Pixel-level Scene Understanding in One Token: Visual States Need What-is-Where Composition

Supplementary Material

Appendix

In the supplemental material, we provide:

A. Additional Implementation Details

Decoder. Following ToBo [42], we employ a decoder composed of self-attention and multi-layer perceptron (MLP) layers. The decoder configuration follows MAE [31], with the detailed settings summarized in Tab. 4a.

Pre-training. The pre-training details are provided in Tab. 4b. Unless otherwise specified, we follow the setup of RSP [37] for a fair comparison. For the main comparison, we report pre-training for 400 epochs; with repeated sampling [24, 34] of 2, this corresponds to running 200 training epochs in practice, where each sample is seen twice per epoch. For ablation studies, we train for 100 epochs and reduce the number of warmup epochs from 40 to 20.

Augmentation. The augmentation details are provided in Tab. 4c. Note that the global crop is sampled from the selected frame, and the local crop is then sampled from the global crop, so that the local crop remains fully contained within the global view. We use global and local crop scales of [0.5, 1.0] and [0.3, 0.6], respectively, following CropMAE [19]. We also apply horizontal flipping in a synchronized manner across global and local views for improved training stability (i.e., if a global view is flipped, the corresponding local view is flipped as well).

Table 4. **Implementation details.** We follow MAE [31] for the decoder, RSP [42] for pre-training, and CropMAE [19] for augmentation, except for synchronized horizontal flipping.

Parameter	Setting
Depth	8
Embedding dimension	512
Number of attention heads	16
MLP ratio	4.0

(a) **Decoder configuration.**

Parameter	Setting
Batch size	1536
Epochs	400
Warmup epochs [21]	40
LR scheduler	Cosine decay [44]
Learning rate	1.5×10^{-4}
Optimizer	AdamW [45]
Adam β_1, β_2	(0.9, 0.95) [8]
Weight decay	0.05
Repeated sampling [34]	2

(b) **Pre-training configuration.**

Parameter	Setting
Augmentation type	RandomCrop + HFlip
Crop aspect ratio	[3/4, 4/3]
Global crop scale	[0.5, 1.0]
Local crop scale	[0.3, 0.6]
Resize	224×224
Interpolation	Bicubic
HFlip probability	0.5 (Synchronized)

(c) **Augmentation configuration.**

B. Additional Related Works

Benchmarking Pixel-level video understanding. Pixel-level video understanding has been studied through a variety of benchmarks that evaluate fine-grained spatiotemporal understanding in dynamic scenes, such as video object segmentation (VOS). Early datasets such as DAVIS [48, 49] and YouTube-VOS [58] established standard VOS benchmarks, with high-quality dense annotations in DAVIS and large-scale data diversity in YouTube-VOS. More recent benchmarks [13–16] further increase scene complexity and motion ambiguity. In particular, MOSE [14] and MOSEv2 [16] introduce crowded and complex scenes with heavy occlusion, frequent disappearance and reappearance, and multiple same-category instances. In parallel, MeViS [13] and MeViSv2 [15] extend this line of work to referring video segmentation based on motion expressions.

While these benchmarks evaluate fine-grained scene understanding with dense pixel-level annotations, our work instead learn such understanding in a compact visual state representation (i.e., [CLS] token). Therefore, our method is not directly aligned with standard VOS evaluation protocols. Instead, we leverage challenging VOS datasets such as DAVIS and MOSEv2 for reconstruction analysis and perceptual straightness, to assess whether the learned representation captures complex scene dynamics.

Visual Representations for Sequential World Modeling. Latent world models [25–29] aim to learn compact states that support long-horizon prediction and decision-making from high-dimensional observations [4]. They rely on visual representations to encode observations into these states [2, 40, 60], since these representations provide the primary interface to the external environment. Accordingly, the representation should both capture fine-grained information from observations and be predictive over time. In this context, CroBo may serve as a suitable visual encoder for sequential world modeling as it encodes what-is-where scene information into a compact state and exhibiting temporal straightness [32, 33, 46]. This further suggests its potential for predictive modeling in representation space, such as next-embedding prediction framework [4, 35, 59].

Locality-aware Image Representation Learning. Recent work suggests that capturing localized information, such as objects and regions, helps fine-grained image understanding. In visual representation learning, this has been explored through dense or local correspondence learning [43, 51, 52, 56], and in vision-language representation learning through region-level alignment between image regions and regional text descriptions [12, 38, 53, 55]. Our work also prioritizes preserving localized scene details but focuses on encoding this fine-grained composition into a single compact visual state for dynamic scene understanding.

C. Additional Reconstruction Visualization

We provide extensive reconstruction visualizations in Fig. 6, extending the analysis in Sec. 3.3 to a broader set of cases. Using the bottleneck token from the reference view, CroBo reconstructs 90% masked target views across CLEVR, DAVIS, MOSEv2, and Franka Kitchen. The results demonstrate that our model faithfully restores object identities and spatial locations. This consistent performance across diverse scene types further confirms that the bottleneck token effectively captures pixel-level **what-is-where** scene composition.

D. Additional Perceptual Straightness Results

To further probe the perceptual straightness of CroBo's learned representations, we present additional qualitative analyses on three representative scenarios in Fig. 7. We visualize representation trajectories using PCA and compare CroBo with DINOv2 [47] and CropMAE [19]. Together, these examples cover near-linear motion, periodic rotation, and sequential manipulation, illustrating how the learned representation behaves under distinct temporal structures.

Straight taxiing airplane (MOSEv2, `4xz71muo`). This video shows an airplane slowly taxiing along a runway, with smooth camera motion that tracks the airport scene layout. The motion is nearly linear from a perceptual standpoint, with minimal viewpoint change and steady progression across the scene. Our model produces a smooth and nearly linear trajectory that closely follows this temporal progression, indicating that the representation preserves the underlying motion in a consistent and ordered manner.

Rotating radar antenna (MOSEv2, `7bh6bqw6`). This video shows a radar antenna rotating clockwise for five turns, with slight camera motion. Our model effectively captures this periodic motion and the accompanying appearance changes, producing a coherent and smoothly evolving trajectory that reflects the underlying cyclic structure. Notably, the trajectory exhibits a repeating C-shaped pattern that matches the Lissajous-like curves obtained when circular motion is projected onto a 2D plane. DINOv2 also captures a reasonably structured trajectory, whereas CropMAE exhibits more entangled and irregular patterns, suggesting weaker consistency in representing the periodic motion.

Opening microwave (Franka Kitchen Demo). This example shows the first two seconds of a Franka Kitchen demo, where the robot arm moves left, grasps the microwave handle, opens the door, and then reaches toward a kettle, while the camera remains fixed. Our model produces a smooth and interpretable trajectory that follows this sequential manipulation process. In particular, the turning point at the left L-shaped corner of CroBo's trajectory corresponds to the moment when the robot hand grasps the microwave handle, suggesting that the learned representation captures a perceptually meaningful transition in the scene. By contrast, the trajectories of competing methods exhibit irregular, zigzagging patterns with frequent local direction changes, making them less structured and harder to interpret.

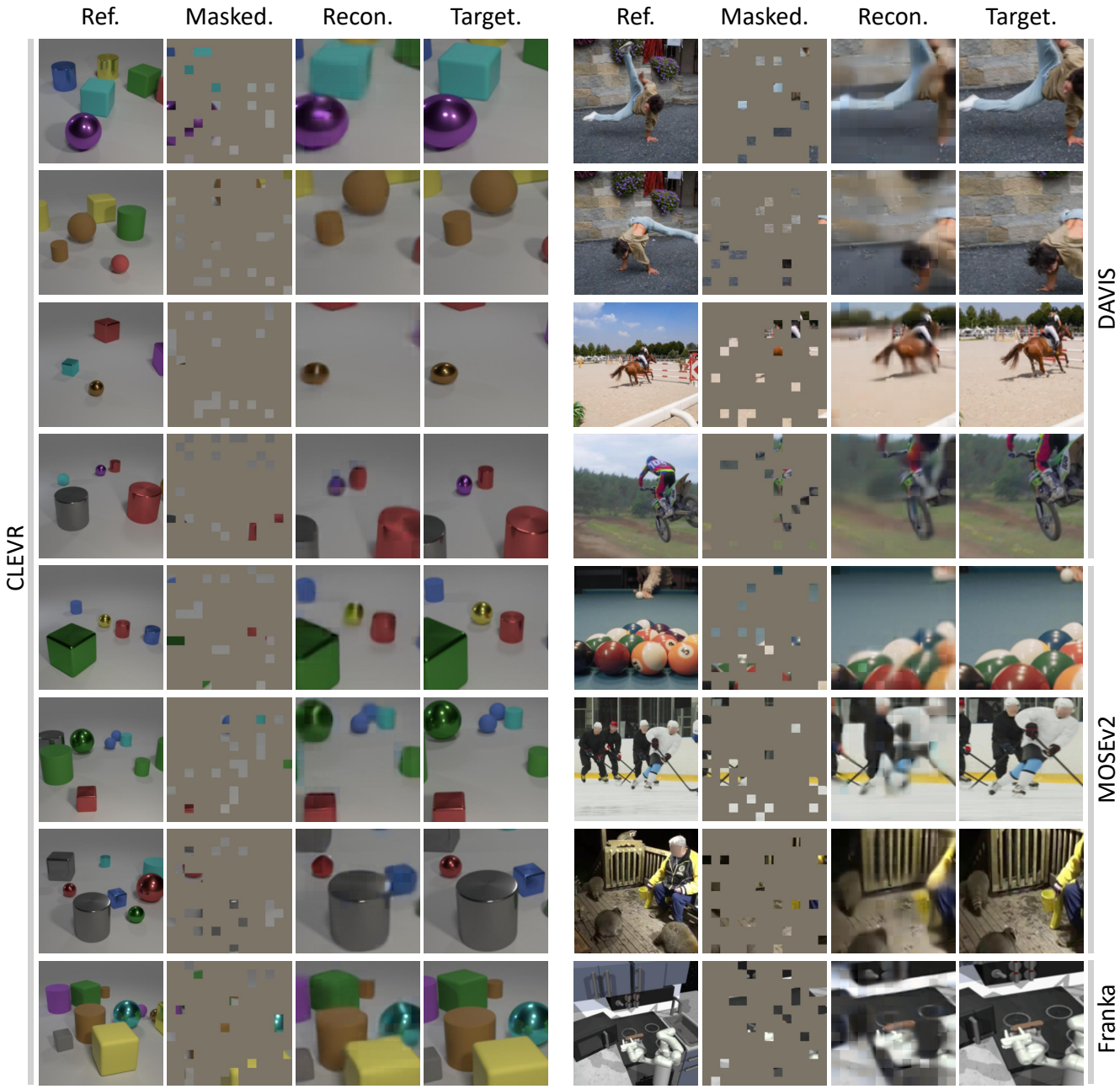

Figure 6. **Reconstruction results of CroBo.** We present image reconstruction examples on CLEVR [39], DAVIS [49], MOSEv2 [16] and Franka Kitchen [23]. Given the bottleneck token extracted from the reference view as context, CroBo reconstructs a heavily masked (90%) target view spatially cropped from the reference view. These reconstructions indicate that the bottleneck token captures sufficient information to recover the overall scene structure, including object identities, spatial locations, and their relationships.

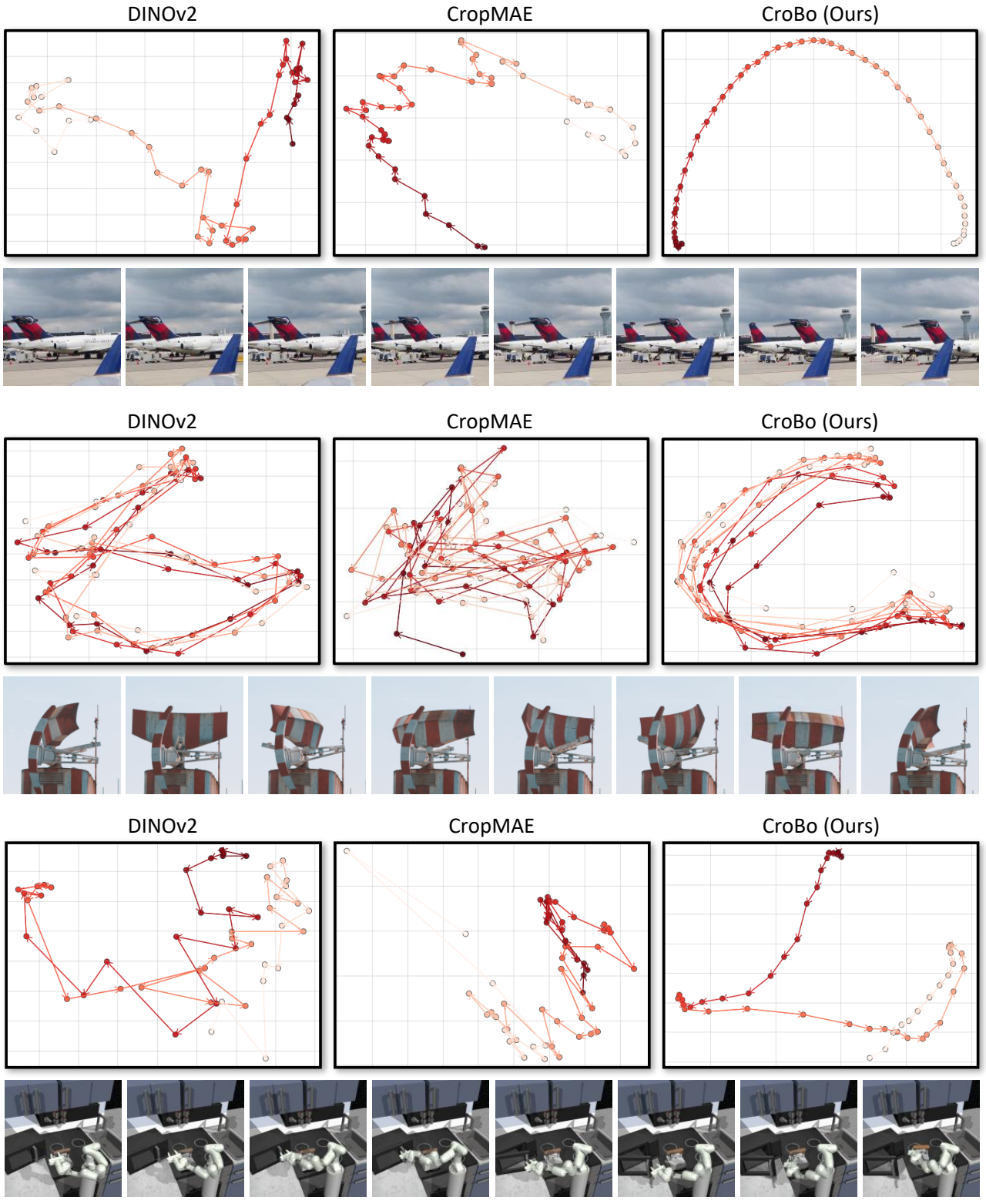

Figure 7. **Perceptual straightness of representation trajectories in video.** We visualize representation trajectories across video frames using PCA, where each point denotes a frame and colors indicate temporal order. CroBo produces smooth and locally linear trajectories that follow the natural evolution of the scene, whereas DINOv2 and CropMAE exhibit more irregular and entangled trajectories. Videos are drawn from MOSEv2 (first and second rows) [16] and Franka Kitchen (third row) [23].