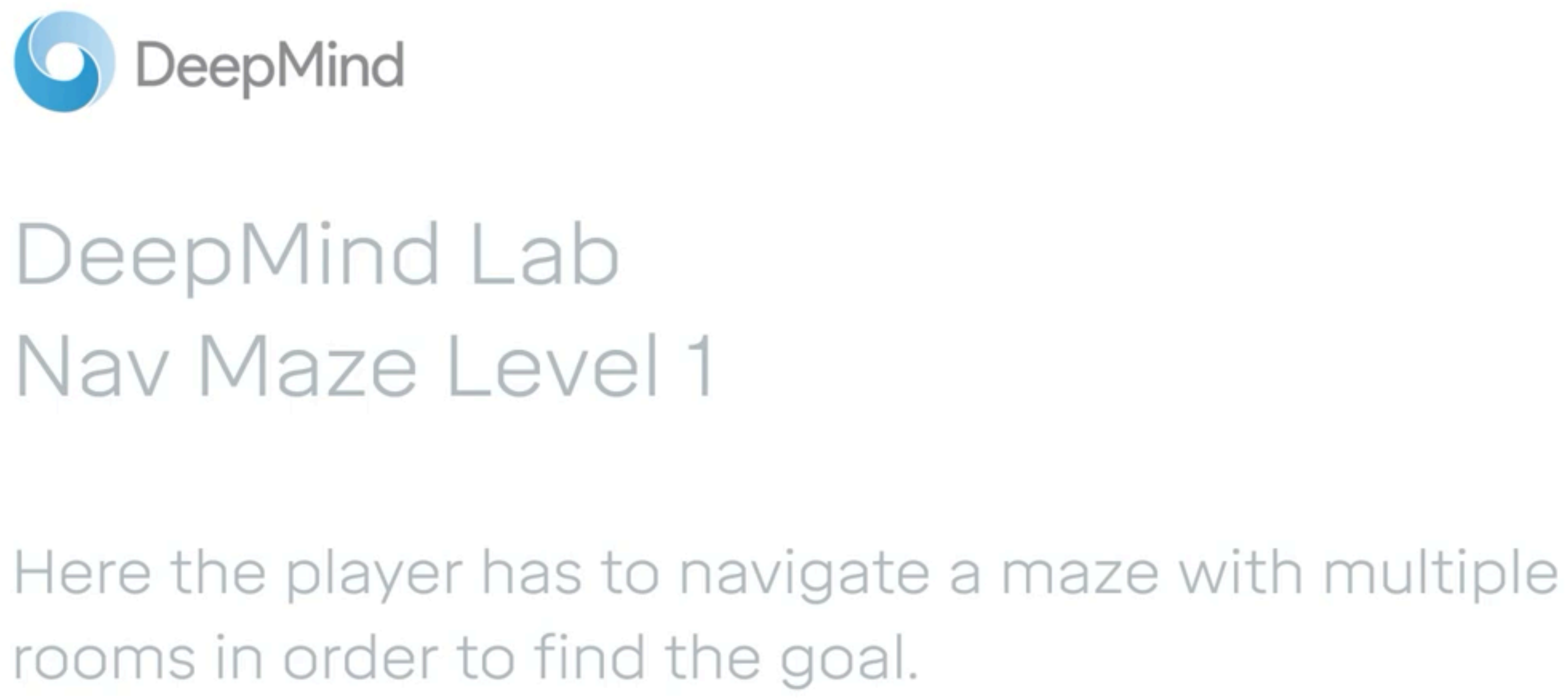
# Learning Structured Video Representations without Supervision

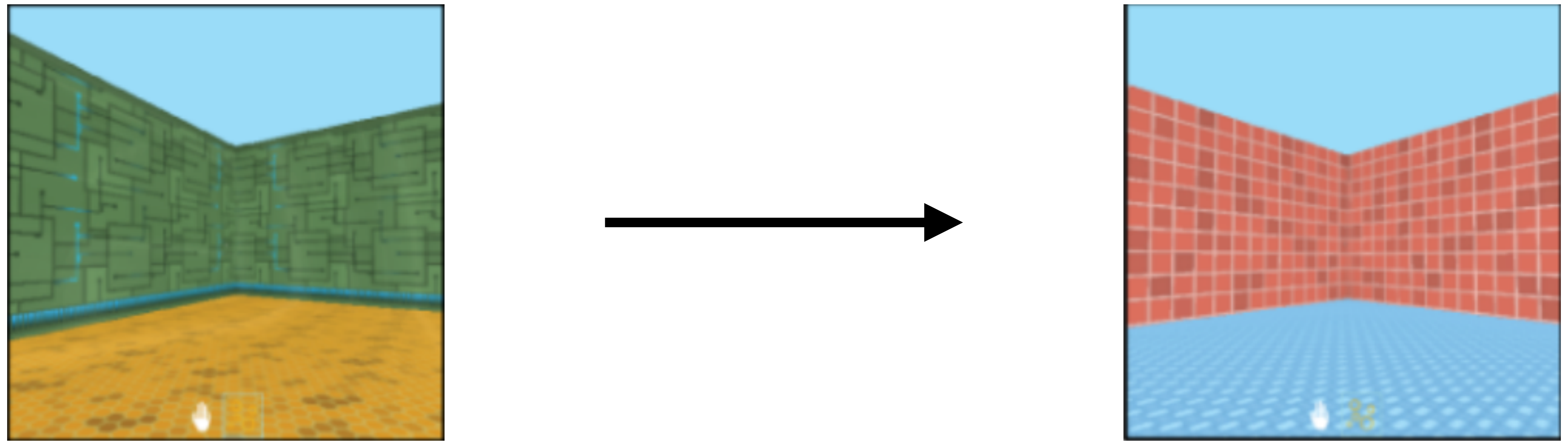
Andrii Zadaianchuk, 2025

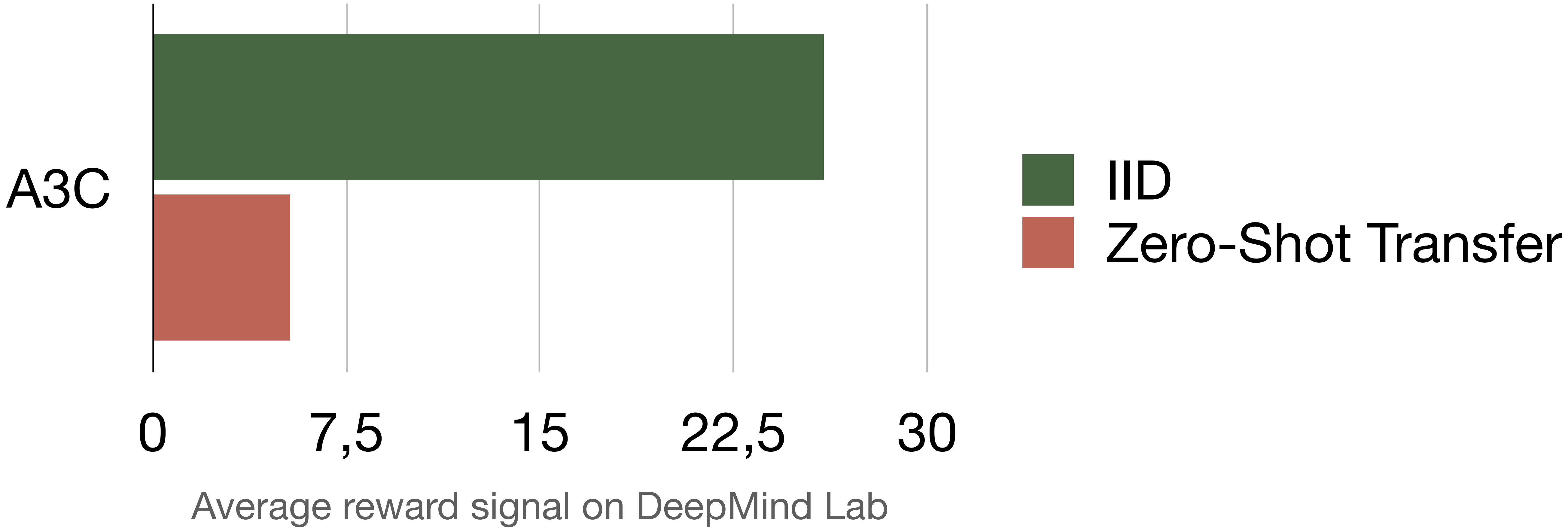
# Why do we need stuctured representations?




Player has to navigate a maze with multiple rooms in order to find the goal.

# What about Zero-Shot Transfer?



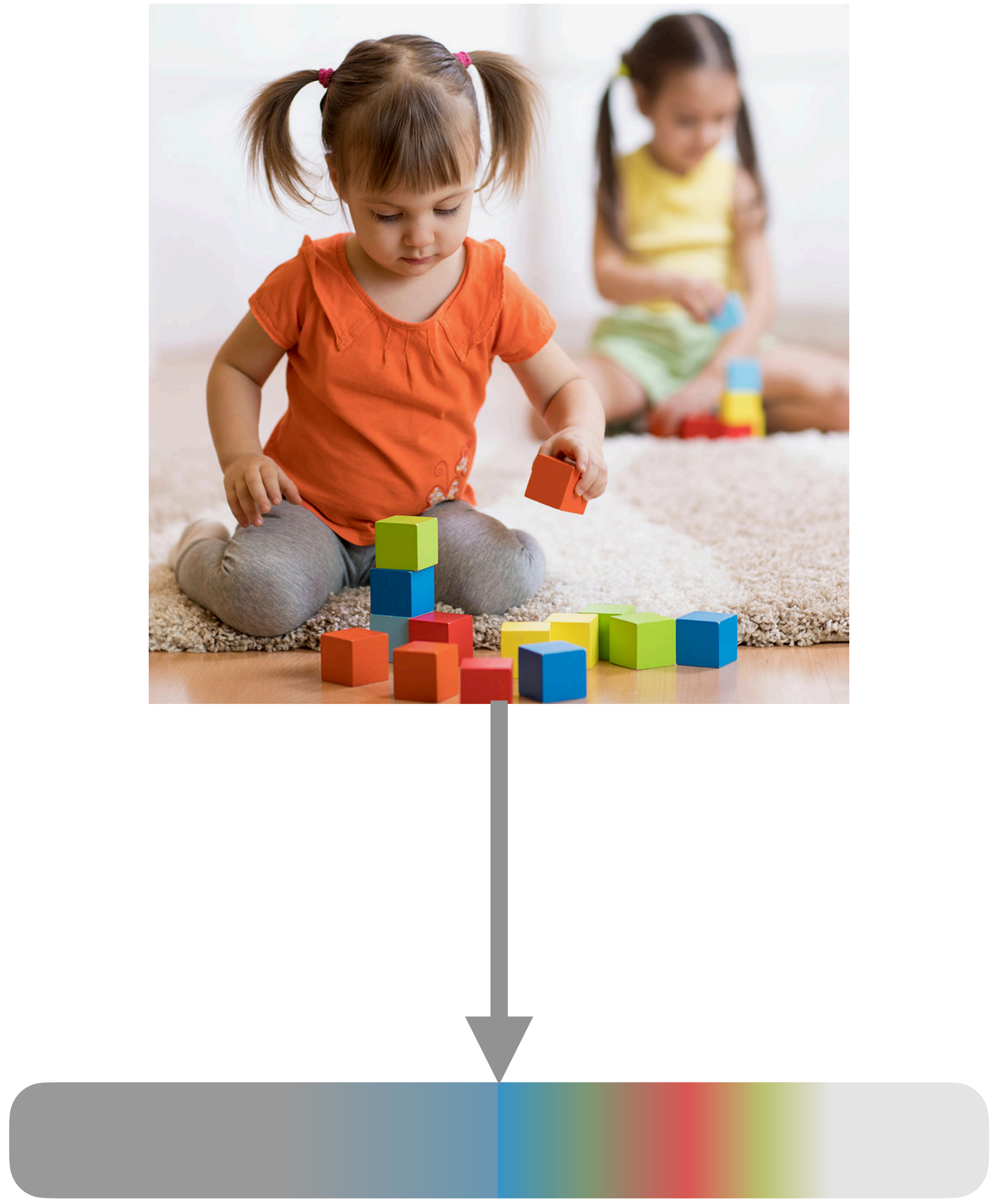



How to represent the world around us robustly?



DARLA: Improving Zero-Shot Transfer in Reinforcement Learning [Higgins et al. 2017]

# Representation Learning from Pixels



Single vector

# Binding Problem in Distributed Representation

a)

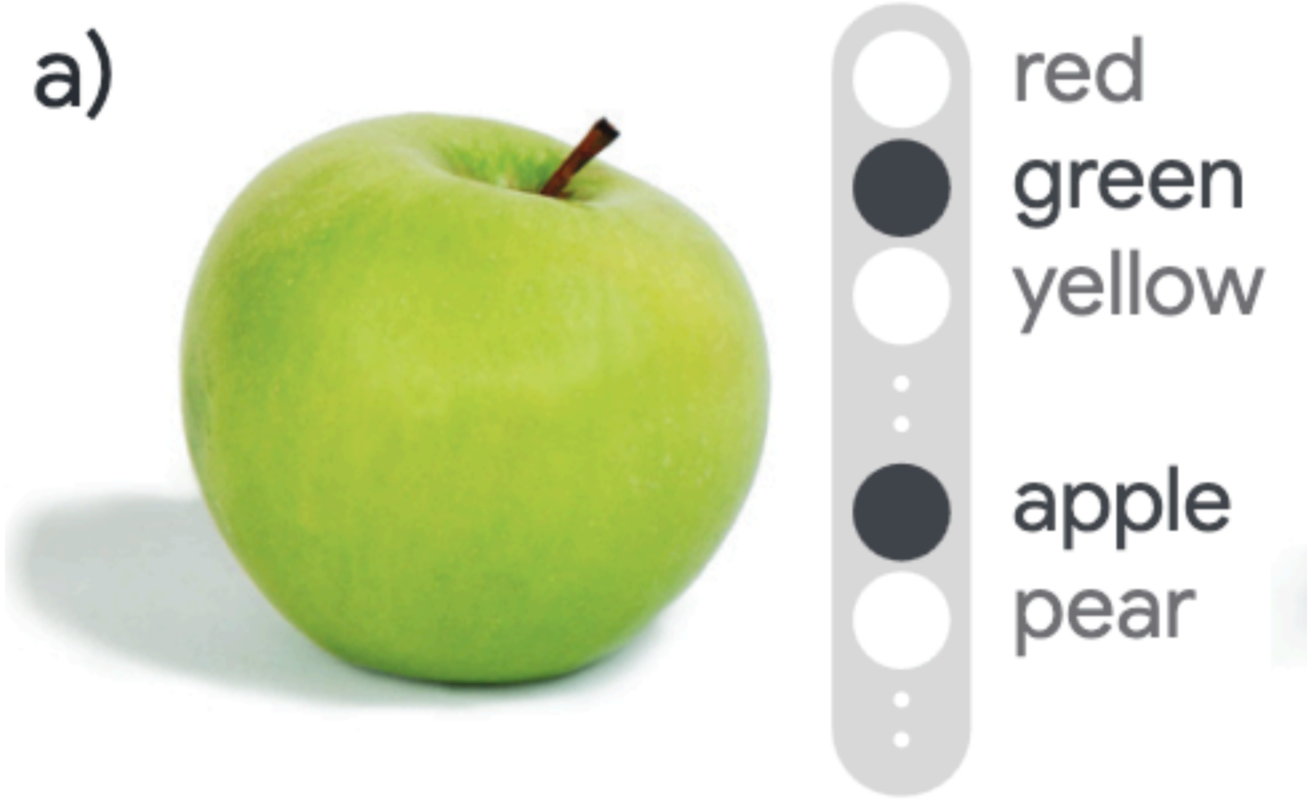





Figure: Greff et al. (2020)

# Binding Problem in Distributed Representation

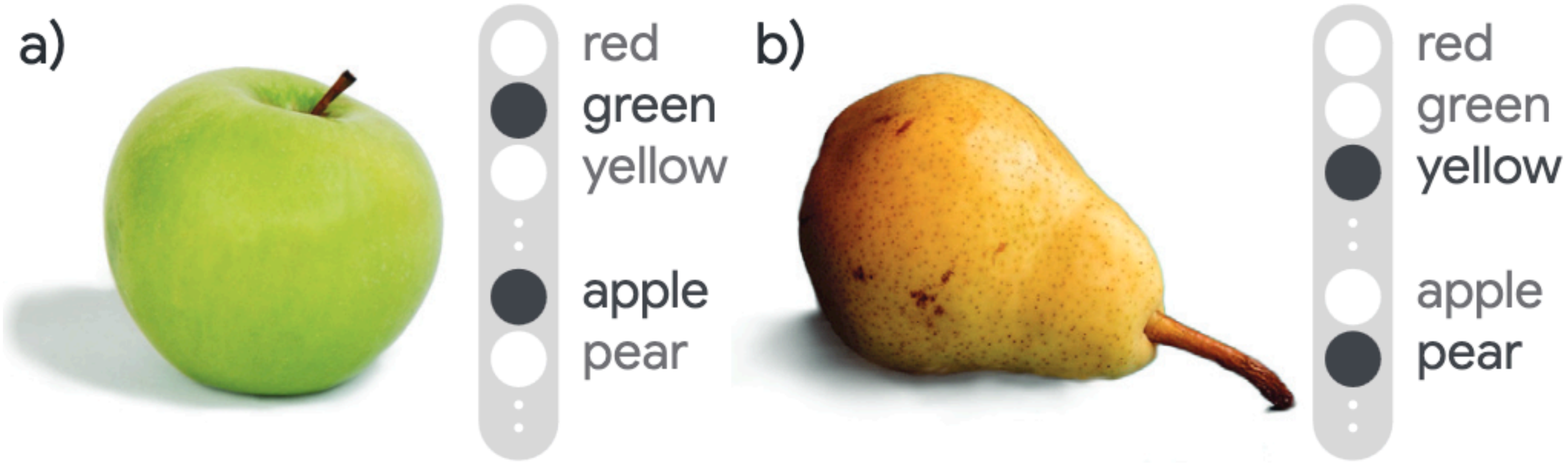





Figure: Greff et al. (2020)

# Binding Problem in Distributed Representation

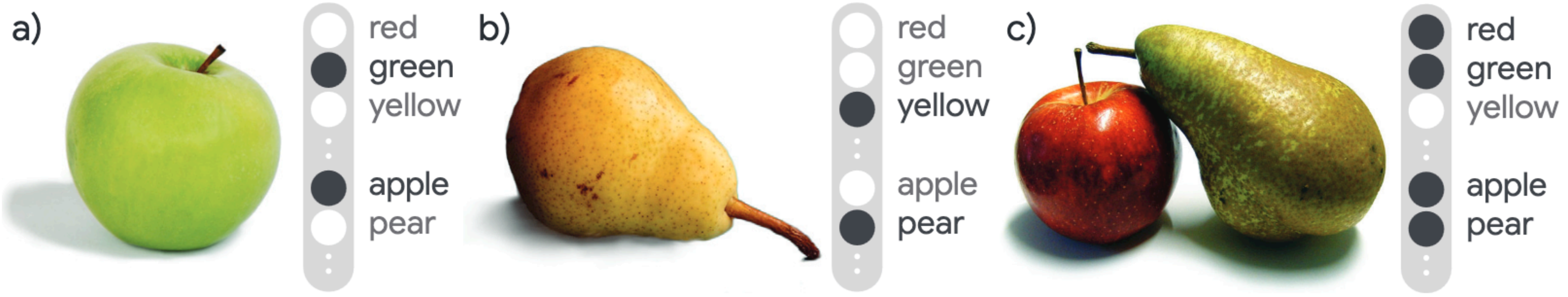
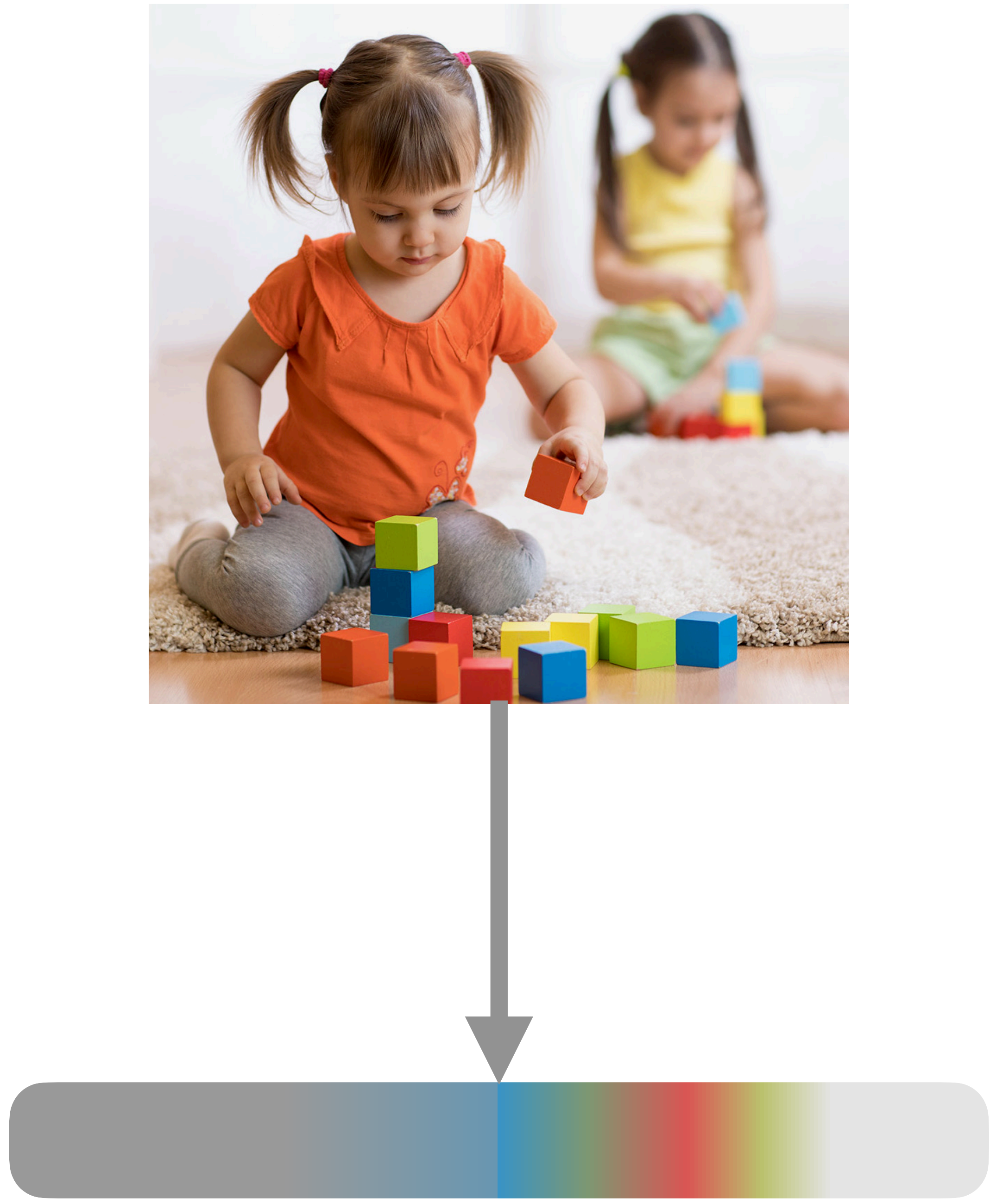


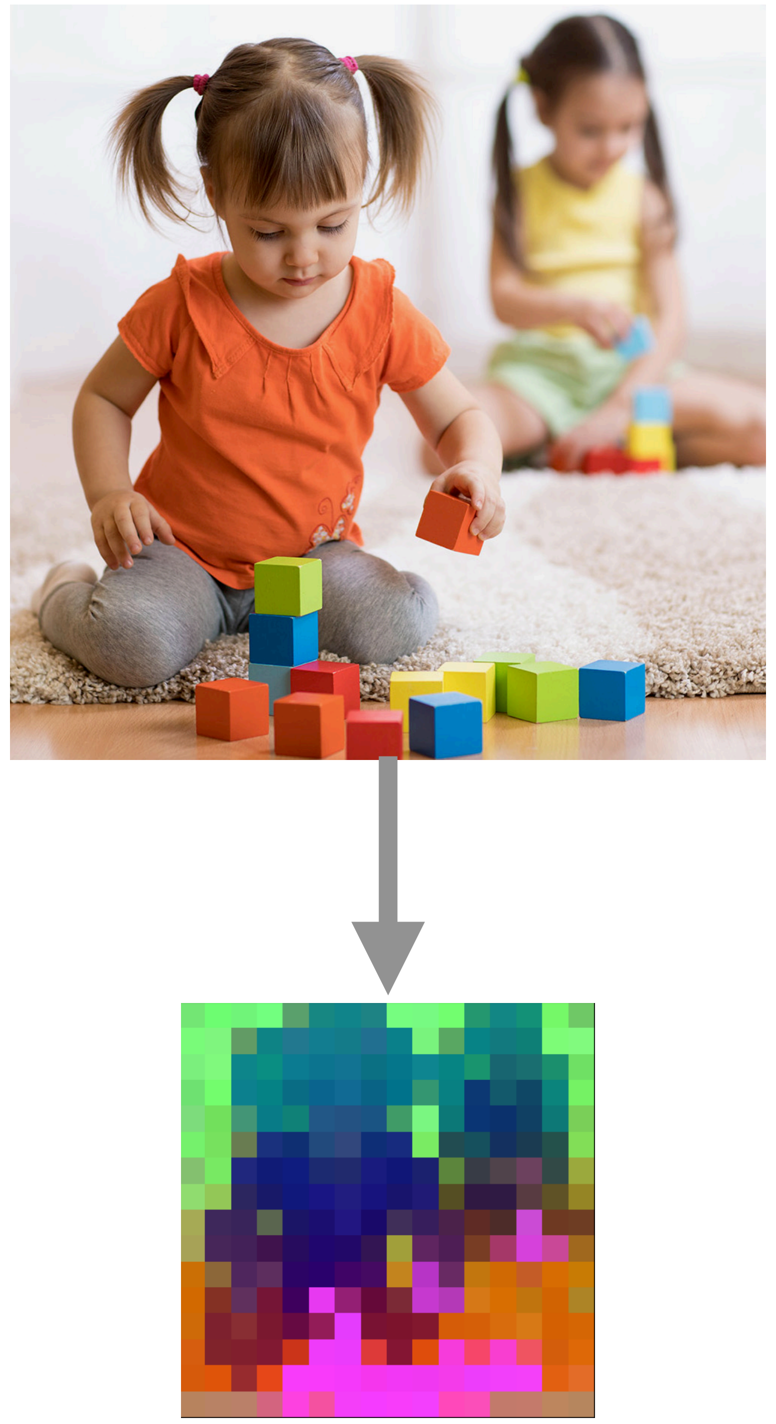



Figure: Greff et al. (2020)

# Representation Learning from Pixels



Single vector



Dense grid of features

?

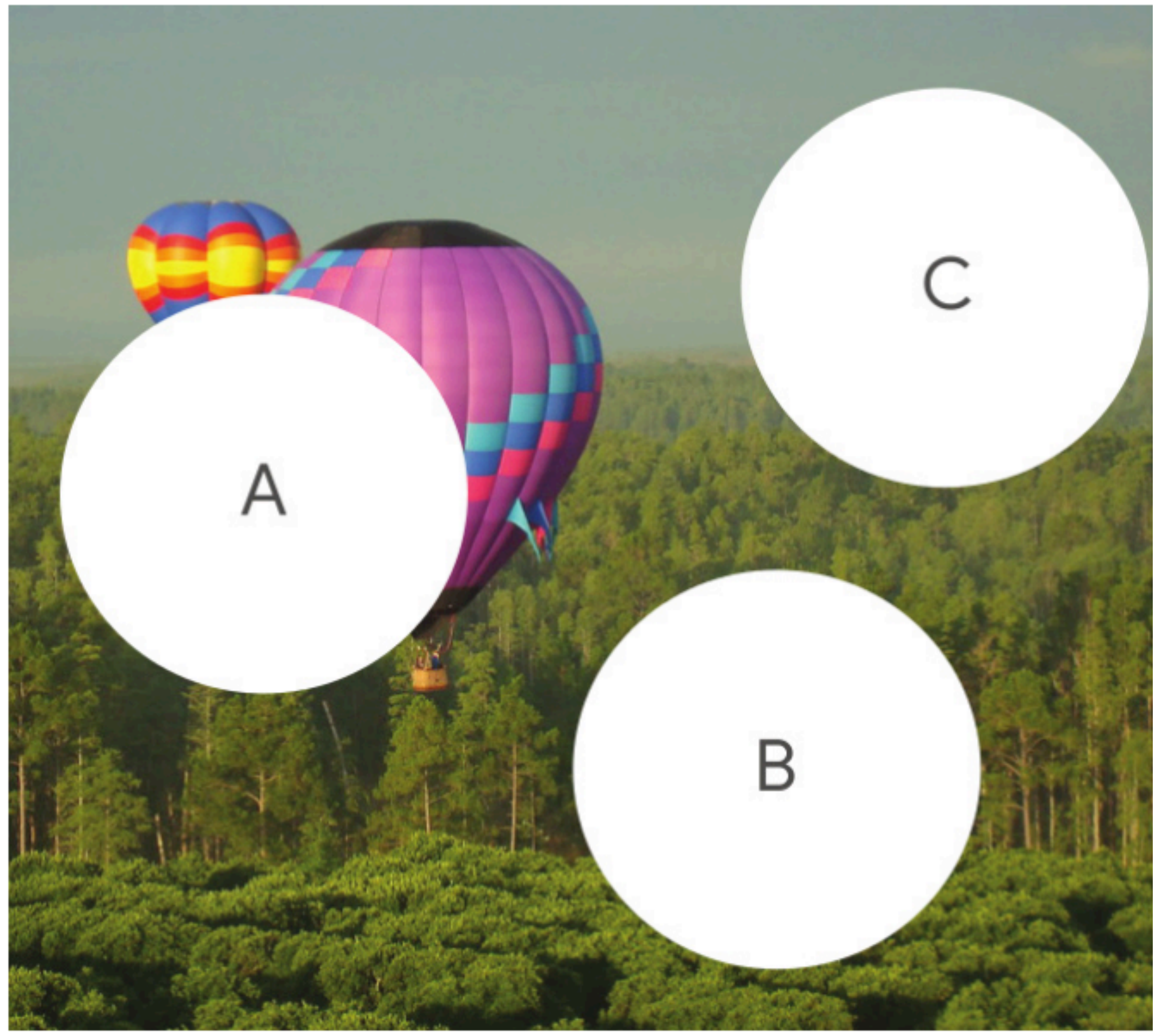
What if we can learn representations that are **structured** similarly to the original scene?

# Do we need structured representations if we have scale?

An image showing 12 plain tea cups from the same set, all identical in style and design.

Here is the updated image showing exactly 12 plain tea cups, all identical in design and style.
Let me know if there's anything else you'd like adjusted!

# How humans structure information about scene?




We group:

- regions that are largely independent of their context
- regions that exhibit strong internal predictive structure

**Objects** are good candidates for both!

# Objects are building blocks of the visual scene?

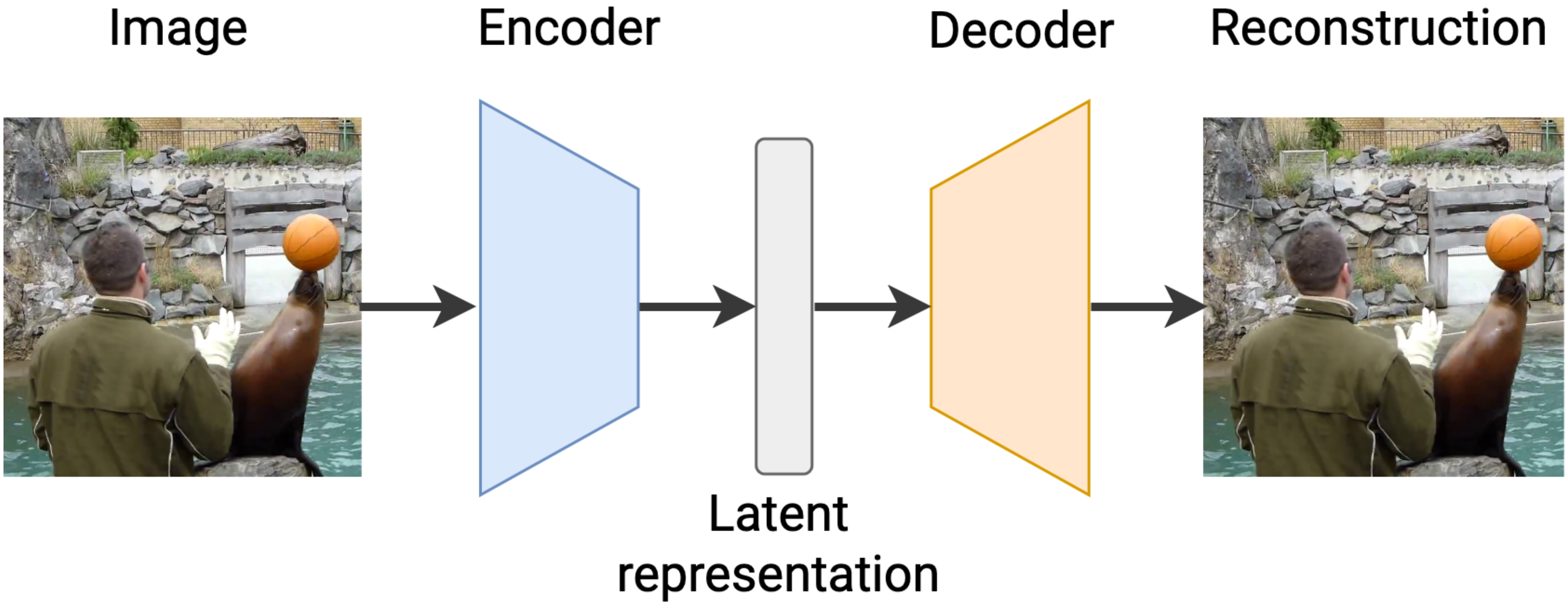
- Instance segmentation and tracking
- Visual reasoning and planning
- Combinatorial generalization

# Unsurevised Object-Centric Representations

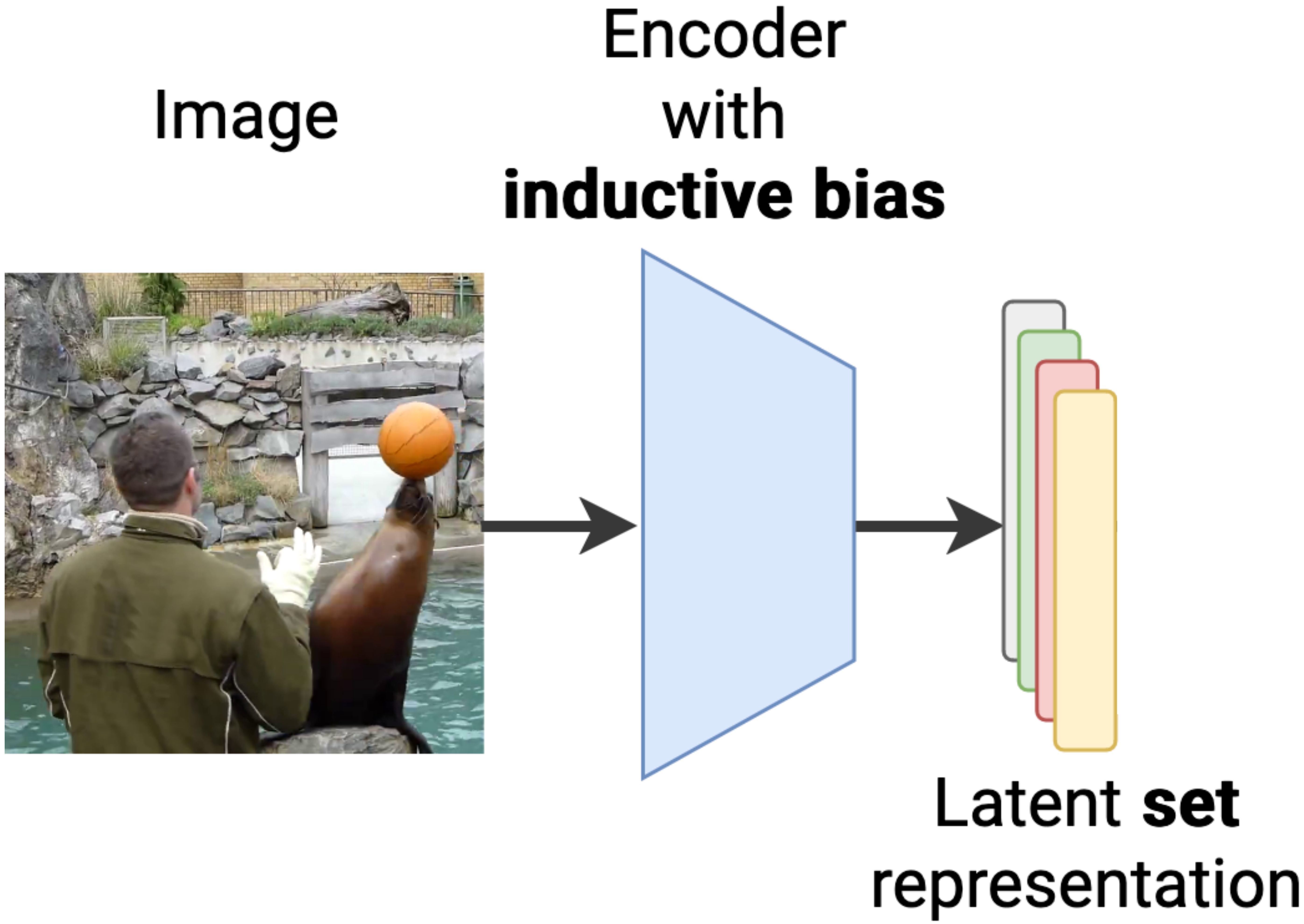
Object-centric representations

- Different objects are represented by different vectors
- Those vectors are grounded on particular image segments
- Trained end-to-end with **architectural inductive biases** and **self-supervision objectives**

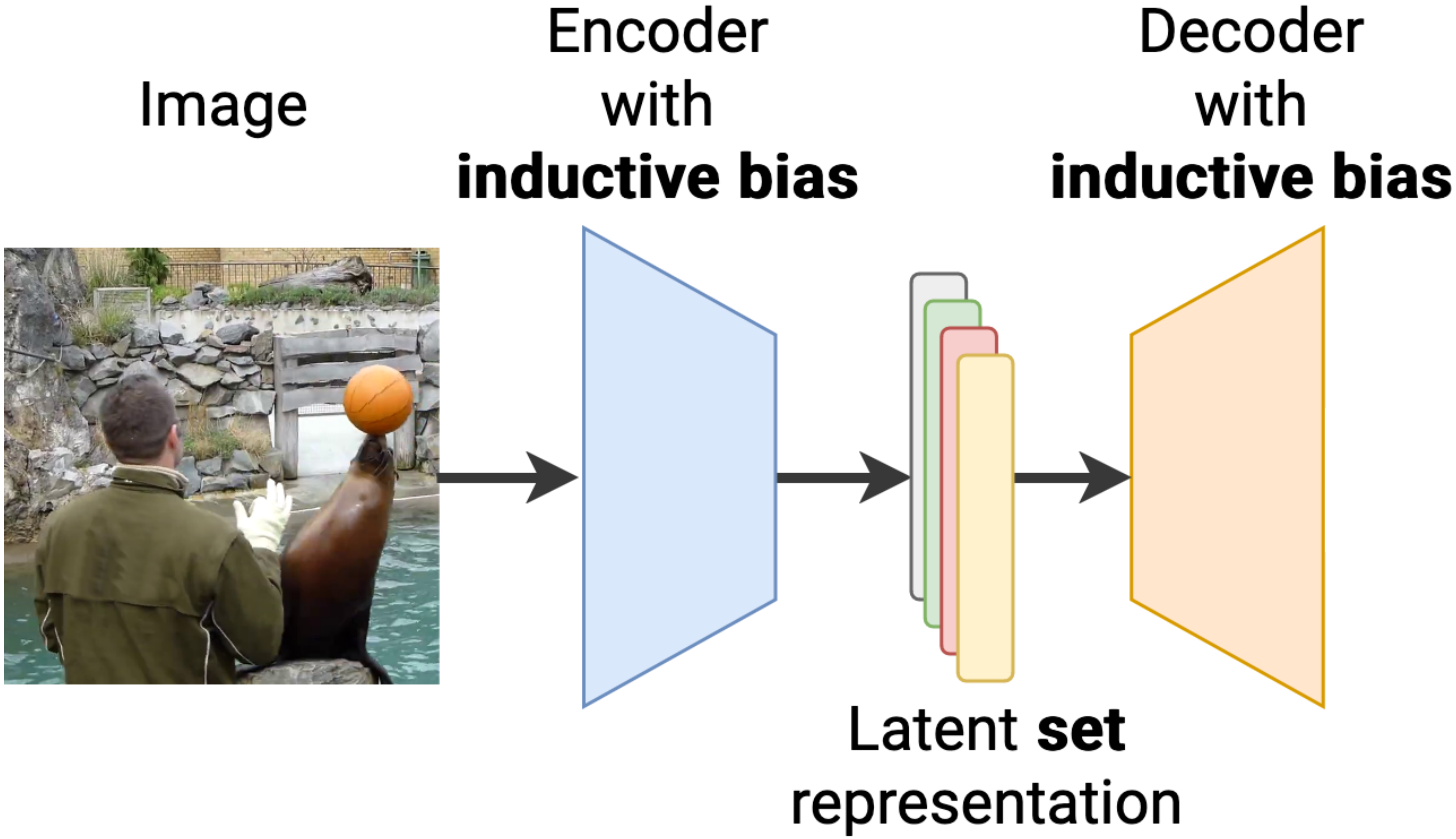
# Unsupervised Representation Learning



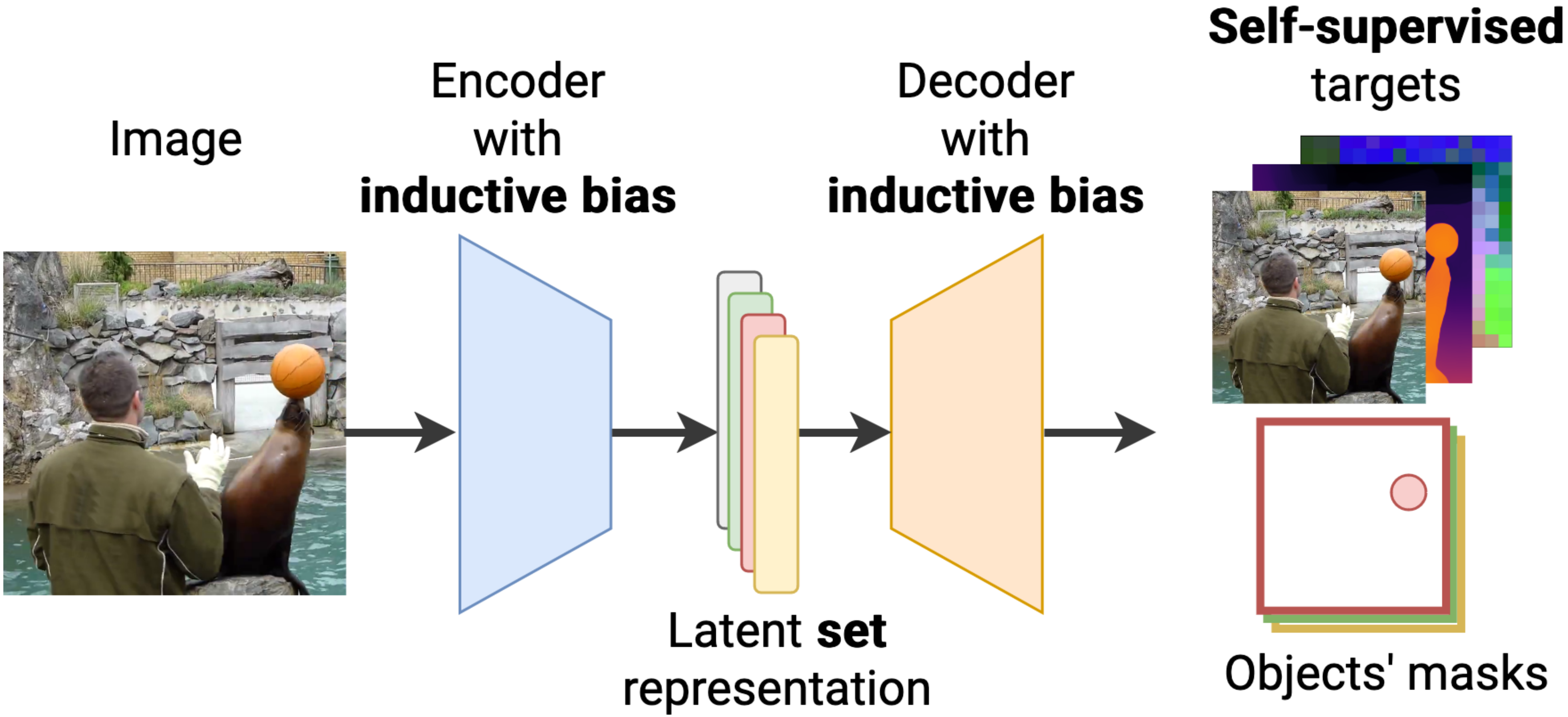
# Object-Centric Representation Learning



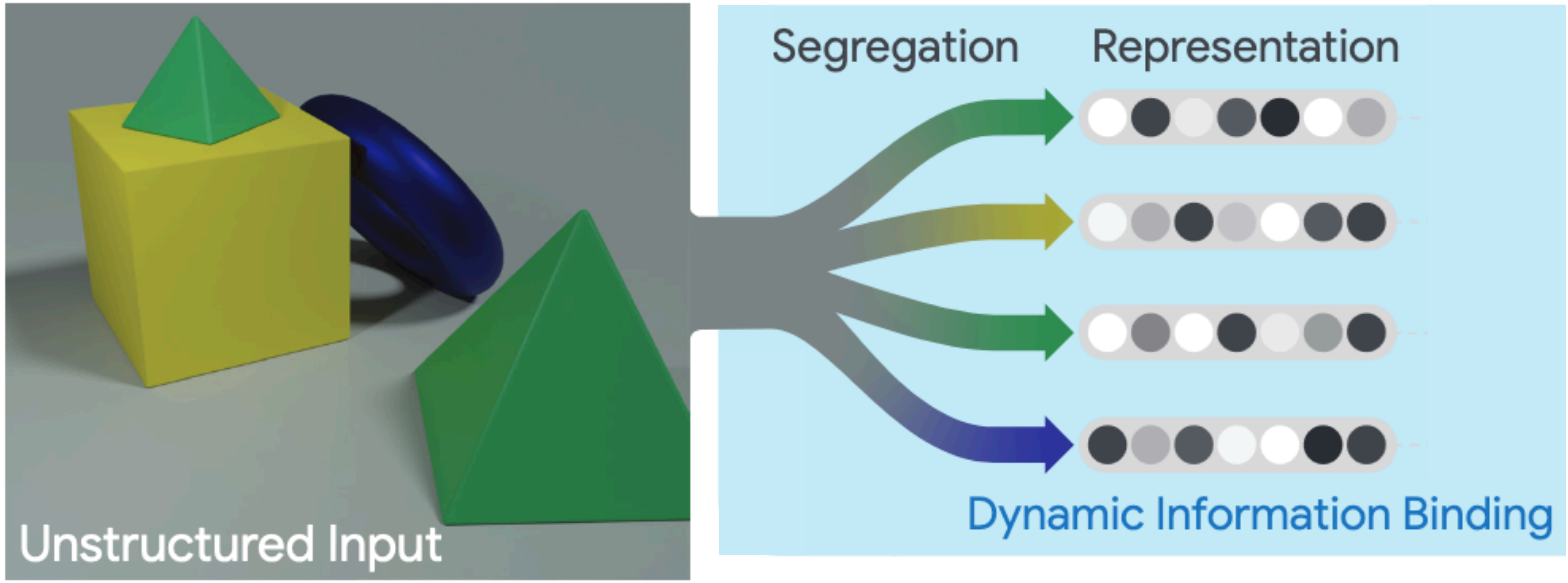
# Object-Centric Representation Learning



# Object-Centric Representation Learning



# Scene Decomposition Into Objects



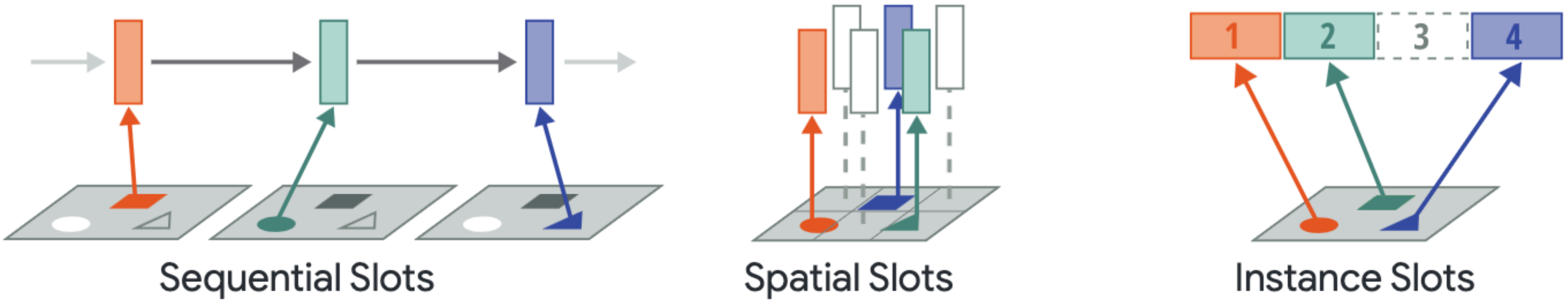

- Dense pixels / features should be separated into discrete set of vectors or *slots*

- *Routing problem:* which vector is responsible for which object?



Figure: Greff et al. (2020)

# Different Ways to Decompose



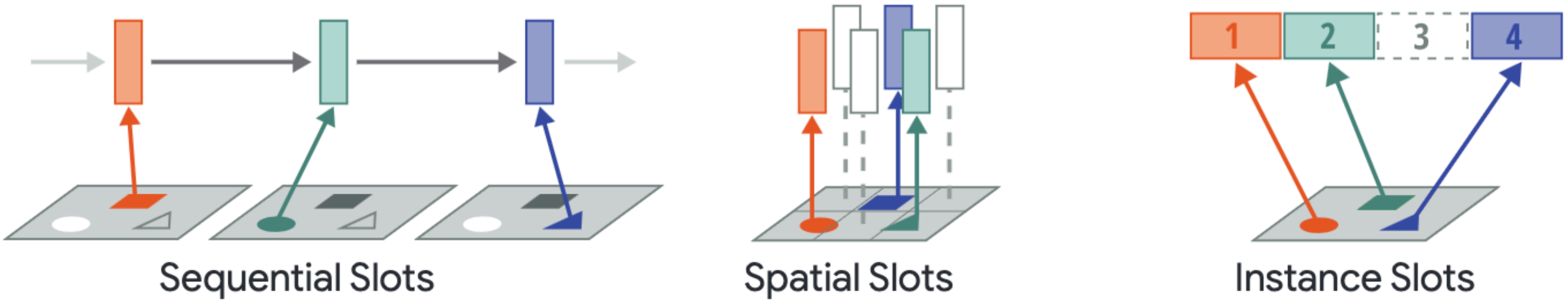

**Encoder inductive biases** could be categorised in terms of encoder outputs named *slots*:

- Sequential slots → ordered **sequence** of vectors
- Spatial slots → sparse **grid** of vectors
- Instance slots → permutation-invariant **set** of vectors



Figure: Greff et al. (2020)

# Different Ways to Decompose




AIR [Eslami et al., 2016]

SQAIR [Kosiorek et al, 2018]

MONet [Burgess et al., 2019]

SPAIR [Crawford & Pineau, 2019]

SPACE [Lin et al., 2020]

SCALOR [Jiang et al., 2020]

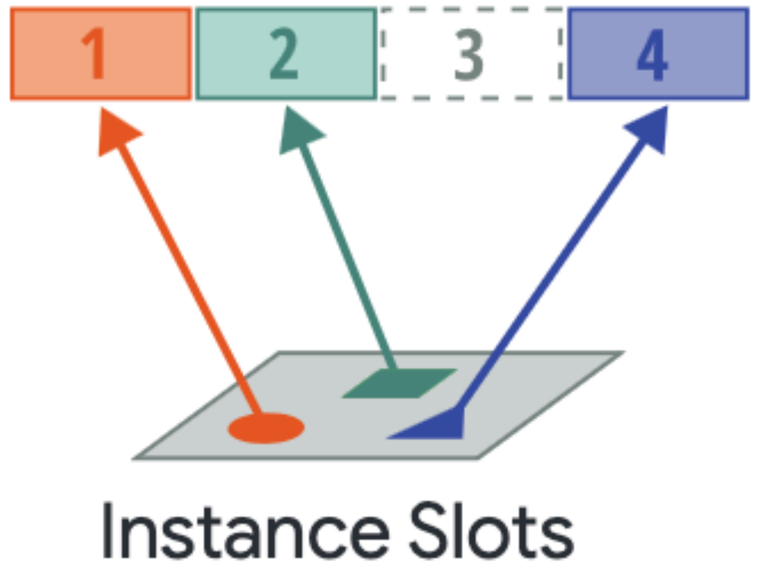
**SA** [Locatello el al., 2020]

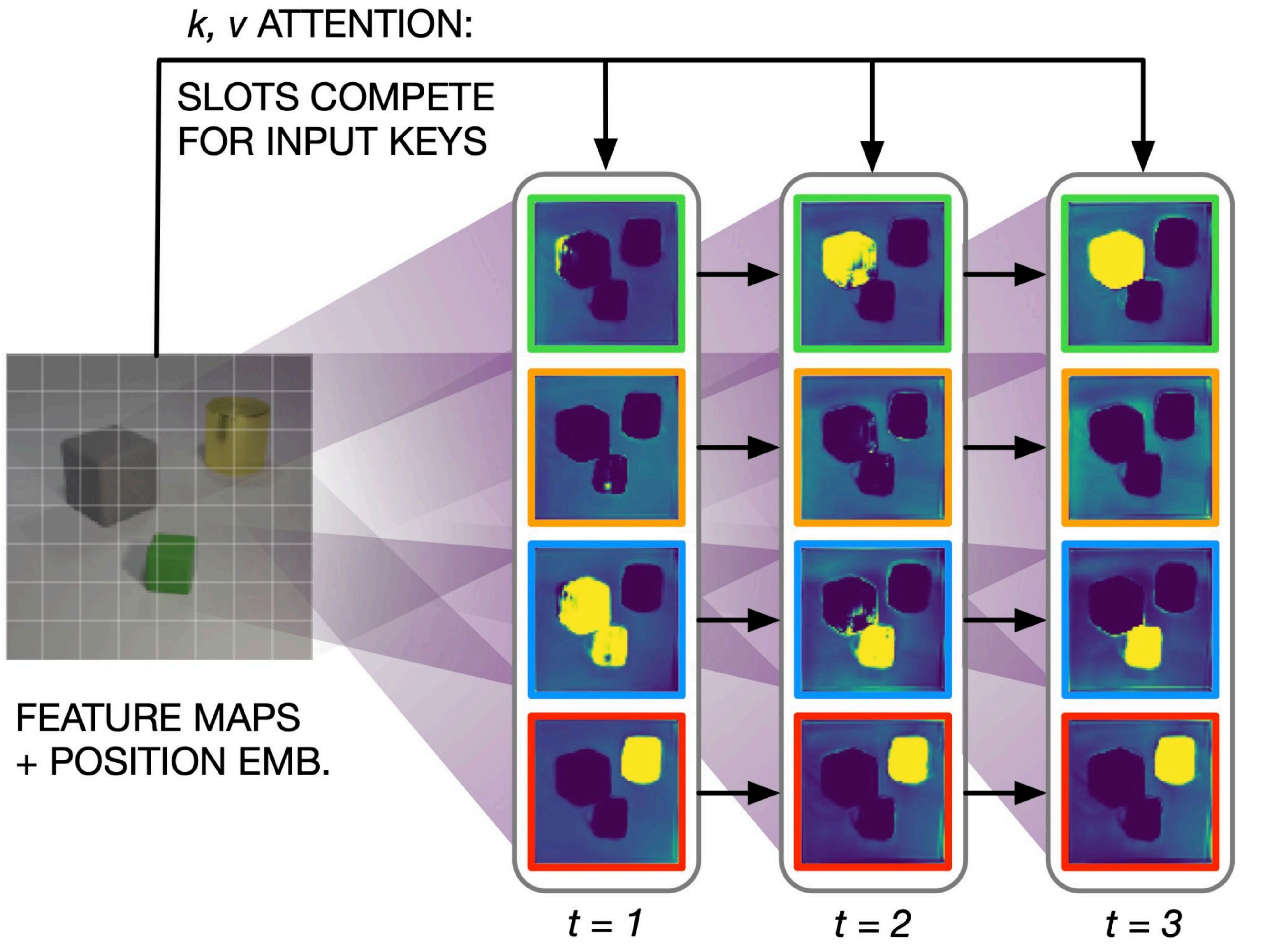
DINOSAUR [Seitzer el al., 2023]



Figure: Greff et al. (2020)

# Instance Slots: Slot Attention Encoder



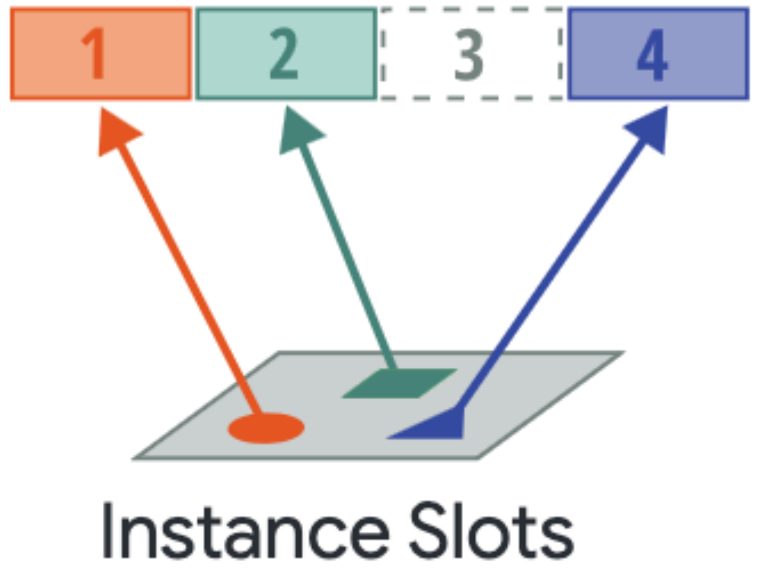

Slot Attention Pseudocode

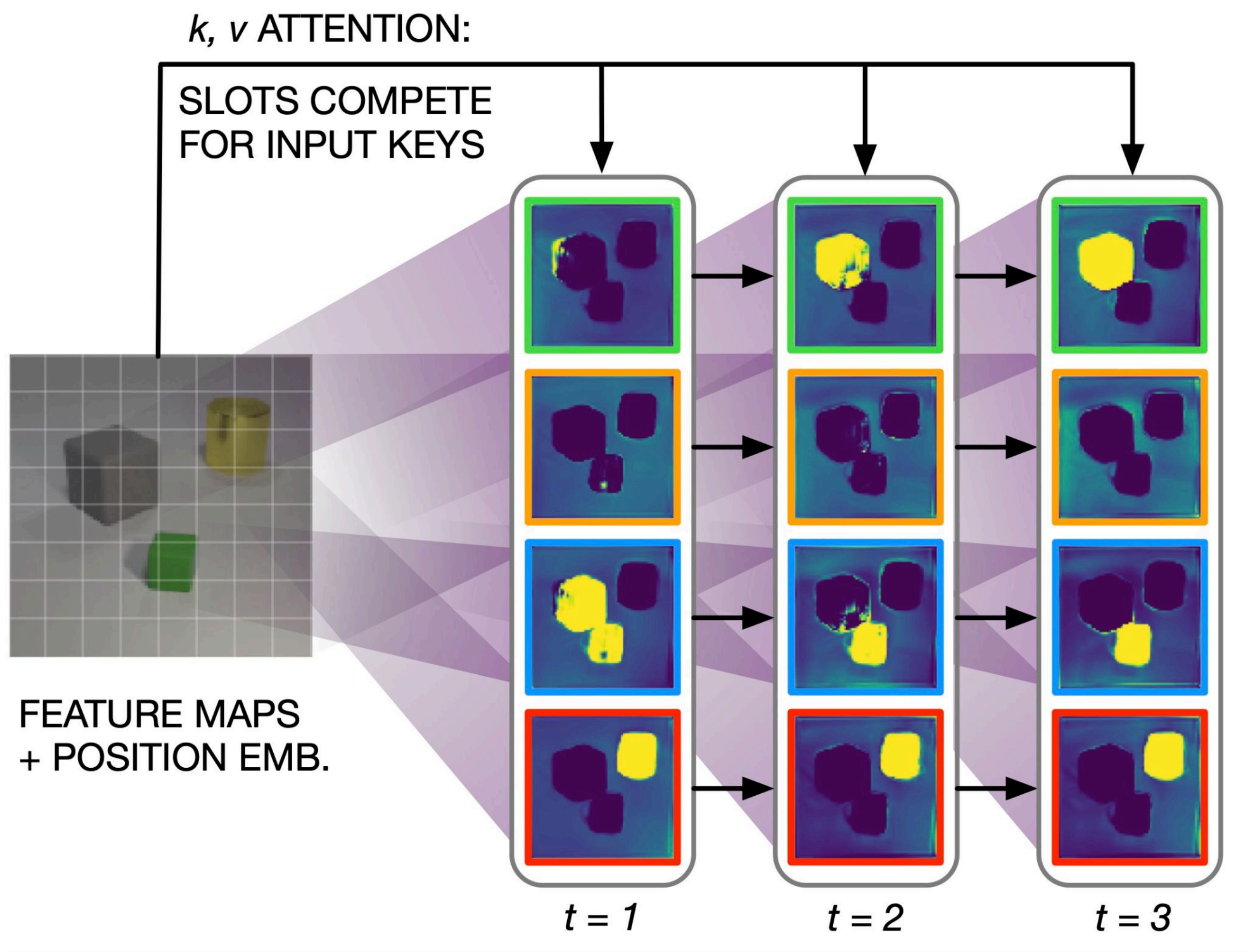
```
# inputs: feature maps + position embedding
slots ~ normal(mean, std)
for t = 0 … T:
    scores = dot(k(inputs), q(slots))
    weights = softmax (scores / t, axis='slots')
    updates = weighted_mean (weights, v(inputs))
    slots = gru(slots, updates) # GRU update
```

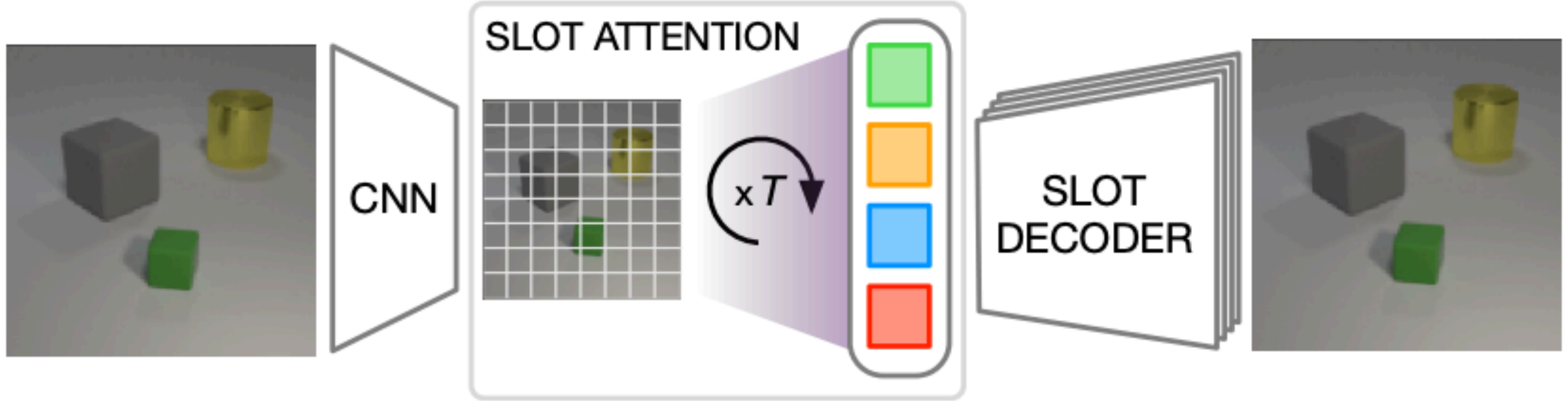


Object-Centric Learning with Slot Attention [Locatello et al.]

# Instance Slots: Slot Attention Training



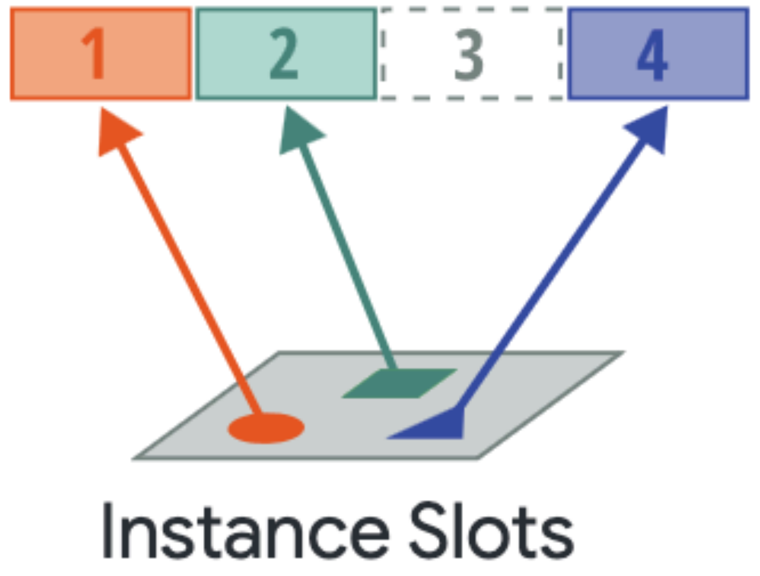



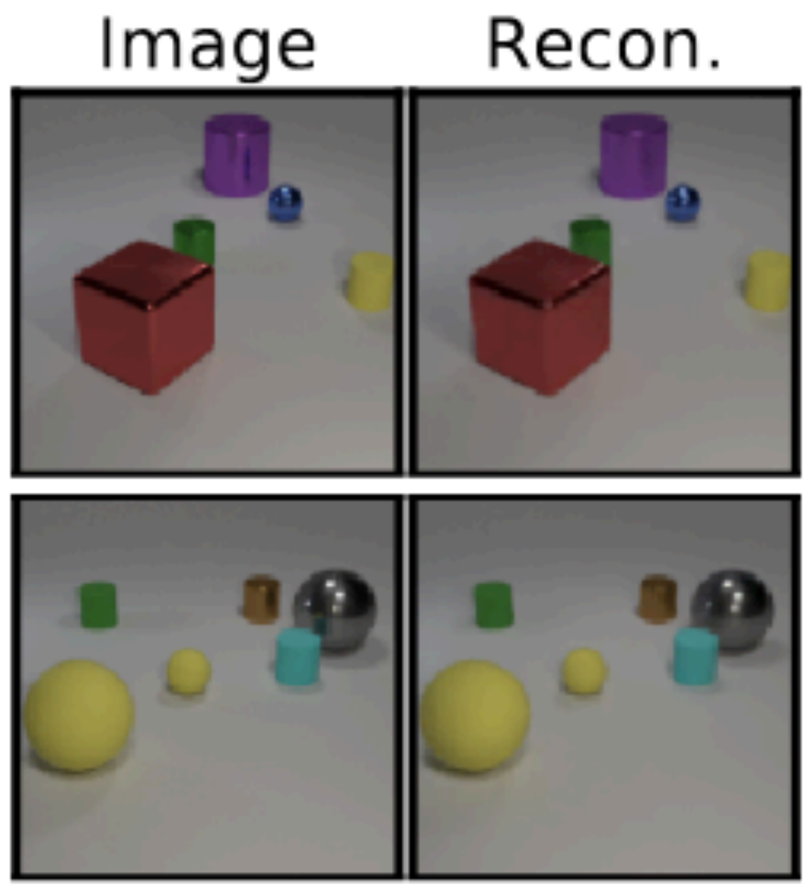






Object-Centric Learning with Slot Attention [Locatello et al.]

# Instance Slots: Slot Attention Results






Object-Centric Learning with Slot Attention [Locatello et al.]

# Discovering Object-Centric Structure from the Real-World Video Data

# Object-Centric Learning for Real-World Data

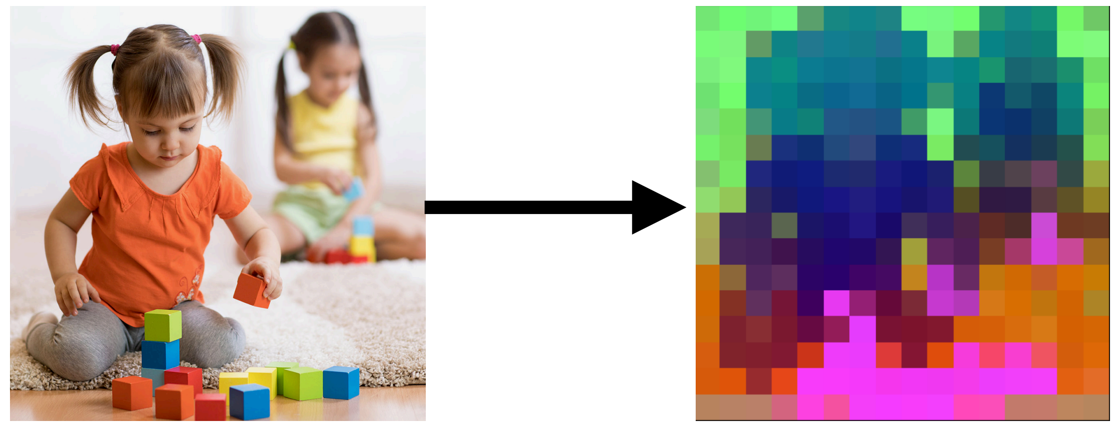

**Image reconstruction** as the target **is not enough** for grouping real-world scenes

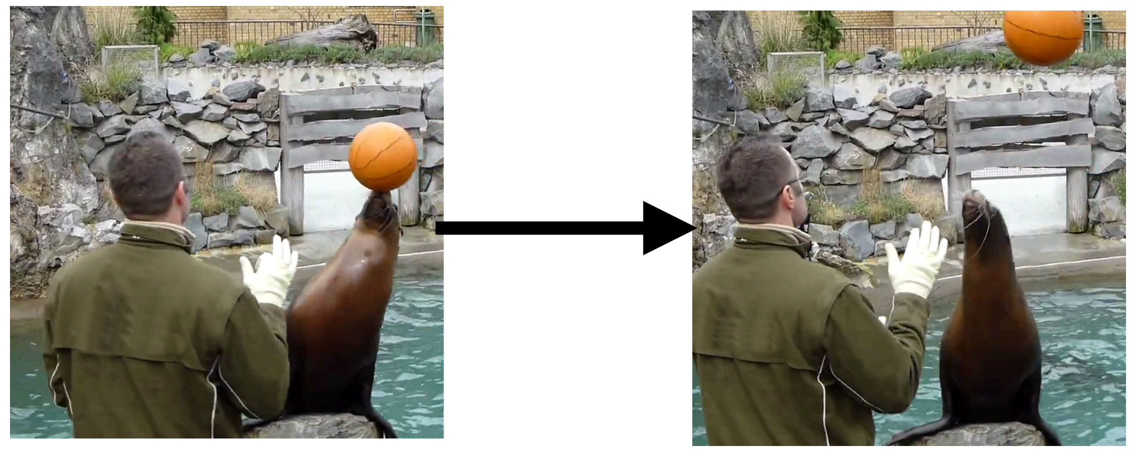
# Self-supervised Object-Centric Objectives

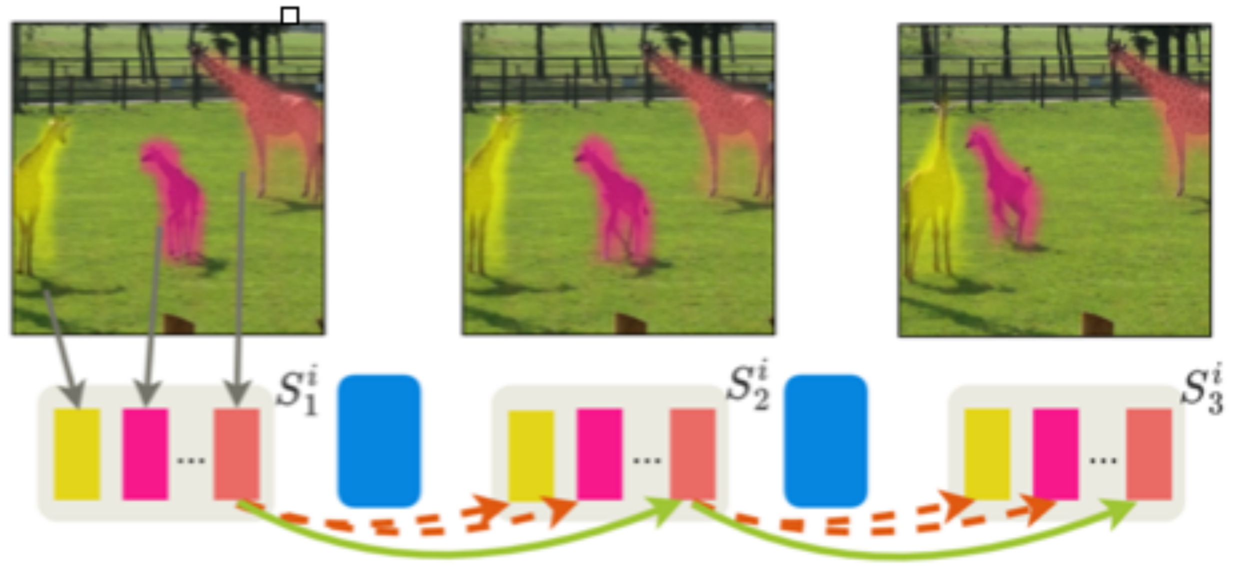
Semantics Reconstruction

Motion Prediction

Identity Preservation





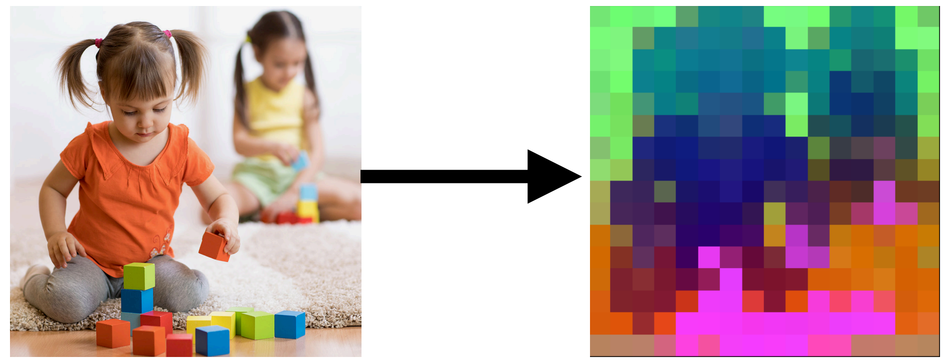


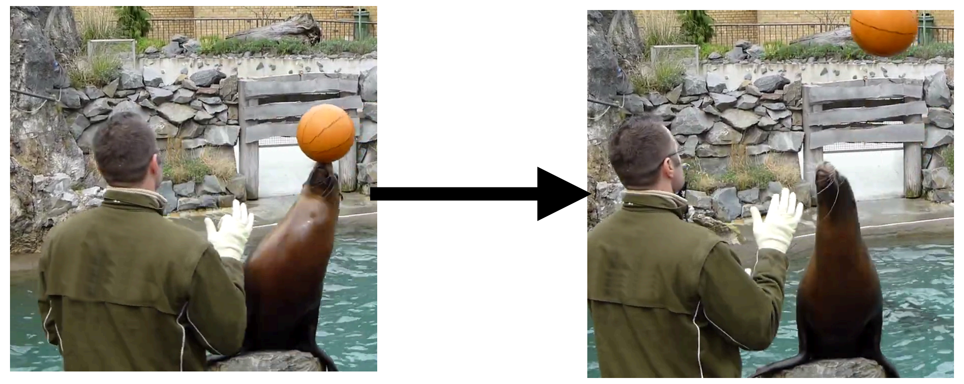
# Self-supervised Object-Centric Objectives

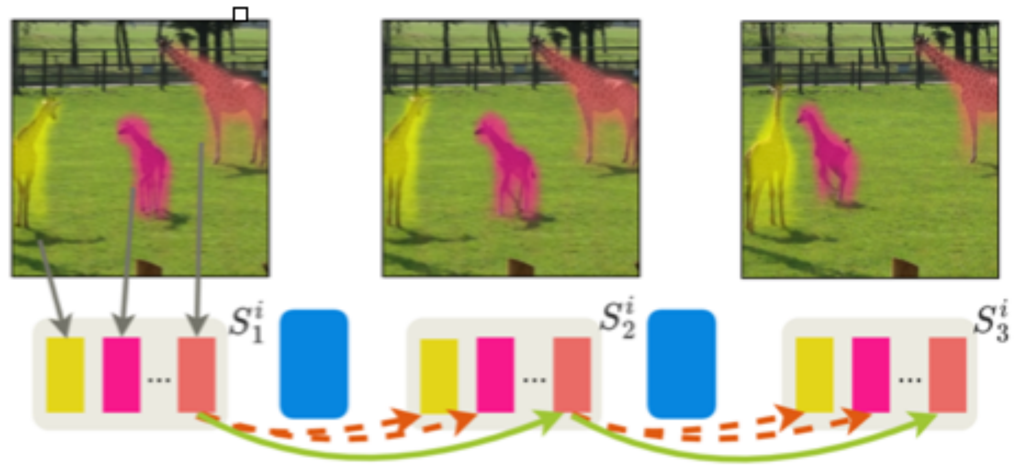
Semantics Reconstruction

Motion Prediction

Identity Preservation







DINOSAUR
[Seitzer el al., 2023]

VideoSAUR
[Zadaianchuk et al., 2023]

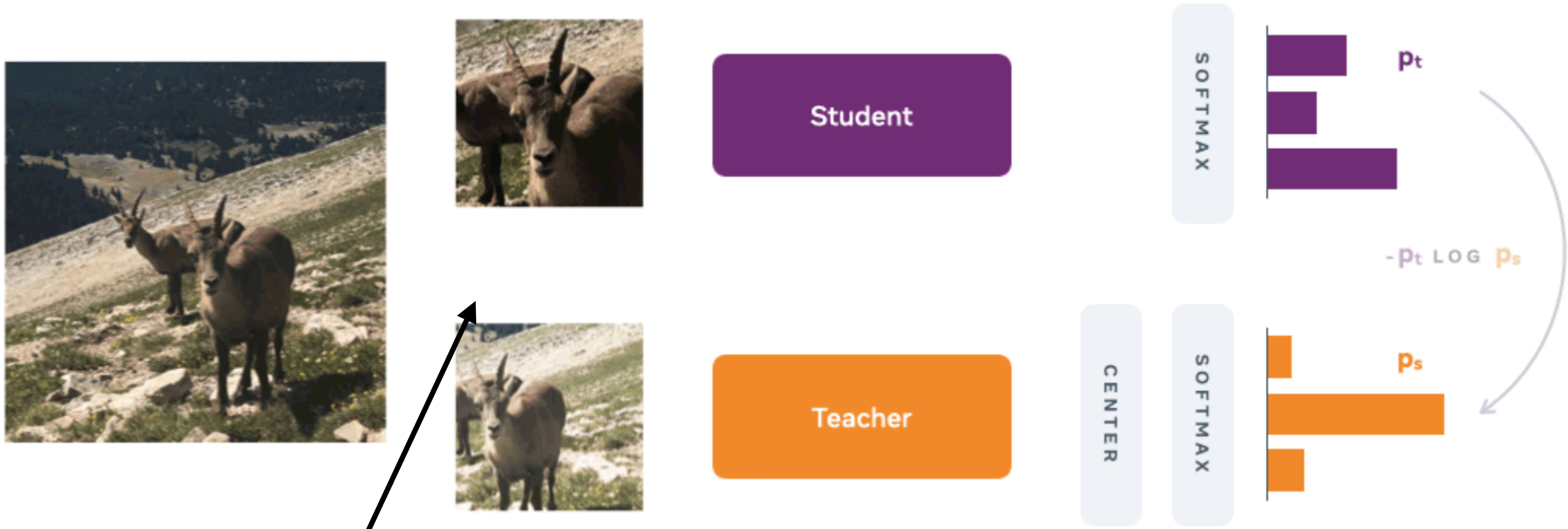
SlotContrast
[Manasyan et al., 2024]

# Self-supervised Semantic Features



Emerging Properties in Self-Supervised Vision Transformers [Caron el al.]

# Self-supervised Semantic Features




Cross

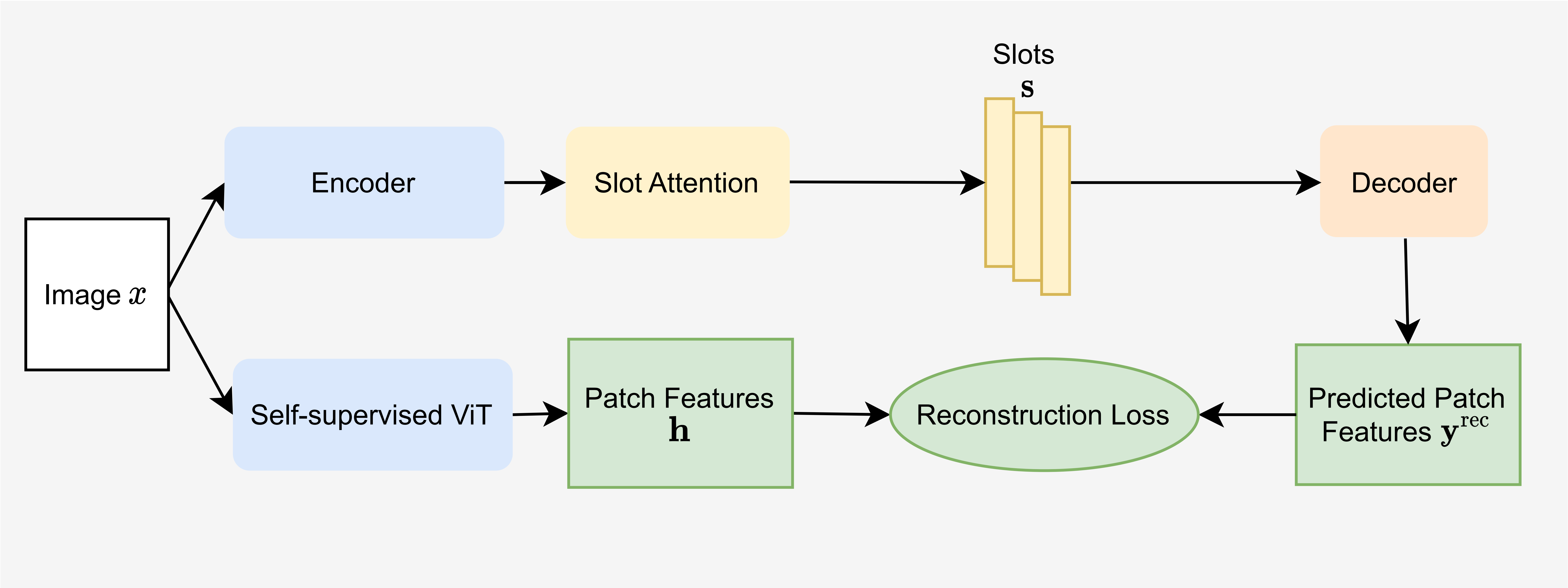
Entropy

Multi-crop augmentations strategy:

many small crops for student &
larger crop for teacher

Emerging Properties in Self-Supervised Vision Transformers [Caron el al.]

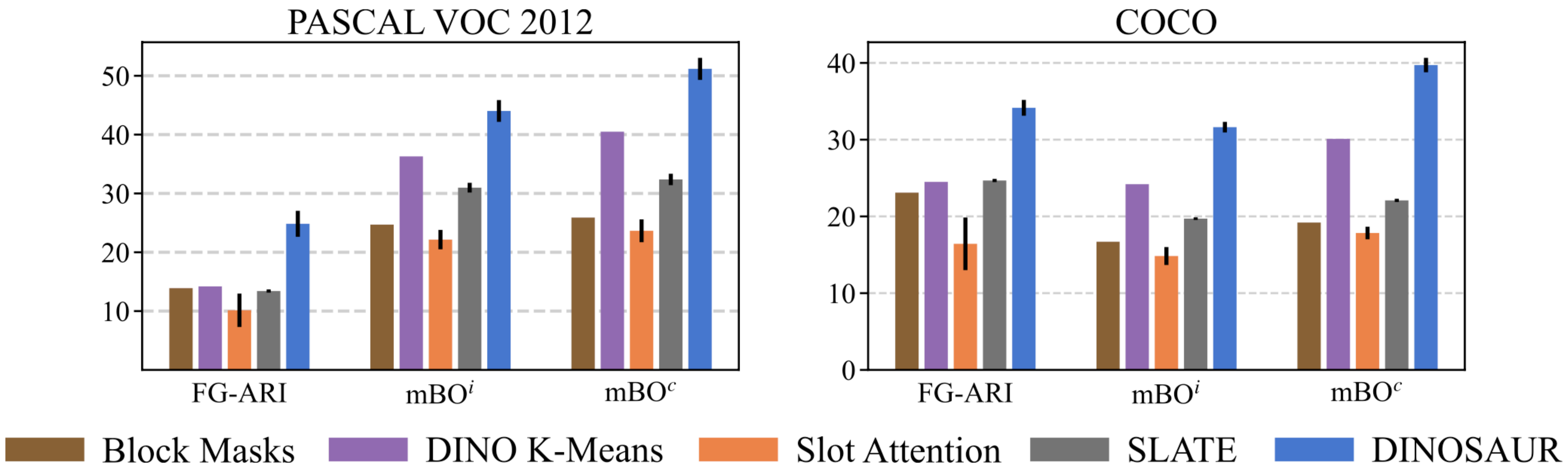
# DINOSAUR: Self-supervised Features as Targets






*e.g. DINO by Cathilde Caron at el.

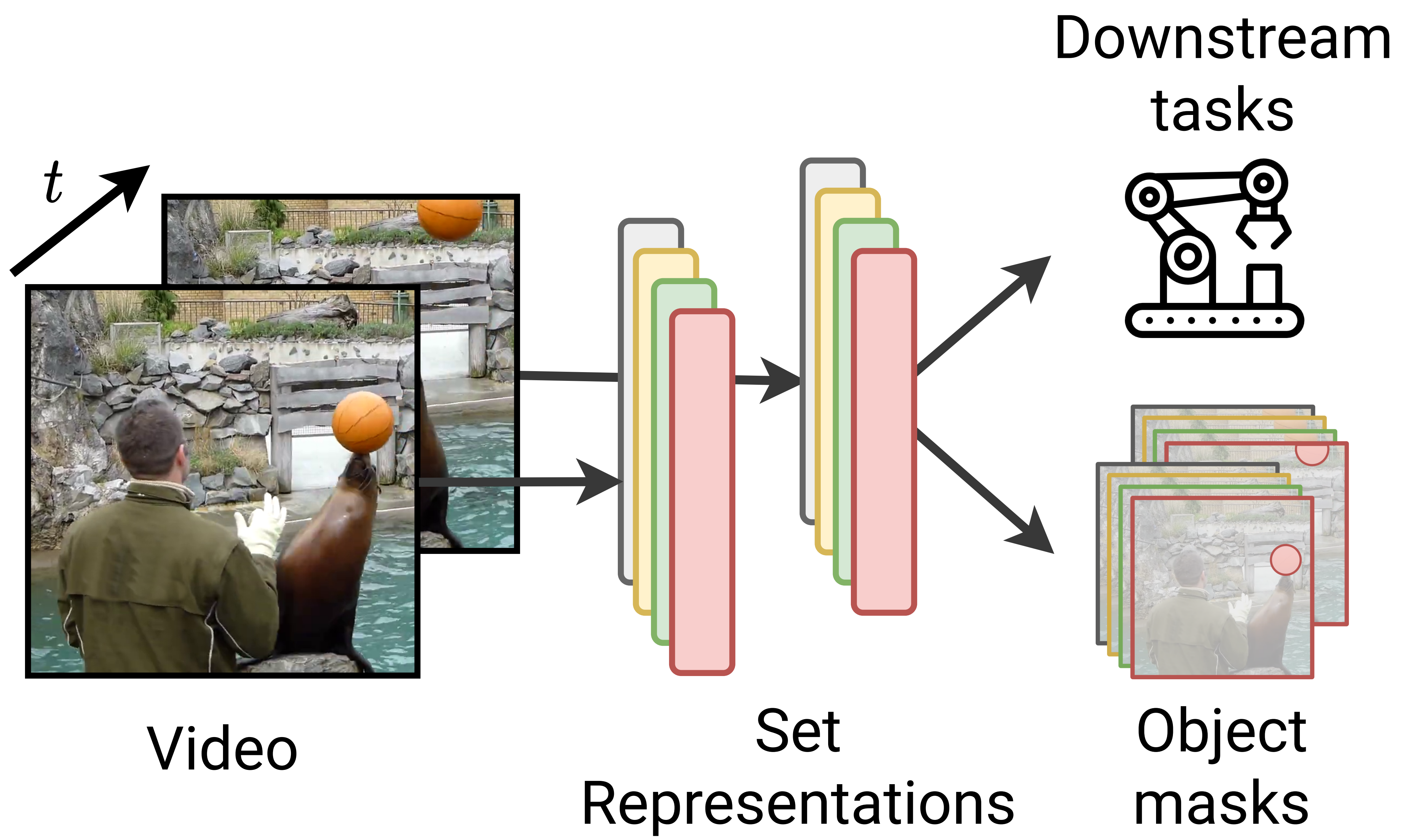
# DINOSAUR Results

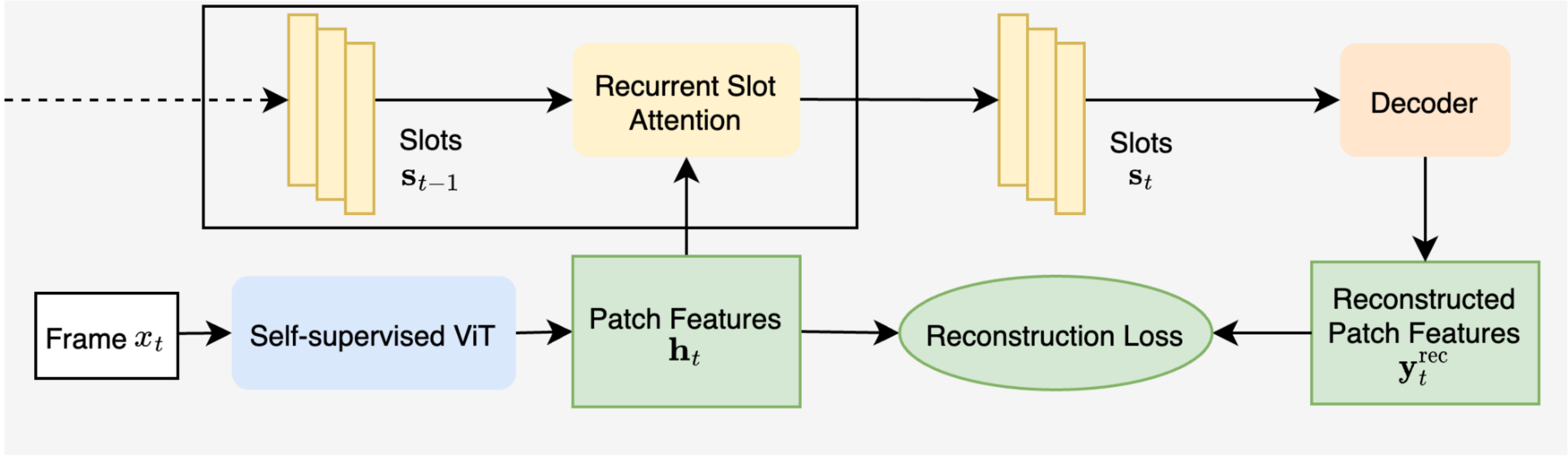



“Bridging the Gap to Real-World Object-Centric Learning”, ICLR 2023
Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, …, Francesco Locatello

# Can we extract even better targets from videos?



# Recurrent Slot Attention for Slot Consistency



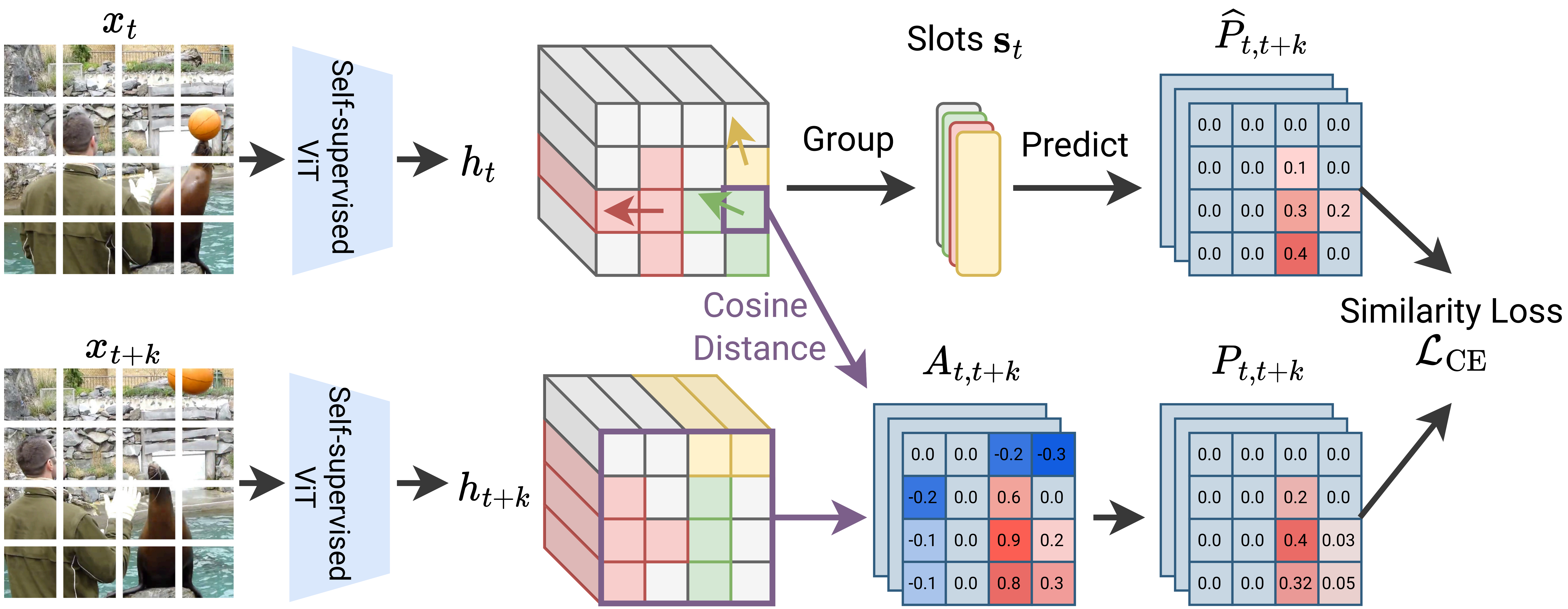

- DINOSAUR training objective: reconstruction of the current frame features
- Recurrent SA from SAVi[9]: connects slots from different frames via initialisation of SA iteration

How can we facilitate video object discovery from the **temporal structure of the video**?



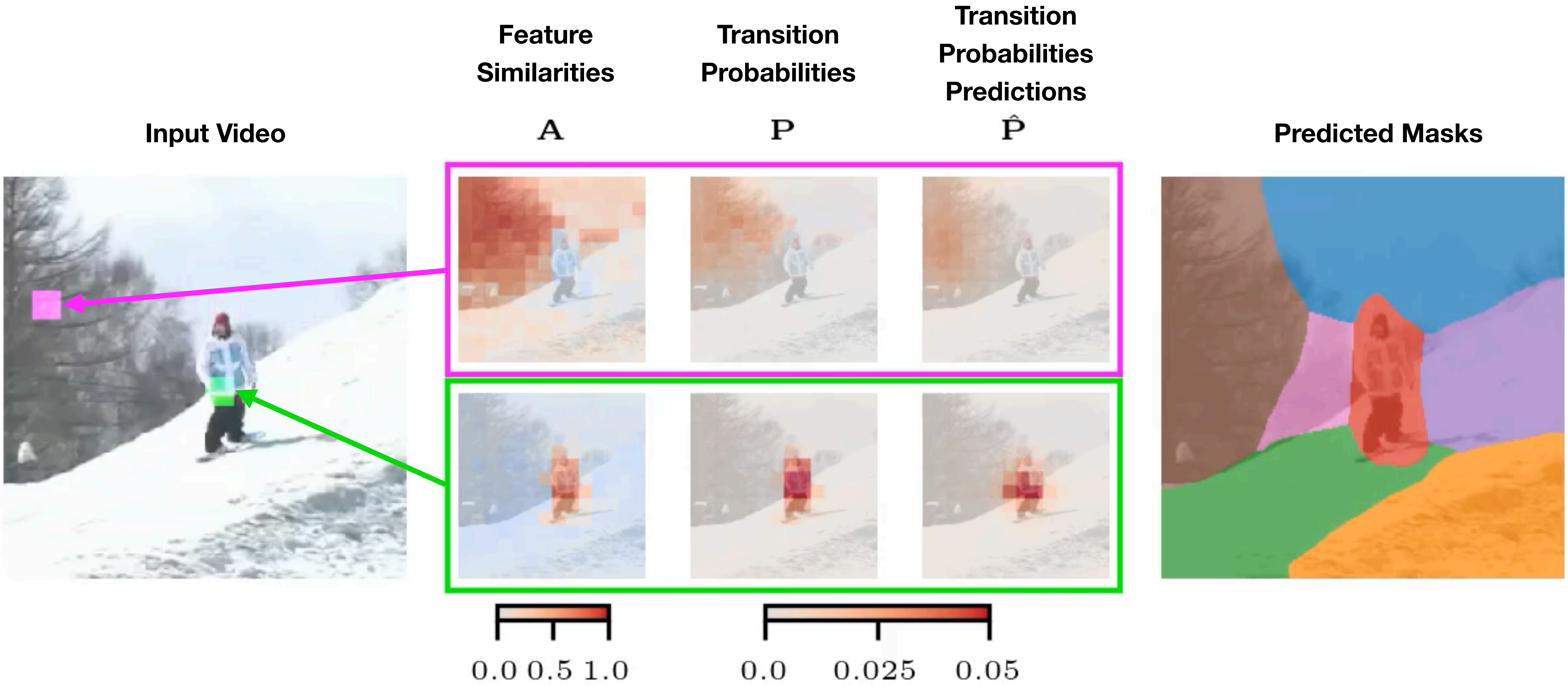
[9] Kipf et al. Conditional Object-Centric Learning from Video. ICLR 2022

# Temporal Features Similarity Prediction

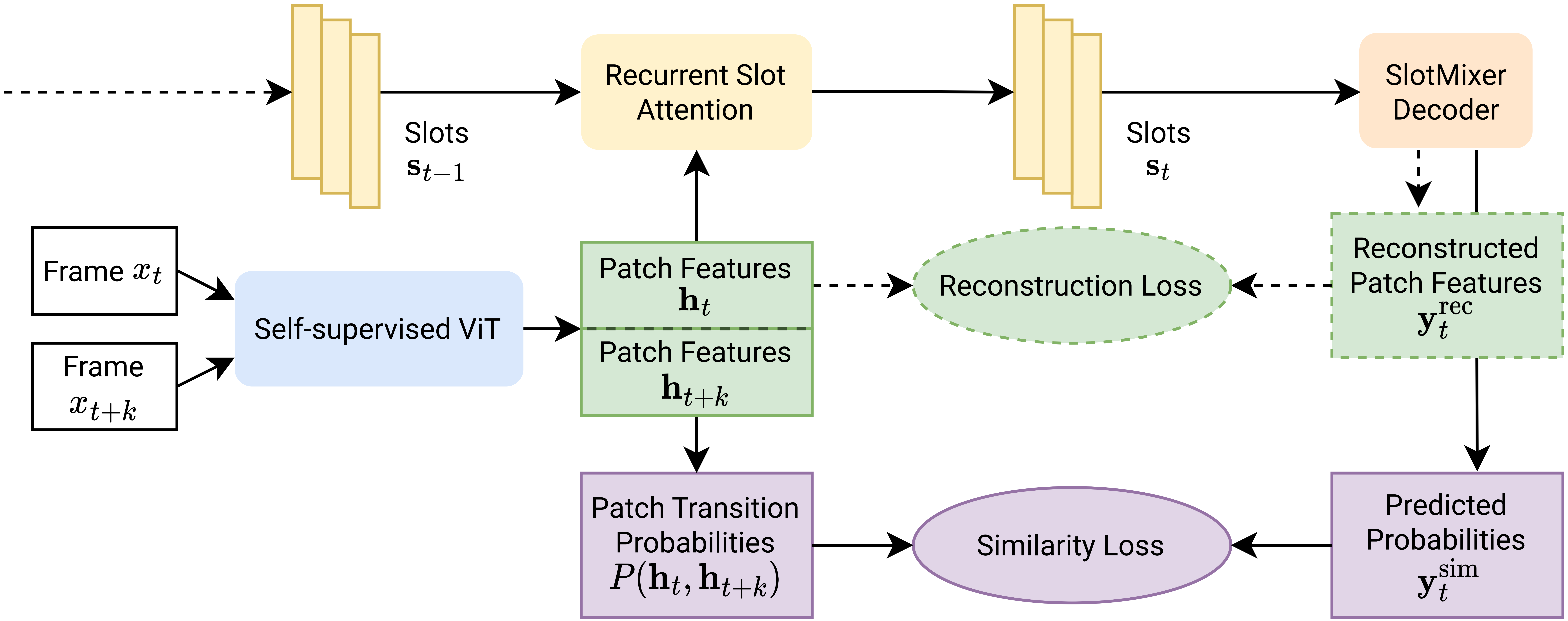



- Successful prediction of the temporal similarity requires **combining semantics and motion information**

# Temporal Features Similarity Prediction

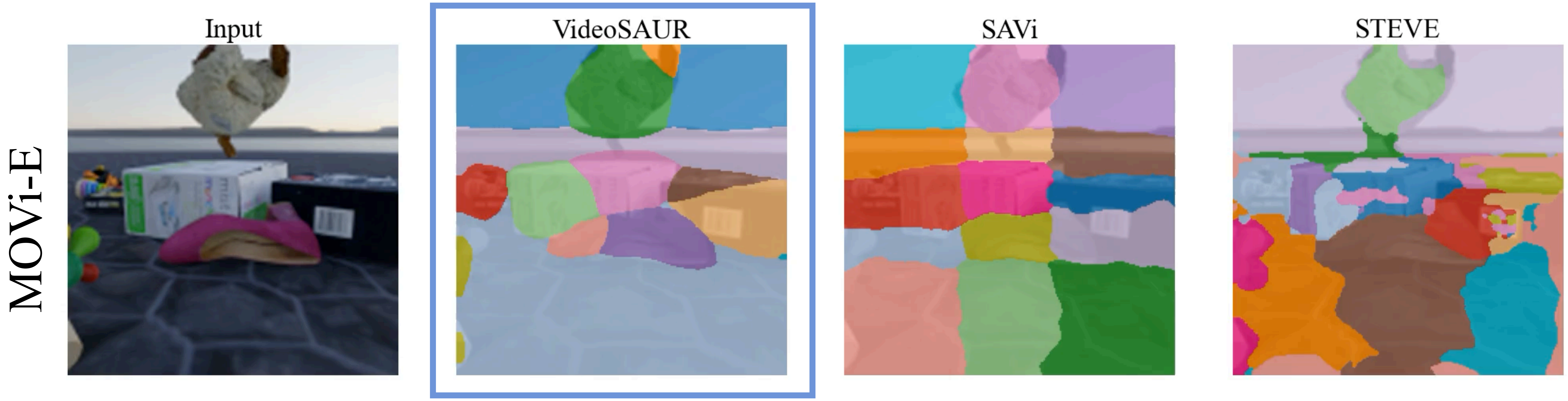


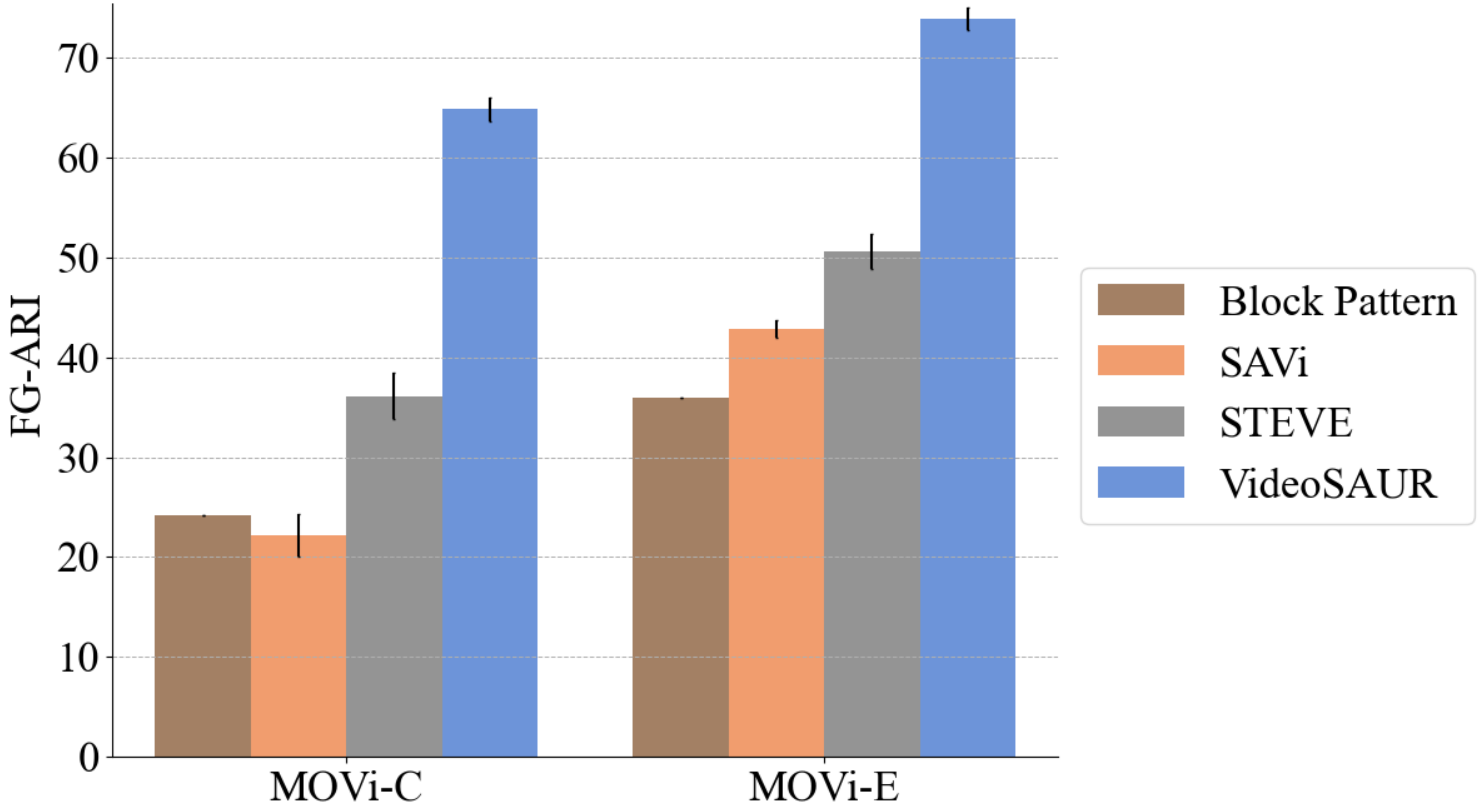
# VideoSAUR



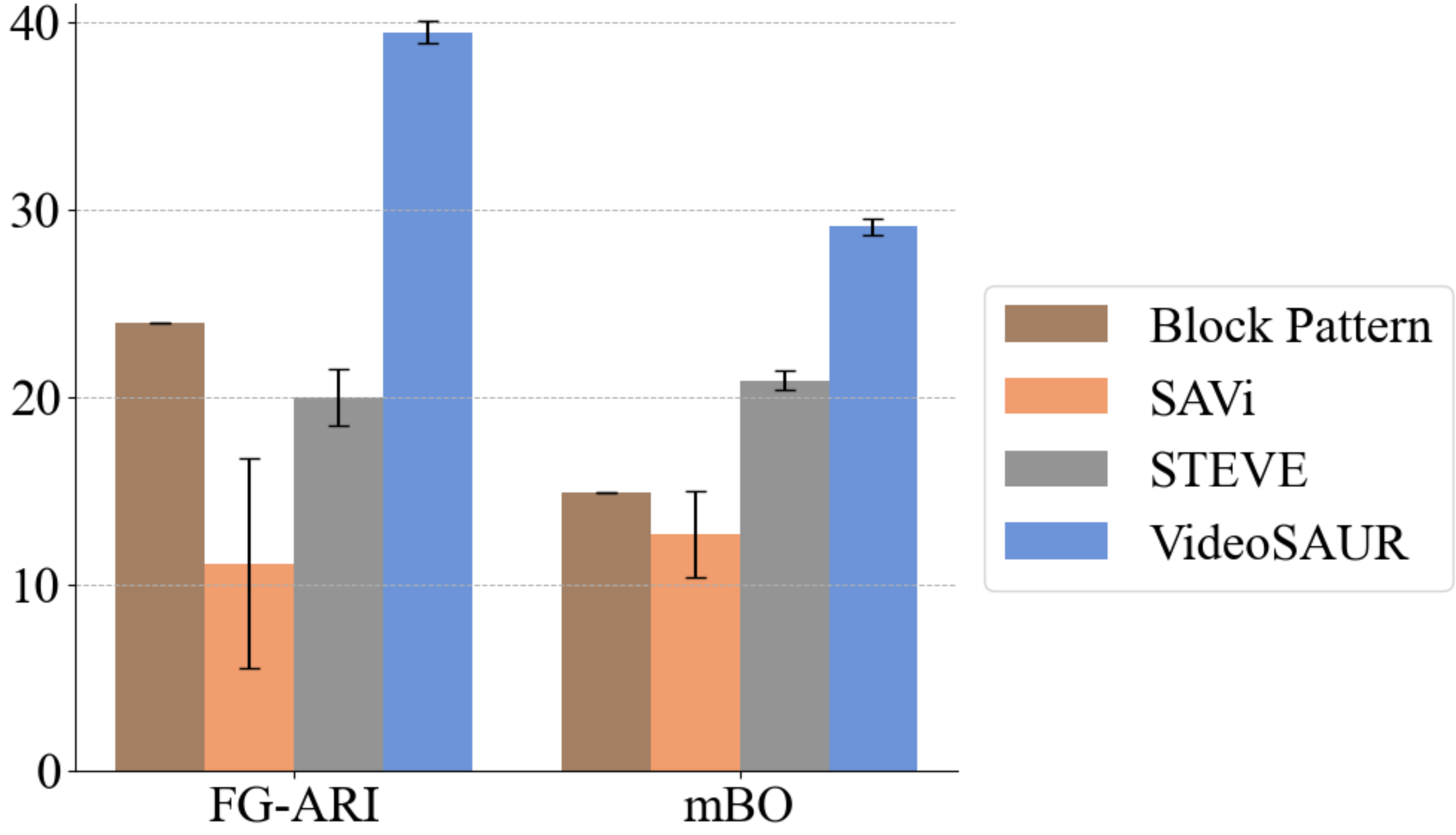
Slots
$\mathbf{s}_{t-1}$
Recurrent Slot Attention
Slots
$\mathbf{s}_t$
SlotMixer Decoder
Frame $x_t$
Frame $x_{t+k}$
Self-supervised ViT
Patch Features $\mathbf{h}_t$
Patch Features $\mathbf{h}_{t+k}$
Reconstruction Loss
Reconstructed Patch Features $\mathbf{y}_t^{\text{rec}}$
Patch Transition Probabilities $P(\mathbf{h}_t, \mathbf{h}_{t+k})$
Similarity Loss
Predicted Probabilities $\mathbf{y}_t^{\text{sim}}$

# VideoSAUR Results on Synthetic Videos



# VideoSAUR Results on Real-World Videos

# Qualitative VideoSAUR Results


VideoSAUR with DINO-v2 features

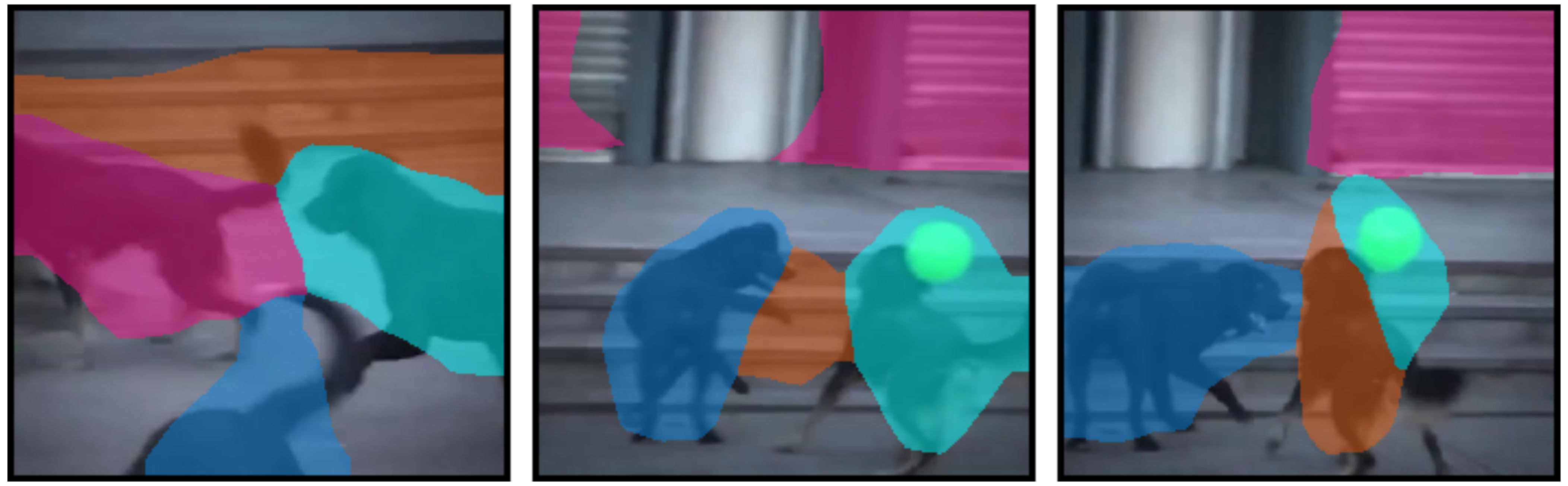
# What about long-term consistency?

**We need to maintain a consistent slot for an object throughout a video sequence.**

**No slot (ID) switches**

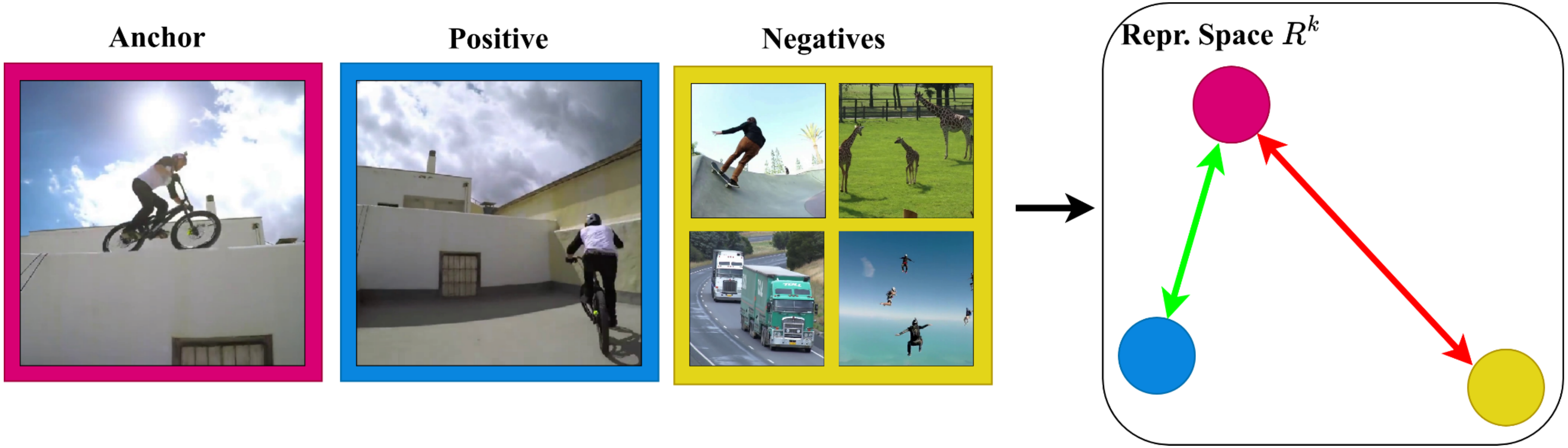

**Reassign original slot:** Reappeared objects should get their original slot (**object permanence**).

**Assign new slots:** Newly appearing objects should use unused slots.



**Preserve slot assignments:** Do not reuse a slot of a disappeared object.

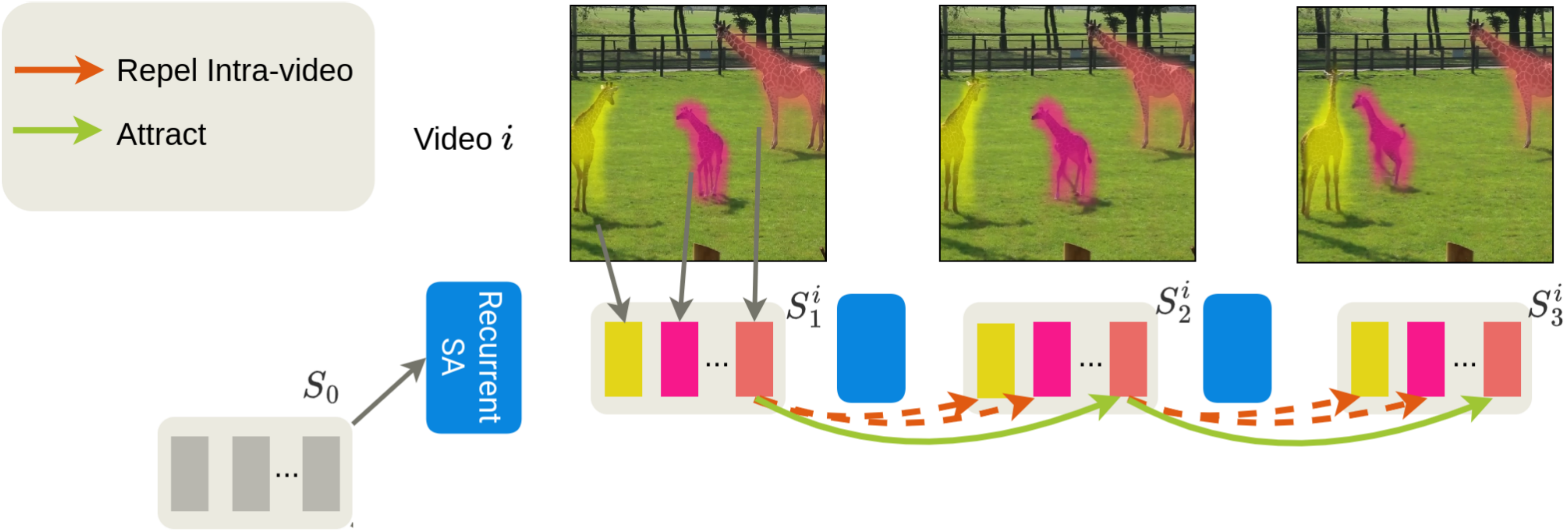
# Video-level Contrastive Learning



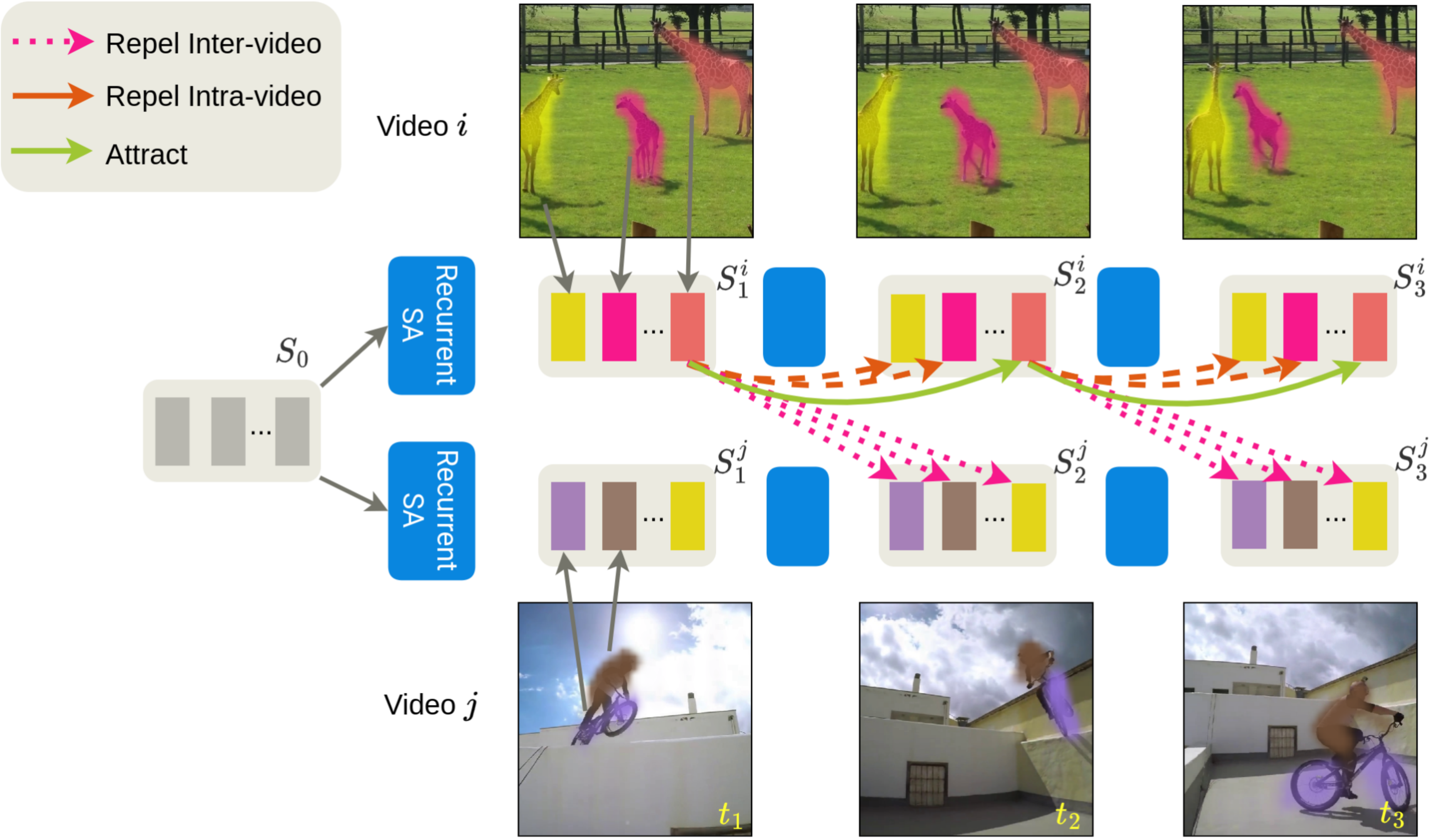

- Attract video frames of the same video
- Rebel frames from different videos in the dataset

How can we incorporate similar **contrastive objective** on more granular **slot representation level**?

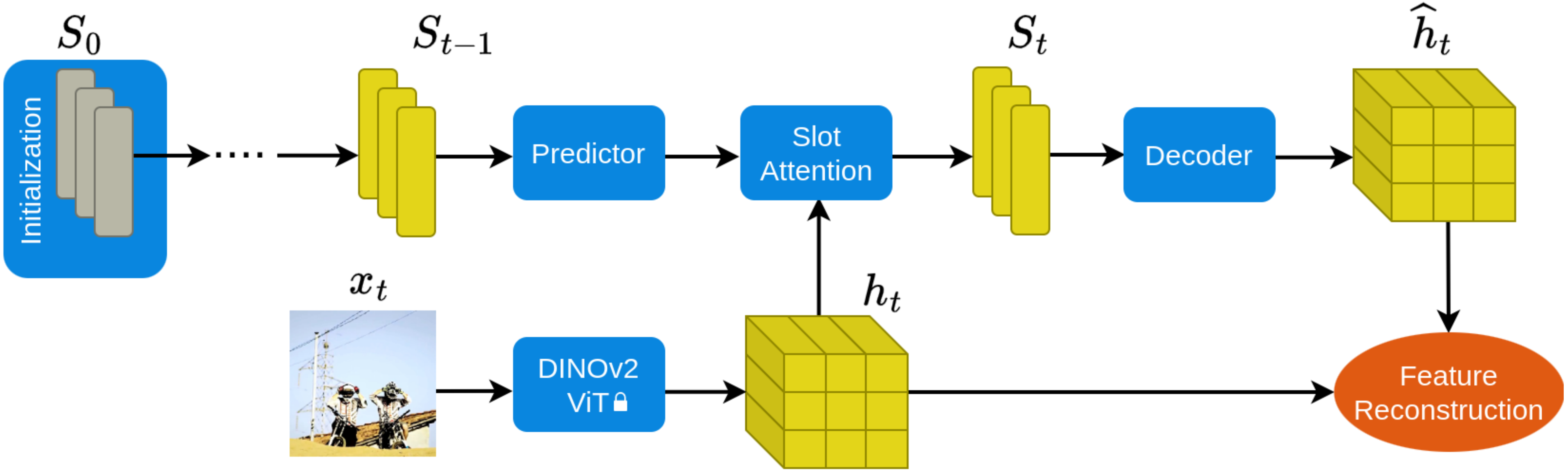
# Slot-Slot Contrastive Loss (Intra-Video)



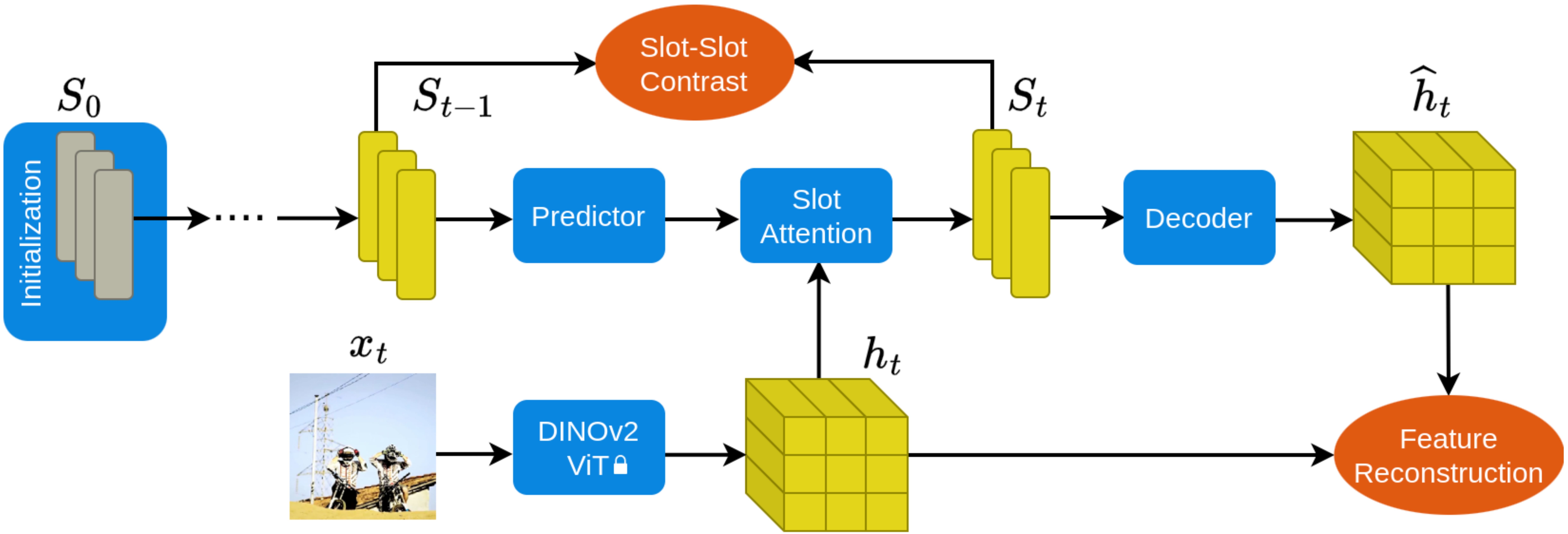
# Batch-level Slot-Slot Contrast



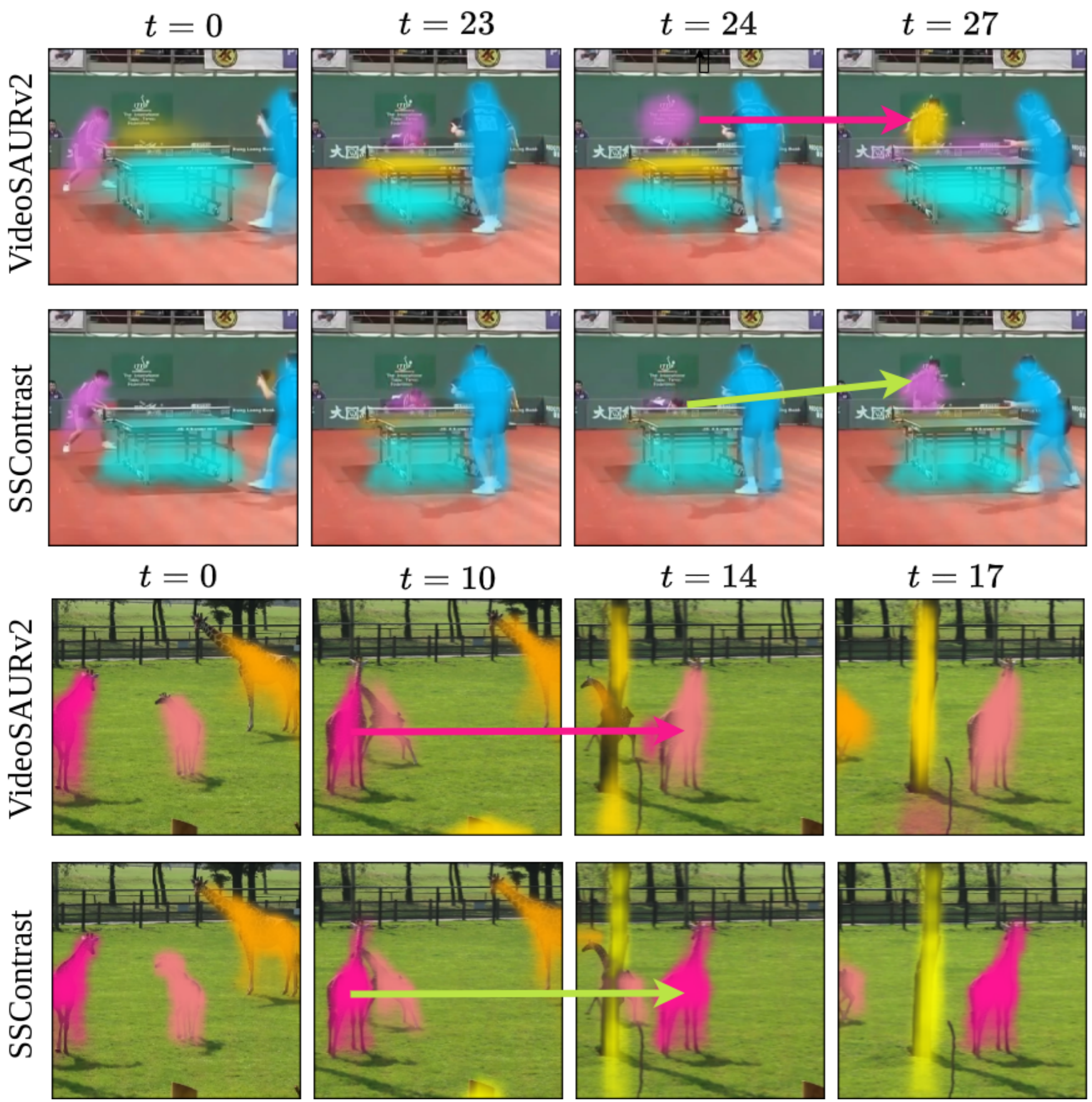
# Video Object-Centric Learning Architecture

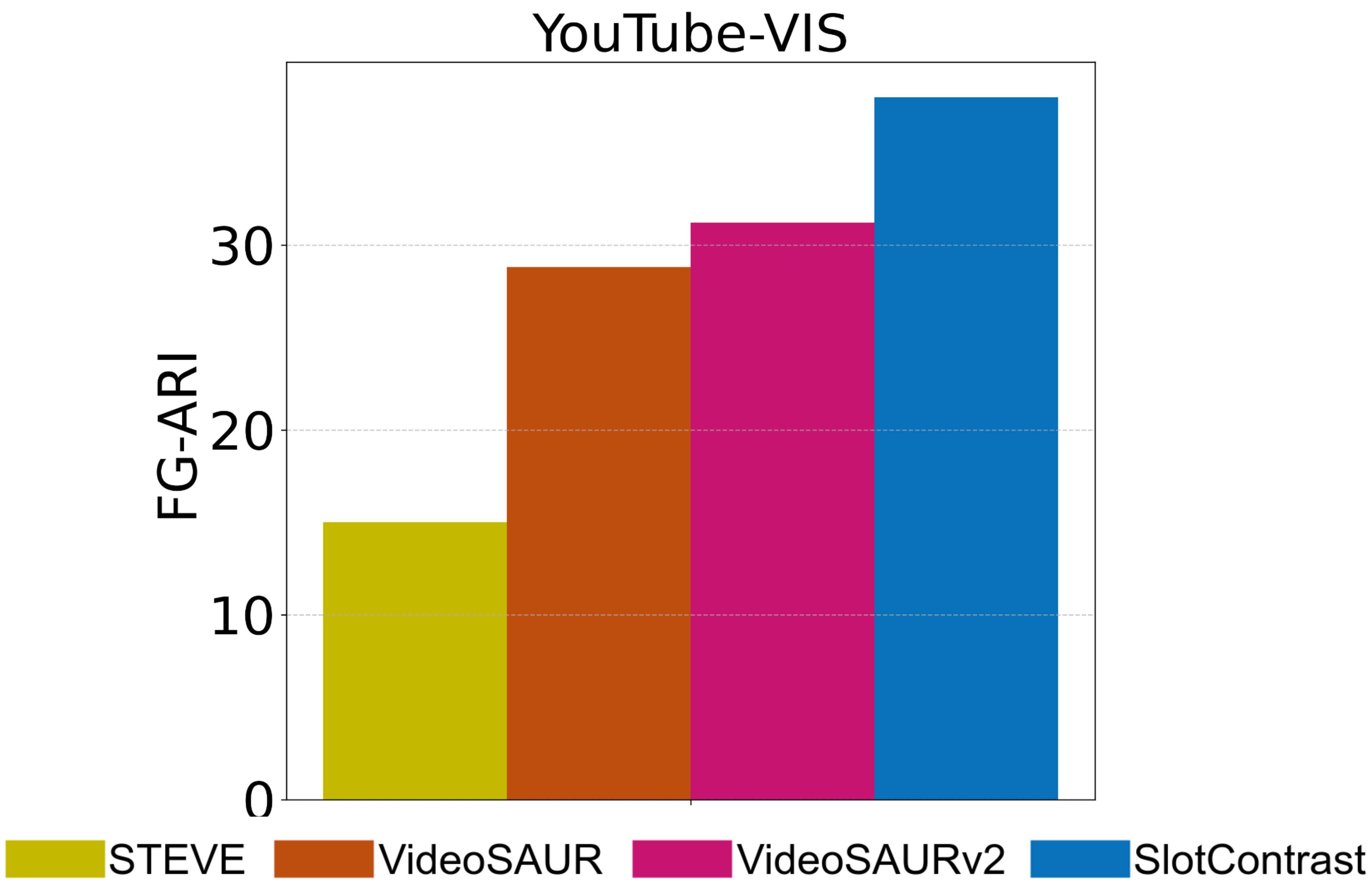


# Slot Contrast Architecture



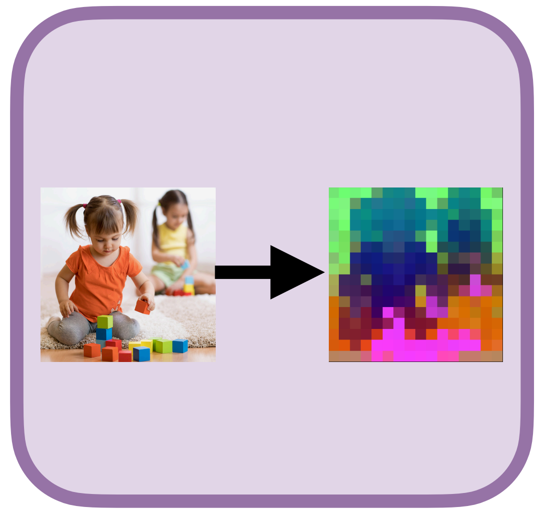
# Object Discovery on Real-World Videos



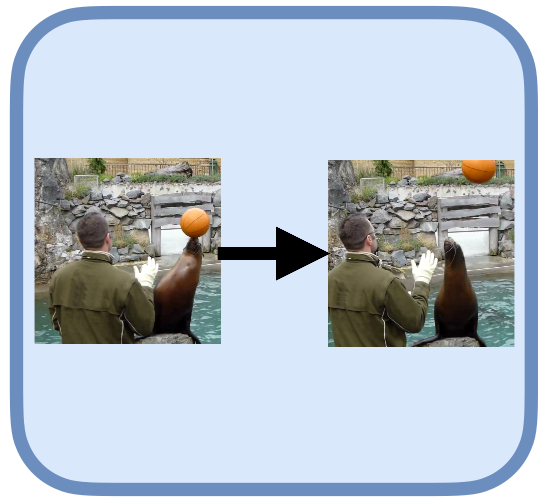



# Object Discovery on Real-World Videos

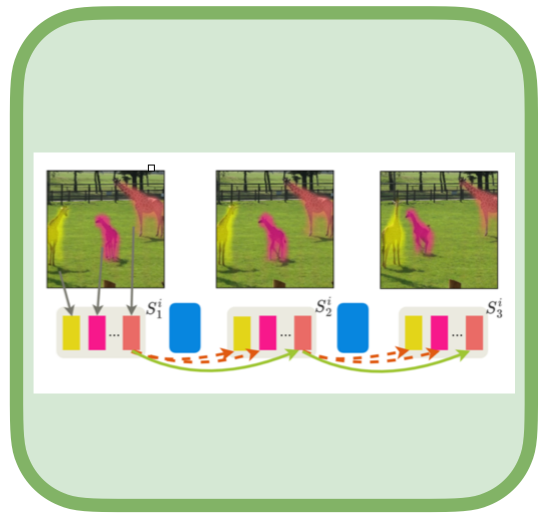
# Summary



- **Semantics reconstruction objective** allows to scale object-centric representations to **real-world images**



- **Temporal similarity prediction** further scales object-centric representation to **real-world videos**



- **Slot Contrast loss** further improves **long-term consistency** of learned representations