SAVi++: Towards End-to-End Object-Centric Learning from Real-World Videos

Anonymous Author(s)
Affiliation
Address
email

Abstract

The visual world can be parsimoniously characterized in terms of distinct entities with sparse interactions. Discovering this compositional structure in dynamic visual scenes has proven challenging for end-to-end computer vision approaches unless explicit instance-level supervision is provided. Slot-based models leveraging motion cues have recently shown great promise in learning to represent, segment, and track objects without direct supervision, but they still fail to scale to complex real-world multi-object videos. In an effort to bridge this gap, we take inspiration from human development and hypothesize that information about scene geometry in the form of depth signals can facilitate object-centric learning. We introduce SAVi++, an object-centric video model which is trained to predict depth signals from a slot-based video representation. By further leveraging best practices for model scaling, we are able to train SAVi++ to segment complex dynamic scenes recorded with moving cameras, containing both static and moving objects of diverse appearance on naturalistic backgrounds, without the need for segmentation supervision. Finally, we demonstrate that by using sparse depth signals obtained from LiDAR, SAVi++ is able to learn emergent object segmentation and tracking from videos in the real-world Waymo Open dataset.

1 Introduction

The natural world consists of distinct entities—people, dogs, cars, trees, etc.—and its complexity emerges from the combined, mostly independent, actions of the entities. This compositional structure must be appreciated to predict future states of the world and to effect particular outcomes. People have an intrinsic understanding of objects: objects have spatiotemporal coherence, they interact when in close proximity, and they possess persistent, latent characteristics that determine their behavior over extended periods of time [25, 42]. Just as object-centric representations are critical to human understanding, they have the potential in machine learning to greatly improve sample efficiency, robustness, visual reasoning, and interpretability of learning algorithms [15, 34]. For example, consider the challenge faced by an autonomous vehicle operating in diverse surroundings (Figure 1). Generalization across situations requires learning about recurring entities like cars, traffic lights, and pedestrians, and the rules that govern interactions among these entities.

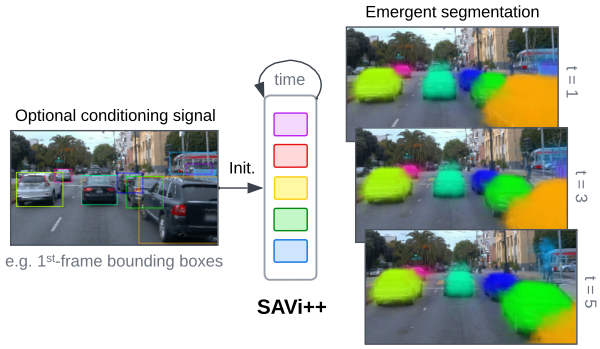


Figure 1: Emergent segmentation and tracking in SAVi++.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

In human brains, the ability to organize edges and surfaces into unitary, bounded, and persisting object representations is not innate; it develops through experience from infancy and without explicit instruction via a 'core system of object representation' [42], i.e., a form of cognitive inductive bias. In deep learning, such an inductive bias has been proposed in *slot-based* architectures which segregate knowledge about individual objects into nonoverlapping but interchangeable pools of neurons. The resulting representational modularity can facilitate causal reasoning and prediction for downstream tasks [15, 40].

A grand challenge in computer vision has been to discover the compositional structure of real-world dynamic visual scenes in an unsupervised fashion. By unsupervised, we mean no segmentation information is provided that specifies which pixels belong together as part of a single object. Initial efforts focused on single-frame, synthetic RGB images [13, 14, 35, 46], but extending this work to video and more complex scenes proved challenging. A key insight to further progress was the realization that a color-intensity pixel array is not the only source of visual information readily available, at least not to human perceptual systems. The human perceptual system extracts motion and depth cues early in the processing stream [9–11, 20, 37]. These cues are correlated with object identities, and can therefore bootstrap the formation of object-centric representations [41].

The recently introduced *Slot Attention for Video* (SAVi) model [30] leveraged *optical flow* (frame-to-frame motion) as a prediction target to obtain object-centric representations of dynamic scenes involving complex 3D scanned objects and real-world backgrounds. However, motion prediction alone is insufficient to learn about the distinction between static objects and the background. Further, in real-world application domains such as self-driving cars, cameras themselves are subject to movement, which globally affects frame-to-frame motion as a prediction signal in non-trivial ways.

In the present work, we describe an enhanced slot-based model for video, referred to as *SAVi++* (Figure 2), which obtains qualitative improvements in object-centric representations by exploiting *depth* signals readily available from RGB-D cameras and LiDAR sensors. SAVi++ is the first slot-based, end-to-end trained model that successfully segments complex objects in naturalistic, real-world video sequences without using direct segmentation or tracking supervision.

A summary of our contributions is as follows:

- We introduce SAVi++: an object-centric slot-based video model that makes several key improvements to SAVi [30] by utilizing *depth prediction* and by adopting best practices for model scaling in terms of *architecture design* and *data augmentation*.
- On the multi-object video (MOVi) benchmark containing synthetic videos of high visual and dynamic complexity [16], we find that SAVi++ is able to handle videos containing complex shapes and backgrounds, and a large number of objects per scene. Improving on SAVi, our approach accommodates both static and dynamic objects and both static and moving cameras.
- Finally, we demonstrate that SAVi++ trained with sparse depth signals obtained from LiDAR enables emergent object decomposition and tracking in real-world driving videos from the Waymo Open dataset [43].

2 Related work

Object-centric learning A growing body of research is addressing the problem of end-to-end learning of object-centric representations from raw perceptual data without direct supervision. Slot-based neural networks such as IODINE [14], MONet [4], and Slot Attention [35] rely on a factorized latent space and independent per-object decoders as inductive bias to enable object discovery in a simple auto-encoding setup. Architectures with stronger inductive biases using fixed object size, presence, or propagation priors have been explored in works such as SQAIR [32] and SCALOR [23], but generally these methods have faced challenges scaling to more complex real-world data when relying on auto-encoding alone. Our work primarily builds on recent advances in object-centric generative models for video sequences [24, 30, 46, 48, 54]. Different from our approach, these methods have so far been unable to scale to complex real-world multi-object video data. An alternative class of methods using contrastive learning for object discovery [19, 29, 36, 51], most notably GroupViT [51] and ODIN [19], has recently achieved some success in discovering semantic groupings in real-world images. However, neither GroupViT nor ODIN model dynamics and typically fail to separate semantically similar object instances in close proximity. In our work, we follow a

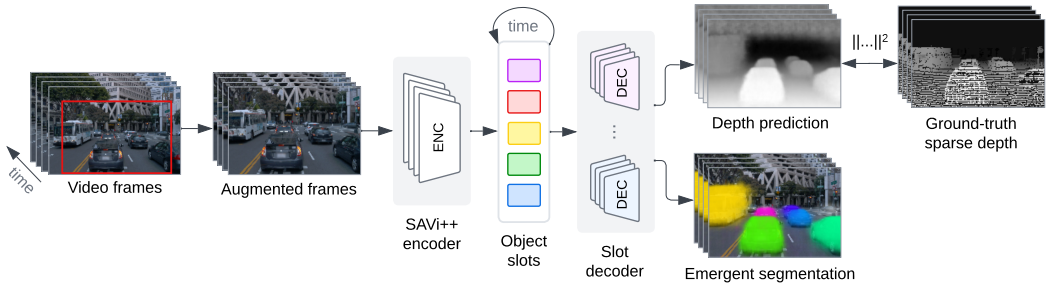


Figure 2: SAVi++ is an object-centric video model based on *Slot Attention for Video* [30], which encodes a video into a set of temporally-consistent latent variables (*object slots*). Input frames and prediction targets are augmented using random crop augmentations. Augmented frames are passed through the improved SAVi++ encoder and mapped onto object slots using an attention mechanism [35]. Slots are updated recurrently for each frame and subsequently decoded independently into a depth map and per-slot alpha masks. SAVi++ is trained using (sparse) depth targets, leading to emergence of temporally-consistent object segmentation in the decoded alpha masks.

generative approach, but instead of tasking the decoder to generate complex visual RGB pixel data, we utilize depth information to bootstrap object-centric learning without direct supervision.

Object discovery in driving scenes A range of recent methods [1, 17, 45, 49] use a multi-stage pipeline of 1) obtaining *pseudo ground truth* (PGT) segmentation or detection labels via some heuristic, and 2) training a model in a supervised fashion on PGT labels. While this class of methods achieves some success in discovering and tracking objects in real-world driving scenes, it crucially hinges on the quality of the PGT labels, requiring carefully engineered task-specific heuristics to extract objects. Earlier methods solely use clustering heuristics to extract approximate segmentation masks directly from motion trajectories for moving objects [3, 38]. In our work, we instead demonstrate that object segmentation and tracking can emerge in an end-to-end setting on complex real-world data without relying on PGT label generation.

Cross-modal learning For self-supervised object-centric learning from visual data, a range of target modalities and training signals have been explored in the literature. By using motion cues from optical flow as prediction targets, several recent methods [30, 52] were able to overcome limitations of purely RGB pixel-level generative models, which frequently failed in the presence of complicated textures [27]. However, this advantage is primarily limited to discovery of moving objects. Utilizing depth targets from a simulator [2] or from sparse LiDAR [17, 45, 49], has been explored in an effort to overcome these limitations. Different from prior works utilizing multi-stage pipelines and hand-crafted heuristics for extraction of pseudo-labels from LiDAR [17, 45, 49], we directly utilize the (sparse) depth signal as target and demonstrate that this can enable emergent object segmentation and tracking on real-world driving data without any additional regularizers or pseudo-labeling techniques.

Scaling strategies for vision models It is common practice to scale architectural capacity with dataset complexity and size, while making use of strong data augmentation when addressing various supervised computer vision tasks [7, 8, 18, 31]. Nonetheless, self-supervised methods for end-to-end object discovery have primarily been relying on overly simplistic and low-capacity backbone architectures [14, 35, 54], likely due to the simplicity of datasets and tasks considered in prior work. By scaling object-centric methods to larger, visually more complex datasets, we find that utilizing stronger visual backbone architectures—in combination with data augmentation—can provide substantial benefits. For simplistic datasets with lower visual complexity (and same number of examples), we found anecdotal evidence in preliminary experiments for the opposite effect: both architecture scaling and data augmentation can negatively affect object discovery performance, likely explaining why prior works have not explored these strategies.

3 Methods

We begin by providing a brief introduction to Slot Attention for Video (SAVi), which is the starting point for our exploration. With SAVi++, we introduce several simple yet crucial improvements, which allow us to bridge the gap to complex real-world data. Our framework is summarized in Figure 2.

3.1 Background

Slot Attention for Video (or SAVi) is a recent state-of-the-art architecture for learning object-centric representations from video with minimal supervision. We briefly highlight some of its key components below and refer the reader for complete details to Kipf et al. [30].

SAVi can be viewed as an autoregressive encoder-decoder video model with a structured latent state composed of K object slots. At a given time-step, an *encoder* first encodes the observed video frame to yield high-level image features that are useful for learning about objects. This is followed by Slot Attention [35] (the *'corrector'*), which updates the slots using these features and encourages individual slots to specialize to different parts of the observation. The content of each slot is decoded separately using a *decoder*, which additionally outputs a pixel-level alpha mask to indicate how the decoded values for each slot should be combined. Together, the mask and decoded slots determine the output of the model at the current time-step from which a loss is computed, e.g., to train the model to predict frame-to-frame motion (optical flow) for this frame. Slots for the next time-step (for the corrector to update) are obtained by applying a *predictor*, which can model interactions between slots and learn about object dynamics to predict their future state.

In addition to optical flow prediction, SAVi relies on one additional cue to succeed. Just as parents direct infants to locations in the visual world where interesting objects occur [12, 53], SAVi assumes spatial cues to a subset of objects in the initial frame of a video. This is implemented by the *slot initializer*, which initializes the slots used in the initial video frame. The initialization may be learned in an unconditional setting (i.e., learn the initial slot states) or obtained by conditioning the initial state on high-level cues such as bounding boxes of objects of interest in the first video frame. This direction of attention or *input conditioning* was necessary for SAVi to succeed in decomposing more complex visual scenes.

3.2 SAVi++

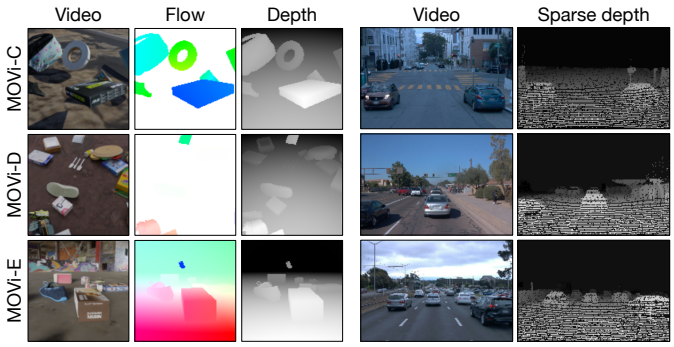
As SAVi relies on optical flow prediction as its main training signal for object discovery, its application is primarily limited to settings where all objects in a scene have independent motion. In addition, SAVi struggled to generalize to scenes with a moving camera, even though the optical flow field encodes information about (static) scene geometry in this case.

Here, we identify two key directions for improving SAVi and bridging its capabilities to real-world video data, while preserving its core foundation for learning object representations from video: (1) exploiting depth as a prediction signal, which is readily available in many real-world settings, and (2) utilizing model scaling strategies in terms of encoder improvements and data augmentation, which, despite being commonly used for classic vision problems, are generally underutilized for object-centric learning. Our improved approach, called SAVi++, successfully segments complex objects in naturalistic, real-world, video sequences without using direct segmentation or tracking supervision.

Exploiting Depth Information Training object-centric models solely using RGB image or video frame reconstruction proves challenging in the presence of complex visual textures, frequently leading to failure modes such as clustering by color or into object-agnostic spatial regions [14, 17]. In SAVi, optical flow was proposed as a prediction signal to mitigate this issue, while still operating on visual RGB inputs [30]. However, relying solely on optical flow as a prediction target for learning about objects has a clear disadvantage: static objects, which make up the vast majority of visual entities we encounter on a daily basis, are not captured in this signal unless the observer or the entire scene is in motion. As a consequence, SAVi fails to represent objects that are at rest, and similarly struggles with scenes observed from a moving camera, as optical flow can prove challenging to model in this case.

Here, we explore depth as a target signal, used in conjunction with flow or even in isolation. Depth estimation has received little attention in slot-based models, yet does not suffer from the limitation of optical flow in datasets with static objects and camera movement. We thus hypothesize that depth may greatly benefit obtaining emergent object decompositions of complex videos. In terms of practical applicability, we note how depth is a readily-available signal in many real-world settings thanks to the prevalence of RGB-D cameras and LiDAR in settings like self-driving cars [43]. Even in the absence of depth sensing capabilities, this signal can be cheaply estimated from multi-camera systems [33].

In our implementation, we represent the depth signal in image space, which we encode using a log transformation $\log(1 + d)$, where d is the distance of a pixel to the camera (see Figure 3a). This



(a) MOVi datasets. (b) Waymo Open dataset.

	Complex texture	Moving objects	Static objects	Moving camera	#frames (train)
MOVi-C	✓	✓	✗	✗	234k
MOVi-D	✓	✓	✓	✗	234k
MOVi-E	✓	✓	✓	✓	234k
Waymo Open	✓	✓	✓	✓	159.6k

(c) Dataset details and statistics.

Figure 3: We consider three synthetic Multi-Object Video (MOVi) datasets [16] and the large-scale real-world driving dataset Waymo Open [43]. All datasets contain complex textures and moving objects. The MOVi datasets increase in complexity from MOVi-C (moving objects only) over MOVi-D (+static objects) to MOVi-E (+ moving cameras). Waymo Open contains all these characterisics.

log-transform puts a stronger emphasis on close-by objects and—in early experiments—we found this form of normalization crucial for reliably training object-centric models using depth targets. SAVi++ is then trained to minimize the squared difference between the decoder output and this target signal. In case of multiple available targets, such as depth and flow, we concatenate the target images along the channel dimension and predict them using an otherwise unchanged model.

For sparse targets such as depth obtained from LiDAR, we ignore any points in the image space for which no signal is present in the computation of the loss. For LiDAR specifically, we obtain the x, y, z coordinates of all the LiDAR points in the self-driving car (SDC) world and compute the distance of each of the points from the LiDAR sensor. We then use the camera and LiDAR calibration parameters to project the LiDAR point distances from the SDC domain to the camera frame. This projection represents a very sparse approximation of the ground-truth depth signal (Figure 3b).

Scaling strategies Visual complexity present in real-world videos necessitates a different class of encoders than those used for simple synthetic datasets. Inspired by successful visual backbone architectures for set-based supervised object detection models [6, 26], we use a more capable encoder that utilizes the ResNet34 [18] architecture followed by a transformer encoder [47] (with 4 layers, unless otherwise mentioned). To avoid computation of batch and/or temporal statistics, we replace the typical batch normalization in ResNet34 with group normalization [50]. We use a stride 1 convolution and use no max-pooling in the ResNet root block. This results in an overall backbone stride of 8 (as opposed 32), which was found to be important for retaining object decomposition capabilities. Please see the appendix for further architectural details.

Drawing inspiration from training schemes commonly used for real-world vision models [44], we further apply Inception-style cropping as data augmentation. In particular, we randomly crop a region of each frame with aspect ratio $\in [0.75, 1.33]$ such that enough of the frame is retained after cropping. The same crop is applied consistently across all frames and the resulting video is resized to the original resolution. Flow fields and depth maps are adjusted accordingly to keep them accurate and spatially aligned with the video frame.

4 Experiments

The goal of our experimental evaluation is twofold: 1) on synthetic video data of varying complexity we would like to analyze the potential advantages of utilizing a depth signal and model scaling strategies for learning emergent segmentation and tracking, and 2) we would like to investigate whether these improvements enable bridging the gap to complex real-world video data.

Section 4.1 covers both qualitative and quantitative comparisons of SAVi++ against baselines on the synthetic MOVi datasets. In Section 4.2, we perform an ablation study on SAVi++. Finally, in Section 4.3 we demonstrate and analyze results for a SAVi++ model applied to real-world driving videos from the Waymo Open [43] dataset.

Datasets As basis for our experiments, we use videos of different scene and camera complexities (Figure 3c). We use three synthetic Multi-Object Video (MOVi) datasets (Figure 3a) introduced in Kubric [16], which are created by simulating rigid body dynamics. We narrow our investigation to MOVi datasets with complex naturalistic backgrounds and 3D-scanned everyday objects (variants C, D, and E). MOVi-C is generated using a static camera, and all objects (max. 10) are initialized to move independently. MOVi-D introduces more objects, some of which are dynamic (1-3) and the majority rests statically in the scene (10-20). Finally, MOVi-E introduces random, linear camera movement. Each video contains 24 frames sampled at 12 frames per second (fps).

We also train and evaluate SAVi++ in a real-world driving setting using the Waymo Open dataset (Figure 3b). Waymo Open is comprised of high resolution video data of 1280×1920 original resolution from a multi-camera system collected by Waymo vehicles [43]. The dataset consists of 798 train and 202 validation scenes of 20s video each, sampled at 10 fps. We subsample the dataset at 5 fps both for training and validation. The dataset also includes LiDAR signals that we use to compute sparse depth maps as discussed in Section 3.

Training setup For all our experiments, unless stated otherwise, we resize frames to a height of 128 pixels while keeping the aspect ratio fixed, resulting in a 128×128 resolution for MOVi datasets, and a resolution of 128×192 for Waymo Open. We train SAVi++ for 500k steps (300k for Waymo Open) on Tensor Processing Unit (TPU) accelerators with a batch size of 64 using Adam [28].

We train on randomly sampled sub-sequences of only 6 frames using 24 slots for MOVi and 11 slots for Waymo Open. See appendix for further training details and hyperparameters.

4.1 SAVi++ improves object-centric learning on complex synthetic video data

We investigate whether the key changes introduced to SAVi [30], which constitute our improved SAVi++ model, allow us to overcome limitations of SAVi and address the most challenging synthetic multi-object video (MOVi) benchmarks introduced in Kubric [16].

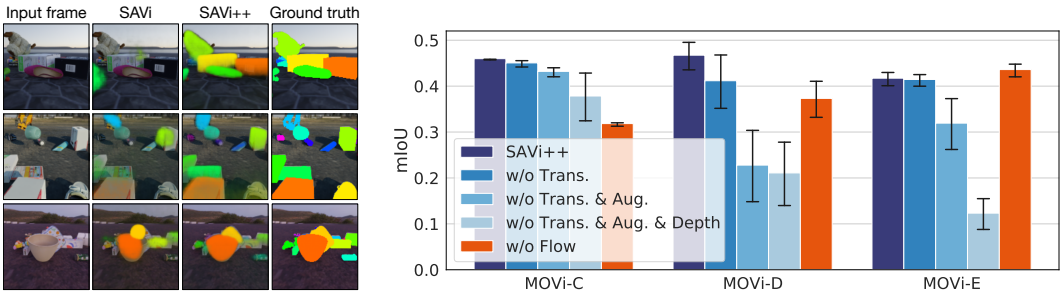
Setup We train all models independently on each dataset variant. Both SAVi and SAVi++ are trained in a conditional setting where we initialize slots using ground-truth bounding box information in the first frame. We report the same segmentation metrics as in prior work, i.e. Foreground Adjusted Rand Index (FG-ARI) [21, 39] and Mean Intersection over Union (mIoU). FG-ARI is a permutation-invariant clustering similarity metric frequently used for evaluating scene decomposition quality. It compares discovered segmentation masks with ground-truth masks while ignoring any pixels that belong to the background. It is sensitive to temporal consistency of masks, but insensitive to their ordering. The mIoU metric is a standard segmentation metric, here adapted for video as in [5]. We note that this implementation is sensitive to the correct ordering of masks, i.e. it also measures whether models used the conditioning signal (here, first-frame bounding boxes) correctly.

Baselines Besides comparing to SAVi [30], the most representative prior method for the task we are interested in, we compare against a range of baselines aimed at establishing the difficulty of the unsupervised, bounding box-conditioned video object segmentation task: 1) a bounding box copy (*BBox copy*) baseline, which simply repeats the first-frame boxes throughout the video, 2) a learned *BBox propagation baseline* that does not receive visual inputs, to test for easily exploitable biases in the datasets, 3) k-Means clustering baselines, that cluster the flow and/or depth signal across the video sequence (initialized using the ground-truth object centers in the first frame), and 4) a label propagation baseline, that uses visual features to propagate the initial boxes (rendered as rectangular masks) across the video, based on Contrastive Random Walks (CRW) [22]. See appendix for further details.

Results Quantitative results can be seen in Table 1 and qualitative results on MOVi-E in Figure 4a. The *BBox copy* method serves as a trivial baseline which a learning-based approach should outperform. While the original SAVi model does so on MOVi-C, it clearly fails to model the more complex MOVi-D and -E datasets. The BBox copy baselines is—perhaps unsurprisingly—strongest on MOVi-D, where most objects are static. SAVi++ outperforms this baseline on all datasets, indicating that it learns non-trivial segmentation and tracking capabilities. Indeed, this advantage does not solely come from fitting certain biases in the datasets, as a learned BBox propagation baseline (using the same predictor as in SAVi++) that does not receive visual input, fails to generalize to unseen evaluation videos. It is worth noting that neither of the MOVi tasks can easily be solved by simply clustering the target signals, as the results for the k-Means baselines demonstrate.

Table 1: MOVi results in terms of mean score ± standard error (5 seeds) from evaluating SAVi++ and baseline models on validation set video sequences of increased length (24 frames). *: we use the official implementation of CRW [22], which does not report FG-ARI.

	mIoU↑ (%)			**FG-ARI↑ (%)**		
Model	**MOVi-C**	**MOVi-D**	**MOVi-E**	**MOVi-C**	**MOVi-D**	**MOVi-E**
BBox copy	12.3	42.8	32.9	11.8	68.0	54.7
BBox propagation	22.9 ± 0.1	26.7 ± 0.8	24.1 ± 1.1	9.6 ± 0.5	24.9 ± 3.7	18.4 ± 3.9
K-Means (depth)	7.1 ± 0.3	6.0 ± 0.4	5.4 ± 0.3	26.3 ± 1.0	30.9 ± 0.7	32.2 ± 0.6
K-Means (flow)	10.7 ± 0.5	7.4 ± 0.4	6.0 ± 0.3	26.5 ± 1.0	30.9 ± 0.8	33.1 ± 0.7
K-Means (flow+depth)	10.6 ± 0.6	6.7 ± 0.4	5.3 ± 0.3	26.6 ± 1.0	35.9 ± 1.0	34.8 ± 0.7
CRW [22]	27.8 ± 0.2	45.3 ± 0.0	**47.5** ± 0.1	*	*	*
SAVi [30]	37.8 ± 5.2	21.1 ± 6.9	12.3 ± 3.4	78.2 ± 0.6	52.6 ± 10.2	24.8 ± 6.1
SAVi++ (ours)	**48.3** ± 0.2	**49.8** ± 0.5	**47.0** ± 1.7	**85.1** ± 0.1	**86.5** ± 0.5	**82.3** ± 1.3



(a) MOVi-E qualitative results.

(b) Ablation study

Figure 4: **Left**: Qualitative results of SAVi++ compared to SAVi [30] on the synthetic MOVi-E dataset with camera motion. **Right**: SAVi++ ablation study on MOVi-C, D, and E. Bars reflect validation set mIoU (mean ± standard error for 5 seeds). We ablate: 1) the transformer encoder (*w/o Trans.*), 2) data augmentation (*w/o Trans. & Aug.*), and 3) depth targets (*w/o Trans. & Aug. & Depth*). We further report results for training without flow, while only using depth targets (*w/o Flow*).

Compared to CRW it can be seen how SAVi++ yields markedly better mIoU on MOVi-C and D, while performance on MOVi-E is similar. Note that, unlike SAVi++, CRW is merely capable of propagating pixel-level annotations across frames in a video and does not by itself produce instance-level object segmentations or corresponding object-representations that could be used for down-stream tasks. Finally, comparing SAVi++ and SAVi directly, we see that SAVi++ overcomes the primary limitations of SAVi on the harder MOVi-D and -E datasets, both quantatively (Table 1) and qualitatively (Figure 4a), while also improving performance on MOVi-C.

Discussion It is evident that a small number of critical changes to SAVi [30], namely utilizing depth targets, a stronger architecture, and data augmentation, can have dramatic consequences on the ability of this slot-based model to learn emergent object segmentation and tracking in complex video sequences. The difference between SAVi++ and SAVi is especially evident for the more complex datasets in our study (e.g., improving the mIoU score on MOVi-E from 12.3% to 47.0%; see also Figure 4a). These results demonstrate that SAVi++ is better suited for various data complexities in terms of object dynamics and camera movement, which are likely to exist in real-world data.

4.2 Ablation study

In this section, we report results of an ablation study to gauge the contribution of the different components of SAVi++. The three main ingredients of SAVi++ are 1) the use of depth as training target, 2) the extra capacity added to SAVi by including a transformer encoder, and 3) the use of data augmentation. Figure 4b shows a systematic ablation of each of those components. Removing the transformer encoder reduces object segmentation quality, yet the degradation in performance is relatively limited. While data augmentation only has a mild effect on the simpler MOVi-C dataset, it

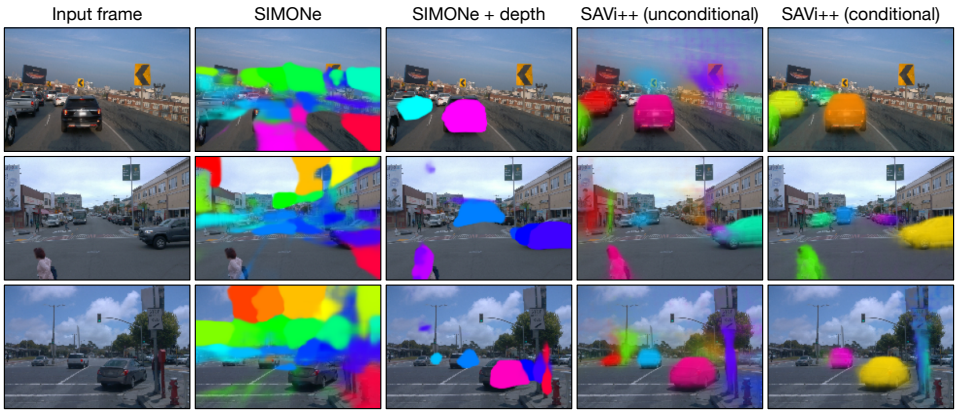


Figure 5: Qualitative segmentation comparison on the Waymo Open validation set. Naive application of a SIMONe [24] baseline model to this dataset results in failure, while adapting SIMONe to predict (sparse) depth maps yields rough (but frequently misaligned) segmentation masks. SAVi++ generally produces highly accurate segmentation masks, while its unconditional results are promising. Here, we hide masks that occupy more than 1300 pixels on average per frame to ease interpretability.

makes a substantial difference on the more challenging datasets, MOVi-D and E. Finally, removing depth targets reduces performance further and is particularly catastrophic on MOVi-E.

In fact, we find that training *solely* using depth targets without relying on predicting optical flow as well (see *w/o Flow* in Figure 4b) still allows the model to accurately segment and track objects, especially on the more complex MOVi-D and E datasets. This result is particularly strong on MOVi-E where jointly predicting optical flow presents a difficult task for scenes with camera movement.

4.3 SAVi++ enables emergent segmentation on real-world driving data

In the previous section, we found that solely using depth as a training target can be sufficient to learn emergent object segmentation and tracking. This finding provides a strong motivation for scaling this class of methods to real-world data, where the availability of optical flow relies on approximate and potentially inaccurate flow estimation methods, whereas depth can be accurately measured using technologies like LiDAR. To investigate this possibility, we use the Waymo Open dataset [43], which includes videos obtained from cameras mounted on cars in various traffic environments.

Setup To obtain a depth signal, we project 3D LiDAR points into the camera frame, resulting in a very sparse depth image for each time step (see Figure 3b for examples). We exclude pixels that do not have a valid LiDAR point when computing the L2 loss in image space. We train SAVi++ with 11 slots on 6 frames and evaluate the model on sequences of 10 frames. Due to the absence of ground-truth segmentation labels in Waymo Open, we quantitatively measure performance compared to ground-truth bounding boxes using two metrics. The Center-of-Mass (CoM) distance measures the average Euclidean distance between the centroid of the predicted segmentation masks and the centers of the ground-truth bounding boxes. We report the centroid distance normalized by the maximum achievable distance in the video frame. The Bounding Box mIoU (B. mIoU) is analog to mIoU using predicted and ground-truth bounding boxes. The former are obtained by training a readout MLP to predict bounding boxes from the slot representations. See appendix for further details.

Baselines We quantitatively compare to the subset of previous baselines that work with (sparse) depth. Further, we report qualitative results for SAVi++ in the unconditional setting, i.e. without providing first-frame bounding boxes to the model to initialize slots, and compare to SIMONe [24] as a representative object-centric video model baseline from the literature.

Results Quantitative results can be seen in Table 2 and qualitative results in Figures 5–6. We find that SAVi++ markedly outperforms the BBox copy and propagation baselines, as well as the clustering baseline. The qualitative results better reflect the significance of SAVi++'s performance as well as its potential utility for object-centric representation learning from real-world videos.

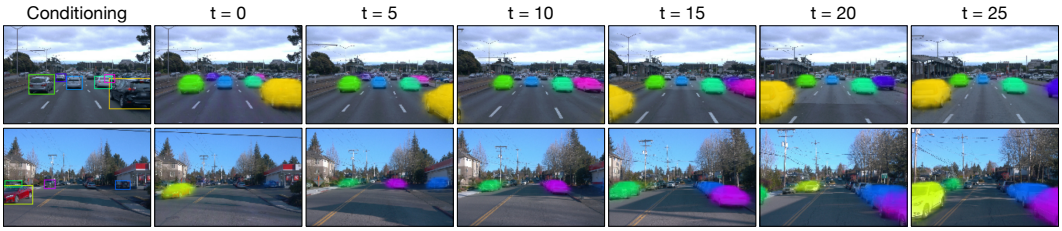


Figure 6: Waymo Open qualitative results of SAVi++ (conditional) over long sequences.

We additionally experimented with removing the bounding box conditioning in SAVi++ in the initial frame. Removing this conditioning signal and using a learned initialization together with a simplified encoder also yielded good object decompositions (see SAVi++ (unconditional) in Figure 5). Compared to using plain SIMONe [24], we observe that SAVi++ (unconditional) performs markedly better. Interestingly, modifying the non-autoregressive SIMONe baseline similar to SAVi++ by predicting sparse depth instead of RGB also showed improvement, yet its segmentations appear worse. This gives further evidence that using depth is suitable for learning object-centric representations from real videos.

Table 2: Waymo Open results (mean $\pm$ standard error in %, 3 seeds) from evaluating models on sequences of 10 frames. SAVi++ HR is a variant trained on higher-resolution (256×384) video frames.

Model	**CoM**$\downarrow$	**B. mIoU**$\uparrow$
BBox Copy	5.0	44.3
BBox Prop.	5.1 ± 0.1	38.5 ± 0.5
K-Means (depth)	13.0 ± 0.1	–
SAVi++	$\mathbf{4.1} \pm 0.3$	$\mathbf{49.5} \pm 0.8$
SAVi++ HR	$\mathbf{3.6} \pm 0.0$	$\mathbf{51.0} \pm 0.2$

We show qualitative results for longer sequences in Figure 6 and in video-format in the supplementary material. It is worth noting that SAVi++ was only trained on 6 frames and did not receive any tracking supervision. Interestingly, we find that objects are often consistently tracked until the moment they leave the scene. At this stage, slots are freed up again and tend to bind to previously unexplained or new objects. This behaviour indicates that our reported tracking metrics are an underestimation of the capabilities of the model, as such re-binding is not accounted for. It is, however, conceivable that re-binding events could be identified post-hoc if one were to use the representations learned by SAVi++ for downstream tasks, which is an interesting avenue for future work.

4.4 Limitations

With SAVi++, we demonstrated the first proof of concept that an emergent object-centric decomposition of real-world complex videos is possible with an end-to-end slot-based approach. Yet, there is still a lot of room for improvement. We focused our exploration on the conditional setup where we provided cues in the form of bounding boxes of objects in the first frame. Our preliminary results with unconditional SAVi++ suggest that this information may not be necessary and could be removed in future research. In both the conditional and the unconditional setting, the segmentation and tracking performance, though impressive given the minimal amount of supervision the model receives, still qualitatively lags behind supervised approaches. Improving the temporal consistency of object tracks, especially in the unconditional setting, is a promising direction for future work.

5 Conclusion

We demonstrate that object tracking and segmentation can emerge from utilizing information about scene geometry in the form of depth signals in complex video data with slot-based neural architectures. We utilize a series of synthetic multi-object video benchmarks with increasing complexity to find a simple yet effective set of changes to an existing state-of-the-art object-centric video model (SAVi), allowing us to bridge the gap from synthetic to complex real-world driving videos.

Our work marks a first step towards building end-to-end trainable systems that learn to perceive the world in an object-centric, decomposed fashion without relying on detailed human supervision. While many open challenges remain, this result evidences that object-centric deep neural networks are not inherently limited to simple synthetic environments, and we are excited about the potential for this class of methods to radically reduce the need for human supervision in building scalable perceptual systems for the real world.

References

[1] Zhipeng Bao, Pavel Tokmakov, Allan Jabri, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert. Discovering objects that can move. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[2] Daniel M Bear, Chaofei Fan, Damian Mrowca, Yunzhu Li, Seth Alter, Aran Nayebi, Jeremy Schwartz, Li Fei-Fei, Jiajun Wu, Joshua B Tenenbaum, et al. Learning physical graph representations from visual scenes. In *Advances in Neural Information Processing Systems*, 2020.

[3] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. In *European Conference on Computer Vision*, pages 282–295. Springer, 2010.

[4] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. MONet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

[5] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 DAVIS challenge on vos: Unsupervised multi-object segmentation. *arXiv preprint arXiv:1905.00737*, 2019.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 2020.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[9] J. Driver, P. McLeod, and Z. Dienes. Motion coherence and conjunction search: Implications for guided search theory. *Perception and Psychophysics*, 51:79–85, 1992.

[10] J. T. Enns and R. A. Rensink. Influence of scene-based properties on visual search. *Science*, 247:721–723, 1990.

[11] J. T. Enns and R. A. Rensink. Sensitivity to three-dimensional orientation in visual search. *Psychological Science*, 1(5):323–326, 1990.

[12] Teresa Farroni, Mark H. Johnson, Margaret Brockbank, and Francesca Simion. Infants' use of gaze direction to cue attention: The importance of perceived motion. *Visual Cognition*, 7(6):705–718, 2000.

[13] Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Neural expectation maximization. In *Advances in Neural Information Processing Systems*, pages 6691–6701, 2017.

[14] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *International Conference on Machine Learning*, pages 2424–2433, 2019.

[15] Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020.

[16] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[17] Adam W Harley, Yiming Zuo, Jing Wen, Ayush Mangal, Shubhankar Potdar, Ritwick Chaudhry, and Katerina Fragkiadaki. Track, check, repeat: An EM approach to unsupervised tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[19] Olivier J Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelović. Object discovery and representation networks. *arXiv preprint arXiv:2203.08777*, 2022.

[20] Todd S Horowitz, Jeremy M Wolfe, Jennifer S DiMase, and Sarah B Klieger. Visual search for type of motion is based on simple motion primitives. *Perception*, 36(11):1624–1634, 2007.

[21] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[22] Allan Jabri, Andrew Owens, and Alexei A Efros. Space-time correspondence as a contrastive random walk. In *Advances in Neural Information Processing Systems*, 2020.

[23] Jindong Jiang, Sepehr Janghorbani, Gerard de Melo, and Sungjin Ahn. SCALOR: Generative world models with scalable object representations. In *International Conference on Learning Representations*, 2020.

[24] Rishabh Kabra, Daniel Zoran, Loic Matthey Goker Erdogan, Antonia Creswell, Matthew Botvinick, Alexander Lerchner, and Christopher P. Burgess. SIMONe: View-invariant, temporally-abstracted object representations via unsupervised video decomposition. In *Advances in Neural Information Processing Systems*, 2021.

[25] Daniel Kahneman, Anne Treisman, and Brian J Gibbs. The reviewing of object files: Object-specific integration of information. *Cognitive psychology*, 24(2):175–219, 1992.

[26] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. MDETR – Modulated Detection for End-to-End Multi-Modal Understanding. *arXiv preprint arXiv:2104.12763*, 2021.

[27] Laurynas Karazija, Iro Laina, and Christian Rupprecht. ClevrTex: A texture-rich benchmark for unsupervised multi-object segmentation. In *NeurIPS Track on Datasets and Benchmarks*, 2021.

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[29] Thomas Kipf, Elise van der Pol, and Max Welling. Contrastive learning of structured world models. In *International Conference on Learning Representations*, 2020.

[30] Thomas Kipf, Gamaleldin F Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. In *International Conference on Learning Representations*, 2022.

[31] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (BiT): General visual representation learning. In *European Conference on Computer Vision*, pages 491–507. Springer, 2020.

[32] Adam Kosiorek, Hyunjik Kim, Yee Whye Teh, and Ingmar Posner. Sequential attend, infer, repeat: Generative modelling of moving objects. In *Advances in Neural Information Processing Systems*, pages 8606–8616, 2018.

[33] Hamid Laga, Laurent Valentin Jospin, Farid Boussaid, and Mohammed Bennamoun. A survey on deep learning techniques for stereo-based depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[34] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

[35] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, 2020.

[36] Sindy Löwe, Klaus Greff, Rico Jonschkowski, Alexey Dosovitskiy, and Thomas Kipf. Learning object-centric video models by contrasting sets. *arXiv preprint arXiv:2011.10287*, 2020.

[37] K. Nakayama and G. H. Silverman. Serial and parallel processing of visual feature conjunctions. *Nature*, 320:264–265, 1986.

[38] Peter Ochs and Thomas Brox. Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[39] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[40] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634, 2021.

[41] E. S. Spelke. Principles of object perception. *Cognitive Science*, 14:29–56, 1990.

[42] Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007.

[43] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[44] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[45] Hao Tian, Yuntao Chen, Jifeng Dai, Zhaoxiang Zhang, and Xizhou Zhu. Unsupervised object detection with lidar clues. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[46] Sjoerd van Steenkiste, Michael Chang, Klaus Greff, and Jürgen Schmidhuber. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. In *International Conference on Learning Representations*, 2018.

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[48] Rishi Veerapaneni, John D Co-Reyes, Michael Chang, Michael Janner, Chelsea Finn, Jiajun Wu, Joshua Tenenbaum, and Sergey Levine. Entity abstraction in visual model-based reinforcement learning. In *Conference on Robot Learning*, pages 1439–1456, 2020.

[49] Antonin Vobecky, David Hurych, Oriane Siméoni, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, and Josef Sivic. Drive&segment: Unsupervised semantic segmentation of urban scenes via cross-modal distillation. *arXiv preprint arXiv:2203.11160*, 2022.

[50] Yuxin Wu and Kaiming He. Group normalization. In *European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[51] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. *arXiv preprint arXiv:2202.11094*, 2022.

[52] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[53] Chen Yu and Linda B. Smith. Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PLOS ONE*, 8(11):1–10, 11 2013.

[54] Daniel Zoran, Rishabh Kabra, Alexander Lerchner, and Danilo J Rezende. Parts: Unsupervised segmentation with slots, attention and independence maximization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10439–10447, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Limitations section.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Yes, see Societal Impact section in appendix.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Instructions to reproduce experiments are included in the main text and appendix sections. We will release an implementation for SAVi++ by the time of publication. We only used publicly available datasets.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Please see experimental section and appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Please see experimental section and appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We use publicly available data. For Waymo Open all faces and license plates are blurred and made unrecognizable by default.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]