The Applications of Object-Centric Learning

Dr. Yanwei Fu

School of Data Science, Fudan University

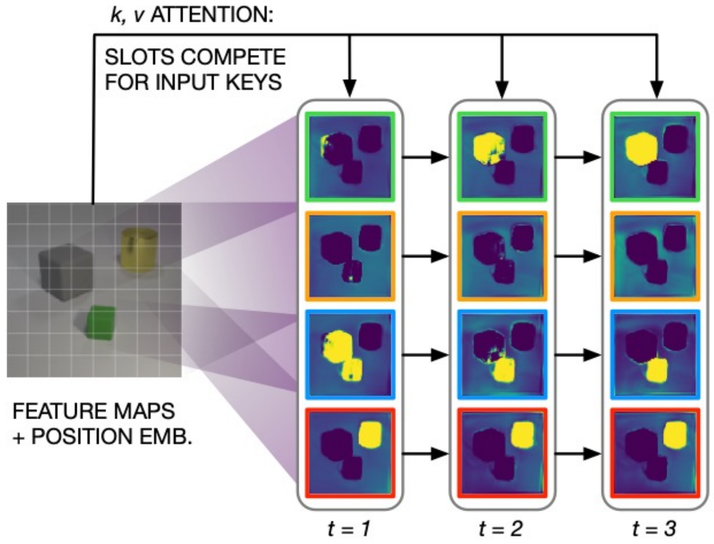
Homepage: http://yanweifu.github.io

CVPR 2024 Tutorial: Object-centric Representations in Computer Vision

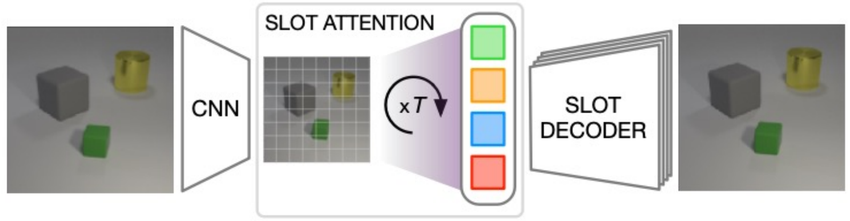
https://object-centric-representation.github.io/object-centric-tutorial-2024/

大数据学院
School of Data Science

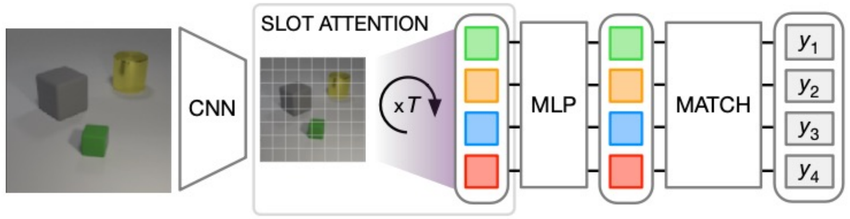
Recap: Slot Attention



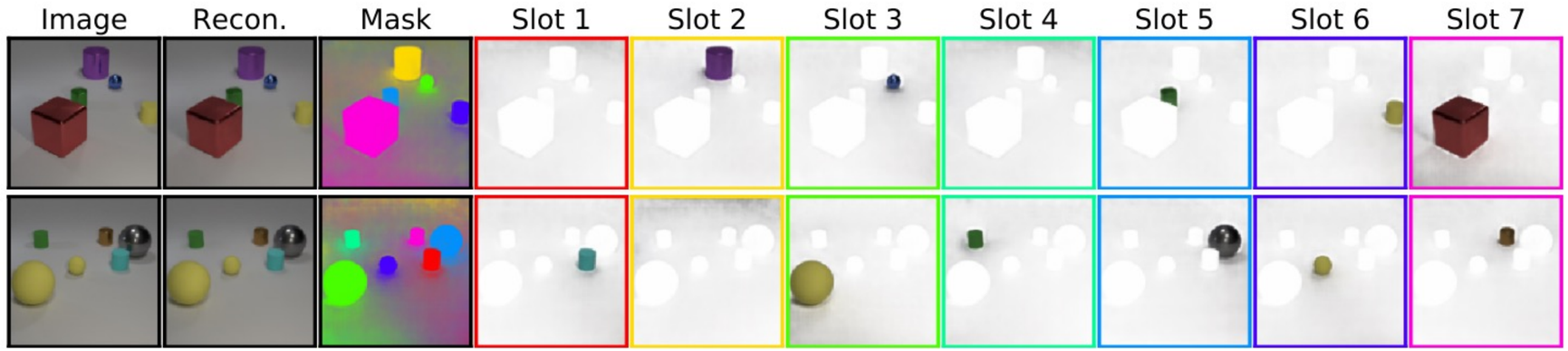
(a) Slot Attention module.



(b) Object discovery architecture.



(c) Set prediction architecture.

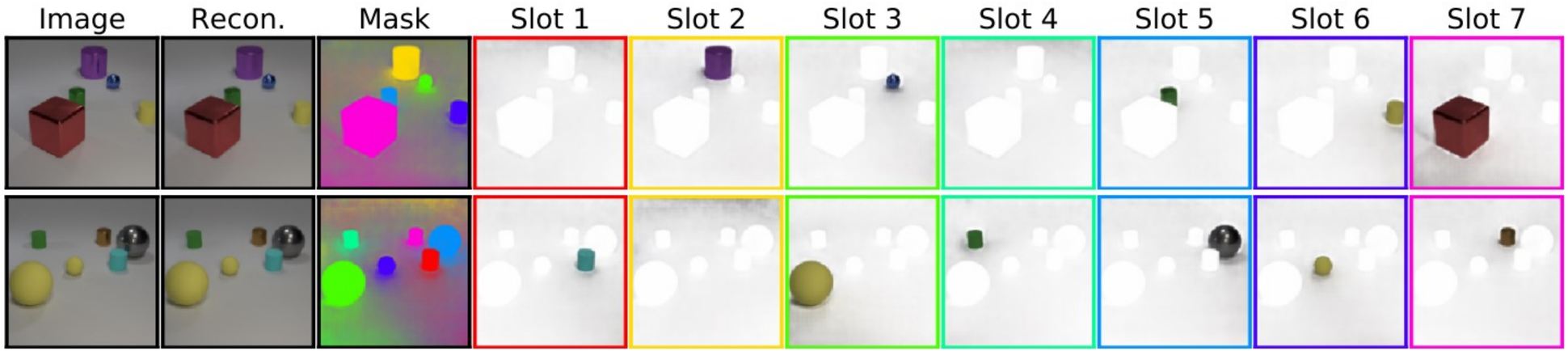


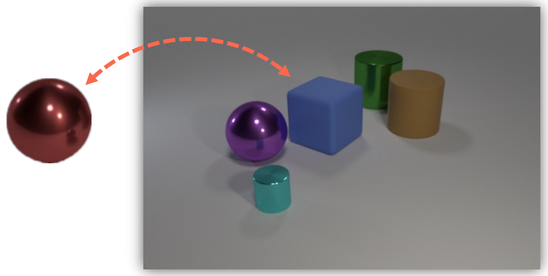
Locatello et al. Object-Centric Learning with Slot Attention. NeurIPS 2020 (image credit)

The Applications of Object-Centric Learning

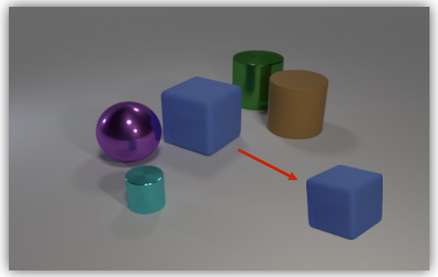
- **Image Manipulation**
- Segmentation
- Embodied AI
- Discussion

Three Ways of Slot based “Image Manipulation”

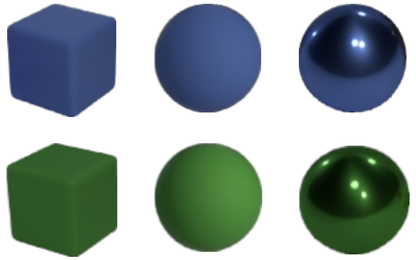




Substituting one slot with another



Moving a slot explicitly



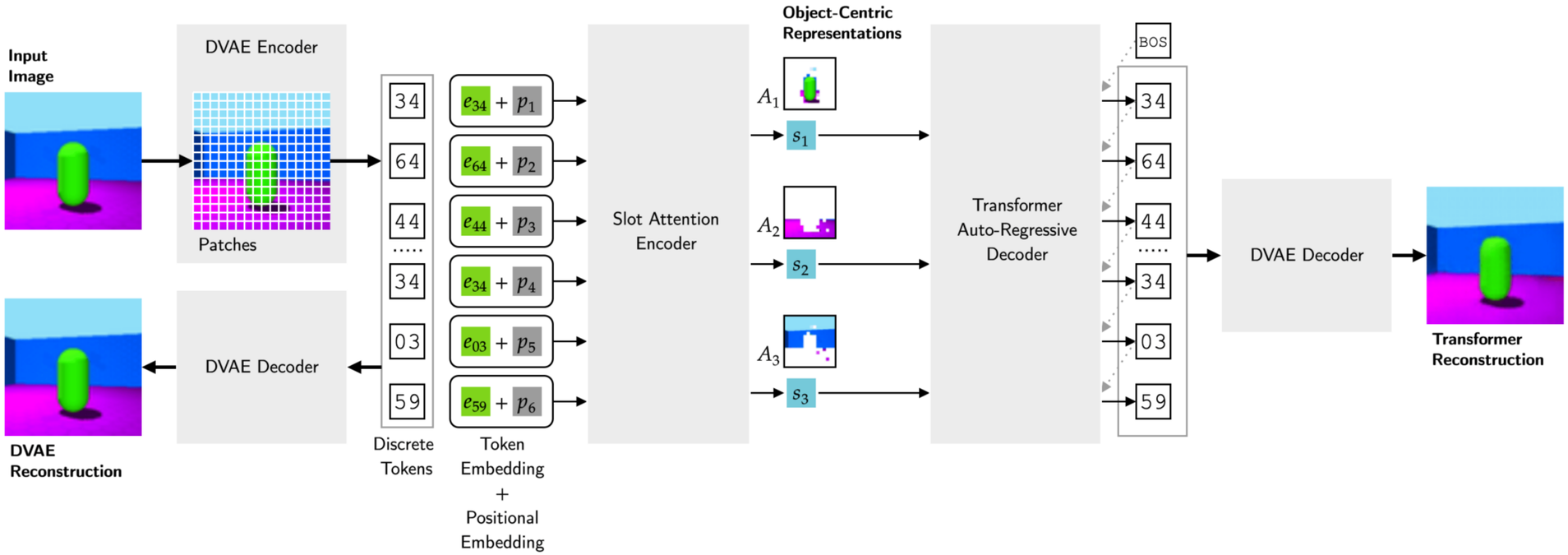
Manipulating the disentangled properties of slots

Salewski, Leonard, et al. “Clevr-x: A visual reasoning dataset for natural language explanations.” CVPR 2022 XAI4CV. (image credit)

Learning Slots for Visual Concept Library

Substituting

SLATE:



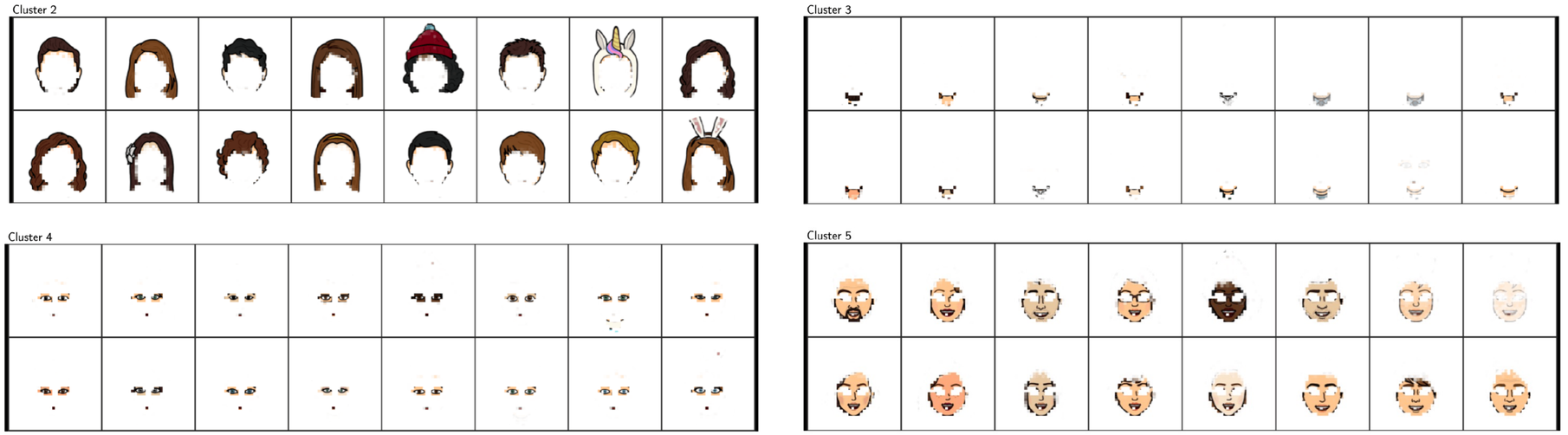
Collect Slots：Gather generated slots

Cluster Slots：K-means clustering on slots

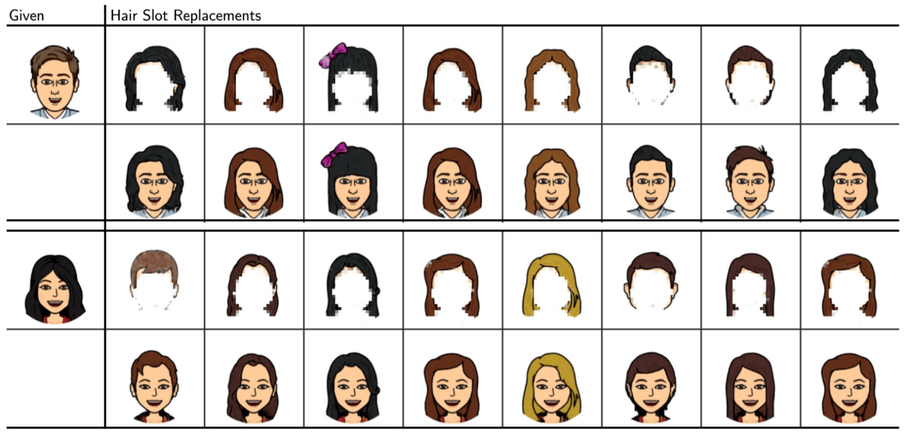
Compose Images：Select slots from the library

Singh, Gautam, Fei Deng, and Sungjin Ahn. "Illiterate dall-e learns to compose." ICLR 2022. (image credit)

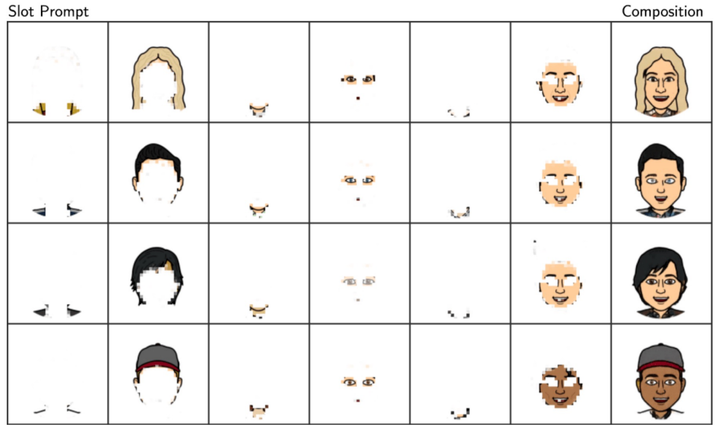
Examples of the Visual Concept Library



Examples of visual concept library. Each cluster corresponds a specific parts of face.



Replacing the hair slot



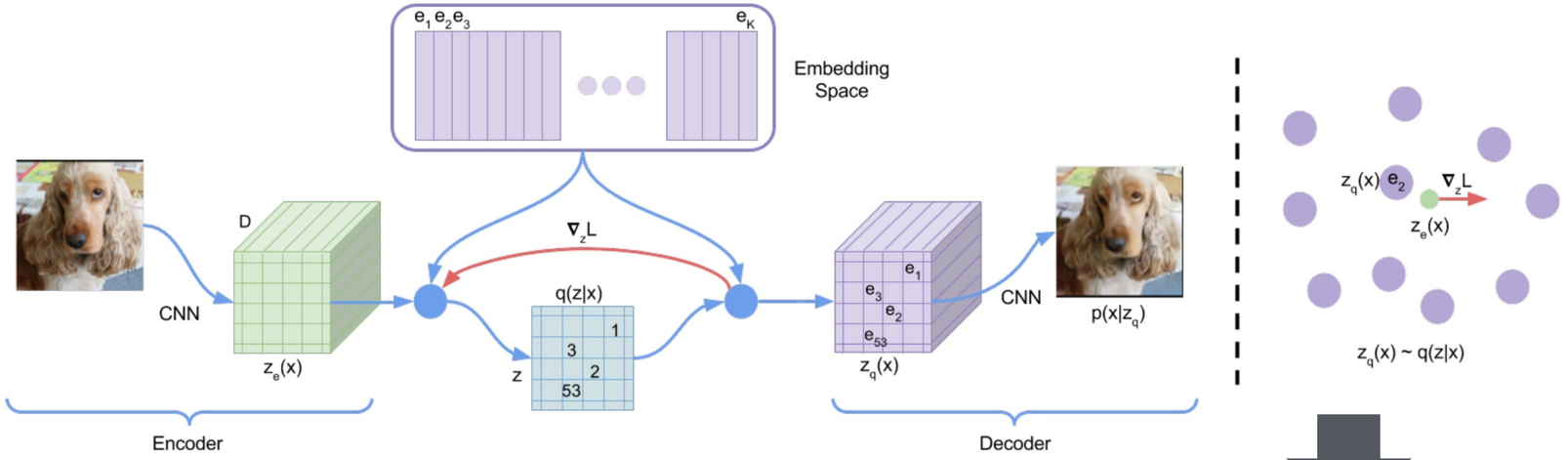
Composition of multiple slots

Singh, Gautam, Fei Deng, and Sungjin Ahn. "Illiterate dall-e learns to compose." ICLR 2022. (image credit)

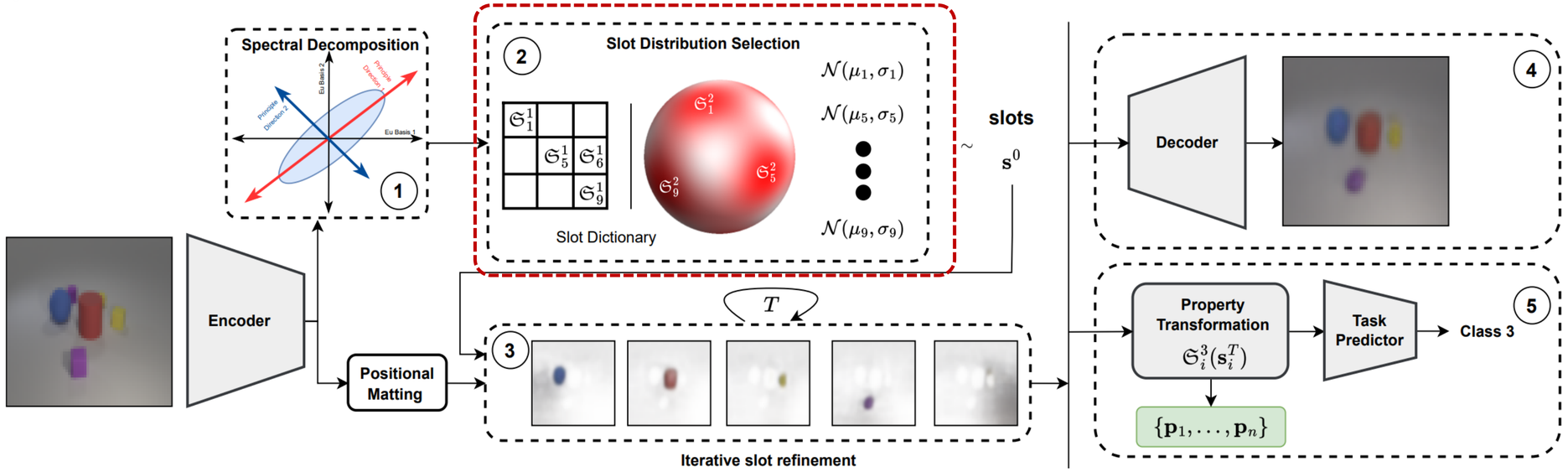
Grounded Slot Dictionary (GSD)

Substituting

VQ-VAE：



CoSA:



Van Den Oord, Aaron, and Oriol Vinyals. "Neural discrete representation learning." NeuIPS 2017. (image credit)
Kori, Avinash, et al. "Grounded Object-Centric Learning." ICLR 2024. (image credit)

Grounded Slot Dictionary (GSD) Binding

Substituting

(a) Embedding-7 $(\mathfrak{S}_7^1)$ (b) Embedding-14 $(\mathfrak{S}_{14}^1)$ (c) Embedding-25 $(\mathfrak{S}_{25}^1)$ (d) Embedding-55 $(\mathfrak{S}_{55}^1)$

Cheeks ***Forehead*** ***Eyes*** ***Facial hair***

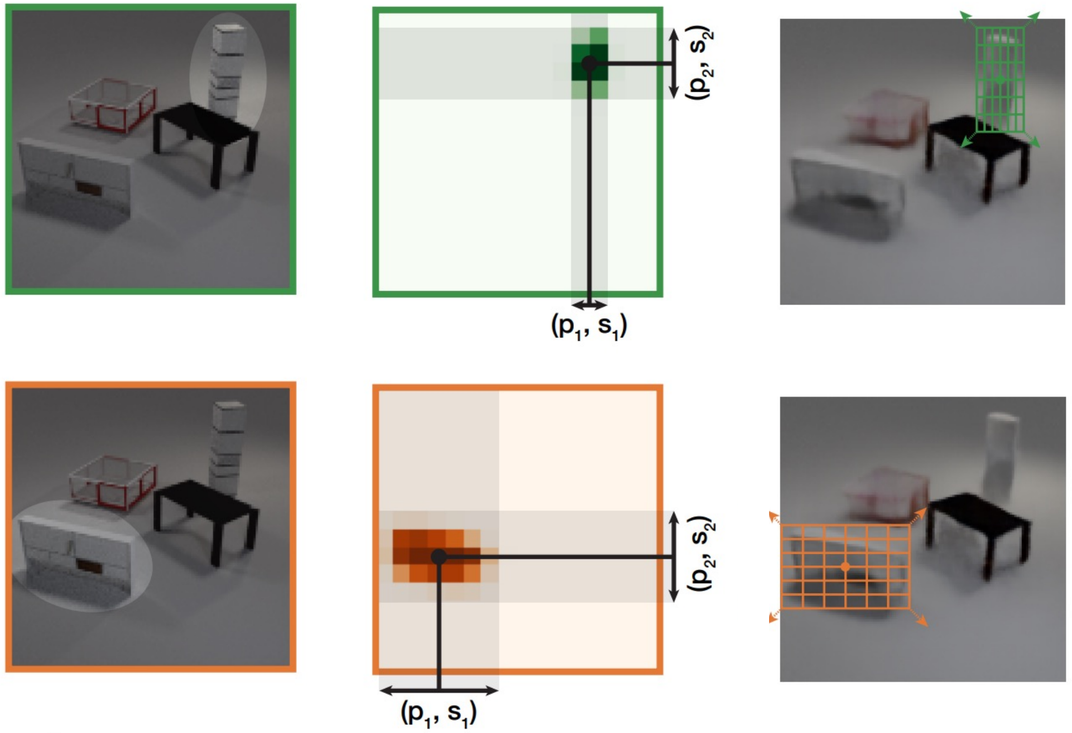
Sampling from an object-centric representation codebook,
We can map the grounded slot dictionary elements to particular object types.

Kori, Avinash, et al. "Grounded Object-Centric Learning." ICLR 2024. (image credit)

Spatial Symmetry into Slot Attention

Moving

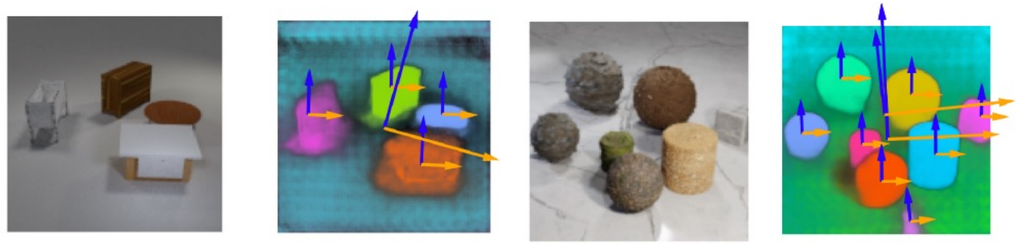
Question: How to identify object with spatial symmetry?



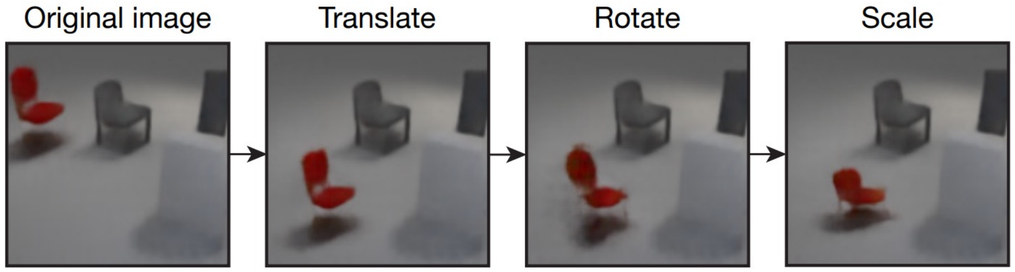
1. Slot attends to an encoded image.

2. Compute slot position and scale from attention masks.

3. Create relative coordinate grids.



slot position, scale and orientation

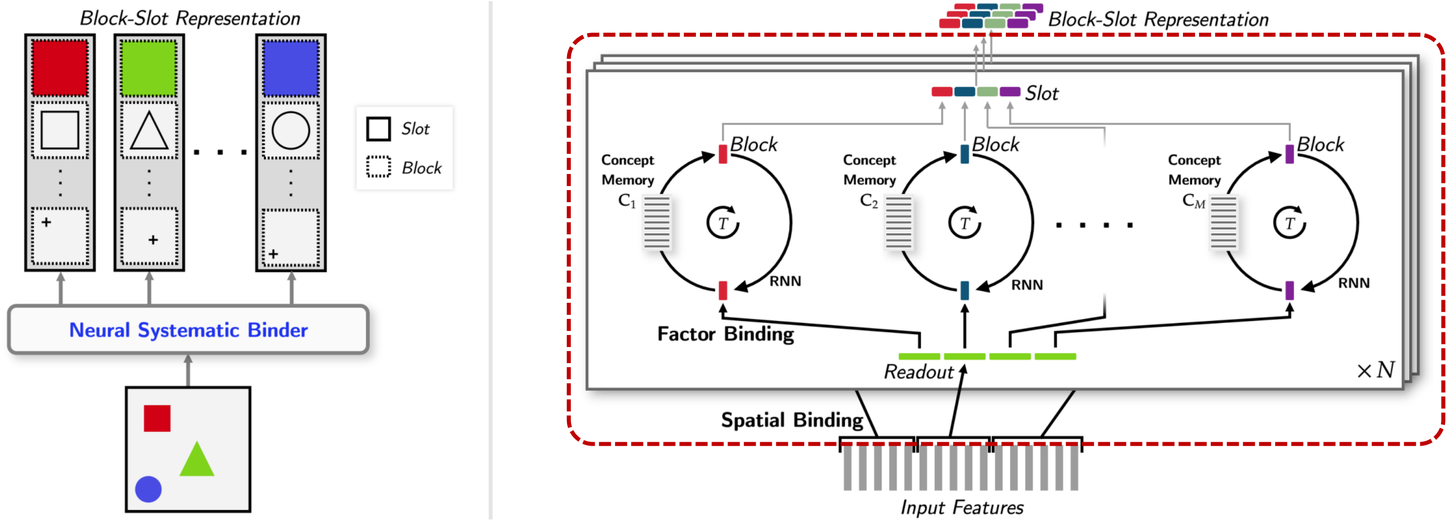


Control the slot position and scale

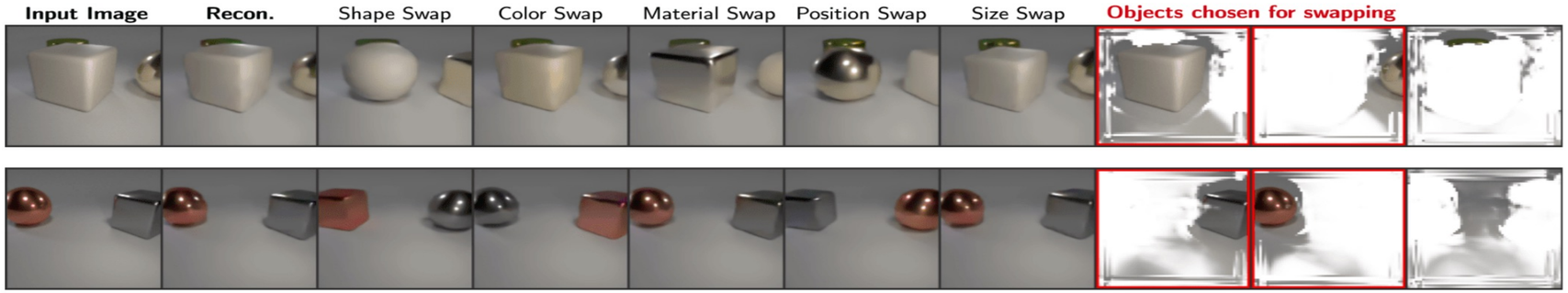
Biza, Ondrej, et al. "Invariant slot attention: Object discovery with slot-centric reference frames.“ ICML 2023. (image credit)

Dividing Slots into Blocks

Question: How to disentangle the property of the slot for manipulation?



SysBinder represents a slot by concatenating multiple blocks.

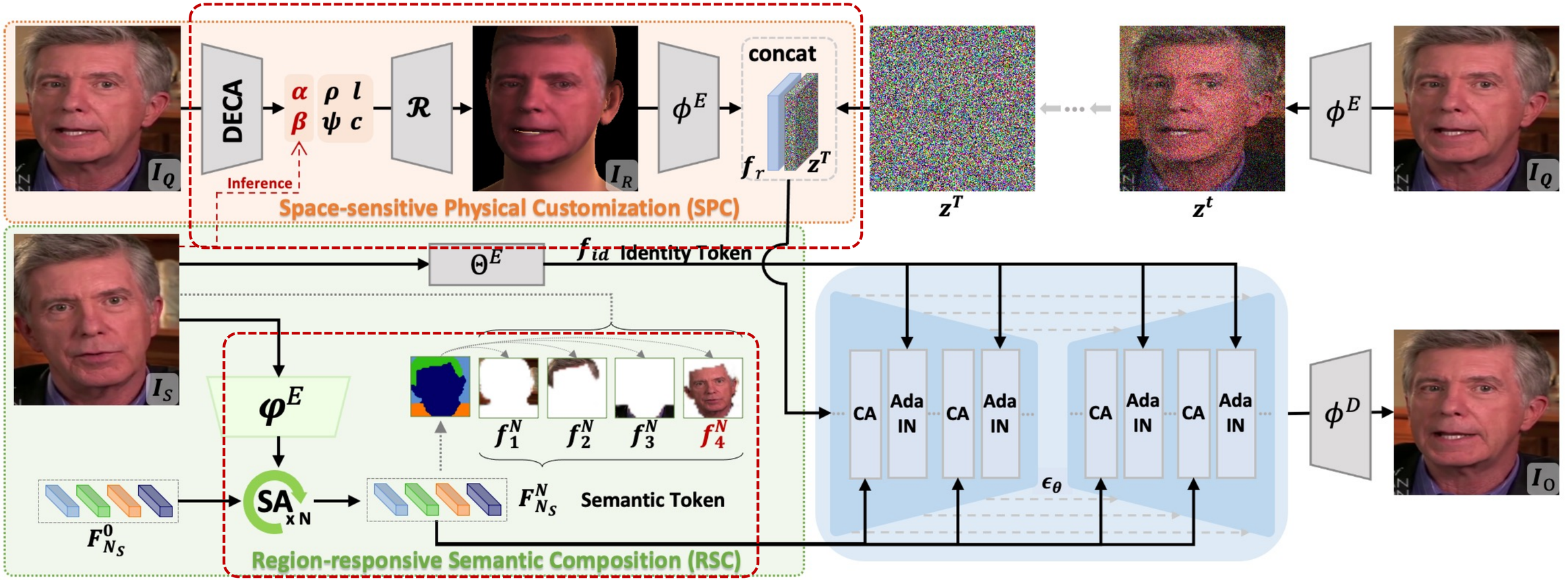


Swap a specific property of two objects

Singh, Gautam, Yeongbin Kim, and Sungjin Ahn. "Neural Systematic Binder." ICLR 2023. (image credit)

Slots as Conditions in Facial Appearance Editing

Manipulating



Pretraining of RSC: decouple face into semantic tokens with slot attention

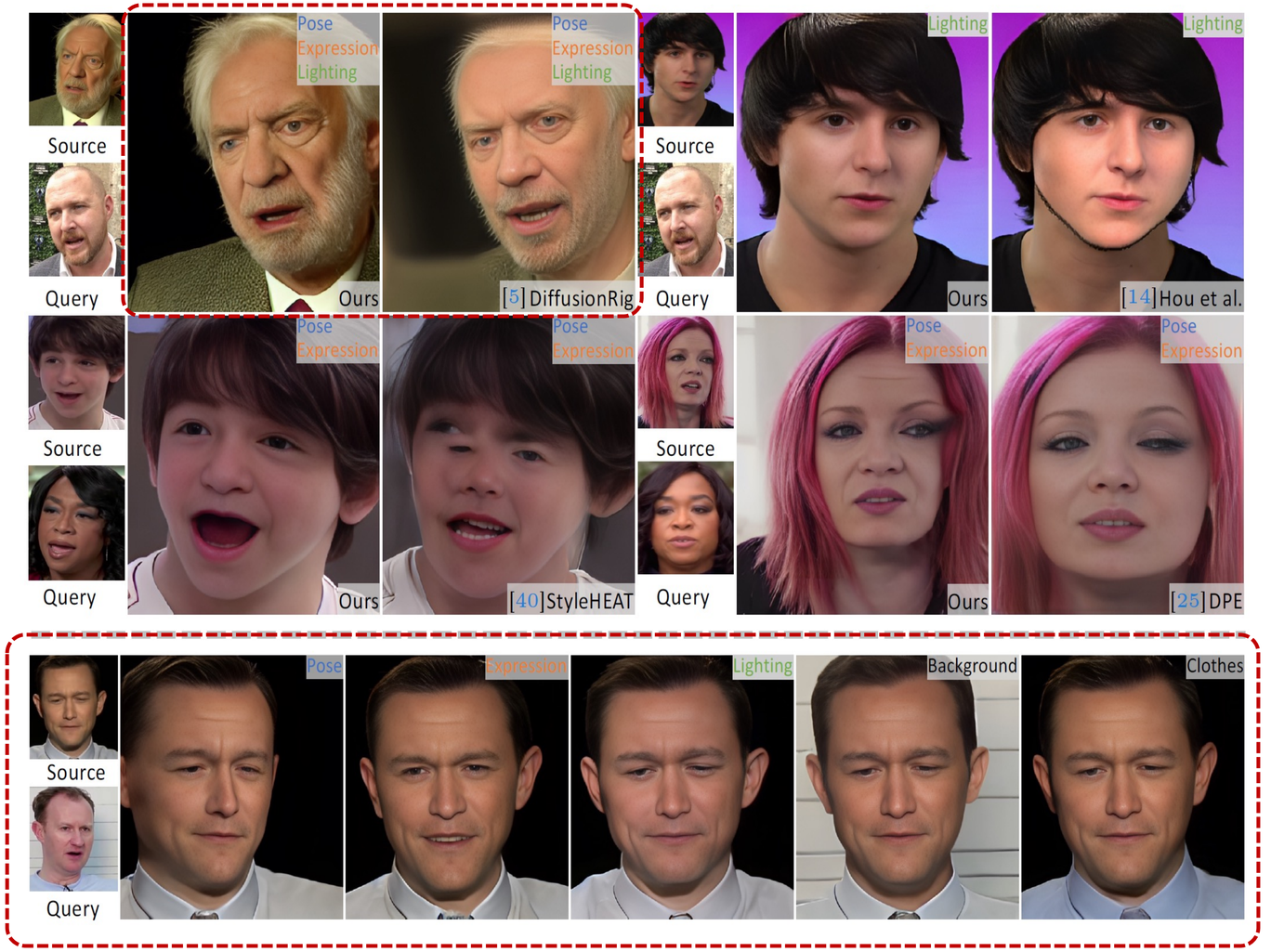
Finetuning of LDM: feed the semantic token as condition to LDM

Real-world face dataset !

Wang, Qilin, et al. “DiffFAE: Advancing High-fidelity One-shot Facial Appearance Editing with Space-sensitive Customization and Semantic Preservation.” arXiv:2403.17664.

Results of Slot based Facial Appearance Editing

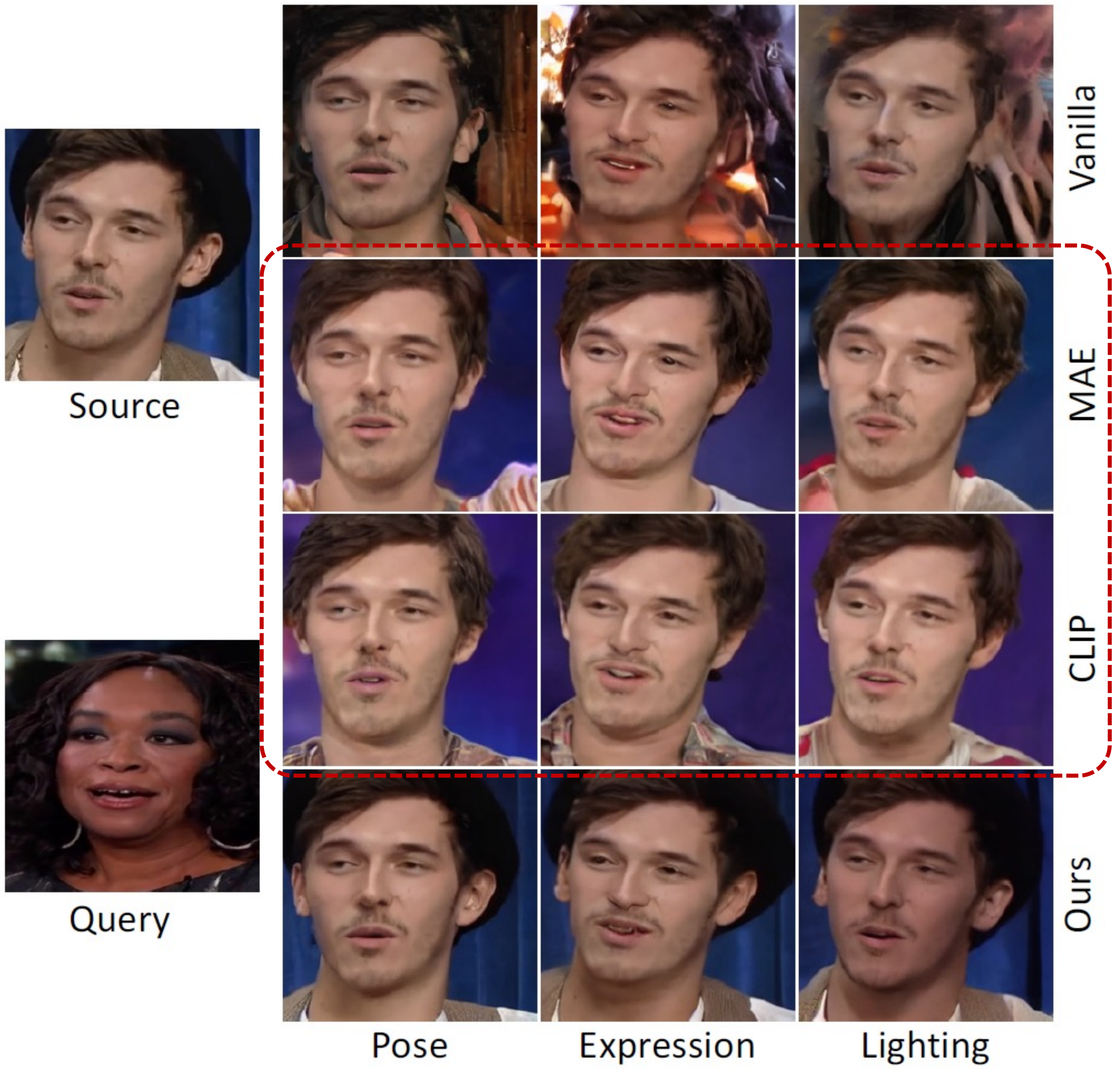
Manipulating



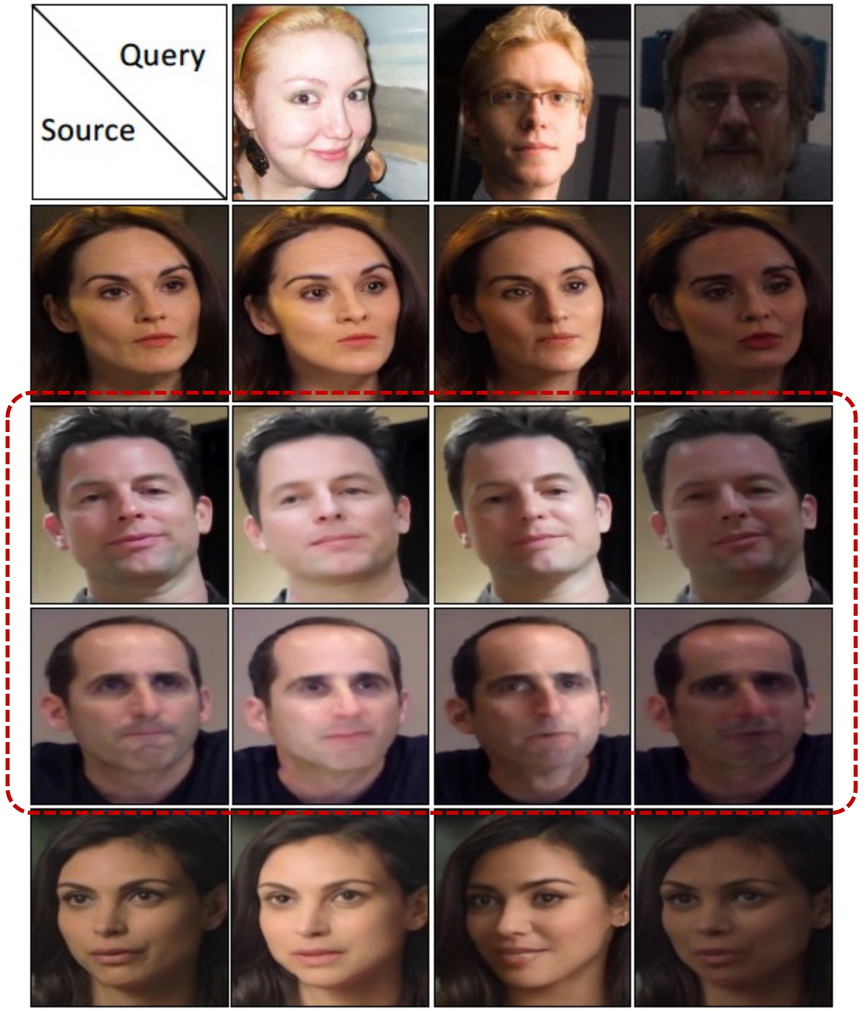
Facial compositional editing of pose, expression, and lighting.

Wang, Qilin, et al. “DiffFAE: Advancing High-fidelity One-shot Facial Appearance Editing with Space-sensitive Customization and Semantic Preservation.” arXiv:2403.17664.

‘Fancy’ Manipulating Slots in Appearance Editing



Semantic tokens extracted by slot attention outperform MAE and CLIP

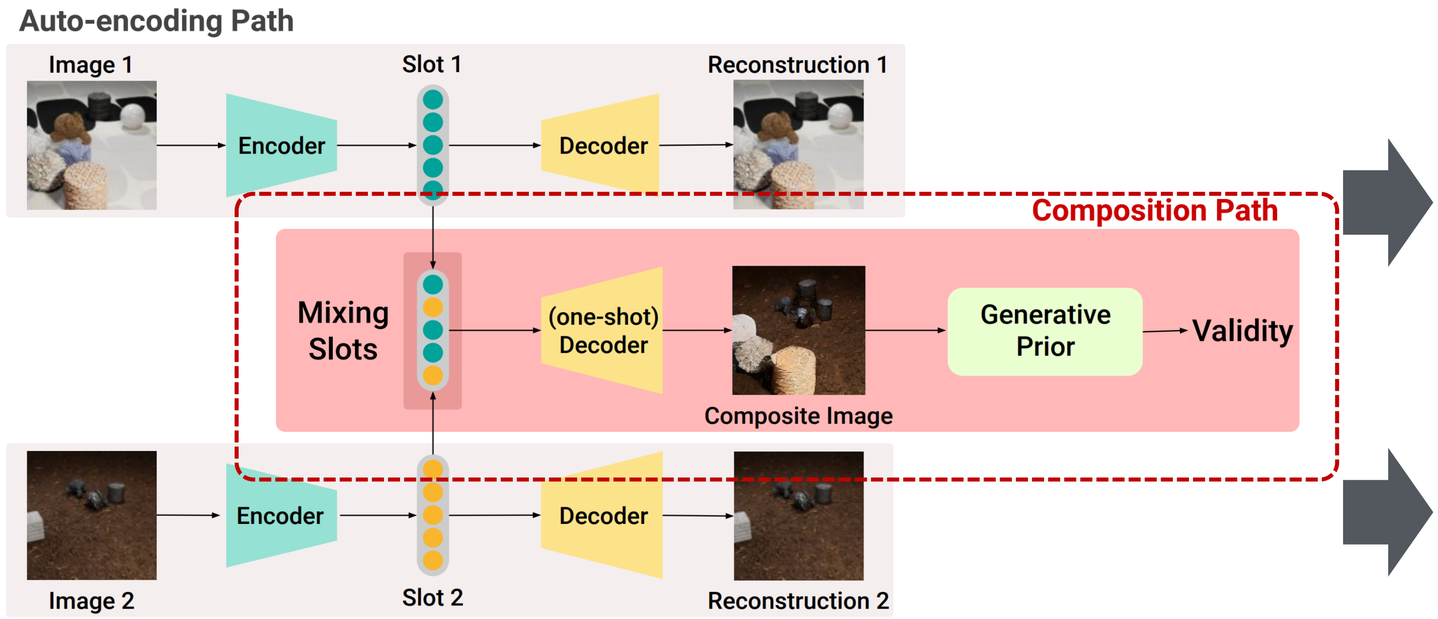


Additional relighting results under various and challenging lighting conditions.

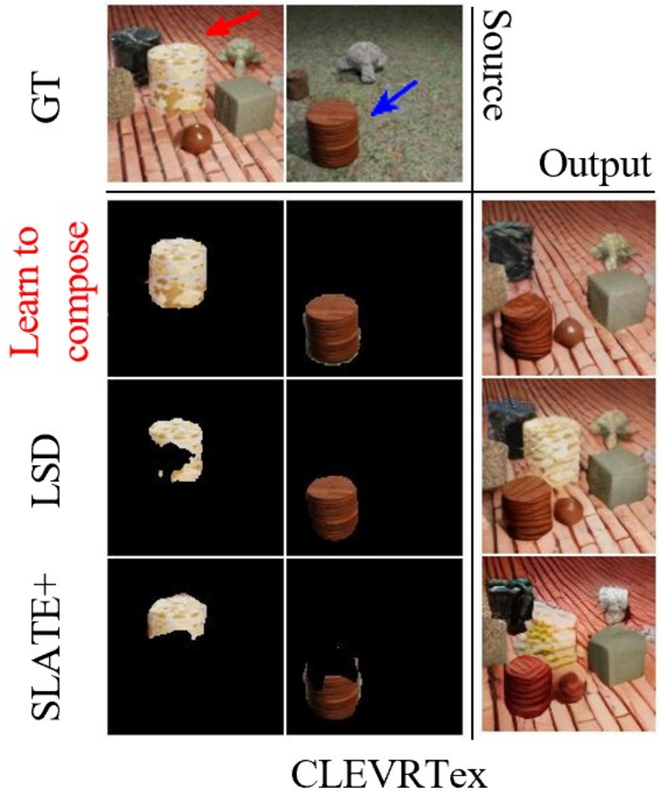
Wang, Qilin, et al. “DiffFAE: Advancing High-fidelity One-shot Facial Appearance Editing with Space-sensitive Customization and Semantic Preservation.” arXiv:2403.17664.

Enhancing Compositionality of Slots

Better Compositionality



Promoting compositionality by blending slots from two separate images



Better object mask and composition result with composition branch

Jung, Whie, et al. "Learning to Compose: Improving Object Centric Learning by Injecting Compositionality." ICLR 2024. (image credit)

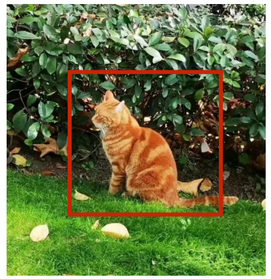
The Applications of Object-Centric Learning

- Image Manipulation
- **Segmentation**
- Embodied AI
- Discussion

The Settings of (Video) “Segmentation”

Only visual information

Self-Supervised



A cat sleeping in the fork of a tree

Weakly-Supervised

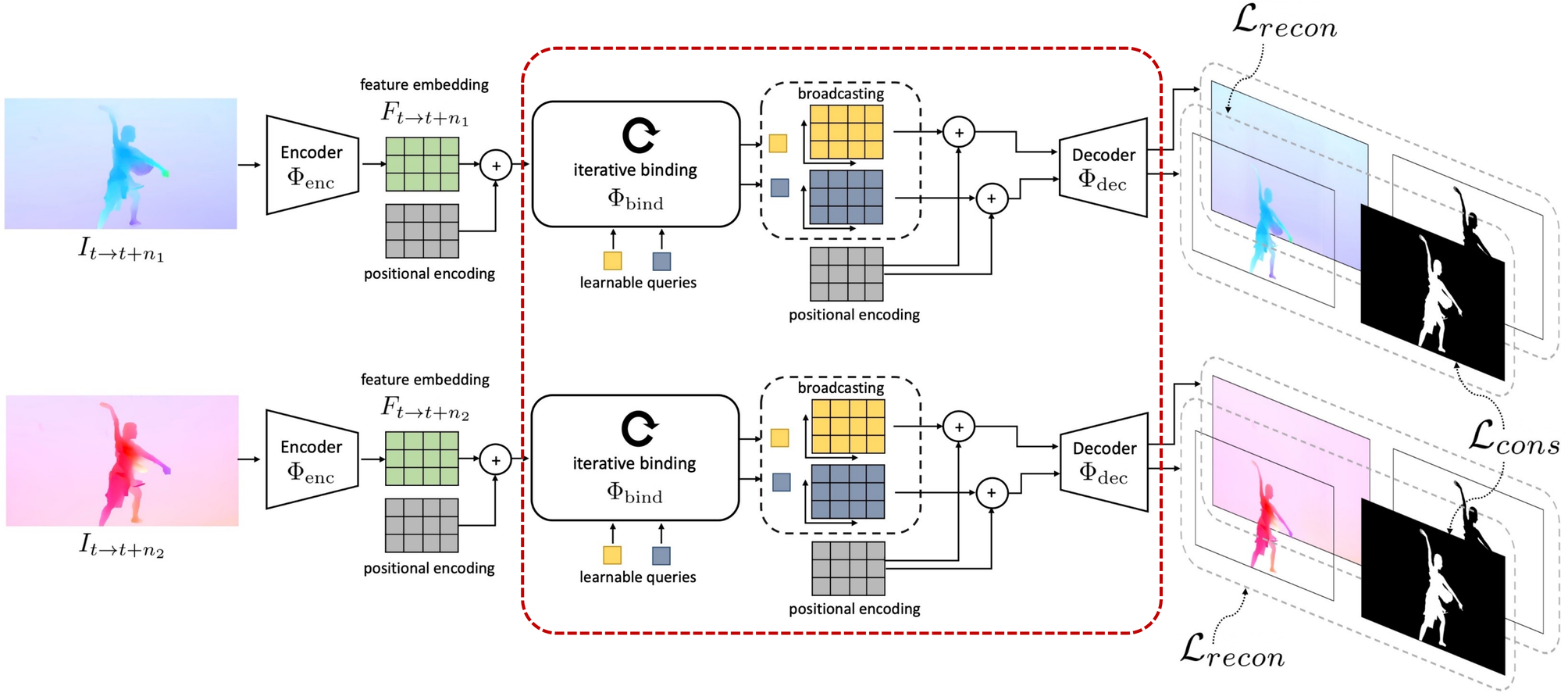
Image/video

masks

Fully-Supervised

Self-Supervised Approach

Motion Segmentation

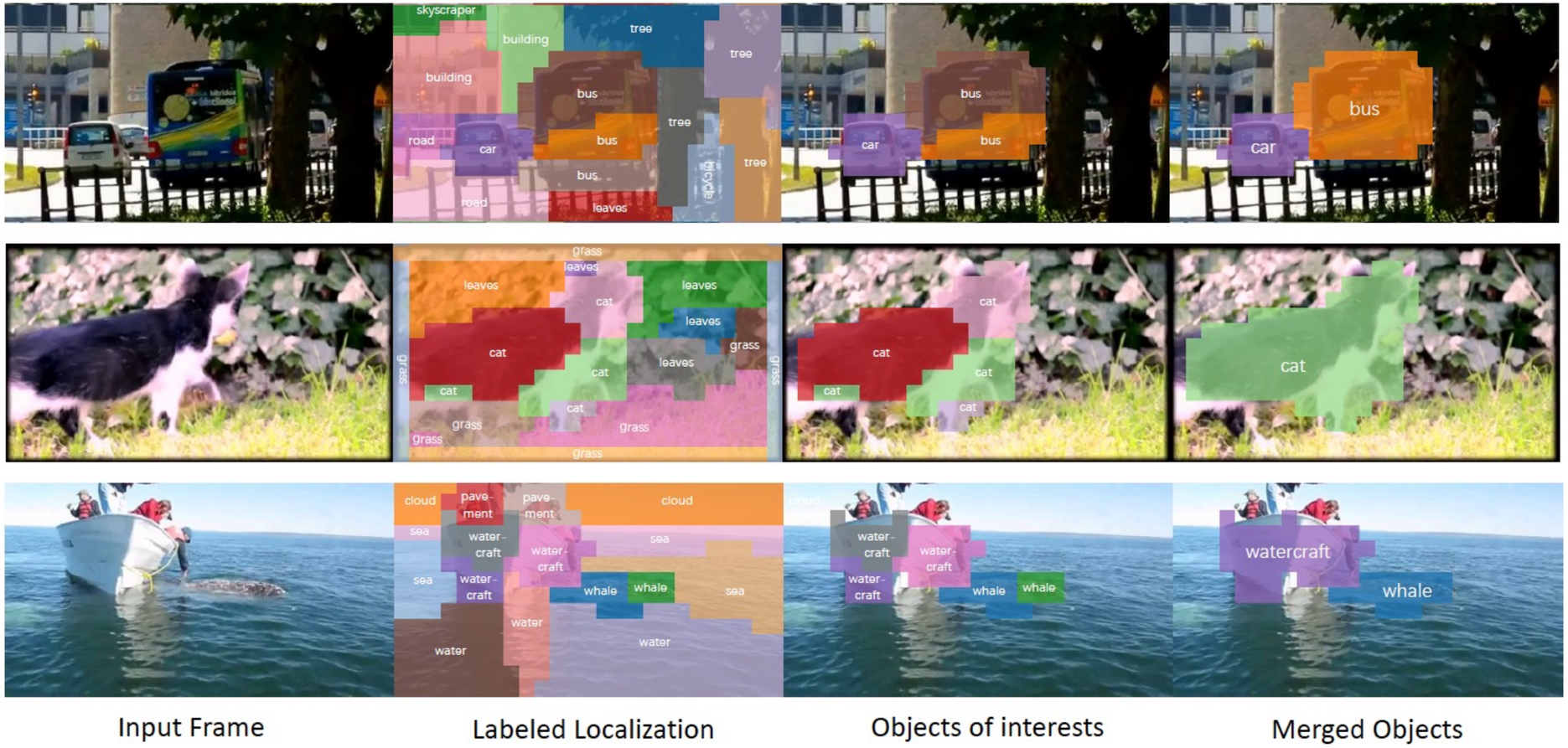


Motion segmentation utilizes slot attention to group the optical flow into foreground and background.

Yang, Charig, et al. "Self-supervised video object segmentation by motion grouping." ICCV 2021. (image credit)

Self-Supervised Approach

Video-level Slot

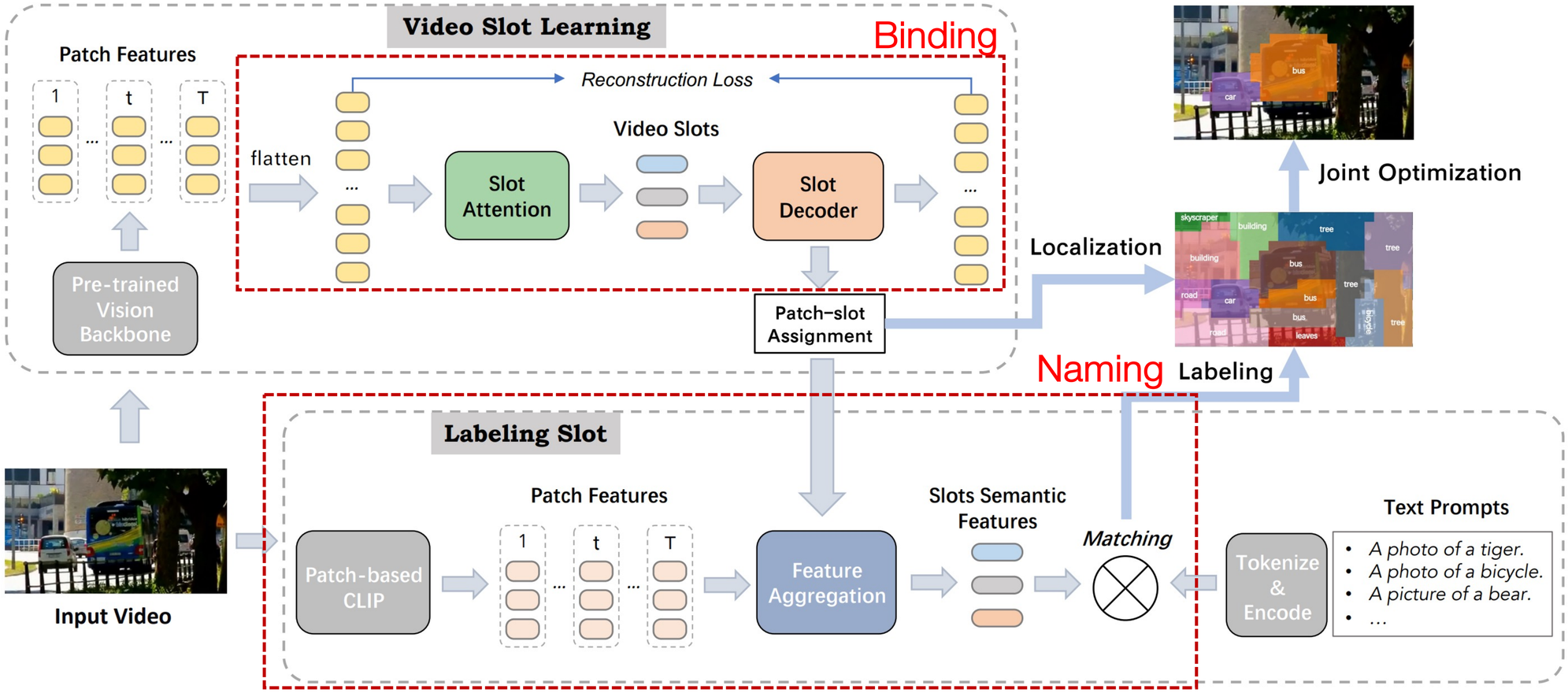


It is an unsupervised approach to localize and name objects in real-world videos.

Fan, Ke, et al. "Unsupervised Open-Vocabulary Object Localization in Videos." ICCV 2023.

Self-Supervised Approach

Video-level Slot



Our method uses a single slot to represent an object throughout the entire video and groups extracted features using a self-supervised vision transformer

Fan, Ke, et al. "Unsupervised Open-Vocabulary Object Localization in Videos." ICCV 2023.

Self-Supervised Approach

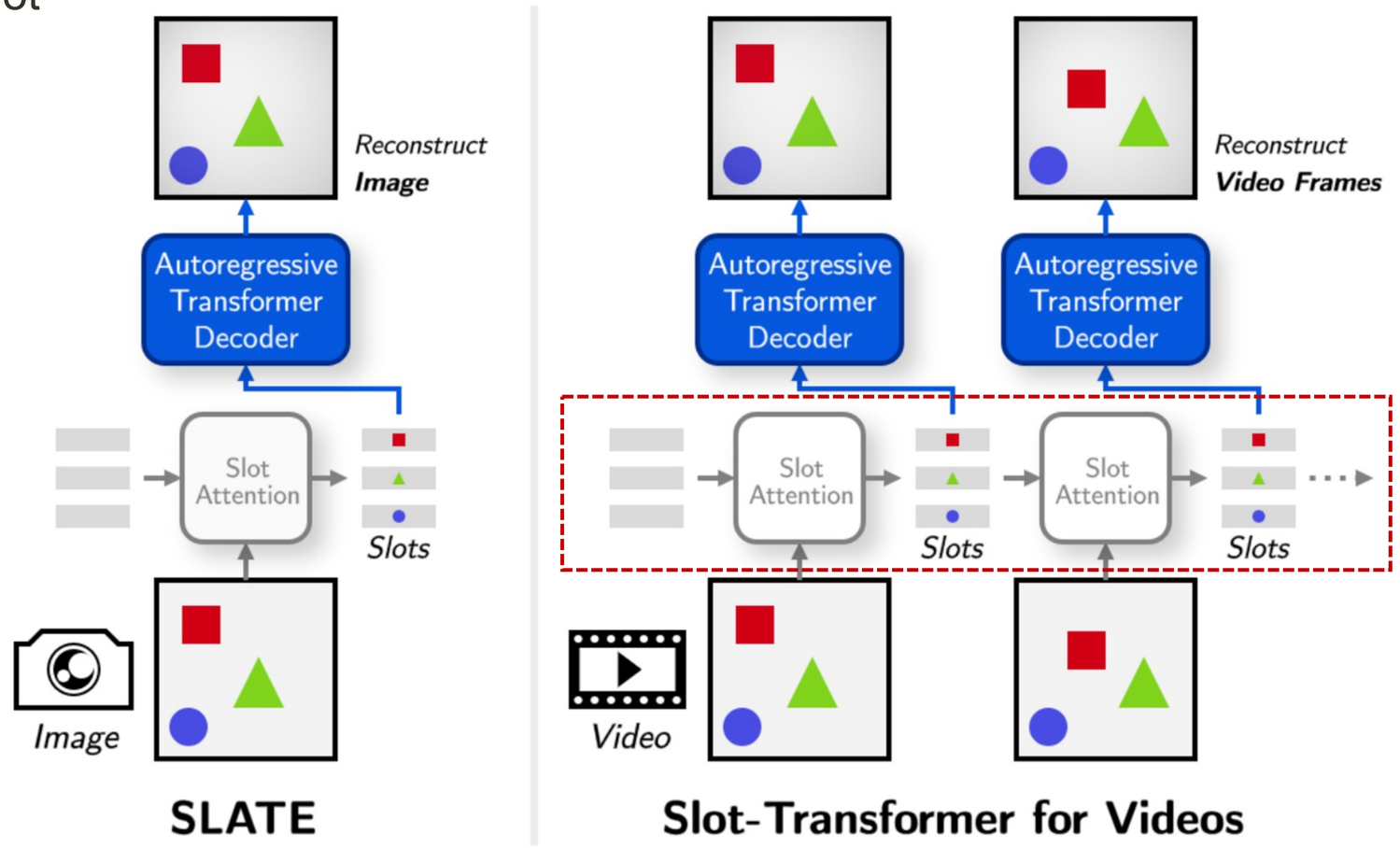
Video-level Slot

Our method ensures temporal consistency by consistently localizing the same object with the same slot across frames.

Fan, Ke, et al. "Unsupervised Open-Vocabulary Object Localization in Videos." ICCV 2023. (image credit)

Self-Supervised Approach

Recurrent Image-level Slot

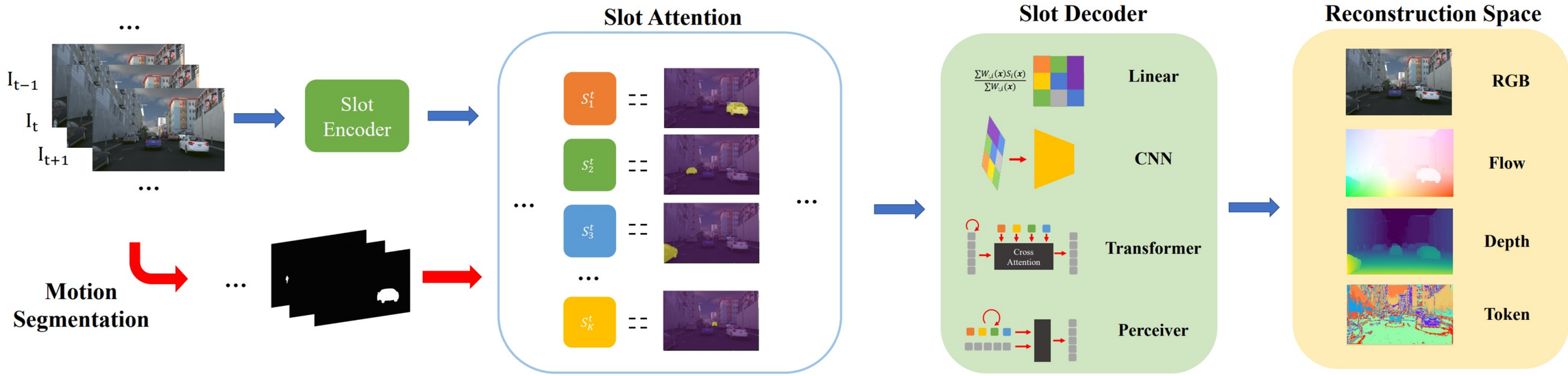


STEVE (**S**lot-**T**ransform**E**r for **V**id**E**os) generalized SLATE to video by recurrent using slots from the previous frame as the initialization for the next frame

Singh, Gautam, Yi-Fu Wu, and Sungjin Ahn. "Simple unsupervised object-centric learning for complex and naturalistic videos." NeurIPS 2022. (image credit)

Self-Supervised Approach

Pseudo Label



MoToK distill the information from the pseudo segmentation masks produced by an off-the-shelf motion segmentation method

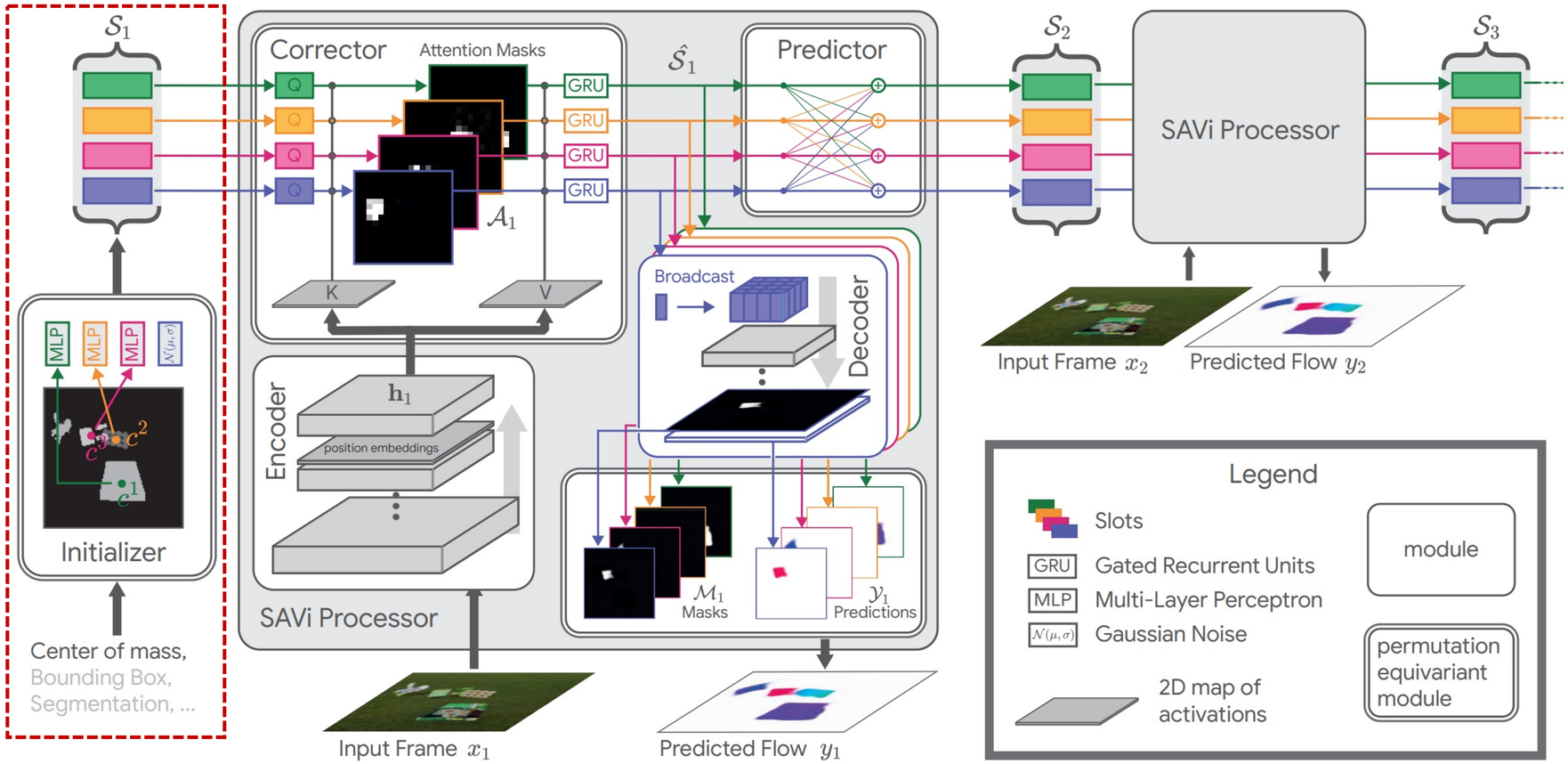
$$\mathcal{L}_{\text{motion}} = \sum_{i=1}^{K} \mathbb{1}_{\{m_i \neq \emptyset\}} \mathcal{L}_{\text{seg}}(m_i, W^t_{:,\hat{\sigma}(i)}),$$

$$\hat{\sigma} = \arg\min_{\sigma} \sum_{i=1}^{K} \mathcal{L}_{\text{seg}}(m_i, W^t_{:,\sigma(i)}),$$

Bao, Zhipeng, et al. "Discovering objects that can move." CVPR 2022.
Bao, Zhipeng, et al. "Object discovery from motion-guided tokens." CVPR 2023. (image credit)

Weakly Supervised Approach

Initial Frame Supervision

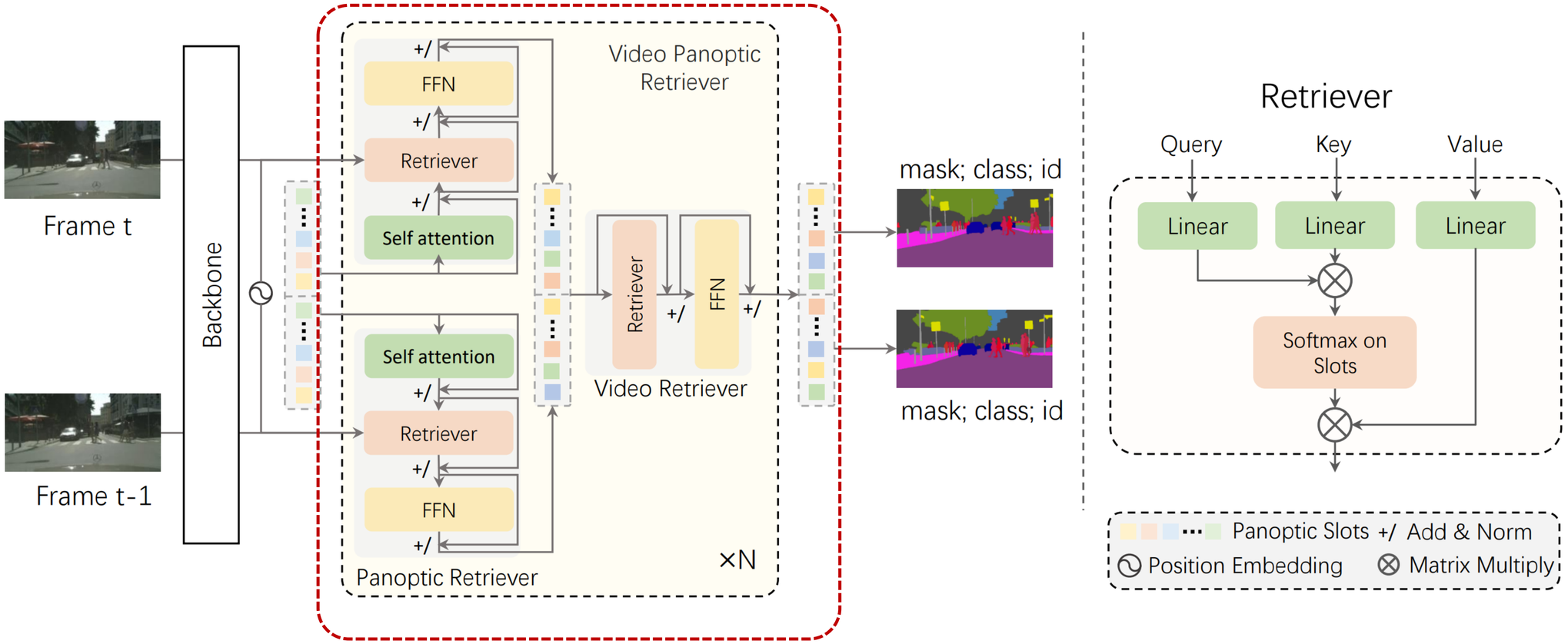


SAVi and SAVi++ track and segment multiple objects with only weak supervision for the video's first frame.

Kipf, Thomas, et al. "Conditional object-centric learning from video." ICLR 2022. (image credit)
Elsayed, Gamaleldin, et al. "Savi++: Towards end-to-end object-centric learning from real-world videos." NeurIPS 2022.

Fully Supervised Approach

Video Panoptic Segmentation

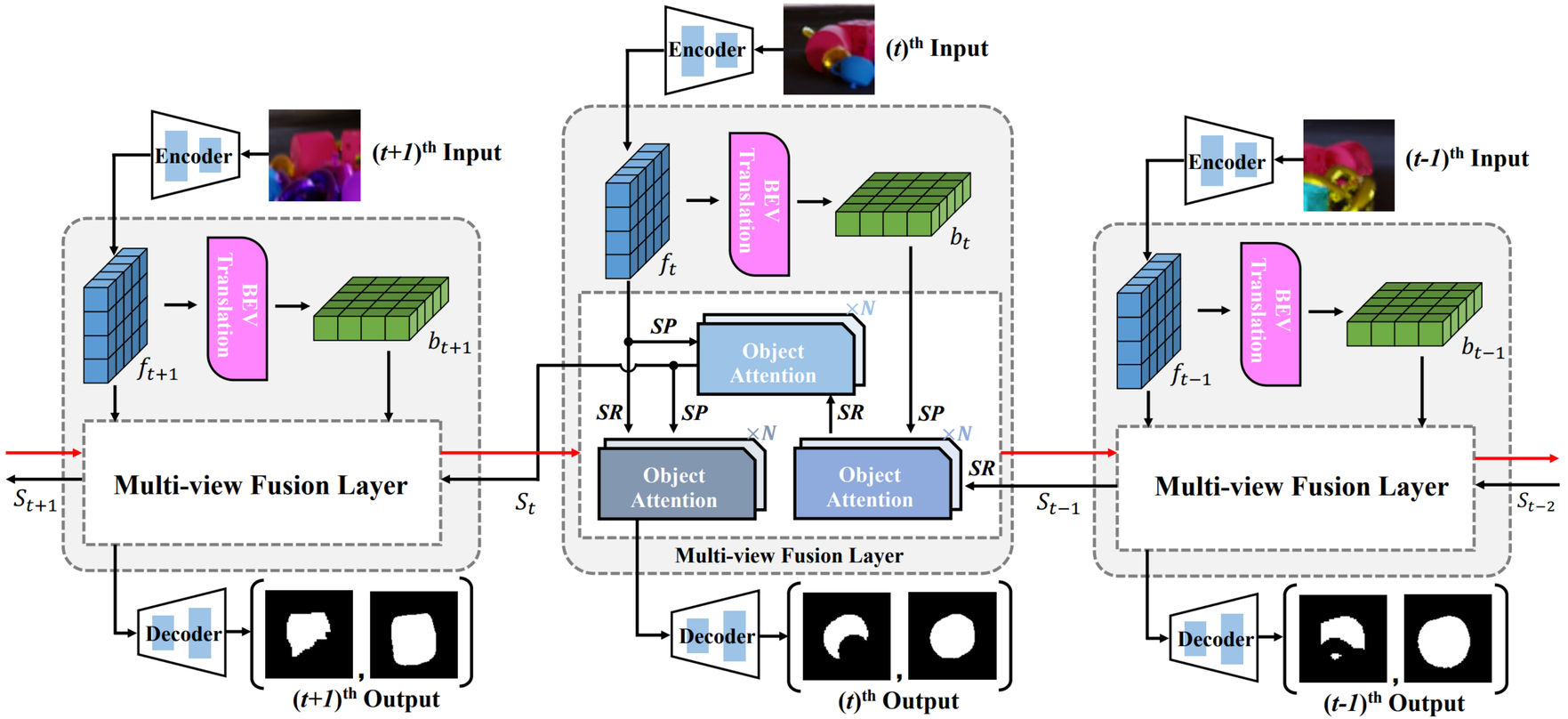


Slot-VPS uses slot attention to extract panoptic slots, which are then decoded into object masks and classes.

Zhou, Yi, et al. "Slot-vps: Object-centric representation learning for video panoptic segmentation." CVPR 2022. (image credit)

Fully Supervised Approach

Amodal Video Segmentation



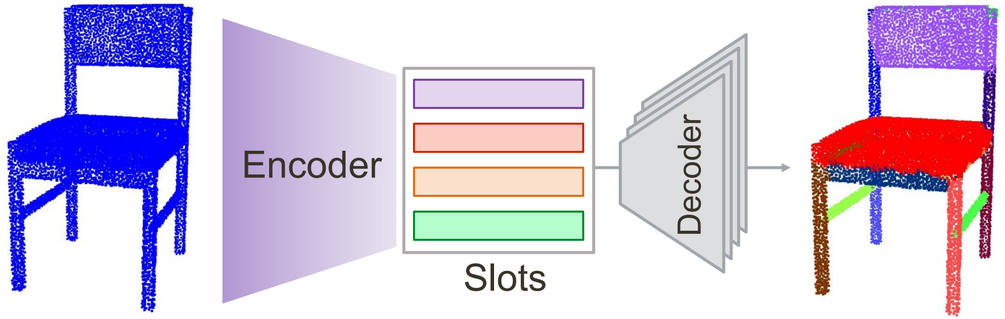
EoRaS utilize slots to recurrently integrate front-view feature and BEV features.

Fan, Ke, et al. "Rethinking amodal video segmentation from learning supervised signals with object-centric representation." ICCV 2023. (image credit)

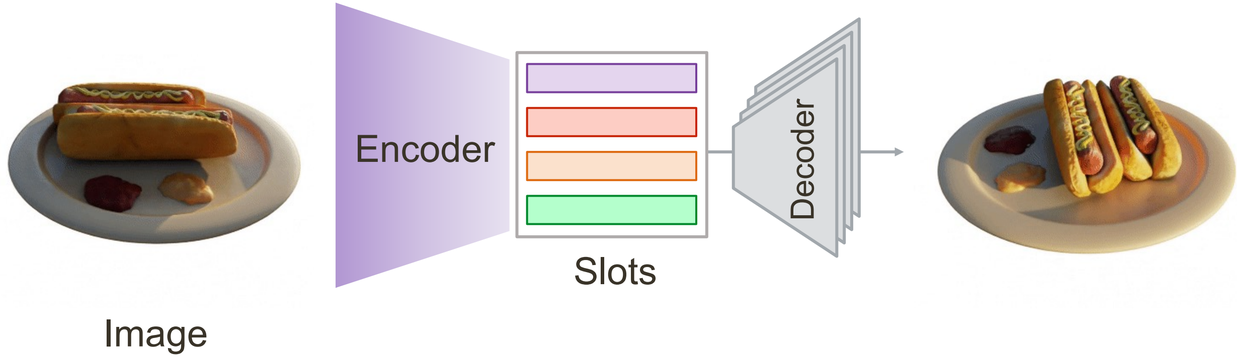
The Applications of Object-Centric Learning

- Image Manipulation
- Segmentation
- **Embodied AI**
- Discussion

Slots from 3D Modeling to Embodied AI



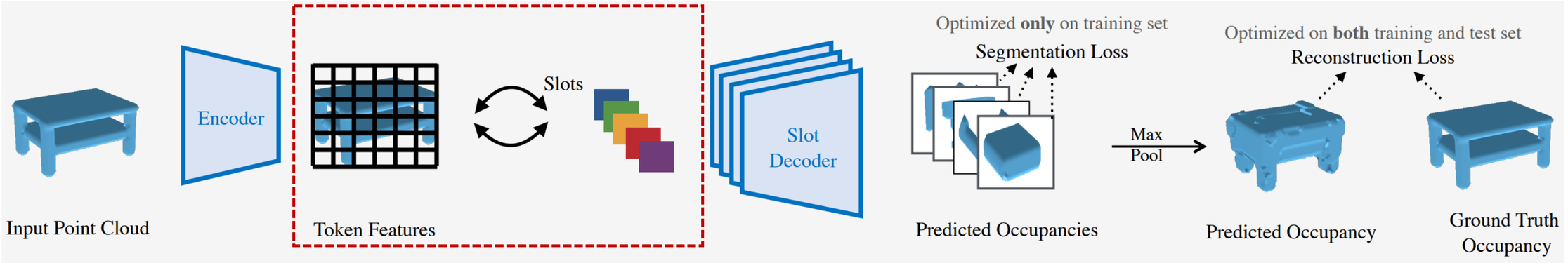
Generic 3D



Lifting 2D to 3D

Decomposing Occupancy Network

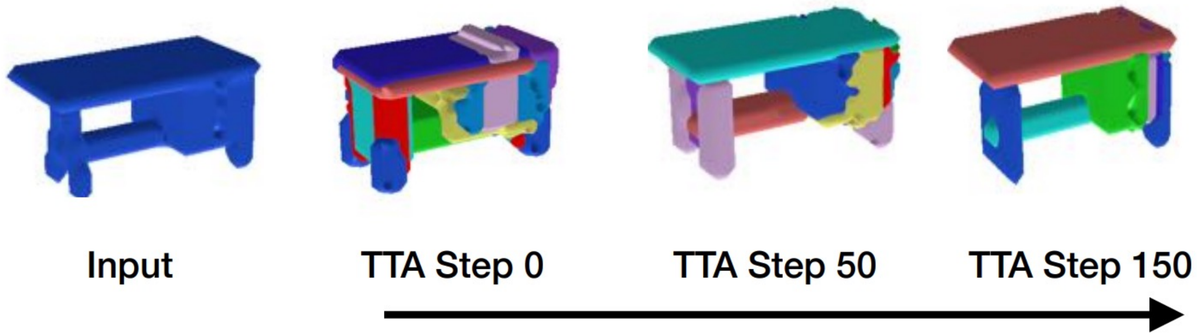
Generic 3D



Pipeline of SlotTTA point cloud reconstruction

Joint optimization of segmentation and reconstruction

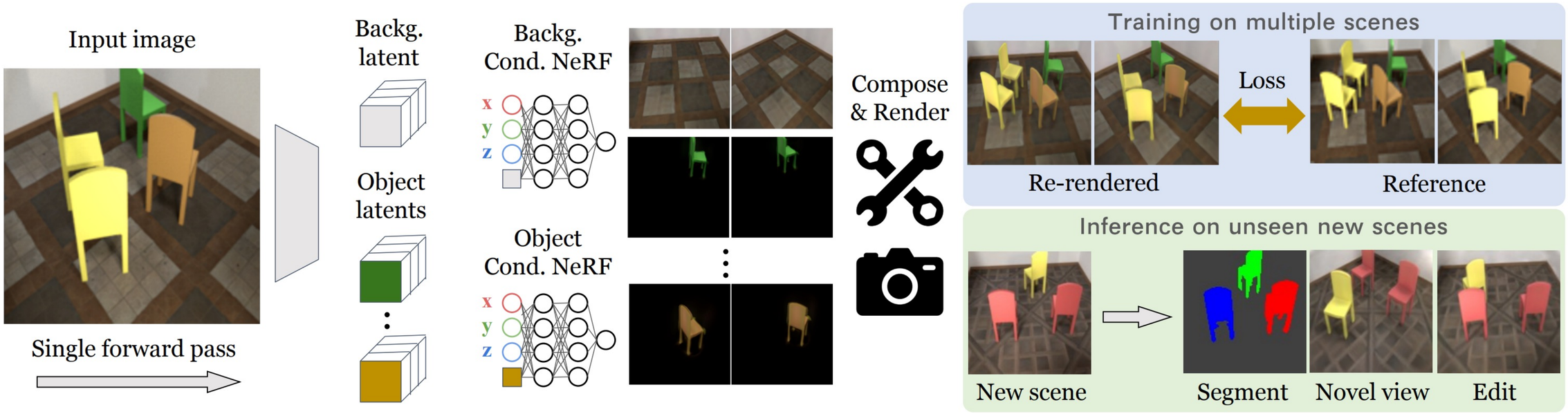
$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \lambda_s l_{seg}(x_i, y_i; \theta) + \lambda_r l_{recon}(x_i; \theta)$$



Prabhudesai, Mihir, et al."Test-time Adaptation with Slot-Centric Models." ICML 2023. (image credit)

Mapping Slot to Neural Radiance Fields

Lifting 2D to 3D



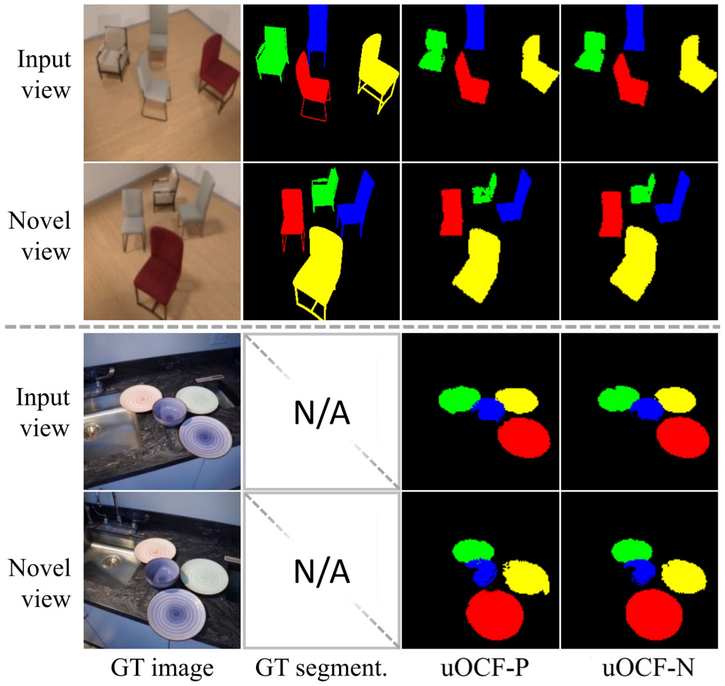
Composition of neural radiance fields

$$\overline{\sigma} = \sum_{i \geq 0} \omega_i \sigma_i, \ \overline{\mathbf{c}} = \sum_{i \geq 0} \omega_i \mathbf{c}_i, \ \text{where} \ \omega_i = \frac{\sigma_i}{\sum_{j \geq 0} \sigma_j}.$$

Yu, Hong-Xing, Leonidas J. Guibas, and Jiajun Wu. "Unsupervised discovery of object radiance fields." ICLR 2022. (image credit)
Luo, Rundong, Hong-Xing Yu, and Jiajun Wu. "Unsupervised Discovery of Object-Centric Neural Fields." arXiv:2402.07376.

Mapping Slot to Neural Radiance Field

Lifting 2D to 3D



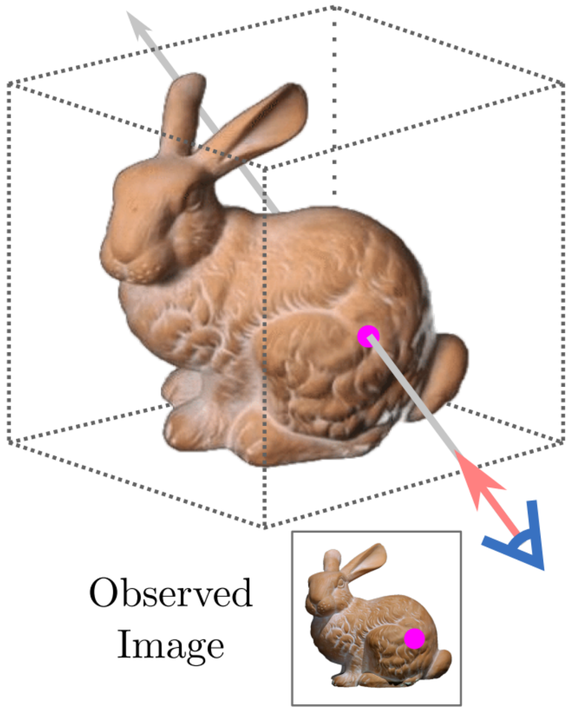
Unseen input image

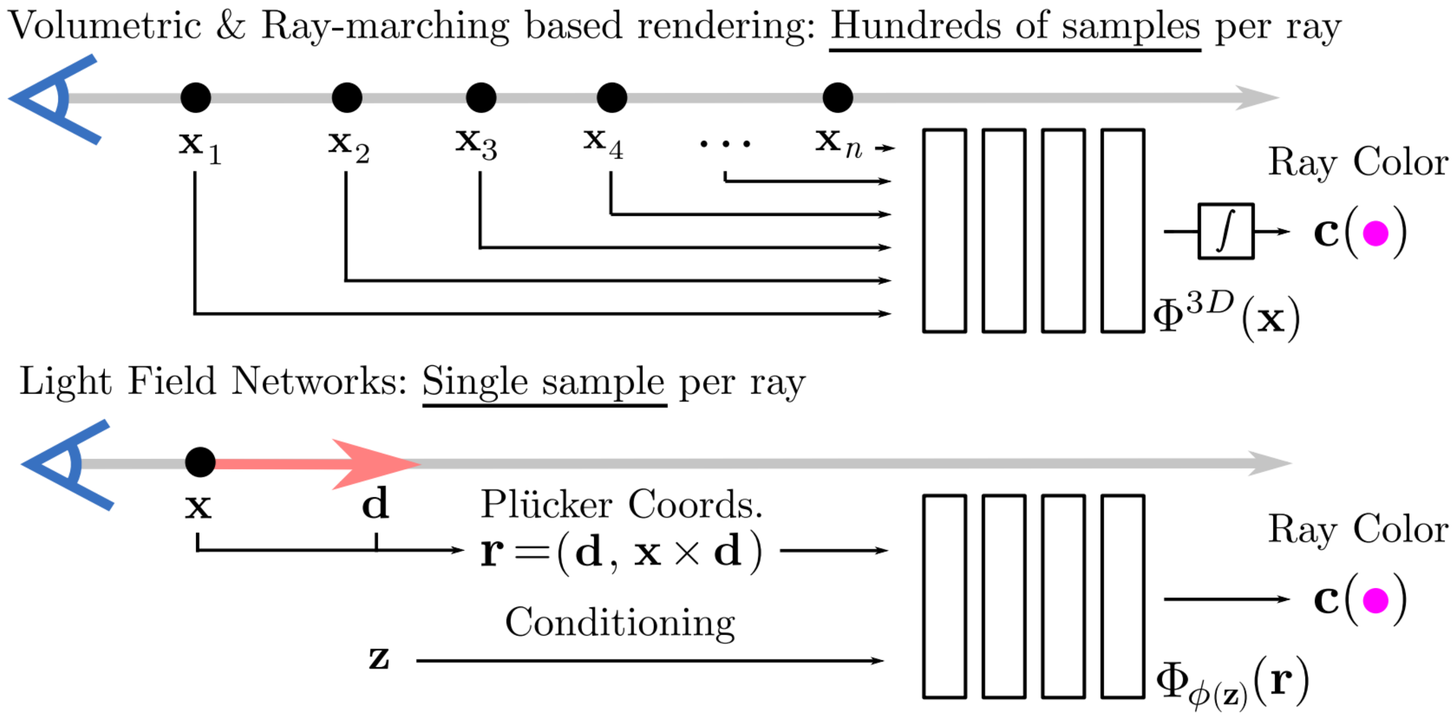
Object-Centric neural Fields enables scene reconstruction and manipulation from novel views

Yu, Hong-Xing, Leonidas J. Guibas, and Jiajun Wu. "Unsupervised discovery of object radiance fields." ICLR 2022.
Luo, Rundong, Hong-Xing Yu, and Jiajun Wu. "Unsupervised Discovery of Object-Centric Neural Fields." arXiv:2402.07376. (image credit)

Mapping Slots to Light Field Network

Lifting 2D to 3D



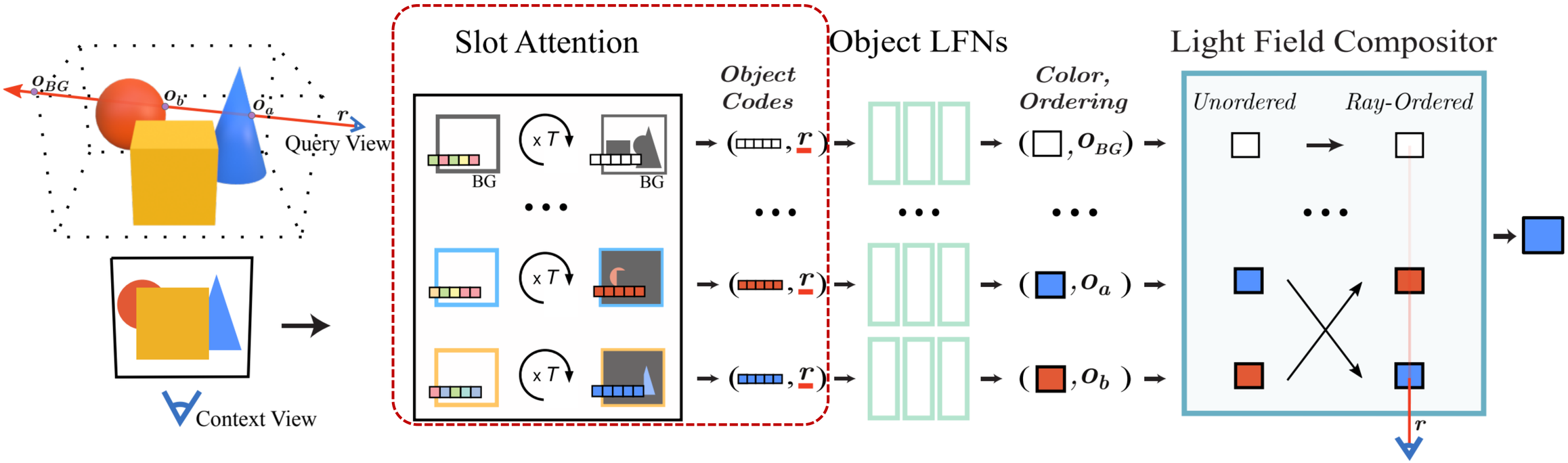


LFNs directly map an oriented ray to the radiance observed along that ray.

Sitzmann, Vincent, et al. "Light field networks: Neural scene representations with single-evaluation rendering." NeurIPS 2021. (image credit)

Mapping Slots to Light Field Network

Lifting 2D to 3D



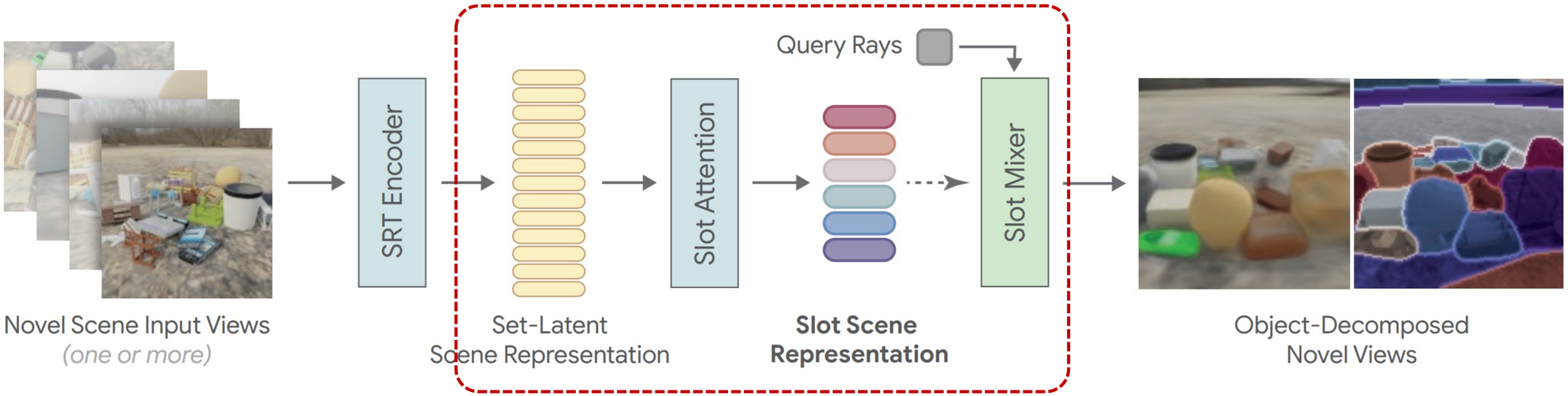
COLF maps each slot to a LFN, and learns ray-visibility weights to composite

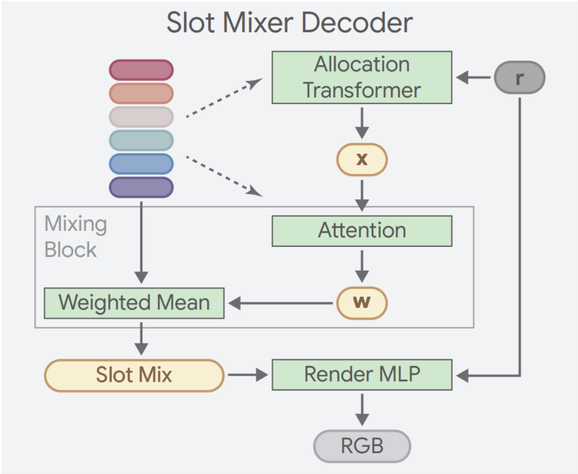
Smith, Cameron Omid, et al. "Unsupervised Discovery and Composition of Object Light Fields." TMLR 2023. (image credit)

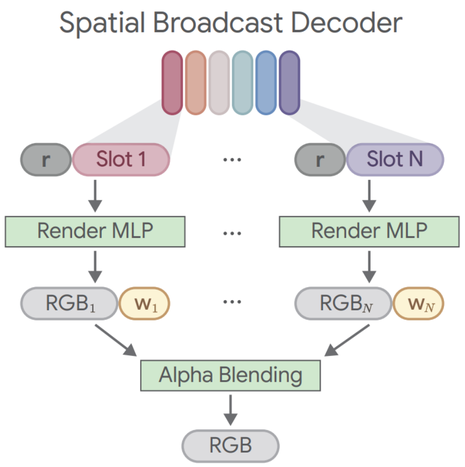
Mapping Slot to Light Field Network

Lifting 2D to 3D

OSRT:



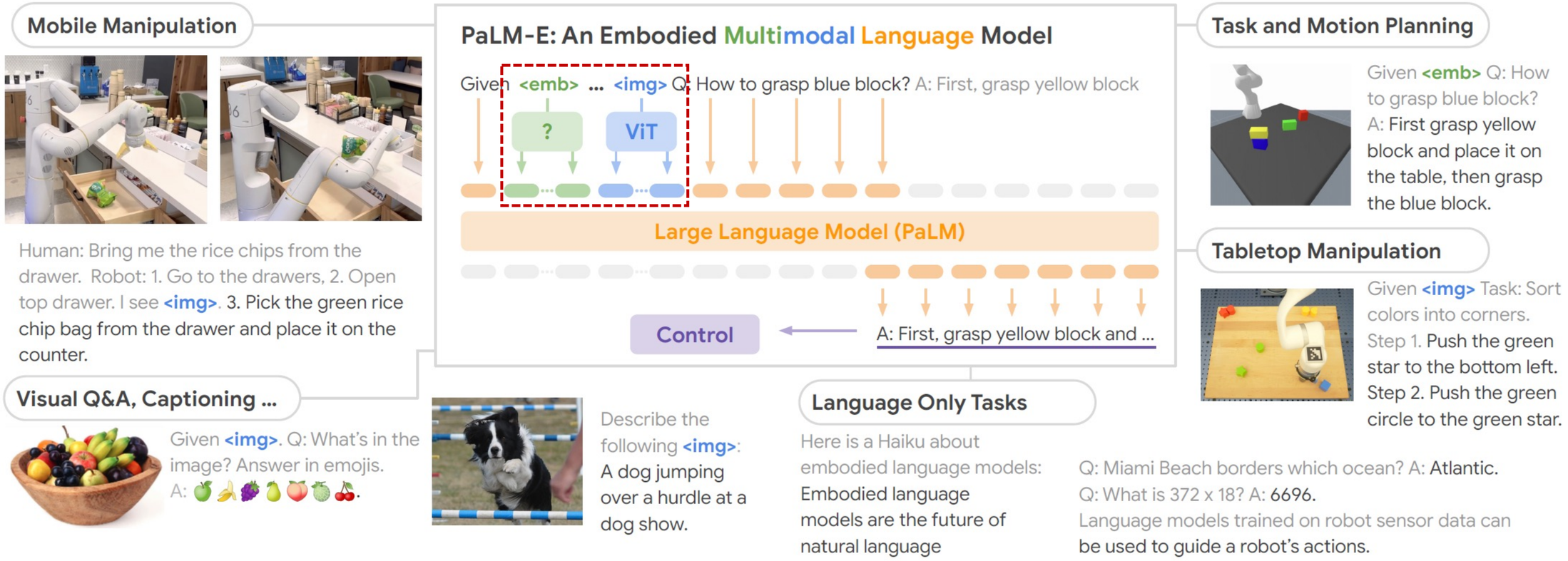




OSRT first mixes the slots, then maps the mixture of slots to a light field network

Sajjadi, Mehdi SM, et al. "Object scene representation transformer." NeurIPS 2022. (image credit)

PaLM-E



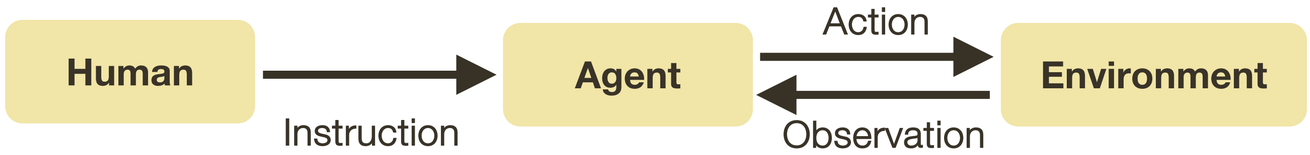
PaLM-E takes object slots learned by ORST as input to train a multimodal language model.

Driess, Danny, et al. "PaLM-E: an embodied multimodal language model." ICML 2023. (image credit)

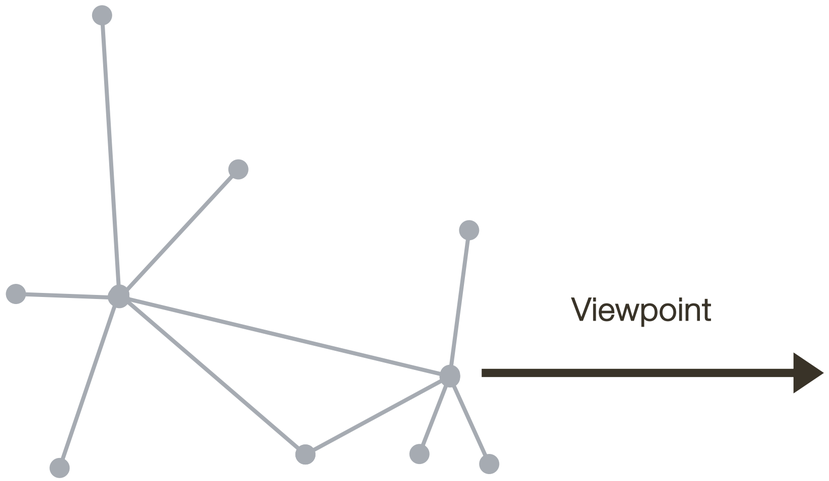
Vision-and-Language Navigation

Problem Formulation

Given a natural language instruction, agent makes decision about the next move automatically based on past and current visual observations.



Room-to-Room Navigation Environment



Viewpoint

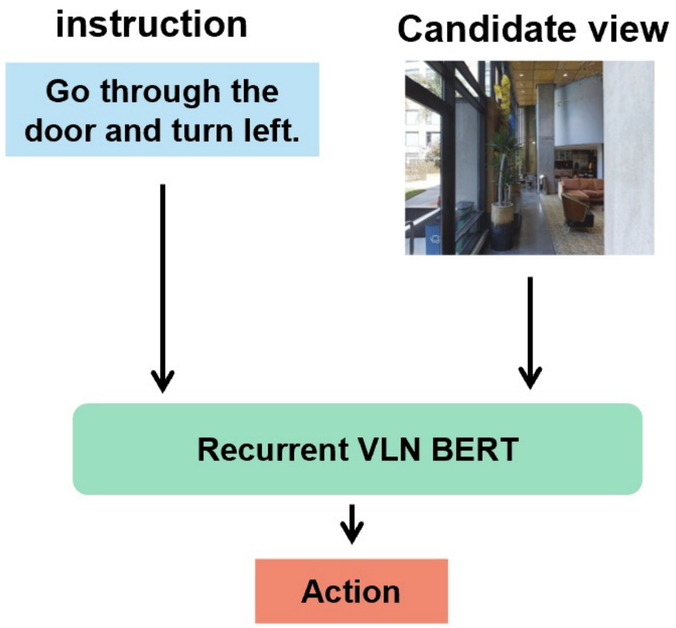
Discretized into 36 views

Zhuang, Yifeng, et al. "Local slot attention for vision and language navigation." ICMR 2022.
Huo, Jingyang, et al. "Geovln: Learning geometry-enhanced visual representation with slot attention for vision-and-language navigation." CVPR 2023.

Vision-and-Language Navigation

Previous Work

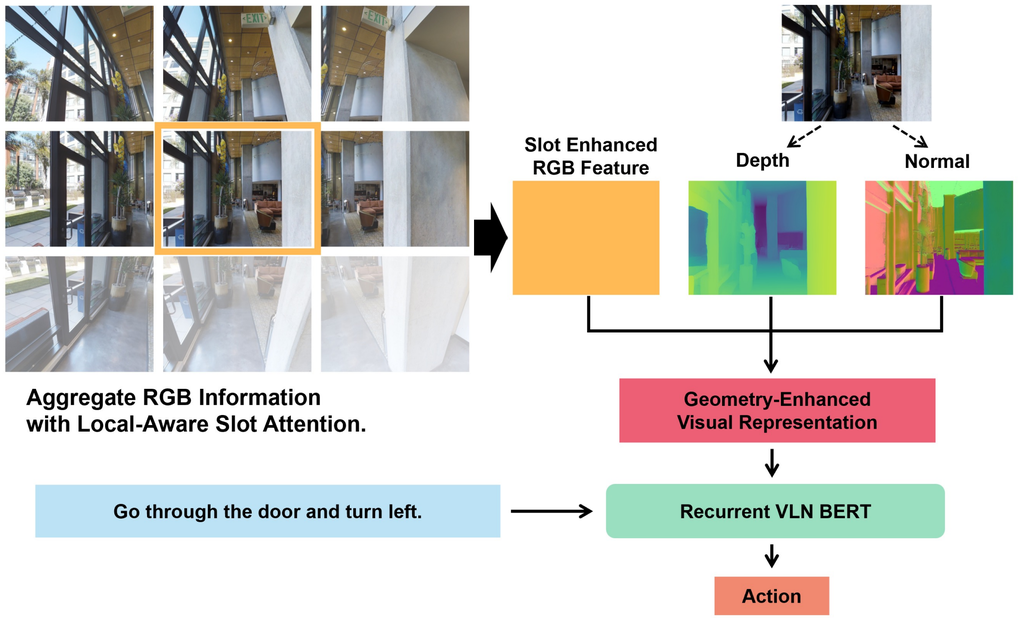
- Utilize RGB images only
- Lack local spatial context around candidate view.



Our Work

Learn range-expanded visual representation with local slot attention

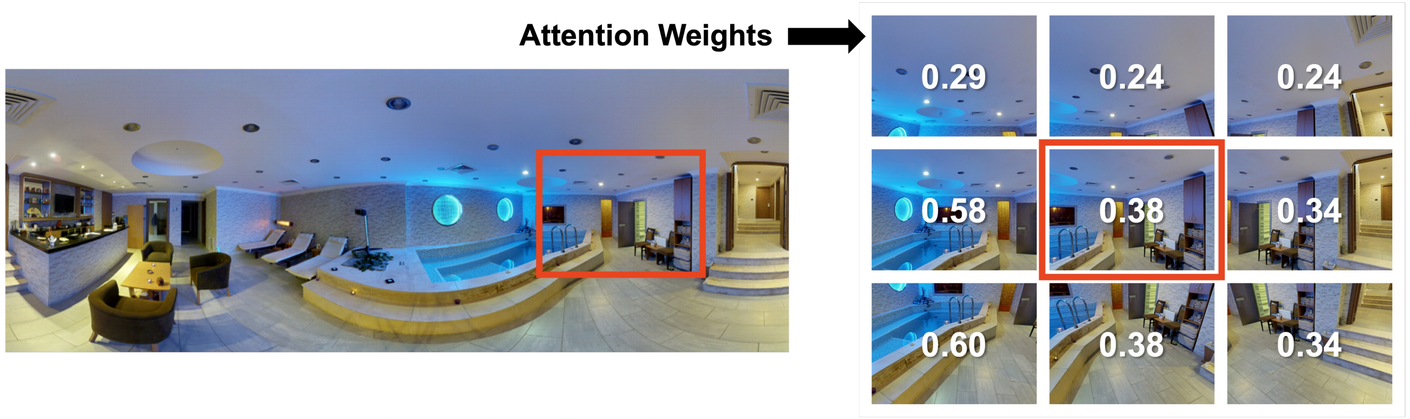
Enhance geometric information with depth maps and normal maps



Zhuang, Yifeng, et al. "Local slot attention for vision and language navigation." ICMR 2022.
Huo, Jingyang, et al. "Geovln: Learning geometry-enhanced visual representation with slot attention for vision-and-language navigation." CVPR 2023.

Vision-and-Language Navigation

How the **local-aware slot attention** module aggregates local observations to candidate views.



Instruction: "Pass the pool then go into the …".

Instruction: "Walk up stairs …".

Zhuang, Yifeng, et al. "Local slot attention for vision and language navigation." ICMR 2022.
Huo, Jingyang, et al. "Geovln: Learning geometry-enhanced visual representation with slot attention for vision-and-language navigation." CVPR 2023.

The Applications of Object-Centric Learning

- Image Manipulation
- Segmentation
- Embodied AI
- Discussions

Discussion: Selecting the Hyperparameter

Slot attention needs to predefine a slot number K

- Fixed K will neglect the variation in object number within dataset
- Wrong K will lead to improper segmentation

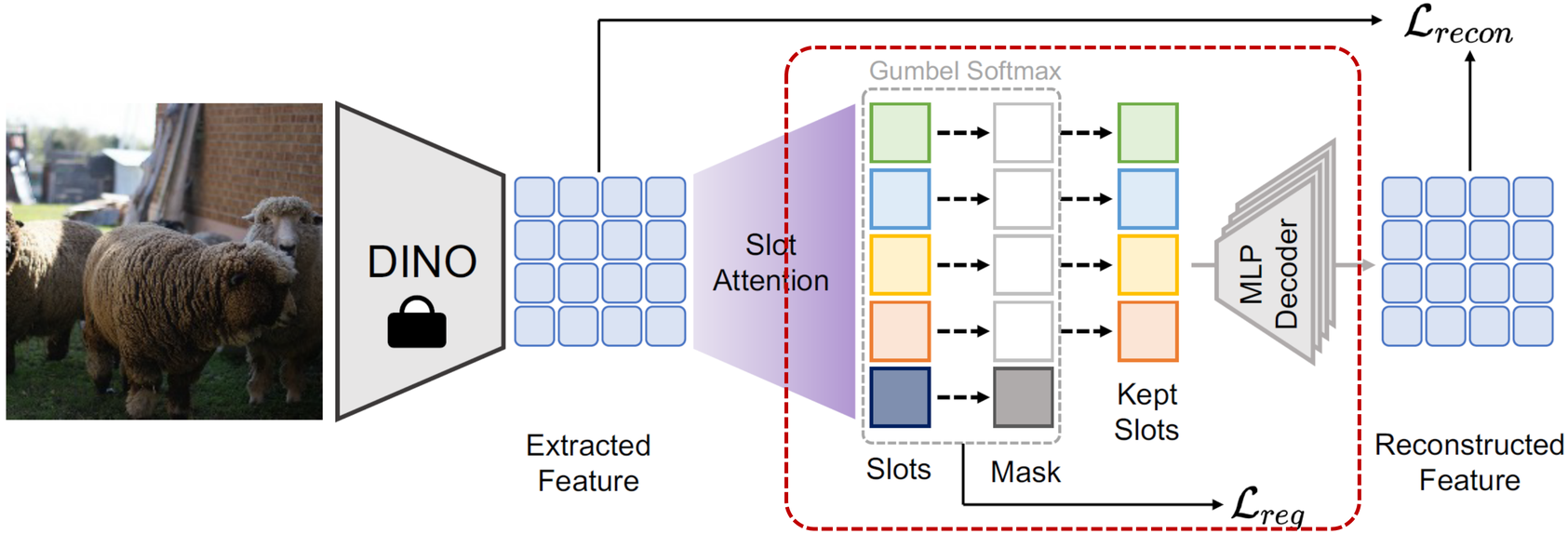
Raw Image | Under-Segmentation | Proper-Segmentation | Over-Segmentation

Question: How to select a slot number for each instance dynamically?

Fan, Ke, et al. "Adaptive slot attention: Object discovery with dynamic slot number." CVPR 2024.

Discussion: How to Select the Hyperparameter?

We regard the dynamic selection problems as a subset selection problem.

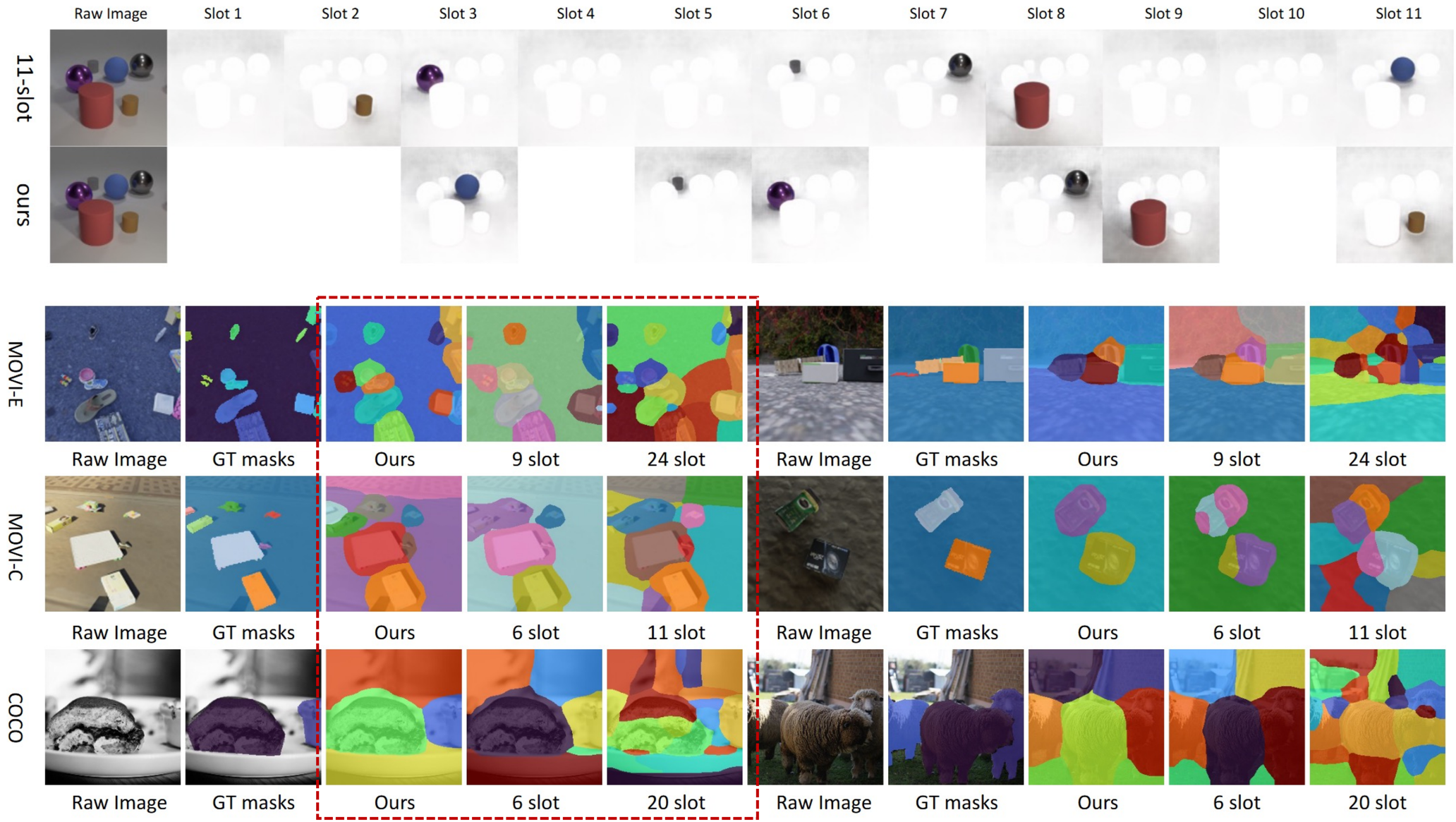


Pipeline of Adaptive Slot Attention

The expectation number of selected slots are utilized as a regularization.

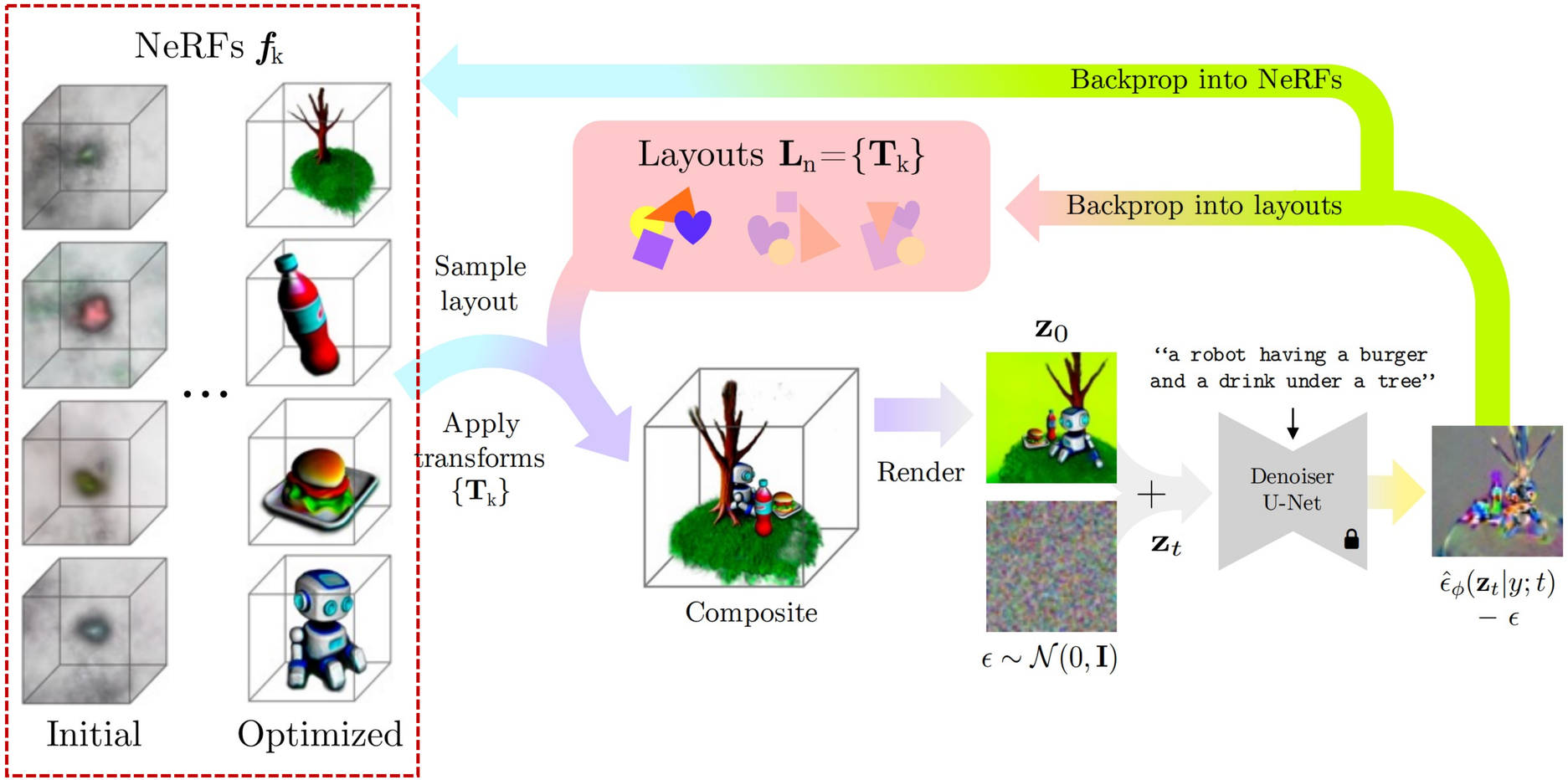
Fan, Ke, et al. "Adaptive slot attention: Object discovery with dynamic slot number." CVPR 2024.

Discussion: How to Select the Hyperparameter?



Fan, Ke, et al. "Adaptive slot attention: Object discovery with dynamic slot number." CVPR 2024.

Discussion: Beyond Slot Attention



Directly learning multiple neural radiance fields
and their layout to decompose

Epstein, Dave, et al. "Disentangled 3D Scene Generation with Layout Learning." arXiv:2402.16936. (image credit)

Discussion: Beyond Slot Attention

Epstein, Dave, et al. "Disentangled 3D Scene Generation with Layout Learning." arXiv:2402.16936. (image credit)

Dr. Yanwei Fu

School of Data Science, Fudan University

Homepage:

http://yanweifu.github.io

yanweifu@fudan.edu.cn

CVPR 2024 Tutorial: Object-centric Representations in Computer Vision

https://object-centric-representation.github.io/object-centric-tutorial-2024/

Paper List: object-centric learning

- [Singh-SLATE22] Singh, Gautam, Fei Deng, and Sungjin Ahn. "Illiterate dall-e learns to compose." ICLR 2022.
- [Oord-VQVAE17] Van Den Oord, Aaron, and Oriol Vinyals. "Neural discrete representation learning." NeuIPS 2017.
- [Kori-Grounded2023] Kori, Avinash, et al. "Grounded Object-Centric Learning." ICLR 2024.
- [Biza-ISA2023] Biza, Ondrej, et al. "Invariant slot attention: Object discovery with slot-centric reference frames.“ ICML 2023.
- [Singh-SysBinder23] Singh, Gautam, Yeongbin Kim, and Sungjin Ahn. "Neural Systematic Binder." ICLR 2023.
- [Wang-DiffFAE2024] Wang, Qilin, et al. “DiffFAE: Advancing High-fidelity One-shot Facial Appearance Editing with Space-sensitive Customization and Semantic Preservation.” arXiv:2403.17664.
- [Jung-LTC2024] Jung, Whie, et al. "Learning to Compose: Improving Object Centric Learning by Injecting Compositionality." ICLR 2024.
- [Wiedemer-PCG2024] Wiedemer, Thaddäus, et al. "Provable Compositional Generalization for Object-Centric Learning." ICLR 2024.

Paper List: object-centric learning

- [Yang-MoSeg2021] Yang, Charig, et al. "Self-supervised video object segmentation by motion grouping." ICCV 2021.
- [Singh-STEVE2022] Singh, Gautam, Yi-Fu Wu, and Sungjin Ahn. "Simple unsupervised object-centric learning for complex and naturalistic videos." NeurIPS 2022.
- [Fan-VOL2023] Fan, Ke, et al. "Unsupervised Open-Vocabulary Object Localization in Videos." ICCV 2023.
- [Bao-Discovering2022]Bao, Zhipeng, et al. "Discovering objects that can move." CVPR 2022.
- [Bao-MoTok2023] Bao, Zhipeng, et al. "Object discovery from motion-guided tokens." CVPR 2023.
- [Xu-Groupvit22] Xu, Jiarui, et al. "Groupvit: Semantic segmentation emerges from text supervision." CVPR 2022.
- [Xu-OVSegmentor23] Xu, Jilan, et al. "Learning open-vocabulary semantic segmentation models from natural language supervision." CVPR 2023.
- [Kipf-Savi2022] Kipf, Thomas, et al. "Conditional object-centric learning from video." ICLR 2022.
- [Elsayed-Savi++2022] Elsayed, Gamaleldin, et al. "Savi++: Towards end-to-end object-centric learning from real-world videos." NeurIPS 2022.
- [Zhou-SlotVPS2022] Zhou, Yi, et al. "Slot-vps: Object-centric representation learning for video panoptic segmentation." CVPR 2022.
- [Fan-EoRaS2023] Fan, Ke, et al. "Rethinking amodal video segmentation from learning supervised signals with object-centric representation." ICCV 2023.

Paper List: object-centric learning

- [Prabhudesai-SlotTTA2023] Prabhudesai, Mihir, et al."Test-time Adaptation with Slot-Centric Models." ICML 2023.
- [Sitzmann-LFN2021] Sitzmann, Vincent, et al. "Light field networks: Neural scene representations with single-evaluation rendering." NeurIPS 2021.
- [Smith-COLF2023]Smith, Cameron Omid, et al. "Unsupervised Discovery and Composition of Object Light Fields." TMLR (2023).
- [Sajjadi-OSRT2022] Sajjadi, Mehdi SM, et al. "Object scene representation transformer." NeurIPS 2022.
- [Yu-uORF2021] Yu, Hong-Xing, Leonidas J. Guibas, and Jiajun Wu. "Unsupervised discovery of object radiance fields." ICLR 2022.
- [Luo-uOCF2024] Luo, Rundong, Hong-Xing Yu, and Jiajun Wu. "Unsupervised Discovery of Object-Centric Neural Fields." arXiv:2402.07376.
- [Driess- PaLME2023] Driess, Danny, et al. "PaLM-E: an embodied multimodal language model." ICML 2023.
- [Zhuang-LSA2022] Zhuang, Yifeng, et al. "Local slot attention for vision and language navigation." ICMR 2022.
- [Huo-Geovln2023] Huo, Jingyang, et al. "Geovln: Learning geometry-enhanced visual representation with slot attention for vision-and-language navigation." CVPR 2023.

Paper List: object-centric learning

- [Fan-AdaSlot2024] Fan, Ke, et al. "Adaptive slot attention: Object discovery with dynamic slot number." CVPR 2024.
- [Epstein-Layout2024] Epstein, Dave, et al. "Disentangled 3D Scene Generation with Layout Learning." arXiv:2402.16936.
- [Wang-Cyclic2024] Wang, Ziyu, Mike Zheng Shou, and Mengmi Zhang. "Object-centric learning with cyclic walks between parts and whole." NeurIPS 2024.
- [Löwe-CAE2022] Löwe, Sindy, et al. "Complex-Valued Autoencoders for Object Discovery." TMLR 2022.
- [Löwe-Rotating2024] Löwe, Sindy, et al. "Rotating features for object discovery." NeurIPS 2024.