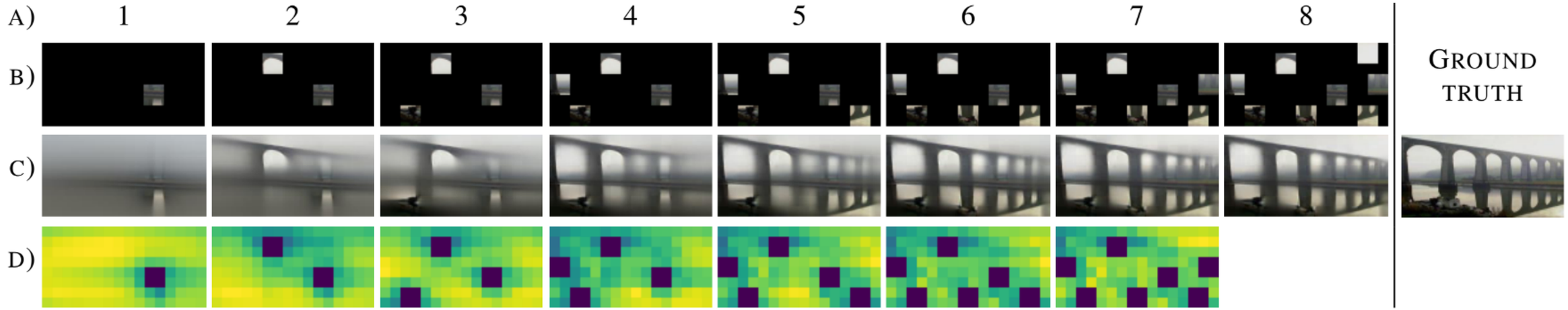


Figure 4: **Glimpse-based reconstruction step-by-step:** The figure shows a glimpse selection process based on AME for 8×32^2 glimpses for a sample 256×128 image. The rows correspond to A) step number, B) model input (glimpses), C) model prediction given, D) decoder attention entropy (known areas are explicitly set to zero). The algorithm explores the image in places where the reconstruction result is blurry.

WACV 2025
TUCSON, ARIZONA • FEB 28 - MAR 4

Beyond Grids: Exploring Elastic Input Sampling for Vision Transformers

Adam Pardyl, Grzegorz Kurzejamski, Jan Olszewski,
Tomasz Trzciński and Bartosz Zieliński

IDEAS NCBR

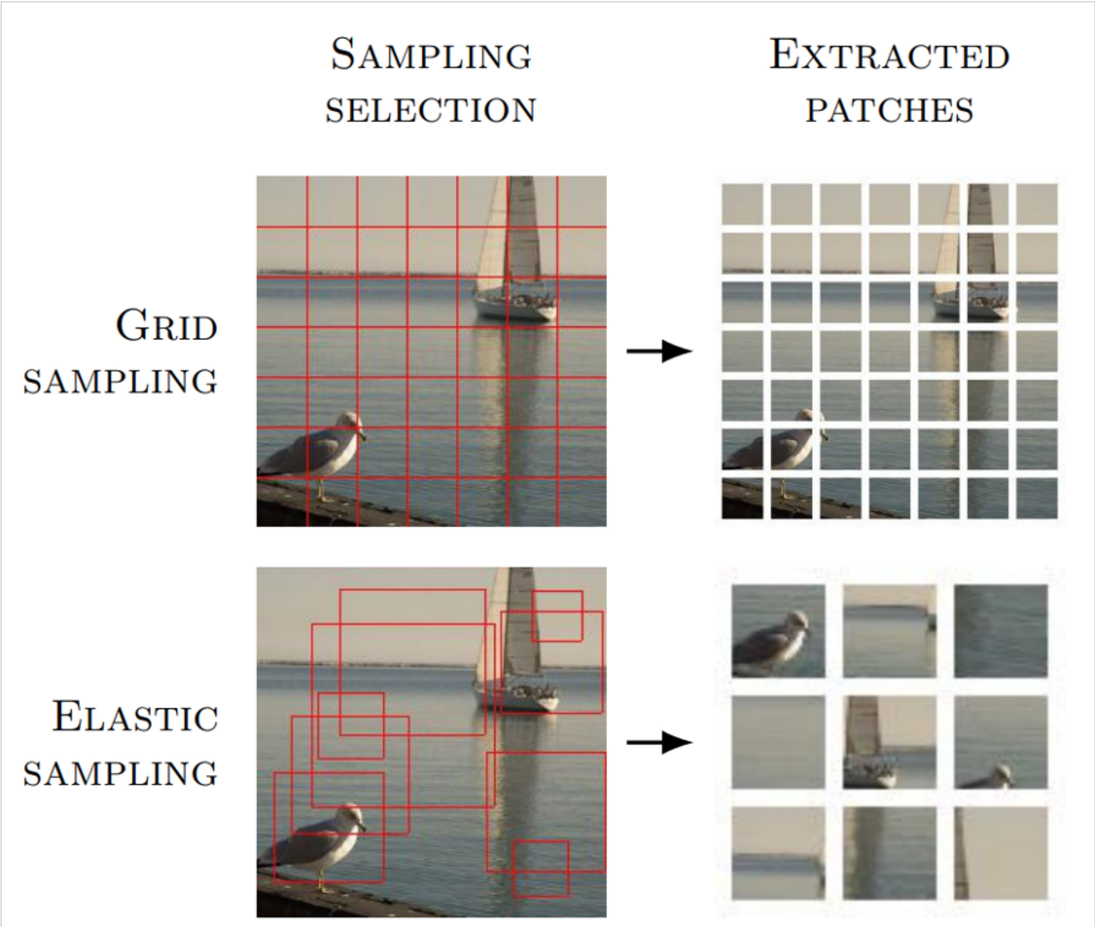
JAGIELLONIAN UNIVERSITY IN KRAKÓW

group of machine gmum learning research

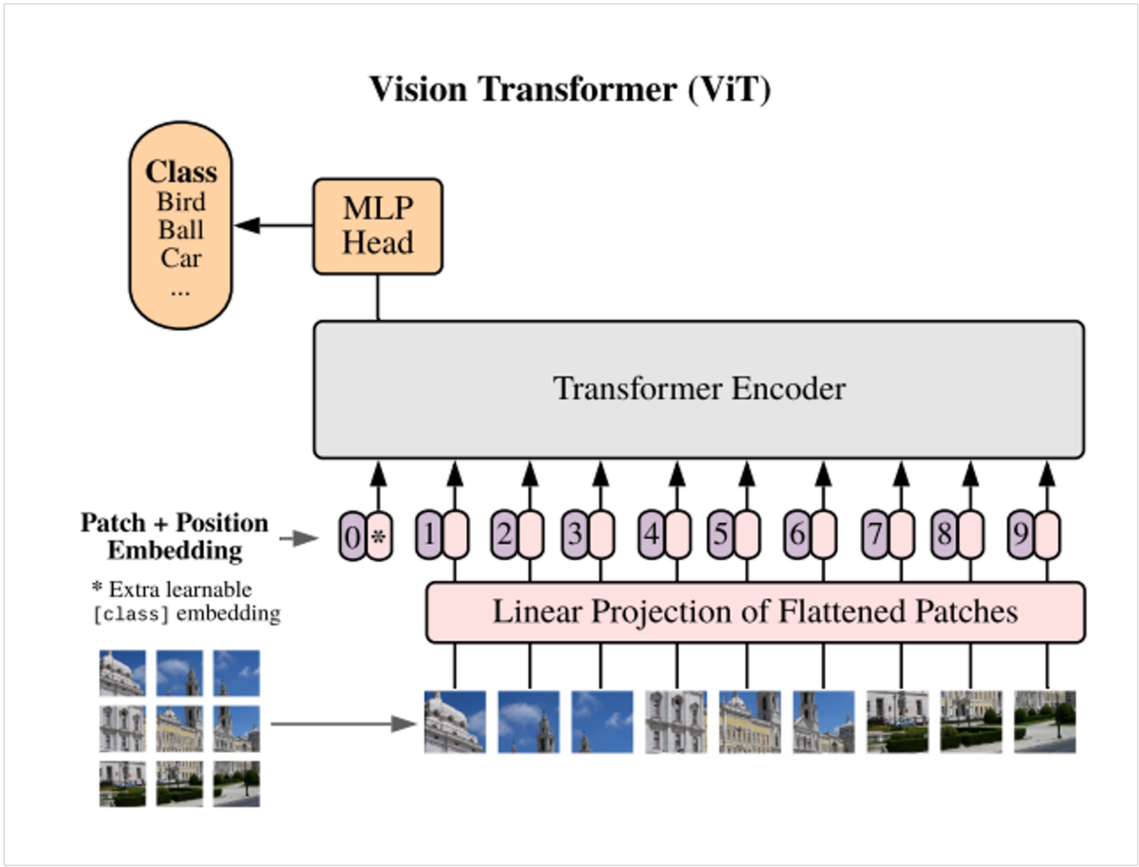
Warsaw University of Technology

UNIWERSYTET WARSZAWSKI

Elastic sampling



Re-think vision transformers

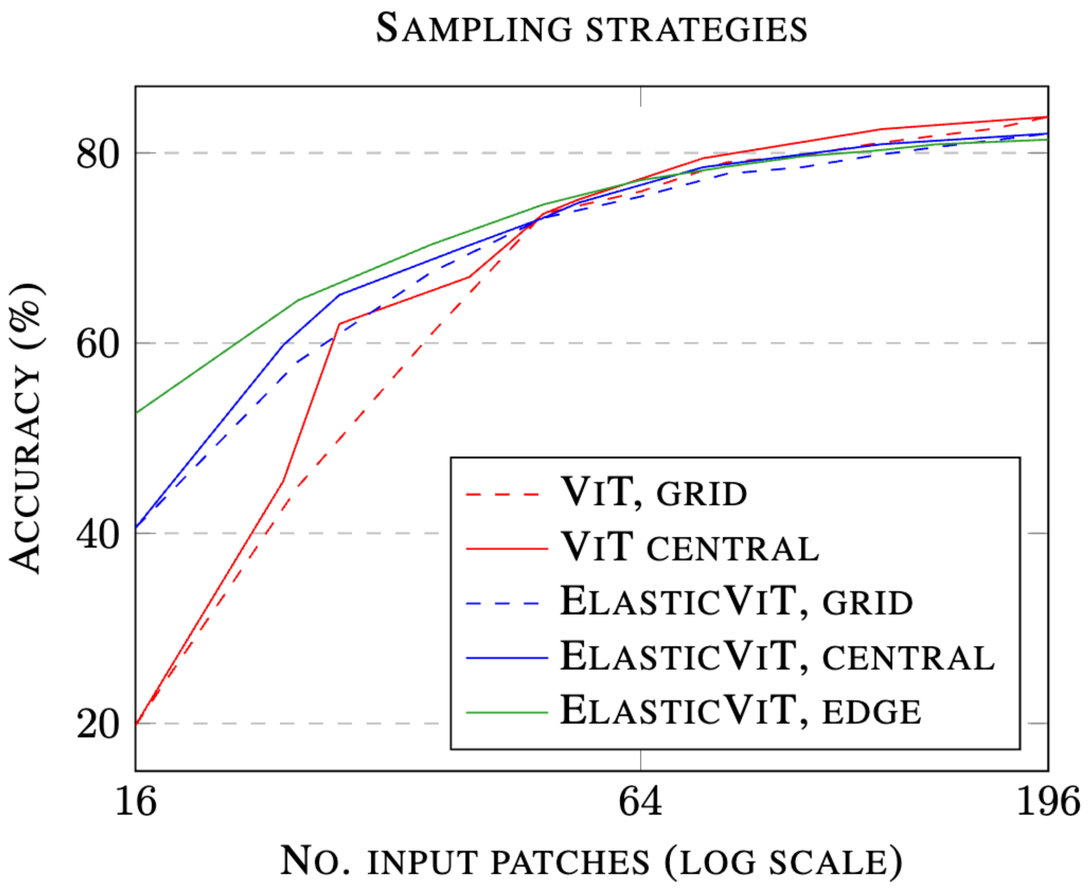


Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." (ICLR 2020)

Grid-free transformer

	Standard ViT	Elastic ViT (ours)
Patch sampling	Fixed grid	Arbitrary patch positions and scales
Positional embedding	1D learned	4D sine-cosine (up-left & low-right corner relative positions in cont. space)
Training regime	Standard augmentations, MixUp, CutMix	Random sampled patches, MixUp, PatchMix (ours)

Adaptive sampling



AdaGlimpse: Active Visual Exploration with Arbitrary Glimpse Position and Scale

Adam Pardyl, Michał Wronka, Maciej Wołczyk, Kamil Adamczewski, Tomasz Trzciński and Bartosz Zieliński

ECCV

IDEAS NCBR

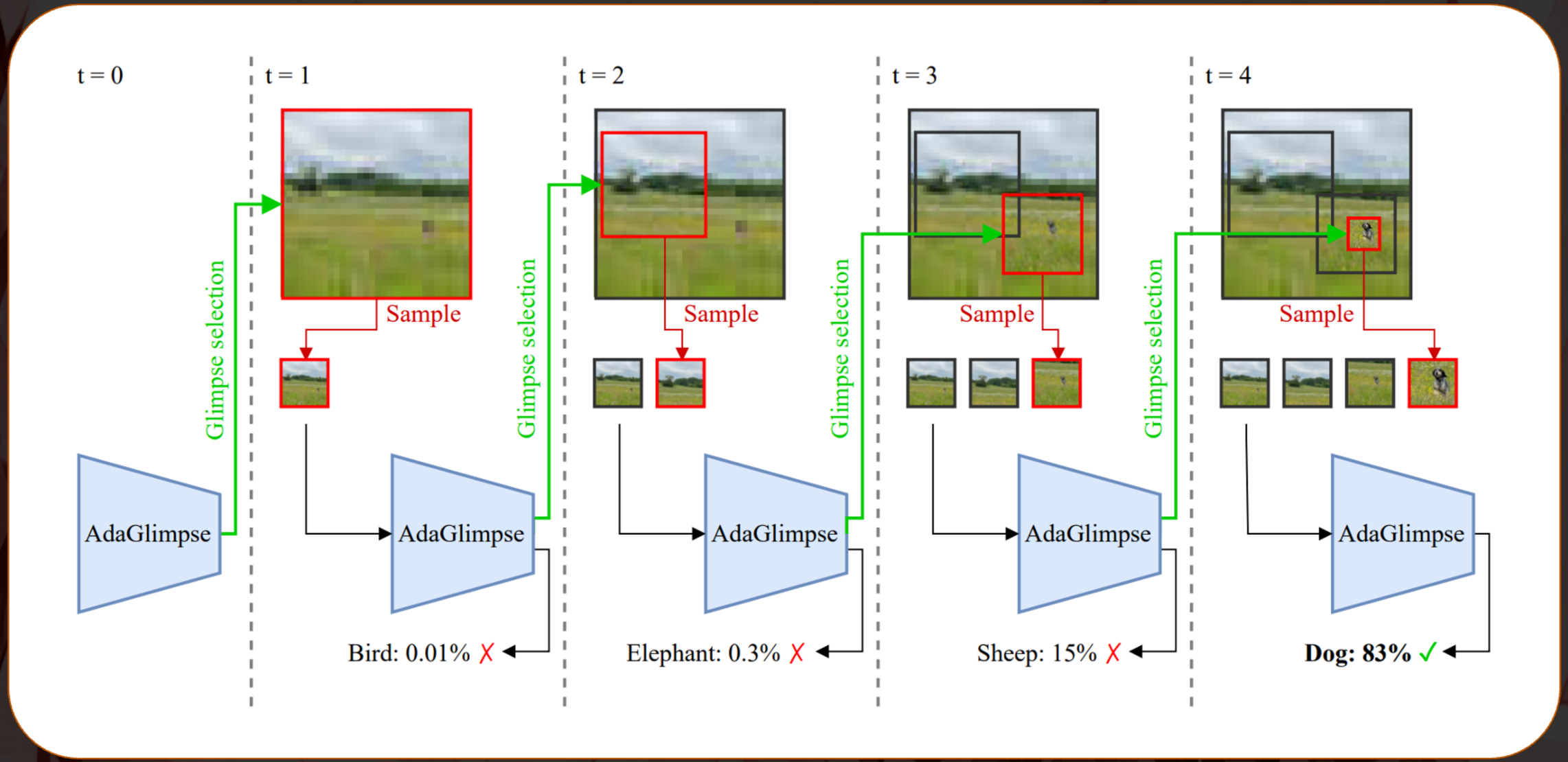
Jagiellonian University in Kraków

group of machine gmum learning research

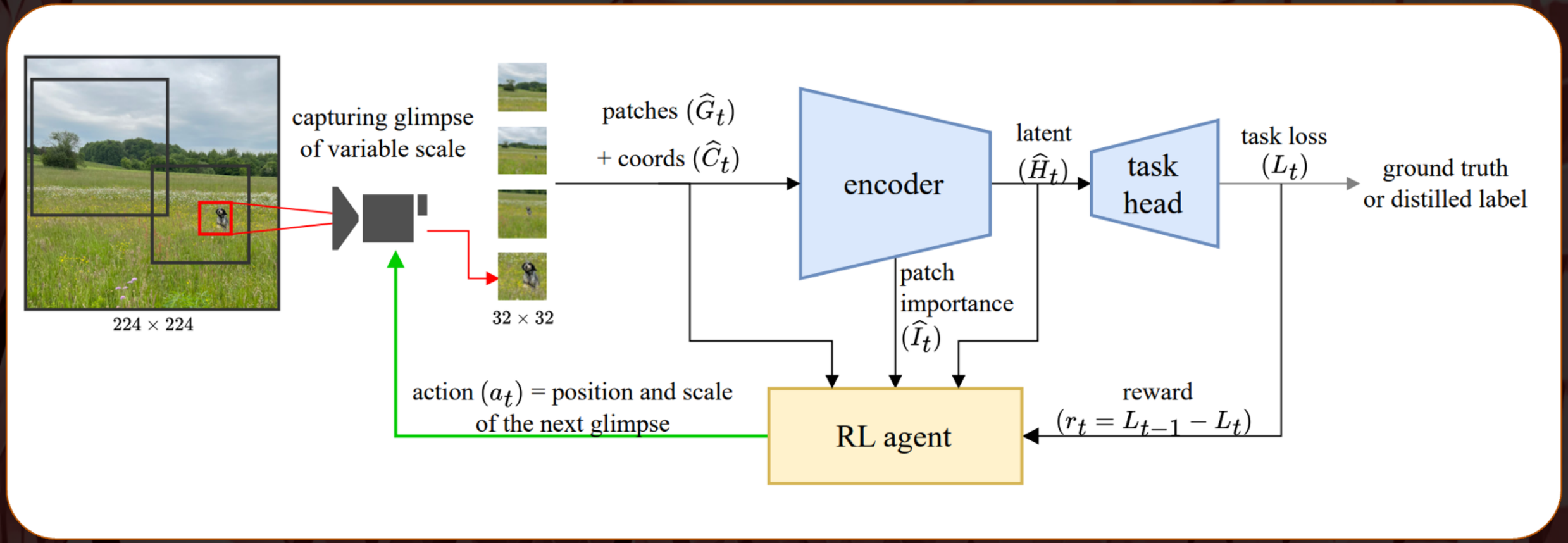
Warsaw University of Technology

Narodowe Centrum Nauki

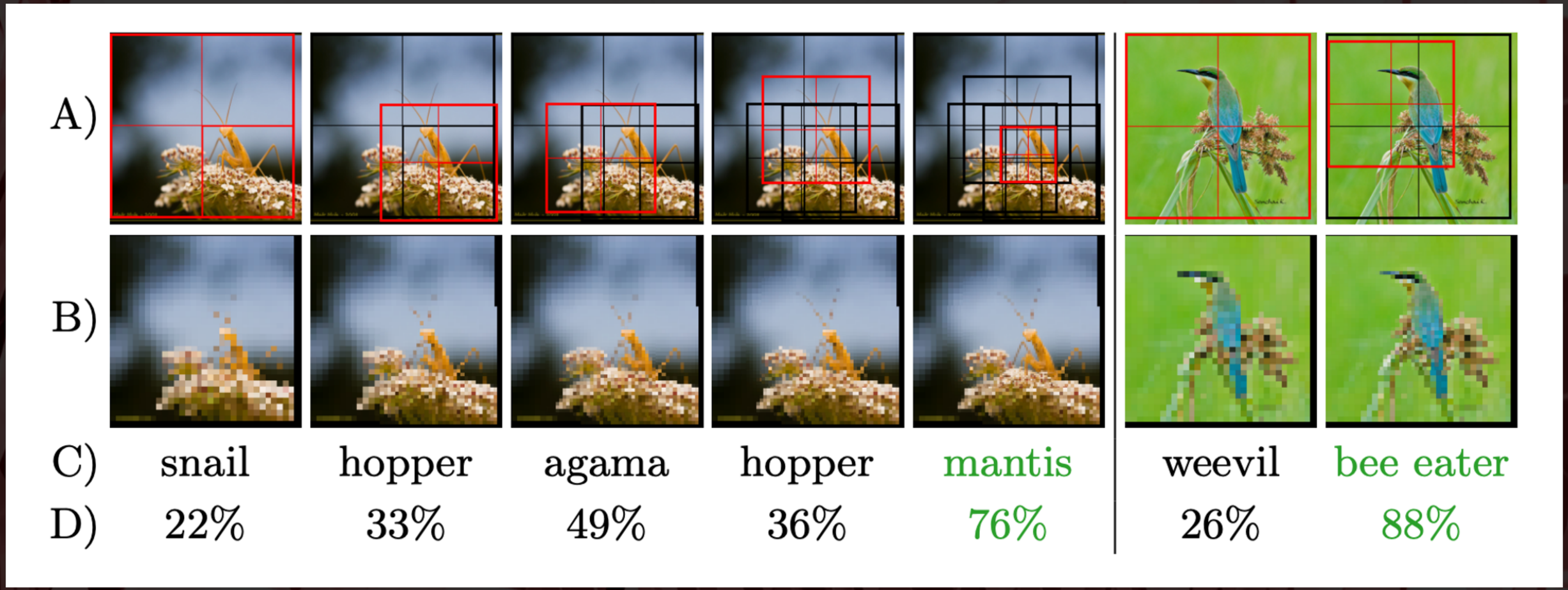
AdaGlimpse

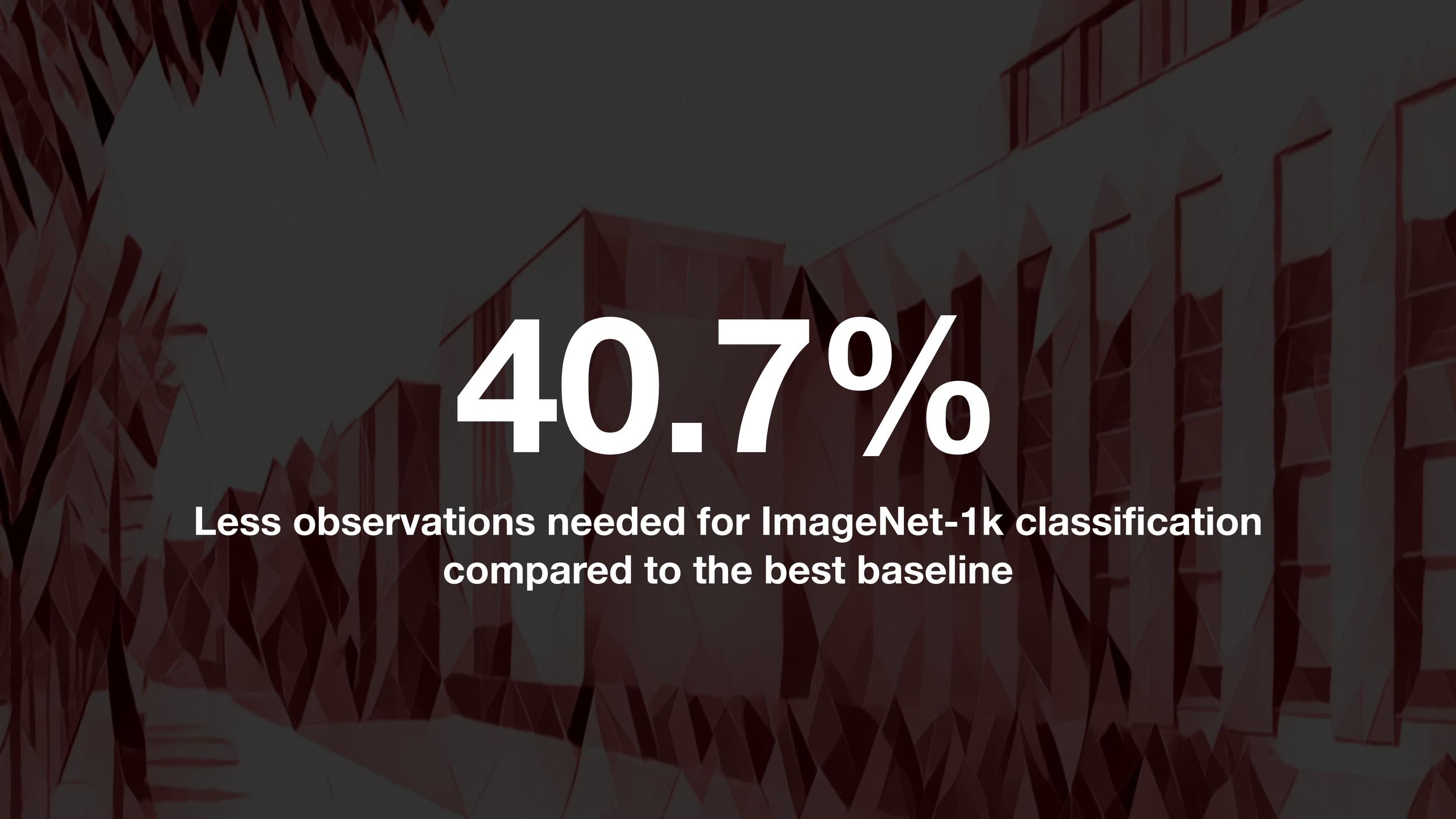

t = 0
t = 1
t = 2
t = 3
t = 4
Glimpse selection
Sample
AdaGlimpse
Bird: 0.01% ✗
Elephant: 0.3% ✗
Sheep: 15% ✗
Dog: 83% ✓

Architecture

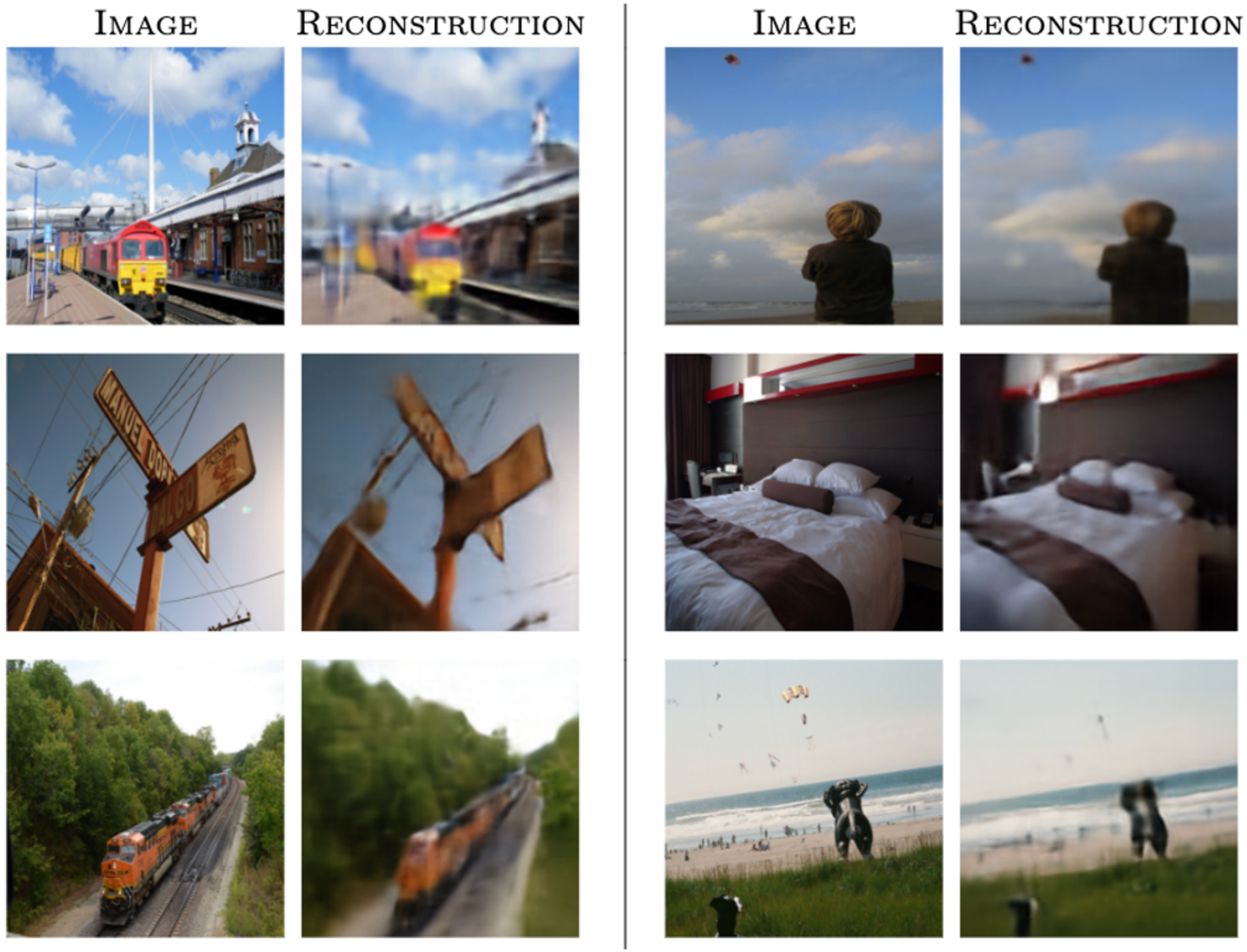


In practice

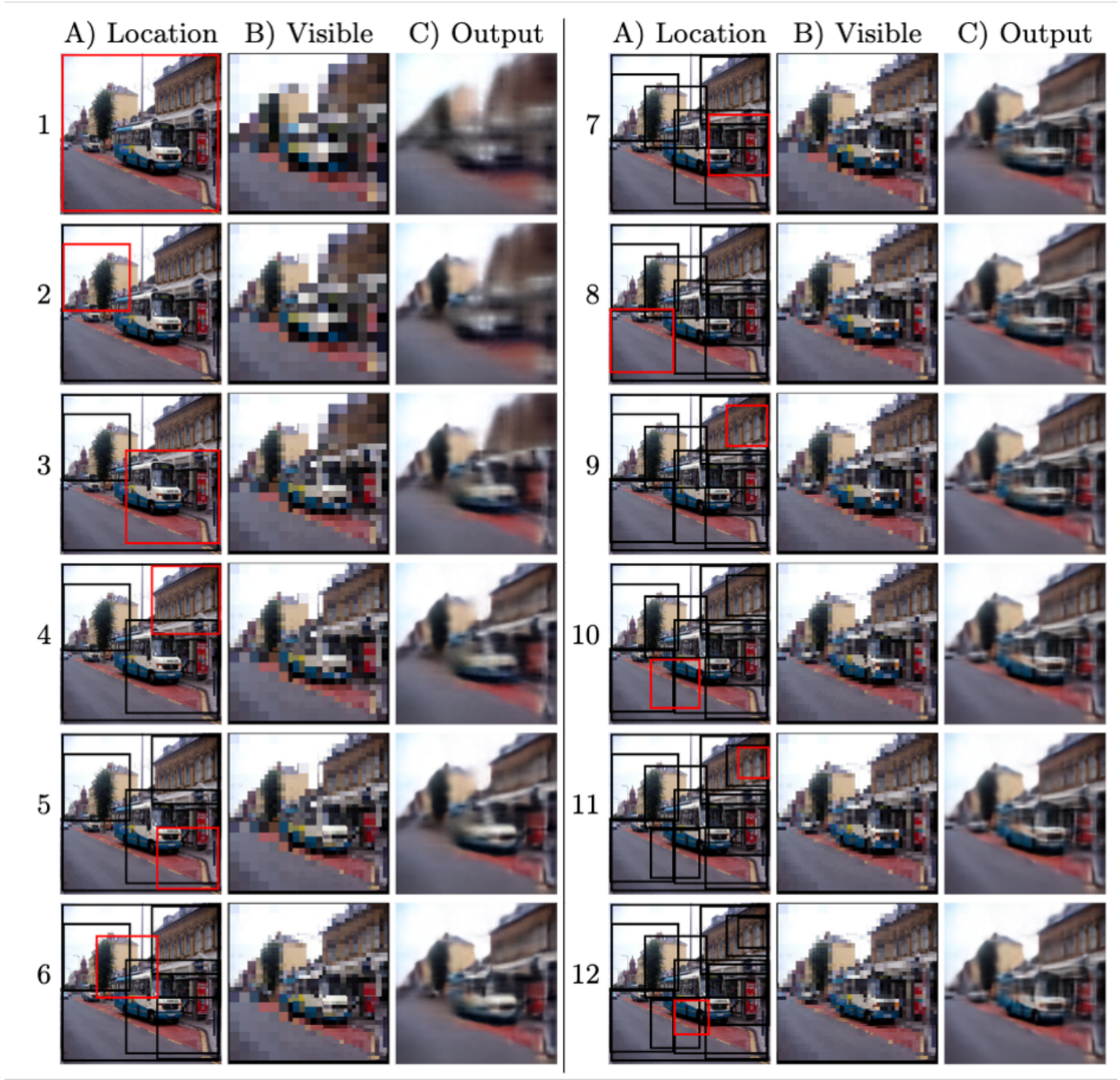

A)
B)
C) snail hopper agama hopper mantis weevil bee eater
D) 22% 33% 49% 36% 76% 26% 88%

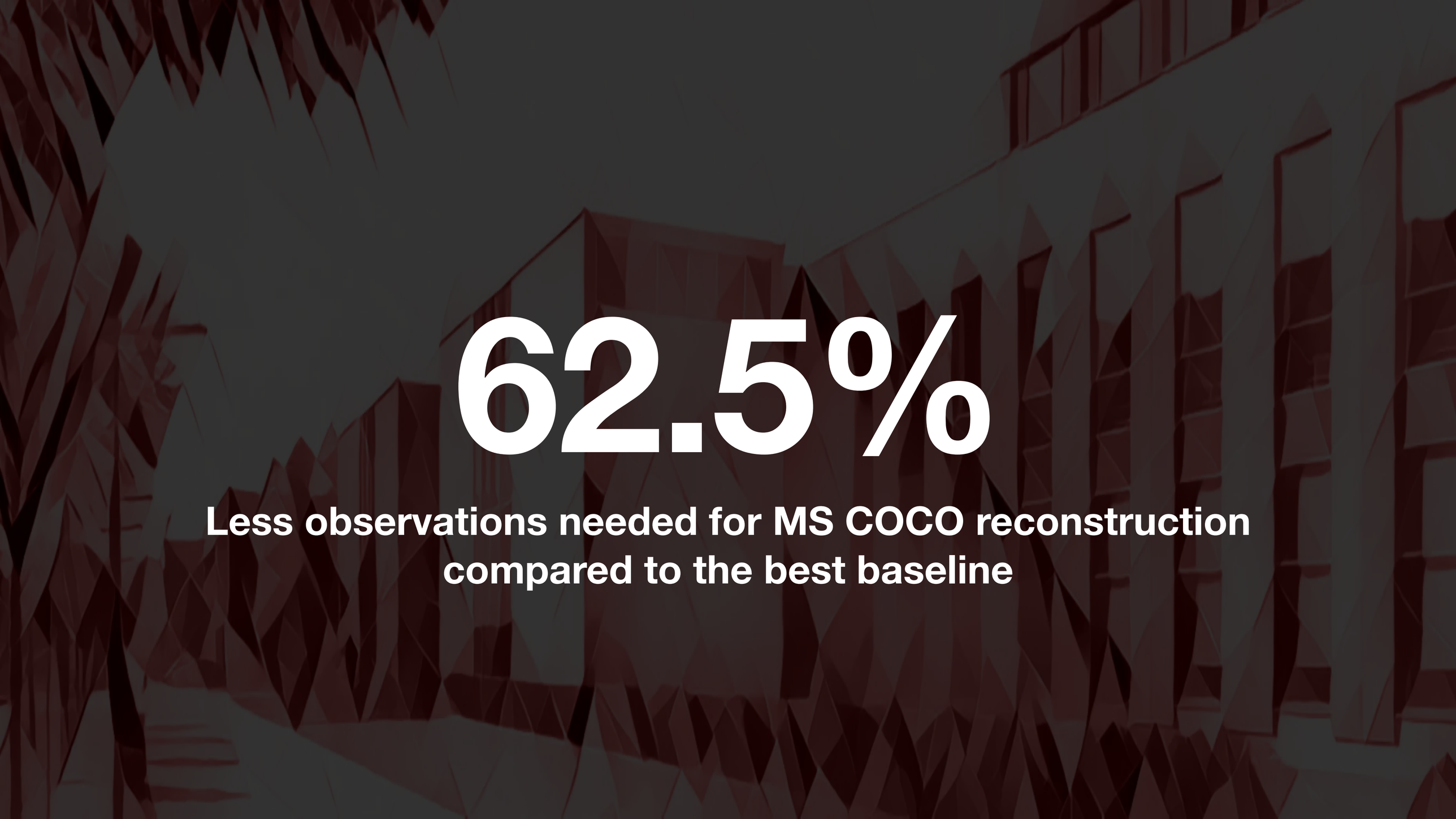

40.7%
Less observations needed for ImageNet-1k classification
compared to the best baseline

Reconstruction examples (6.12% of scene visible)

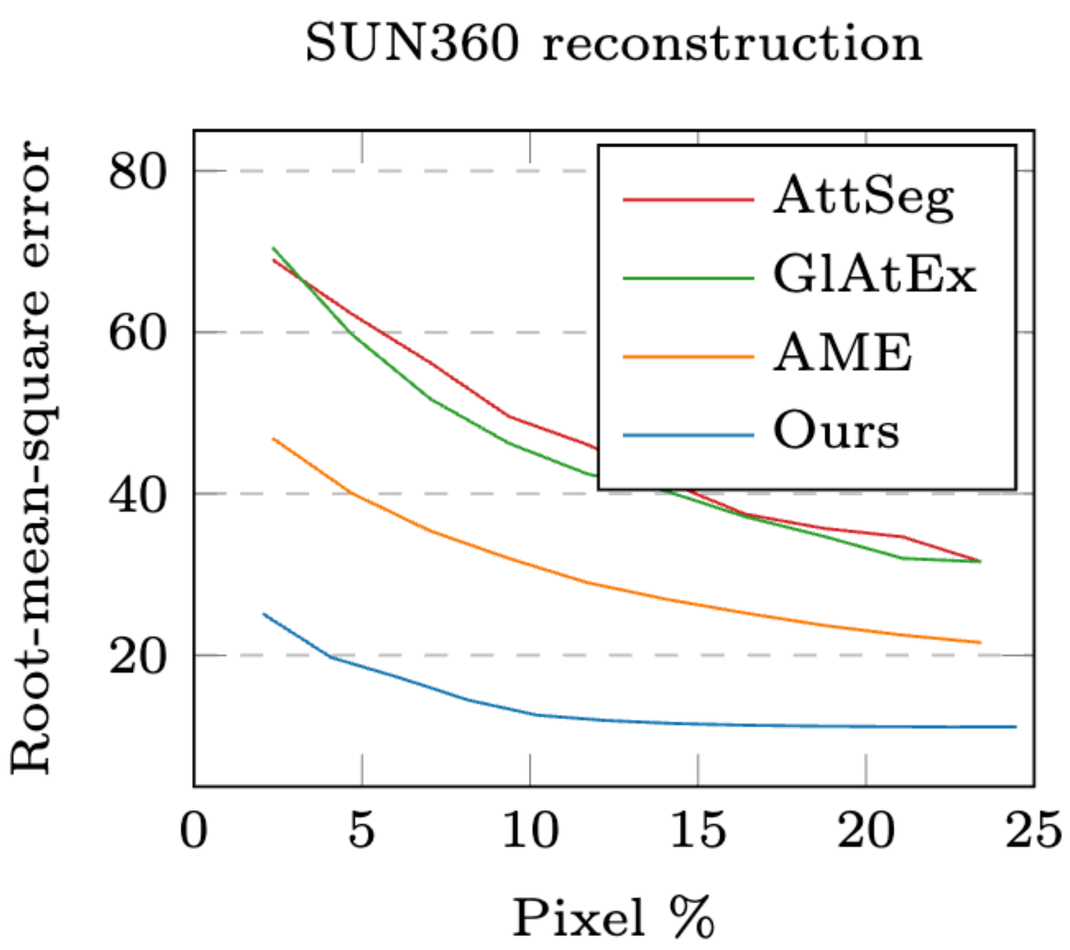


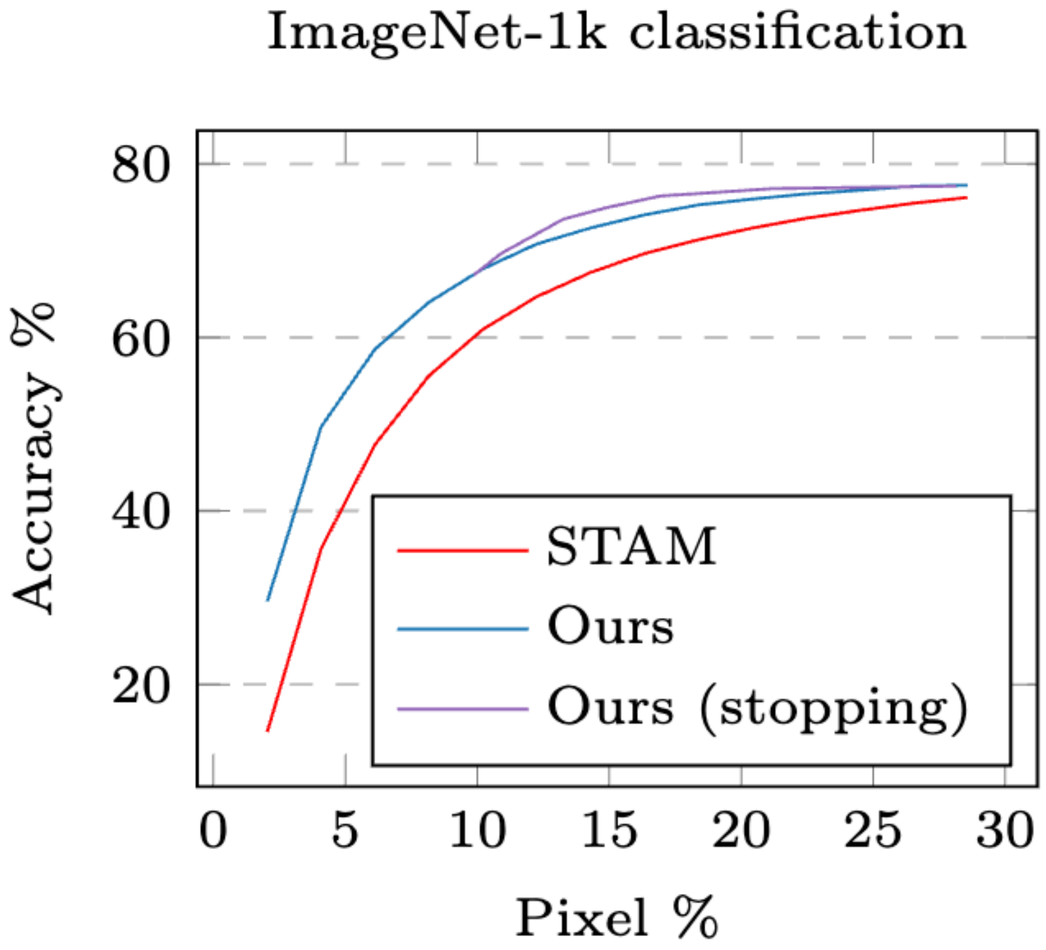
Step-by-step reconstruction (6.12% of scene visible)



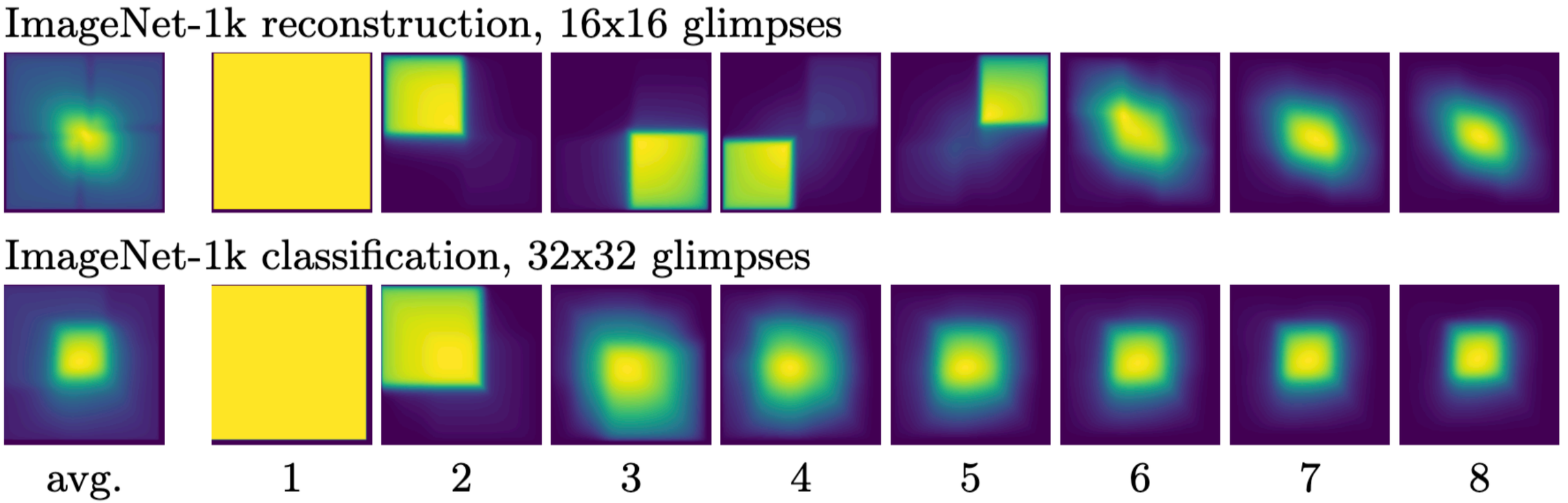

62.5%
Less observations needed for MS COCO reconstruction
compared to the best baseline

Performance by percentage of image pixels observed





Average glimpse image



- consistent with ImageNet center bias

Next step (work in progress):

Where to look next 3.0

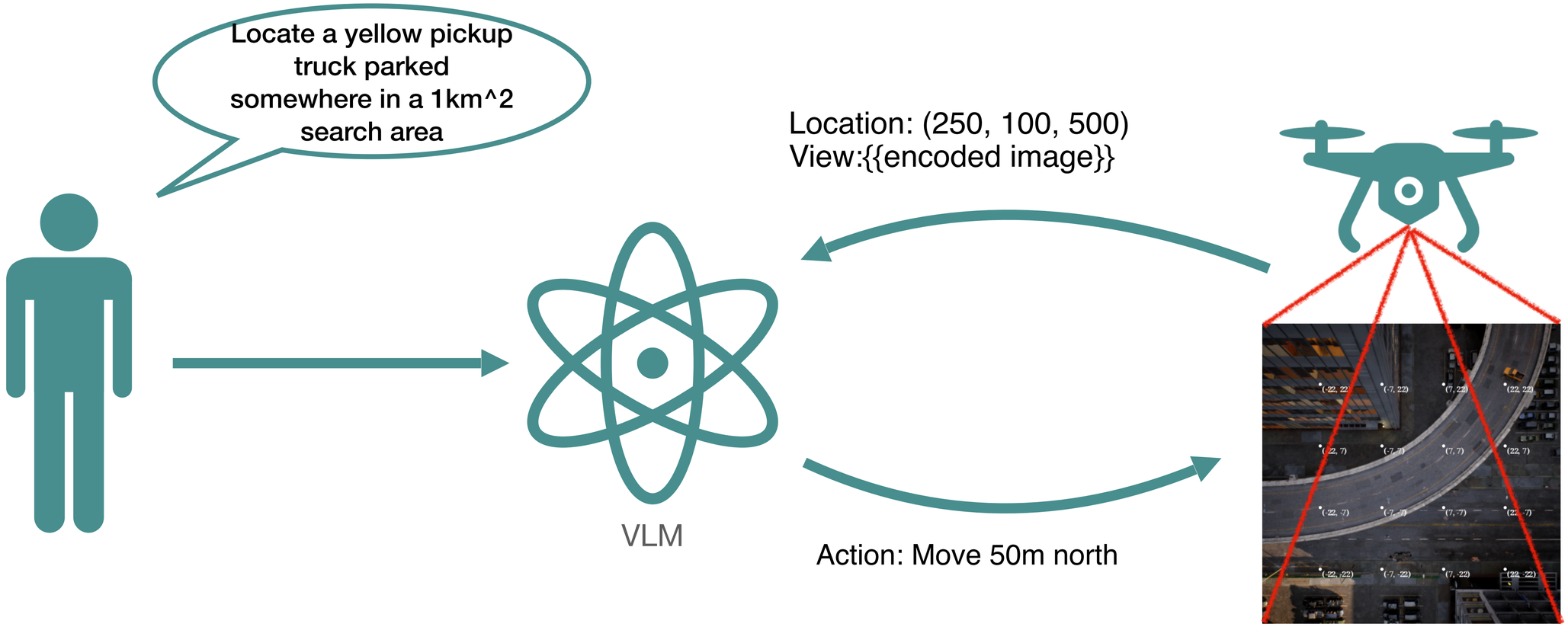
Adam Pardyl, Dominik Matuszek, Maciej Wołczyk, Marek Cygan and Bartosz Zieliński

3D Active Visual Exploration benchmark

Based on Unreal Engine 5

- Free camera movement in 3D space
- Photorealistic simulation
- Procedural scenario generation
- Both urban and nature scenarios for object localisation

Vision-language models for exploration



Example I: The good

Fly to a yellow pickup truck, GPT-4o baseline

GPT-4o: <Reasoning>The yellow pickup truck is just slightly north of my current position. I'll move a little north to get directly above it.</Reasoning>

<Action>MOVE NORTH 7</Action>

Example I: The good

Fly to a yellow pickup truck, GPT-4o baseline

GPT-4o: <Reasoning>I am directly above the yellow pickup truck and as close as possible.Task complete.</Reasoning>

<Action>DONE</Action>

Example II: The bad

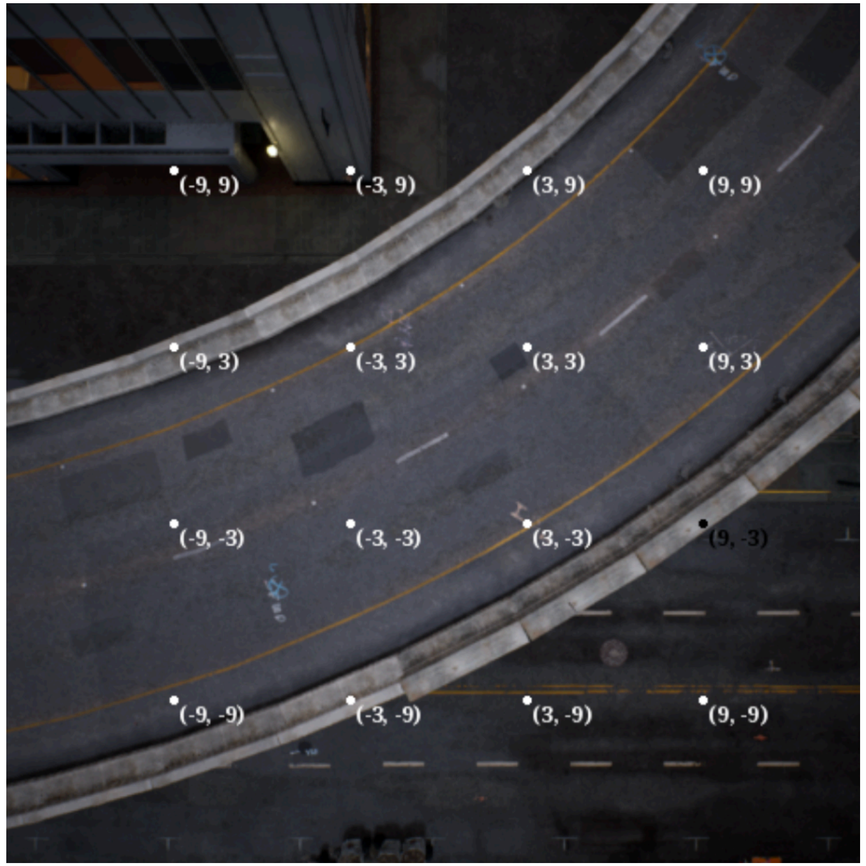
Fly to a yellow pickup truck, GPT-4o baseline

GPT-4o: <Reasoning>The yellow pickup truck is now visible near (-22, -22) below my current position. I need to move down to get closer.</Reasoning>

<Action>MOVE DOWN 20</Action>

Example II: The bad

Fly to a yellow pickup truck, GPT-4o baseline

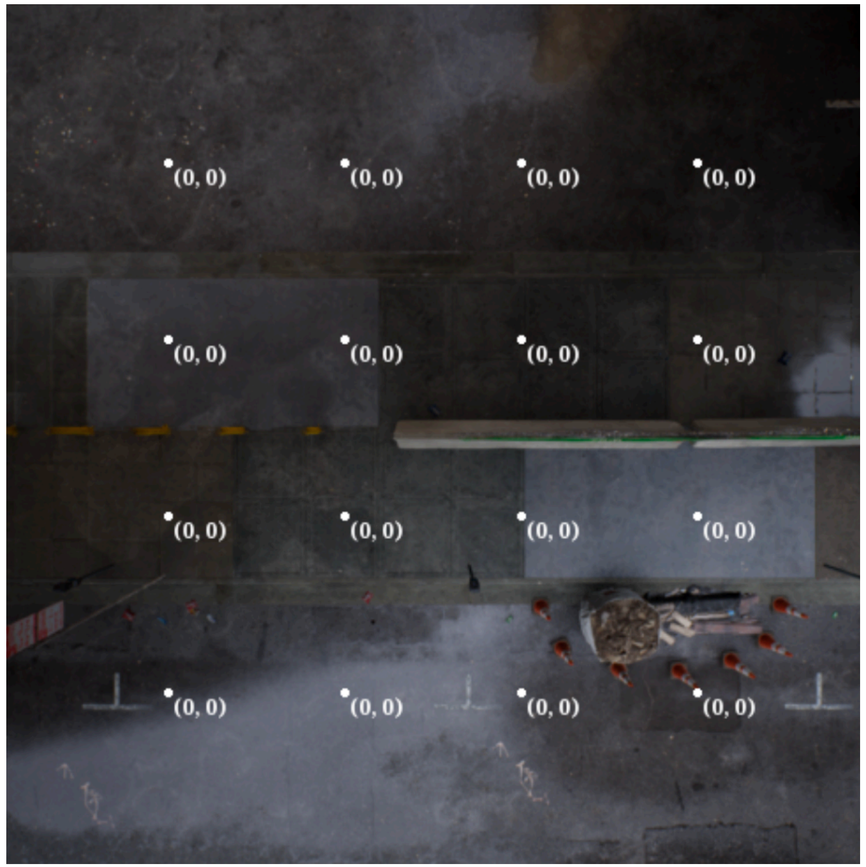


GPT-4o: <Reasoning>The yellow pickup truck is directly below me near (-22, -22). I will move down to get as close as possible.</Reasoning>

<Action>MOVE DOWN 15</Action>

Example II: The bad

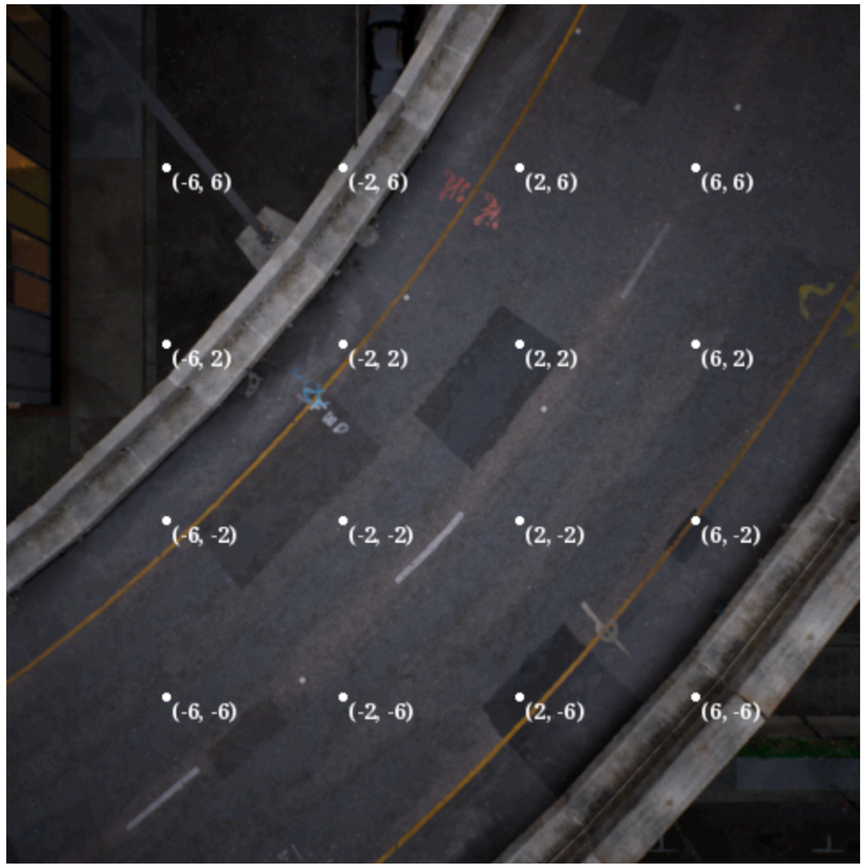
Fly to a yellow pickup truck, GPT-4o baseline



GPT-4o: DONE

Example III: The strange

Fly to a yellow pickup truck, GPT-4o baseline



GPT-4o: <Reasoning>The yellow markings on the road could potentially obscure a yellow vehicle. To gain a better view of the area and identify the yellow pickup, I will move EAST to explore the grid more thoroughly.</Reasoning>
<Action>MOVE EAST 6</Action>

Challenges

Vision-language exploration

- Standard VLMs struggle with 3D spatial reasoning
  - Tested on GPT-4o, LLaVa-v1.6 and InternVL-8B
- Representing multi-step exploration for a VLM is not trivial
  - Open source VLMs often have issues with multiple images in context
- VLM models are currently too slow for UAV/robotic hardware
  - Out of scope of this project

AdaGlimpse

Active Visual Exploration with Arbitrary Glimpse Position and Scale

Adam Pardyl, ML in PL Conference 2024

group of machine gmum learning research

JAGIELLONIAN UNIVERSITY IN KRAKÓW

Warsaw University of Technology

IDEAS NCBR

NARODOWE CENTRUM NAUKI

Paper available at:
(+ references)

AdaGlimpse: Active Visual Exploration with Arbitrary Glimpse Position and Scale

Adam Pardyl[1,2,3], Michał Wronka[2], Maciej Wołczyk[1], Kamil Adamczewski[1], Tomasz Trzciński[1,4], and Bartosz Zieliński[1,2]

[1] IDEAS NCBR
{adam.pardyl, maciej.wolczyk, kamil.adamczewski, tomasz.trzcinski, bartosz.zielinski}@ideas-ncbr.pl
[2] Jagiellonian University, Faculty of Mathematics and Computer Science
michal.wronka@student.uj.edu.pl
[3] Jagiellonian University, Doctoral School of Exact and Natural Sciences
[4] Warsaw University of Technology

Abstract. Active Visual Exploration (AVE) is a task that involves dynamically selecting observations (glimpses), which is critical to facilitate comprehension and navigation within an environment. While modern AVE methods have demonstrated impressive performance, they are constrained to fixed-scale glimpses from rigid grids. In contrast, existing mobile platforms equipped with optical zoom capabilities can capture glimpses of arbitrary positions and scales. To address this gap between software and hardware capabilities, we introduce AdaGlimpse. It uses Soft Actor-Critic, a reinforcement learning algorithm tailored for exploration tasks, to select glimpses of arbitrary position and scale. This approach enables our model to rapidly establish a general awareness of the environment before zooming in for detailed analysis. Experimental results demonstrate that AdaGlimpse surpasses previous methods across various visual tasks while maintaining greater applicability in realistic AVE scenarios.

Keywords: Active visual exploration · Vision transformers · Reinforcement learning

1 Introduction

Common machine learning solutions for computer vision tasks, such as classification, segmentation, or scene understanding, usually presume access to complete

Check out our other talks during ML in PL!

Friday:

Session 2 / Lecture Hall B / 10:35

Deep learning for effective analysis of high content screening
Adriana Borowa

Session 4 / Lecture Hall A / 14:30

Efficient fine-tuning of LLMs: exploring PEFT methods and LORA-XS insights
Klaudia Bałazy

Session 5 / Lecture Hall B / 14:30

Current trends in intrinsically interpretable Deep Learning
Dawid Rymarczyk

Neural rendering: the future of 3D modeling
Przemysław Spurek

JAGIELLONIAN UNIVERSITY IN KRAKÓW

group of machine gmum learning research

Saturday:

Session 7 / Lecture Hall A / 12:00

AdaGlimpse: Active Visual Exploration with Arbitrary Glimpse Position and Scale
Adam Pardyl

Session 8 / Lecture Hall B / 12:00

Augmentation-aware Self-supervised Learning with Conditioned Projector
Marcin Przewięźlikowski

gmum.net