

Lab CudaVision  
Learning Vision Systems on Graphics Cards (MA-INF 4308)

# CudaLab Project

---

09.07.2025

PROF. SVEN BEHNKE, ANGEL VILLAR-CORRALES

Contact: [villar@ais.uni-bonn.de](mailto:villar@ais.uni-bonn.de)

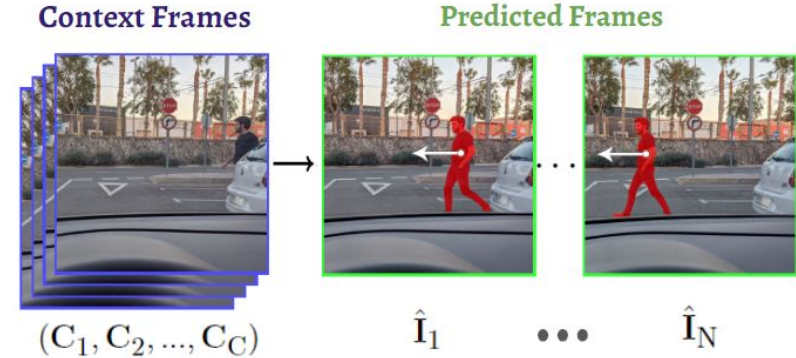
# Evaluating Image Representations for Video Prediction

---

# Future Frame Video Prediction

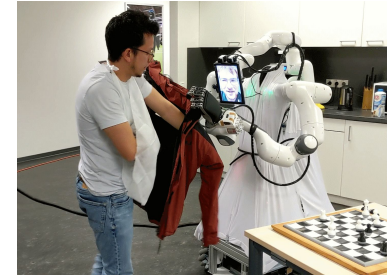
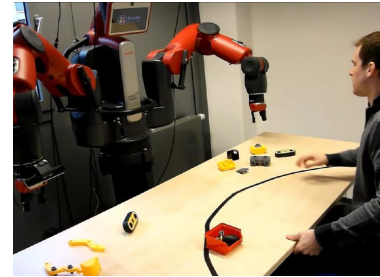
## Video Prediction Problem:

Given **C** consecutive seed/context frames of a video, generate next plausible **N** frames.



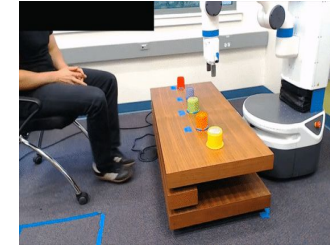
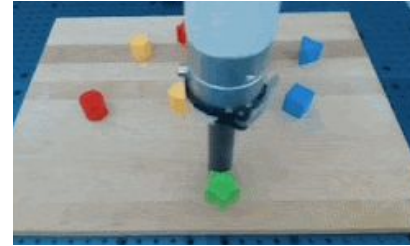
## Motivation:

- Human-robot collaboration
- Action planning
- Representation learning



# Object-Centric Representation Learning

- Scenes contain moving & interacting objects
  - Humans reason over objects — not pixels
- Can we process images as objects instead of pixels?



## Object-Centric Representations:

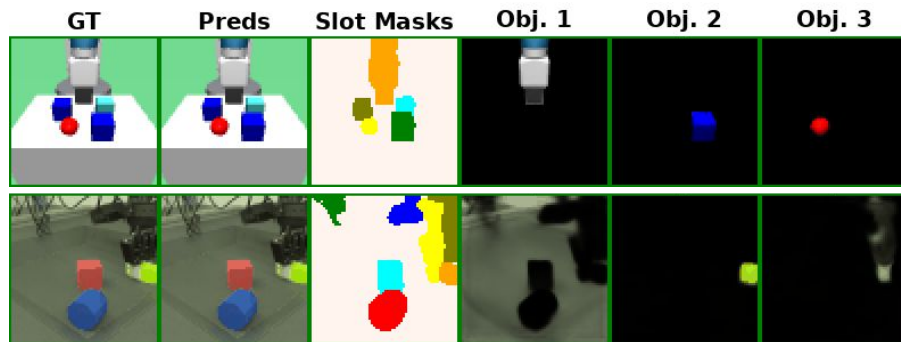
- Encode scene objects
  - Robust, interpretable and generalizable
  - Well aligned with human perception
- Can serve as basis for planning and reasoning!



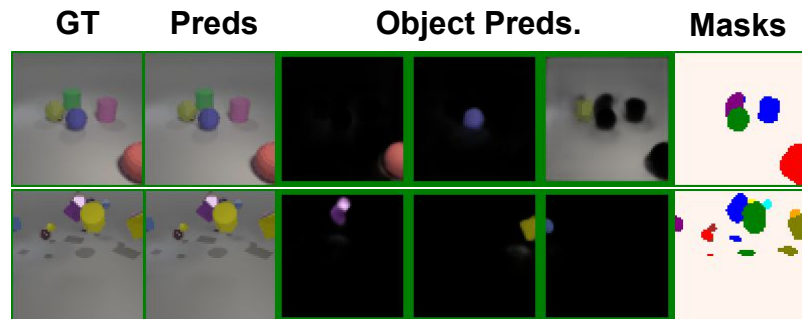
# Object-Centric Video Prediction

- Practical approach to video prediction
  1. Discretize frames into objects
  2. Predict future object states
  3. Reconstruct frames from object states

Learning robot behaviors

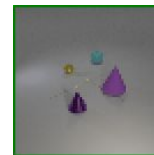


Learning object dynamics and interactions



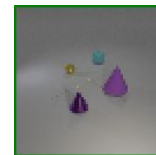
Original Caption

'the **large purple rubber cone** is picked up and placed to ( 2 , 3 ) . the **small gold metal snitch** is picked up and placed to ( -1 , 1 ) .'



Changed Objects

'the **medium cyan rubber sphere** is picked up and placed to ( -2 , -2 ) . the **medium purple metal cone** is sliding to ( 2 , 3 )'



# In this Project...

---

- Train transformer-based video prediction models
- Evaluate different scene representations, including holistic and object-based
- Analyze the effect of different representations

# Model

---

# Goal Inspiration

## OBJECT-CENTRIC VIDEO PREDICTION VIA DECOUPLING OF OBJECT DYNAMICS AND INTERACTIONS

Angel Villar-Corralles<sup>1</sup> Ismail Wulhan<sup>2</sup> Sven Behke

Autonomous Intelligent Systems, University of Bonn, Germany

### ABSTRACT

We present a framework for object-centric video prediction, i.e., parsing a video sequence into objects, and modeling their dynamics and interactions in order to predict the future object states from which video frames are rendered. To facilitate the learning of intra- and inter-object temporal object representations and forecasting of their states, we propose two novel object-centric video prediction (OCVP) transformer models, which decouple the processing of temporal dynamics and object interaction. We show how OCVP predicts future object-centric, video prediction models use no different datasets. Furthermore, we observe that OCVP models learn consistent and interpretable object representations. Animations and code to reproduce our results can be found in our project website.

**Index Terms**—Object-centric video prediction, scene parsing, object-centric learning, frame-frame prediction, transformers

### 1. INTRODUCTION

Humans perceive the world by getting cues into background and multiple foreground objects that can interact with each other [1]. Some modeling approaches that are equipped with inductive biases for such decompositions have the ability to obtain modular and structured representations with desirable properties such as sample efficiency, improved generalization, and robustness to out-of-distribution data [2].

In recent years, supervised approaches to decompose images or video sequences into their constituent objects have achieved impressive results on downstream tasks, such as object discovery [3] or object tracking [4]. Despite these recent advances in object-centric decomposition, modeling temporal dynamics and object interactions from visual observations alone remain challenging.

To model spatio-temporal object dynamics, we propose a framework for object-centric video prediction. Our approach, depicted in Fig. 1, uses a scene parsing module to extract object representation from video frames. Learn a separate prediction module to model temporal dynamics and object interactions using these representations, and render future video frames from predicted object states.

A key component in our framework is the sequence predictor, which models multi-object dynamics to predict future object states. Different predictor designs embody distinct inductive biases, which affect their suitability for object prediction. We investigate modular predictors for object-centric, video prediction and propose two novel object-centric video prediction (OCVP) transformers, which decouple the processing of temporal dynamics and object interaction.

In summary, our contributions are: (1) We present a framework for object-centric video prediction, which decomposes video sequences and can be integrated with a variety of decomposition models.

**Video Prediction:** Future frame video prediction is the task of forecasting future video frames conditioned on past frames. Many different approaches have been proposed for this task, including 3D

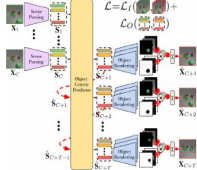


Fig. 1. Overview of our object-centric prediction framework. We decompose the real video frames into object-centric representations and learn an autoregressive object-centric predictor to model the object dynamics and interactions as well as to predict future object states. Predicted object representations are independently rendered into object images and masks, which are combined to generate the subsequent video frames. Our approach is trained by simultaneously minimizing the pixel prediction and video frame prediction mean squared errors.

into object representations and model object dynamics and interaction. (2) We propose two novel object-centric predictive models, which decouple the processing of temporal dynamics and object interaction. (3) Our prediction framework, using our OCVP models outperforms object-centric models for the task of video prediction from video frames. Learn a separate prediction module to model temporal dynamics and object interactions using these representations, and render future video frames from predicted object states.

A key component in our framework is the sequence predictor, which models multi-object dynamics to predict future object states. Different predictor designs embody distinct inductive biases, which affect their suitability for object prediction. We investigate modular predictors for object-centric, video prediction and propose two novel object-centric video prediction (OCVP) transformers, which decouple the processing of temporal dynamics and object interaction.

In summary, our contributions are: (1) We present a framework for object-centric video prediction, which decomposes video sequences and can be integrated with a variety of decomposition models.

**Video Prediction:** Future frame video prediction is the task of forecasting future video frames conditioned on past frames. Many different approaches have been proposed for this task, including 3D

## Are We Done with Object-Centric Learning?

Alexander Rubinstein<sup>1</sup> Ameya Prabhu<sup>1</sup> Matthias Bethge<sup>1</sup> Seung Joon Oh<sup>2</sup>

<sup>1</sup>Tübingen AI Center, University of Tübingen

<sup>2</sup>Project Page <sup>3</sup>OCCLM Colabase

### Abstract

Object-centric learning (OCL) seeks to learn representations which only encode the object, isolated from their objects or background scene in a scene. This approach addresses various aims, including zero-shot generalization (OOD generalization), sample-efficient computation, and modeling of grounded environments. Most research has focused on developing unsupervised methods that separate objects into discrete slots in the representation space, evaluated using unsupervised object discovery. However, with recent sample-efficient segmentation models, we can separate objects in the pixel space and encode them independently. This achieves remarkable zero-shot performance on OOD object discovery benchmarks, is scalable to foundation models, and can handle a variable number of slots out-of-the-box. Hence, the goal of OCL methods to obtain object-centric representations has been largely achieved. Despite this progress, a key question remains: How does the ability to separate objects within a scene contribute to broader OCL objectives, such as OOD generalization? We address this by investigating the OOD generalization challenge caused by spurious background cues through the lens of OCL. We propose a novel, training-free probe called *Object-Centric Classification with Applied Masks (OCCLM)*, demonstrating that segmentation-based decoupling of individual objects significantly outperforms slot-based OCL methods. However, challenges in real-world applications remain. Beyond the toolbox for the OCL community to use scalable object-centric representations, and focus on practical applications and fundamental questions. Our code is available at <https://github.com/alexander-rubinstein/occlm>.

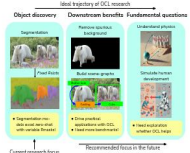


Figure 1. Where Should We Get Object-centric learning (OCL) has focused on developing unsupervised mechanisms to separate the representation space into discrete slots. However, the latest challenges of this task have led to comparatively less emphasis on exploring decomposition approaches and exploring fundamental benefits. Here, we introduce simple, effective OCL methods by separating objects in the pixel space and encoding them independently. We present a case study that demonstrates the downstream advantages of our approach for mitigating spurious correlations. We utilize the tool to develop benchmarks aligned with fundamental goals of OCL and explore the decomposition efficacy of OCL representations.

### 1. Introduction

Object-centric learning (OCL) seeks to develop representations of complex scenes that independently encode each foreground object separately from background cues, ensuring that one object's representation is not influenced by others or the background [2, 12]. This constitutes a fundamental element for many objectives: it supports modular

learning of structured environments [21], enables robust out-of-distribution (OOD) generalization [1, 12, 26, 42, 73], facilitates compositional perception of complex scenes [23], and deepens our understanding of object perception in human cognition [24, 25, 22]. However, despite these broad goals, most research in OCL has centered on advancing “slot-centric” methods that separate objects and encode them into slots, evaluated using unsupervised object discovery as the primary metric [1, 12, 22, 26, 41, 26]. In this paper, we challenge the continued emphasis on developing mechanisms to separate objects in representation space as the main challenge to be addressed in OCL.

We first show that sample-efficient class-agnostic segment-

## On the Benefits of Instance Decomposition in Video Prediction Models

Eliyas Suleyman<sup>1</sup>, Paul Henderson<sup>1</sup>, Nicolas Pugeault<sup>1</sup>

<sup>1</sup>School of Computing Science, University of Glasgow  
2853522@stud.gla.ac.uk, {paul.henderson,nicolas.pugeault}@gla.ac.uk

### Abstract

Video prediction is a crucial task for intelligent agents such as robots and autonomous vehicles, since it enables them to anticipate and act early on time-critical incidents. State-of-the-art video prediction methods typically model the dynamics of a scene jointly and implicitly, without any explicit decomposition into separate objects. This is a challenging and potentially sub-optimal, as in every object in a dynamic scene has their own pattern of movement, typically somewhat independent of others. In this paper, we investigate the benefit of explicitly modeling the objects in a dynamic scene separately within the context of latest transformer video prediction models. We conduct detailed and carefully-controlled experiments on both synthetic and real-world datasets, our results show that decomposing a dynamic scene leads to higher quality predictions compared with models of a similar capacity that lack such decomposition.

### 1. Introduction

Video prediction is the task of predicting future frames based on past frames; it has many applications including autonomous driving [24], [25], weather forecasting from satellite images [26], [27], and even building general world models for robot navigation [28]. Video prediction is challenging, since there are high-dimensional and result from multiple sources: appearances, dynamics and mutual interactions. For example, consider the environment observed while driving a car. To accurately predict the future, we must identify all objects in our field of vision and estimate their likely movement. Different types of objects (e.g. cars, pedestrians, dogs) have very different appearance, but also diverse patterns of movement, and may exhibit complex interactions with other objects.

To reduce this complexity, a natural approach to video prediction is to decompose the scene into several parts, [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], [79], [80], [81], [82], [83], [84], [85], [86], [87], [88], [89], [90], [91], [92], [93], [94], [95], [96], [97], [98], [99], [100]. This enables modeling the appearance and dynamics of each part separately during prediction, thus reducing computational cost and increasing training efficiency. Several works have achieved promising results using such approaches, using different bodies of appearance, motion to separate objects in representation space as the main challenge to be addressed in OCL.

We first show that sample-efficient class-agnostic segment-

ation semantic segmentation models, and [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], [79], [80], [81], [82], [83], [84], [85], [86], [87], [88], [89], [90], [91], [92], [93], [94], [95], [96], [97], [98], [99], [100]. This enables modeling the appearance and dynamics of each part separately during prediction, thus reducing computational cost and increasing training efficiency. Several works have achieved promising results using such approaches, using different bodies of appearance, motion to separate objects in representation space as the main challenge to be addressed in OCL.

In this work, we perform a detailed study of the benefits of explicitly modeling different objects separately during video prediction, when using modern latent transformer models. Rather than introducing an entirely new model, we develop a family of architectures that use ideas from VideoPose [40] and Sotterman [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [78], [79], [80], [81], [82], [83], [84], [85], [86], [87], [88], [89], [90], [91], [92], [93], [94], [95], [96], [97], [98], [99], [100]. This enables modeling the appearance and dynamics of each part separately during prediction, thus reducing computational cost and increasing training efficiency. Several works have achieved promising results using such approaches, using different bodies of appearance, motion to separate objects in representation space as the main challenge to be addressed in OCL.

We find that even with large transformers, object decomposition leads to considerable improvements in handling complex scenes with multiple interacting objects compared to non-object-centric predictors with similar parameter counts and latent dimensions.

• We present the first systematic and comprehensive analysis of the benefits of explicit object decomposition for latent transformer video prediction models.

• To achieve this, we develop a scalable framework for

## Object-centric Video Prediction without Annotation

Karl Schmeckpeper<sup>1,2</sup>, Georgios Georgakis<sup>1,2</sup>, and Kostas Daniilidis<sup>1,2</sup>

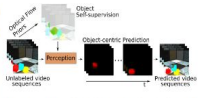


Fig. 1. We seek to model physical interactions with a visual sensor and predict the future states of objects, learning using no ground annotations. Ground-truth object flow priors to generate object-centric representations. We segment the input video frames, then predict the future states of objects, and use these to generate predictions of the next frames. In testing, we predict future video sequences given a single frame from the input.

### 1. INTRODUCTION

Modeling physical interaction is a fundamental agent skill for interacting with the world. This is a challenging task to learn as it requires understanding the scene's dynamics. For object manipulation scenarios, the challenge is exacerbated by the need to understand the environment of the object level, including agent-object and object-object interactions. Addressing this problem using visual sensors offers many advantages. First, high-quality cameras are easily accessible and have low size, weight, and power requirements, allowing them to be included in almost every robotic system.

Second, there is an abundance of existing data samples, allowing powerful deep learning models to be trained. Third, visual observations offer rich information about the environment, including pose, texture, and semantics, that cannot easily be matched by other sensors.

Existing methods typically address this problem by learning an action-conditioned predictive model that infers the changes to the visual scene. These models have been demonstrated mostly through end-to-end deep networks that learn to map inputs and control inputs to future pixels [2]. This paradigm assumes that the model can implicitly learn to visually segment the objects and infer their motion in the scene in spite of the high dimensionality of pixels from the raw image inputs. It does not take advantage of perceptual priors that can be extracted from an observation, forcing the model to function without any aid from existing computer vision methods.

• We present the first systematic and comprehensive analysis of the benefits of explicit object decomposition for latent transformer video prediction models.

• To achieve this, we develop a scalable framework for

There have been efforts to learn pairwise object interactions by treating a visual scene as a collection of objects, which led to the development of object-centric predictive models. These methods typically assume that labeled object information is readily available at each future time step either in the form of object locations [2], or forces applied on the objects [3]. However, many robotic agents are required to operate in unstructured real-world environments that exhibit increased visual variability with no access to dense labels and are often have to generalize to previously unseen objects.

Recent works have demonstrated that utilizing perceptual priors, via powerful computer vision models, reduces sample complexity, enables generalizability across environments, and largely increases performance in visuomotor tasks [4]; [5]; [6]; [7]; [8]; [9]; [10]; [11]; [12]. Inspired by these methods, we make the observation that visual motion is a strong cue for object motion [2] and propose a novel object-centric video predictive model that leverages state-of-the-art perception in the form of object instance segmentation and optical flow, and does not require object annotations. The perception module is trained end-to-end with dynamics and image generation models in order to predict a future frame sequence from a single input frame (see Figure 1). The joint training allows for the perception module to fine-tune on the existing environment, while the dynamics and generation models benefit from the rich future representation encoded in the perception module. This results in an object-centric model that is not restricted only to environments where object level annotations are present, allowing the way towards adaptable predictive models for manipulation tasks. Our contributions include: (i) the introduction of a novel predictive model that does not require object level annotations, (ii) state-of-the-art results on the Shapeways [13] dataset which demonstrate the benefits of

• We present the first systematic and comprehensive analysis of the benefits of explicit object decomposition for latent transformer video prediction models.

• To achieve this, we develop a scalable framework for

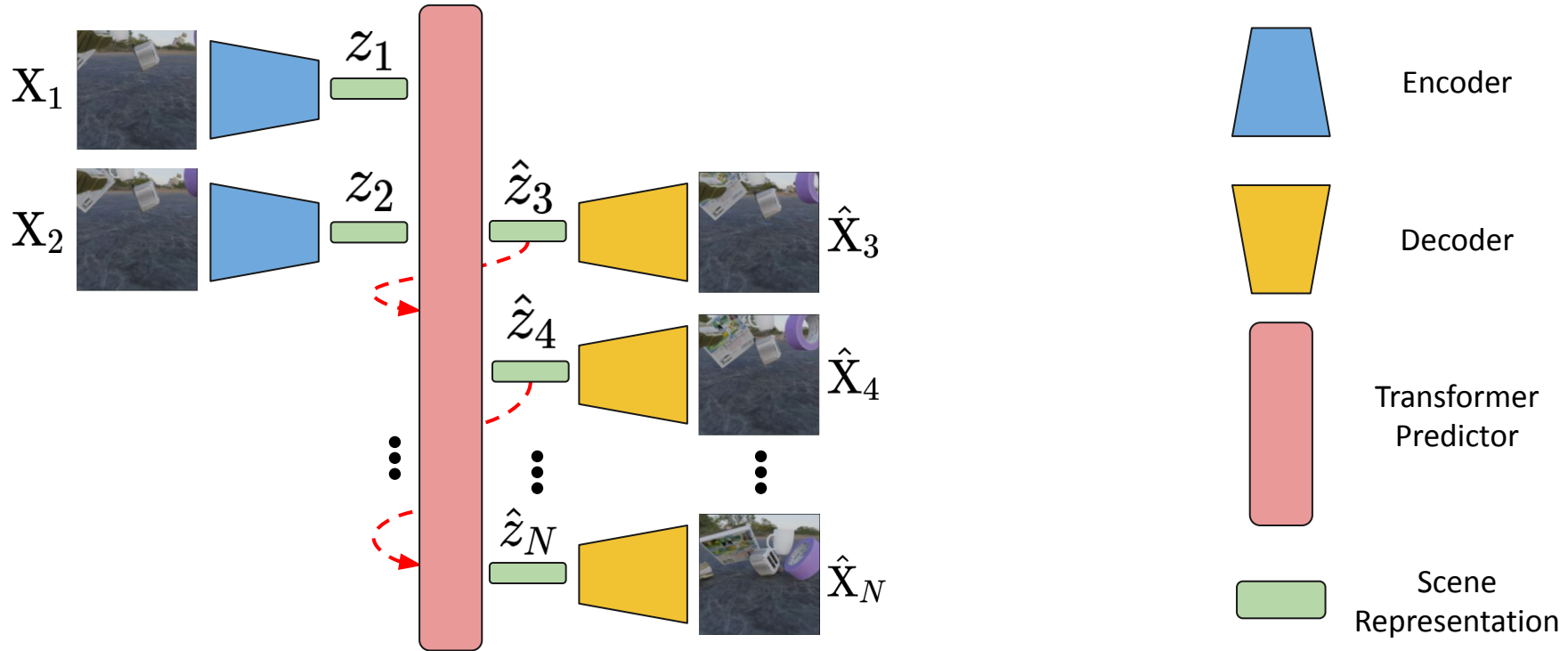
This work has been funded by the DFG SFB 1053/A1 (Autonomous Human Behavior) of the German Research Foundation (DFG).  
1. DFG grant SFB 1053/A1-1  
<https://www.dfg.de/en/funding/sfb1053/a1>

• We present the first systematic and comprehensive analysis of the benefits of explicit object decomposition for latent transformer video prediction models.

• To achieve this, we develop a scalable framework for



# Model

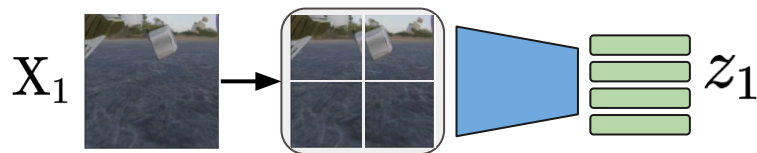


# Encoding

- We will evaluate different representations

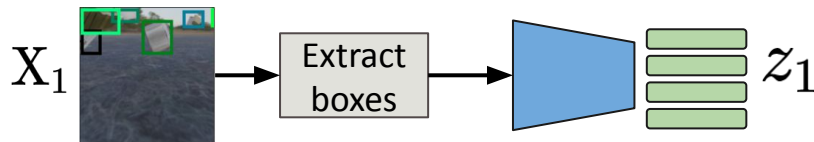
## ViT-like Encoder

Encode image patches with self-attention



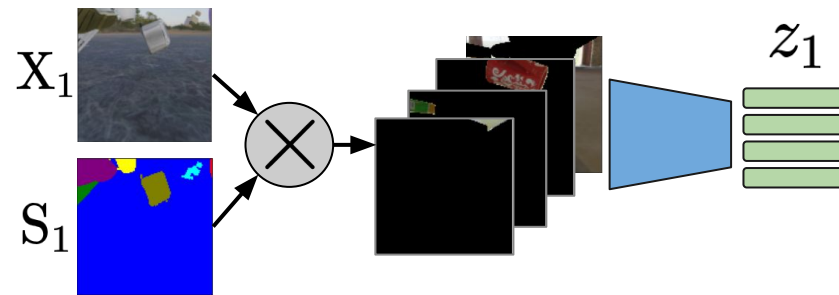
## BBox Encoding

Encoder processes each BBox separately



## Mask-Encoding

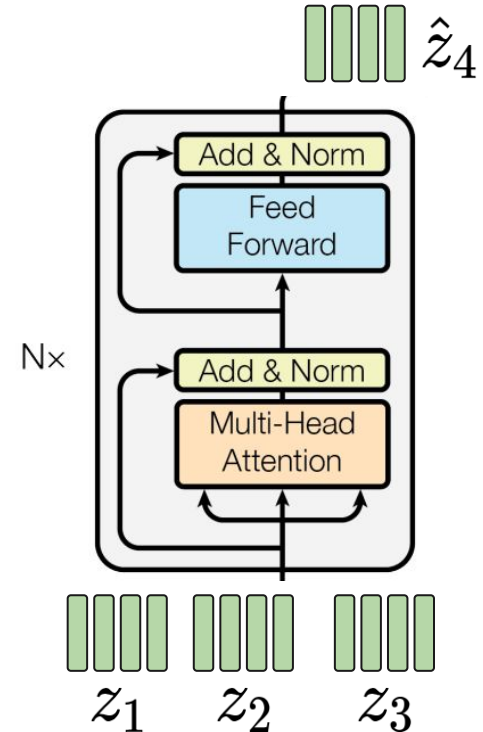
Encode each masked object separately



- Try your own representation/encoding:
  - Concatenate image with segmentation
  - Concatenate image with depth
  - Image + Binary masks
  - ...

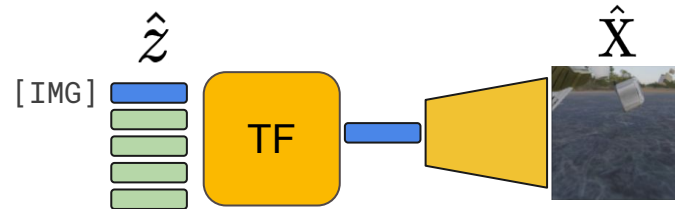
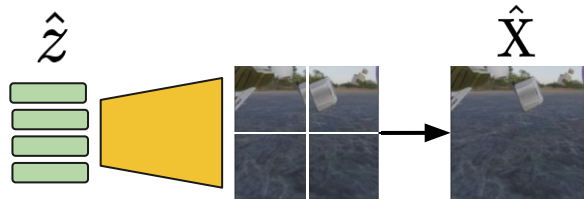
# Predictor

- Autoregressive transformer encoder predictor model
- Process scene representations via self-attention
- Process scene representations via self-attention
- Key modules and variants:
  - Positional encoding
  - Pre-Norm vs. Post-Norm
  - Full Attention vs. Space-time attention (similar to ViViT)



# Decoder

- Map from predicted scene representations to predicted frames
- Different possible architectural designs:
  1. Transformer Decoder:
    - Process tokens with self-attention to reconstruct patches
    - Combine patches to render image
  2. Conv. Decoder:
    - Process tokens with convolutions and upsampling to directly reconstruct the image
    - Many possible design choices



# Training

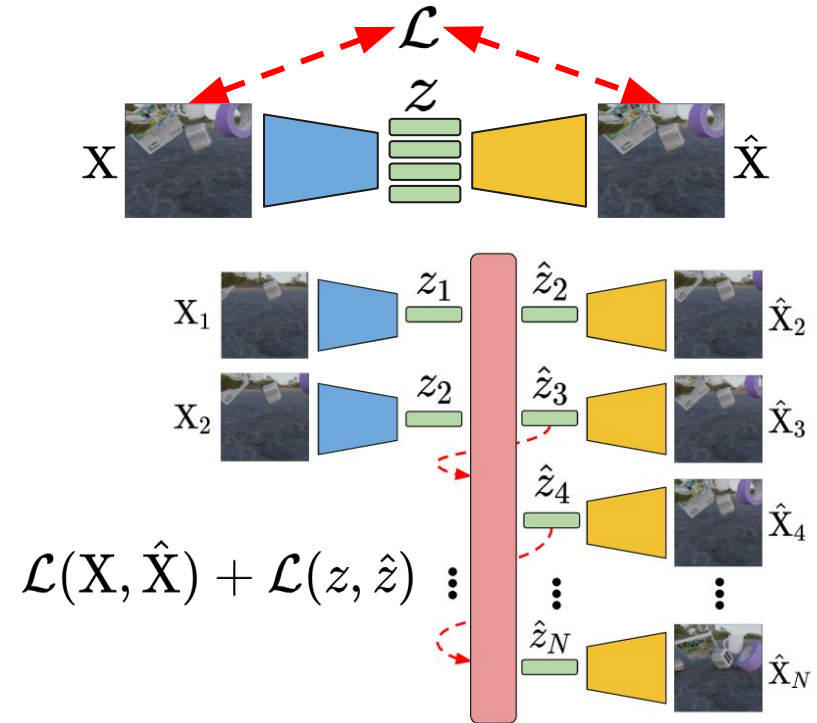
- Two-stage training pipeline

1. Autoencoder Training

- Train only encoder and decoder
- Solve image reconstruction task

2. Predictor Training

- Train only the predictor module
- Encoder and decoder remain frozen
- Solve image and feature prediction task

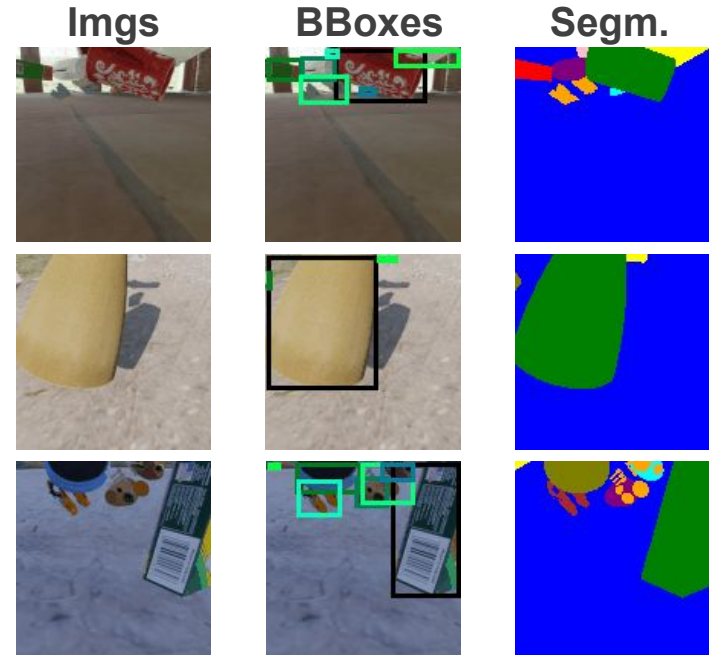



# Datasets

---

# MOVi-C Dataset

- Synthetic Object-Centric Dataset
  - Rendered realistic objects colliding
  - RGB, Semantics, BBoxes, Flow, ...
  - Fixed camera
- Sequences of length 24
  - ~10k training videos
  - ~250 validation videos
- Images are (128, 128)
  - I suggest working with (128, 128) or (64, 64)



 Available in: `/home/nfs/inf6/data/datasets/MOVi/movi_c`

# Dataset Recommendations

---

- It usually helps to ‘have more data’
- Apply spatial augmentations (mirror, rotate)
  - Apply the same augmentation to all frames in the sequence!
- Train using randomly sampled subsequences (e.g  $1 \rightarrow 8$ ,  $12 \rightarrow 20$ , ...)
- Use temporal augmentations (e.g. subsampling) if necessary



# Training & Evaluation

---

# Training on MOVi

---

- Training:
  - Train your models using the 2-stage approach described before
  - Train your models using 5 seed frames to predict the next 5
  - I suggest using image resolution of **(64, 64)** or (128, 128)
- You should train **at least two different model variants** using different representations
  - One should be image-level and the other should be object-level
  - I recommend using: 'Image patches' and 'object segmentations'
  - You are encouraged to try out more representations and change the models

# Evaluation on MOVi

---

- Evaluation:
  - Evaluate your models using 5 seed frames to predict the next 15
  - Compute the following metrics: PSNR, SSIM, LPIPS, FVD for 5 and 15 predictions
  - Compute qualitative evaluations: GIFs, images, ...
- Investigate and answer the following questions:
  - What representation work best?
  - Why do you think so?
  - What are the strengths and weaknesses of each representation and model?
  - What trade-offs do they present?

# Project Goals and Deliverables

---

# Passing Requirements

---

1. Implement the required model, pipelines and utils
2. Train two different model variants using different representations
  - Evaluate your models quantitatively using PSNR, LPIPS, SSIM & FVD
  - Qualitatively evaluate your models via videos and images
  - You must follow the project guidelines up to certain extent
  - Make changes to the model and try out things to achieve good results
3. Create overview notebook
4. Write project report

# Deliverables

---

- Complete codebase
  - Clean and structured
  - Not just a notebook!
- Trained model checkpoints and (tensorboard, WandB, ...) logs
- Overview notebook (.ipynb & .html) showing main functionalities:
  - Load data and display some samples
  - Load pretrained model and display the structure or some stats
  - Display some qualitative results (e.g. results on 5 sequences)
  - Show the quantitative evaluation
- Project report

# Grading

---

- Results and Experiments **55%-60%**:
  - Performing several experiments and obtaining good results
  - **Additional experiments**: more representations, model changes, ablations, ...
  - This grade partly depends on how your results compare to the class
- Codebase & Overview Notebook **20-25%**:
  - Implement all functionalities
  - Modularity and structure
- Report **20%-25%**

# Project Report

---

- Document your work in the project report
- Try to be brief, but readable and informative
- Include figures and tables
- Use *BibTex* for the references
- I expect 8-12 pages, but highly depends on number and size of imgs/tables
- Use the following template
  - <https://www.overleaf.com/read/tmnvhrsdmjrp>



# Important Dates

---

- **09.07:** Starting date
- **05.09-16.09:** Revision session (very flexible dates, just write me an email)
- **22.09:** Draft submission due (optional)
- **30.09:** Final submission

# Questions?





