Reconstruction-Guided Slot Curriculum: Addressing Object Over-Fragmentation in Video Object-Centric Learning

Supplementary Material

A. Datasets

To benchmark our method against existing video object-centric learning approaches, we conducted evaluations on three widely used datasets: YouTube-VIS 2021, MOVi-C, and MOVi-E. To ensure a fair comparison, we adopted the identical data splits and preprocessing methodology as SlotContrast [1]. The YouTube-VIS 2021 dataset [2] is a large-scale, real-world video instance segmentation benchmark derived from the YouTube-VOS dataset [3]. Following our baseline [1], we utilized 2,775 training videos and 210 validation videos at an image resolution of 518×518 pixels. MOVi-C and MOVi-E are synthetic datasets generated using Kubric [4]. MOVi-C comprises scenes with up to 10 scanned 3D objects, realistic backgrounds, and free camera movement. MOVi-E increases the complexity significantly, featuring up to 23 objects per scene and randomized camera trajectories, making it particularly challenging for unsupervised object discovery. Both MOVi datasets include 9,750 training videos and 250 validation videos each, rendered at a resolution of 336×336 pixels. All experiments were conducted using two NVIDIA RTX A6000 GPUs and an Intel Xeon Gold 5220R CPU.

B. Object Identification Recall

In our manuscript, we introduced Object Identification Recall (OIR) to measure how reliably the learned slots identify ground-truth (GT) instances. Specifically, OIR@ρ represents the fraction of foreground GT instances for which at least one slot attains IoU$\geq \rho$ with the instance mask. Let g denote the GT mask of an object at a specific frame, and let $\{m_k\}_{k=1}^{S}$ denote the predicted masks from all S slots, we compute $\mathrm{IoU}(g, m_k) = \frac{|g \cap m_k|}{|g \cup m_k|}$ and report the OIR at threshold ρ (OIR@ ρ):

$$\text{OIR@}\rho = \frac{1}{|\mathcal{G}_{\text{valid}}|} \sum_{g \in \mathcal{G}_{\text{valid}}} \mathbf{1}\left[\max_k \mathrm{IoU}(g, m_k) \geq \rho\right], \quad (1)$$

where $\mathcal{G}_{\text{valid}}$ is the set of GT objects with non-zero area (excluding the background class). OIR measures whether each foreground object is covered by at least one slot or not.

C. Degree of Over-Fragmentation

To precisely assess the over-fragmentation, we introduced the Degree of Over-Fragmentation (DOF). In this section, we illustrate DOF in detail. We say that the k-th slot represents a specific GT object if at least an ρ fraction of the slot area (m_k) lies within the GT mask g:

$$\frac{|m_k \cap g|}{|m_k|} \geq \rho \implies m_k \to g. \quad (2)$$

Let $\mathcal{G}_{\text{det}} = \{\, g \in \mathcal{G}_{\text{valid}} \mid \exists k : m_k \to g \,\}$ denote the set of detected objects. We report the average number of slots that are assigned to each detected object:

$$\text{DOF} = \frac{1}{|\mathcal{G}_{\text{det}}|} \sum_{g \in \mathcal{G}_{\text{det}}} \Big| \{\, k : m_k \to g \,\} \Big|, \quad (3)$$

where lower values (closer to 1) indicate fewer unnecessary splits of the same object. Note that the size of $|\mathcal{G}_{\text{det}}|$ can be different depending on the number of identified GT objects. Also, we only evaluate the DOF on identified GT objects since this isolates the over-fragmentation analysis from object discovery failures; if we include unidentified objects, which have a slot count of zero, it would artificially deflate the average and confound the metric with detection recall (which is already measured by OIR).

Table A1. **Average number of GT objects per slot (Degree of under-fragmentation).** A GT object is matched to a slot if IoU $\geq \rho$. Higher values indicate that single slots contain multiple objects (severe under-fragmentation).

Method	@0.3	@0.5	@0.7
SContrast	1.23	1.21	1.19
Ours	**1.22**	**1.18**	**1.15**

D. Degree of Under-Fragmentation

Complementary to over-fragmentation, we strictly assess the under-segmentation issue using the Degree of Under-Fragmentation (DUF). While over-fragmentation splits one object into many slots, under-fragmentation occurs when a single slot erroneously merges multiple distinct objects. Analogous to Eq. 2, we consider a slot to capture a ground-truth object if their Intersection-over-Union (IoU) exceeds a specified threshold ρ. DUF is then defined as the average count of ground-truth objects assigned to each matched slot. In this metric, a value of 1.0 represents the ideal one-to-one correspondence, while higher values indicate a failure to disentangle distinct instances (i.e., severe under-fragmentation).

The quantitative results are presented in Tab. A1. Notably, SlotCurri achieves the lower DUF scores across all

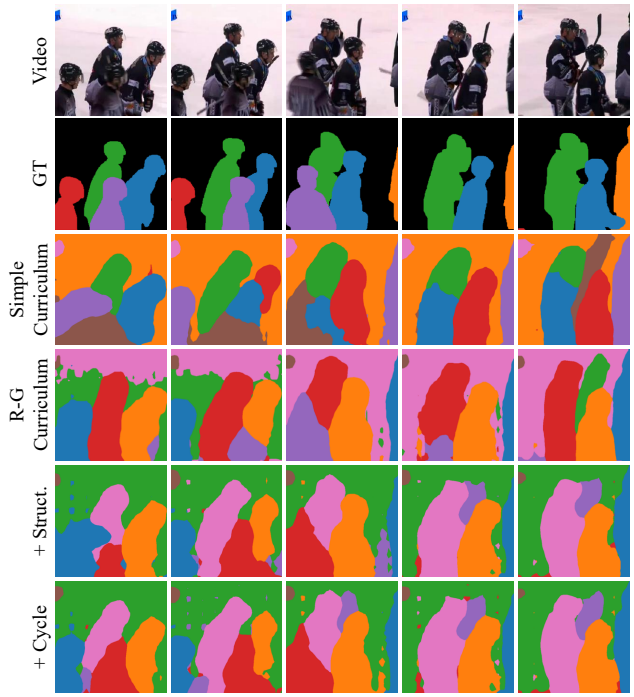

Figure A1. **Qualitative analysis of the impact of each component.**

thresholds compared to the baseline. This indicates that our SlotCurri does not merely shift the error distribution from over- to under-fragmentation but fundamentally improves the distinctness and quality of object disentanglement.

E. Qualitative Ablation Study

This section qualitatively illustrates the impact of each component with the visualizations in Fig. A1. The R-G curriculum mitigates over-fragmentation (4th row), while the SSIM objective sharpens object boundaries (*e.g.*, pink, red, and orange masks in 2nd column of 5th row) and leverages compositional structures to clearly separate distinct entities (*e.g.*, pink and red masks in 1st column of 5th row). Furthermore, cyclic inference effectively incorporates long-range contextual cues, enhancing slot compactness even in early frames (*e.g.*, blue and red masks in 1st column of last row).

F. Analysis of the Accelerated Slot Schedule

Our SlotCurri framework, as defined in Eq. (2), employs an accelerated slot schedule where the number of new slots added increases quadratically at each curriculum stage. This design is based on the hypothesis that the learning process benefits from allocating progressively more representational capacity to later, more complex refinement stages. To validate this hypothesis, we conduct an ablation study comparing our accelerated schedule against two alternatives, all starting with $K_{\text{init}} = 2$ and ending with the same final slot count (K_{final}): (1) Linear Schedule and (2) Decelerated Schedule.

Table A2. **Ablation study on the scheduling strategies on YouTube-VIS.**

Slot Scheduling	YouTube-VIS	
	FG-ARI↑	mBO↑
Decelerated	38.2	32.9
Linear	43.4	34.2
Accelerated	44.8	35.5

To illustrate, for the linear schedule, the total number of new slots is distributed as evenly across the stages, while the decelerated schedule adds the majority of slots in the first expansion stage and progressively fewer in later stages (the inverse of our accelerated approach). As shown in Tab. A2, the choice of schedule is critical. The decelerated schedule, which introduces high capacity too early, yields the smallest gain. This suggests that adding fine-grained capacity before coarse-level semantics have stabilized is detrimental, leading to premature over-fragmentation. The linear schedule performs better, but it is significantly outperformed by our accelerated approach.

This result provides strong evidence for our coarse-to-fine learning hypothesis, which involves two distinct phases. The first stage involves learning to anchor broad, stable, and semantically coherent regions, which requires relatively low slot capacity. Then, the second stage is to partition the fine-grained details within these coarse regions, which is a combinatorially more complex task, requiring a much larger representational budget to capture the explosion of new, smaller entities. Consequently, we observe that the accelerated schedule is effective as it mirrors this learning dynamic.

G. Analysis of the Reconstruction-Guided Slot Spawning Criterion

Table A3. **Analysis of slot spawning criteria on YouTube-VIS.** We compare our proposed criterion (Total Error Mass) against a scale-invariant alternative (Area-Normalized Error).

Slot Splitting Criteria	YouTube-VIS	
	FG-ARI↑	mBO↑
Area Norm	40.4	32.7
Total Error	44.8	35.5

Our reconstruction-guided curriculum allocates new slots by duplicating parent slots responsible for the highest

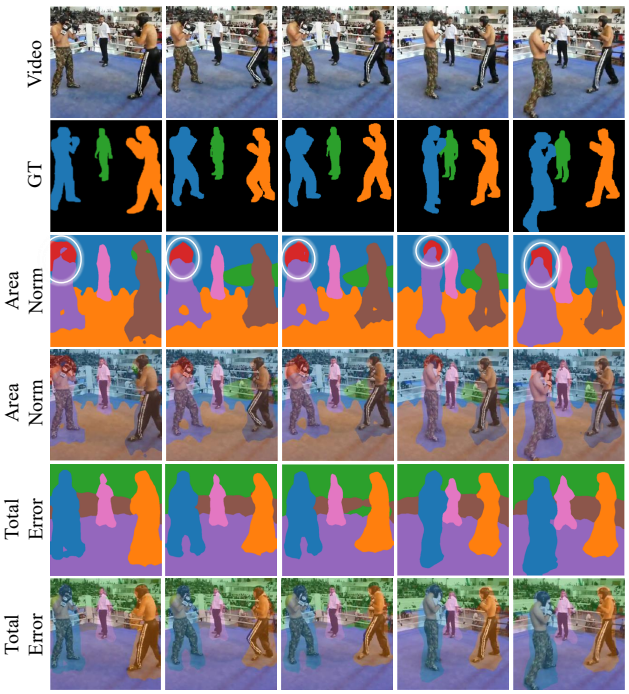

Figure A2. **Qualitative Comparison of Spawning Criteria.**

reconstruction error, $\delta^{(k)}$. This criterion measures the total error mass accumulated by a slot, weighted by its slot's reconstruction weights in the decoder ($\alpha^{(k,t,h,w)}$) across all spatio-temporal positions.

The primary advantage of this formulation is that it provides a usage-aware capacity allocation; slots that are idle, weakly used, or already perfectly reconstructed will have a negligible $\delta^{(k)}$ and are thus ineligible for splitting. This mechanism is crucial as it prevents the model from wasting capacity by unnecessarily fragmenting auxiliary or well-modeled regions, focusing refinement only on slots that are actively contributing to the reconstruction.

An alternative is to form a scale-invariant criterion by normalizing $\delta^{(k)}$ with the slot's total $\asymp$eight mass (*i.e.*, area normalization). While this could, in theory, find high-error slots regardless of their size, it introduces a critical instability. Since the reconstruction weights (α) are continuous, an idle or weakly used slot with a near-zero mass can yield an arbitrarily high normalized error from minor noise. To validate this, we empirically tested this area-normalized variant in Tab. A3, which shows that this variant markedly underperformed on the video benchmark. Also, Fig. A2 illustrates a common failure mode of this area-normalized variant. We observe that semantically distinct regions, such as the head and torso of a person, are often split into different slots. This suggests that normalizing by area can over-emphasize local error in semantically distinct sub-regions (like the head), even when they belong to the same object, leading to undesirable fragmentation.

Our findings suggest that, in our VOCL setup, allocating capacity in proportion to the total error mass is a more robust and effective strategy by naturally balancing the model's focus between reconstruction fidelity and semantic importance. We acknowledge that for domains dominated by small, dense objects (*e.g.,* MOVi-E), hybrid criteria may offer further robustness, which we leave as a promising direction for future work.

H. Chunk-wise Short Cyclic Inference

In the paper, we indicated that our cyclic inference is substantially more lightweight than the heavy encoder or decoder, leading to a negligible increase in inference time. This section further investigates the applicability of our method in streaming video applications where the latency is critical. For long videos, applying the full cycle across all frames introduces a significant computational bottleneck, making it impractical for real-time streaming.

Table A4. **Performance of chunk-wise short cyclic inference on YouTube-VIS.** We compare the full cycle, no cycle (✗), and short cycles constrained to each video chunk of C frames ($C \in \{2, 3\}$) on YouTube-VIS.

Cycle	YouTube-VIS	
#Frames (C)	FG-ARI↑	mBO↑
✗	43.6	35.2
2	43.8	35.2
3	43.9	35.1
full	44.8	35.5

Therefore, we explore whether a short cycle (where the signal propagates only to the next frame or a few subsequent frames before returning) can effectively replace the full cycle. To this end, we experiment by processing the video in chunks of C frames (where C is the chunk size) and constraining the cycle to operate only within these chunks. Specifically, we partition the video into sequential chunks, each containing a maximum of C frames. The cyclic inference is then performed locally within each chunk. As the model processes frames and reaches the end of a chunk, a backward cyclic pass is initiated, propagating information back to the start of that same chunk. This chunk-wise mechanism ensures that the cyclic inference is confined to a short, fixed-duration segment. Consequently, this approach is highly applicable to streaming scenarios, as it (1) strictly bounds the computational latency of the cyclic operation and (2) generates object slots on a per-chunk basis, removing the need to buffer or process the entire video before producing an output. As shown in Tab. A4, while this chunk-wise cycle does not achieve the full performance of a complete cycle, it still demonstrates a clear improvement over the model with no cycle. This finding indicates that even a localized cyclic mechanism offers a practical trade-off, enhancing performance while maintaining viability for

streaming applications.

I. Necessity of Structure Loss in Slot Curriculum Learning

Table A5. **Effectiveness of structure-aware reconstruction loss** ($\mathcal{L}_{\text{SSIM3D}}$). $*$ represents the reproduced results.

Method	$\mathcal{L}_{\text{SSIM3D}}$	YouTube-VIS	
		FG-ARI↑	mBO↑
SlotContrast*	-	36.1	32.7
SlotContrast	✓	36.4	32.7
R-G Curriculum	-	42.9	33.8
R-G Curriculum	✓	44.8	35.5

As discussed in the main manuscript, the commonly used mean-squared-error (MSE) reconstruction loss tends to blur spatial details and obscure the true boundaries between objects. This issue becomes more problematic under our curriculum learning framework, as blurred boundaries lead to ambiguous slot representations that are further subdivided in subsequent stages. To validate this effect, we investigate the impact of incorporating our structure-aware reconstruction loss into the baseline SlotContrast [1]. As shown in Tab. A5, our reconstruction-guided curriculum learning exhibits substantial gains of +2.3 points in FG-ARI and +1.0 points in mBO, when combined with the structure-aware loss. In contrast, applying the same loss to SlotContrast yields only marginal improvements, highlighting the importance of learning clear semantic boundaries within a curriculum learning setup.

J. Analysis of 2D and 3D Structure Loss

Table A6. **Comparison between spatio-temporal (3D-SSIM) and spatial-only (2D-SSIM) structure-aware loss on YouTube-VIS.**

SSIM Objective	YouTube-VIS	
	FG-ARI↑	mBO↑
✗ SSIM	42.9	33.8
+2D-SSIM	44.2	34.3
+3D-SSIM	44.8	35.5

Our framework's performance relies on sharpening semantic boundaries before slot expansion. We achieve this by complementing the standard MSE loss with a structure-aware SSIM objective. However, for video tasks, this structural loss can be applied in two ways: (1) as a 2D-SSIM, computed independently for each frame, or (2) as a 3D-SSIM, computed over spatio-temporal cubes. In this work, we employed a 3D-SSIM objective since 2D-SSIM treats each video frame as an independent image, ignoring the temporal dimension. This objective lacks any incentive to maintain structural consistency across time, which is a critical signal for learning stable object representations across the temporal dimension. To validate this design choice, we trained an identical model where our 3D-SSIM loss was replaced by a standard 2D-SSIM loss applied frame-by-frame. As shown in Tab. A6, the 2D-SSIM variant suffers a performance decrease. This confirms our hypothesis that enforcing structural consistency across time is a key component of our method's success, ensuring that the sharpened boundaries learned in the early curriculum stages are temporally robust.

K. Robustness to Varying Slot Numbers

In unsupervised object-centric learning, the exact number of objects in a scene is typically unknown a priori. Consequently, models are often deployed with a predefined number of slots that exceeds the actual object count to ensure all entities are captured. However, this overestimation of slot capacity frequently leads to a degradation in performance for conventional methods, as excess slots tend to fragment single objects into multiple parts. To evaluate the robustness of our method against such capacity mismatches, we conduct experiments on the MOVi-C dataset by varying the number of slots $K \in \{7, 11, 15\}$.

The results are summarized in Tab. A7. We observe that the baseline, SlotContrast, suffers from a severe performance drop as the slot capacity increases. Specifically, when K is increased from 11 to 15, its FG-ARI plummets by over 7 points (from 69.3 to 61.8), indicating that the model struggles to handle redundant slots and succumbs to over-fragmentation. In contrast, our method demonstrates remarkable stability across all settings. Even when the number of slots significantly overestimates the scene complexity ($K = 15$), SlotCurri maintains a high FG-ARI of 74.8, outperforming the baseline by a large margin (+13.0). This confirms that our curriculum-based strategy effectively suppresses the activation of redundant slots, allowing the model to be robustly deployed without precise knowledge of the scene's object count.

K. Discussion on Failure Modes and Limitations

While SlotCurri effectively mitigates over-fragmentation, we identify two primary areas for future investigation.

Under-fragmentation of Small Objects. First, SlotCurri is less effective on datasets where under-fragmentation is the primary challenge. As shown in Fig. A3, our predictions on the MOVi-E dataset, which contains a large number of small objects, can fail to

Figure A3. **Qualitative results on MOVi-E dataset.** We present visualizations of video frames along with their corresponding GT masks and prediction masks from our SlotCurri. The white circles in the first row highlight the limitation of our work.

Table A7. Results with varying number of slots on MOVi-C.

Method	slot=7		slot=11		slot=15	
	FG-ARI↑	mBO↑	FG-ARI↑	mBO↑	FG-ARI↑	mBO↑
SContrast	74.9	**27.9**	69.3	32.7	61.8	31.2
Ours	**78.7**	27.8	**77.6**	**32.8**	**74.8**	**33.9**

delineate clear boundaries (white circles in the first video) or leave small objects spatially entangled (white circles in the second video). We attribute this limitation to (1) the low spatial resolution (24×24) of the feature maps, which inherently blurs fine-scale structures, and (2) the lack of explicit guidance for distinguishing between similar but spatially overlapping objects. As future work, we plan to leverage overlapping image patches produced by processing the original frames along with the spatially-shifted frames. By exploring the semantic differences between partially overlapping patches, we expect to better capture fine-grained structures, thereby alleviating under-fragmentation.

Refining the Curriculum Schedule. Another limitation of SlotCurri is that it relies on a predefined curriculum, where the number of stages (M) and the iteration timings for slot expansion (*e.g.*, at 10% and 25% of training) are set as hyperparameters. Our extensive experiments demonstrate that this scheduling strategy is robust and highly effective. However, as a fixed schedule, it may require minor manual tuning to achieve optimal performance on new datasets with significantly different characteristics. This presents a promising direction for future research: the development of scene-adaptive slot schedules, as noted in our conclusion. Such a mechanism could, for example, automatically trigger slot expansion based on metrics like the plateauing of reconstruction error, making the framework more robust and general.

L. SlotCurri on Object Dynamics Prediction

Object dynamics prediction aims to forecast the future states of objects based on their historical observations, serving as a critical benchmark for evaluating the temporal consistency and physical meaningfulness of learned representations. To validate the effectiveness of the slot representations acquired by SlotCurri in a practical downstream scenario, we conduct experiments on the object dynamics prediction task. Following standard protocols [1], we train a dynamics predictor module [5] on top of the frozen slots learned by our model.

The quantitative results are summarized in Tab. A8. In

Table A8. Experimental results on object dynamics prediction. SlotFormer (SF) is used to evaluate each pretrained VOCL method.

Method	YouTube-VIS		MOVi-C		MOVi-E	
	FG-ARI↑	mBO↑	FG-ARI↑	mBO↑	FG-ARI↑	mBO↑
Recon	27.4	28.9	50.7	25.9	**70.6**	24.3
SContrast	29.2	29.6	63.8	**26.1**	70.5	**24.9**
SlotCurri (Ours)	**31.4**	**30.1**	**70.0**	25.7	70.2	24.7

real-world scenarios, SlotCurri demonstrates a significant performance improvement over existing baselines, confirming that our curriculum-based approach yields representations that are far more robust to complex dynamics. Notably, on the MOVi-C dataset, our method achieves a remarkably high FG-ARI compared to prior works, indicating that our slots successfully capture distinct object identities. On the MOVi-E dataset, we observe a trend consistent with the VOCL benchmarks: since the primary challenge in MOVi-E stems from under-fragmentation, which differs from the primary target of SlotCurri, the performance gains are naturally less pronounced. Nonetheless, our method demonstrates performance that remains competitive with the state-of-the-art.

M. Additional Results on COCO Dataset

Since SlotCurri operates as a curriculum learning framework that progressively expands the slot capacity, its applicability extends beyond video domains to static image object-centric learning. To verify this generalizability, we conduct experiments on the COCO 2017 dataset. For the static image setting, we adapt the SSIM loss from 3D to 2D and exclude the cyclic inference mechanism, as it relies on temporal dynamics. Consequently, the model is trained using the reconstruction-guided slot curriculum combined with the 2D SSIM loss.

The quantitative results are presented in Tab. A9. Consistent with our observations in video benchmarks, SlotCurri effectively mitigates the over-fragmentation problem in the image domain. This improvement is evidenced by the substantial increase in Image-ARI compared to the reconstruction-based baseline, demonstrating that our curriculum strategy successfully guides slots to capture coherent object instances even without temporal cues.

Table A9. **Quantitative evaluation on the MS COCO dataset.** SlotCurri achieves a significant improvement in Image-ARI by effectively resolving the over-fragmentation issue inherent in the reconstruction-based baseline.

Method	COCO	
	Image-ARI↑	Image-mBO↑
Reconstruction	40.5	28.8
SlotCurri (Ours)	43.4	28.9

N. Additional Qualitative Results

In Fig. A4 and A5, we present additional qualitative comparisons against SlotContrast [1]. Our method, SlotCurri, demonstrates strong object-level grouping, consistently assigning a single slot per semantic entity, even in an unsupervised setting. In contrast, the SlotContrast exhibits clear signs of over-fragmentation. For instance, in the first YouTube-VIS video, the deer (masked with an orange mask in GT) is subdivided into red and brown masks in SlotContrast. Similarly, in the second video, the body and head of bears are assigned to different slots, indicating spatial over-fragmentation. These patterns persist across other videos and are also observed in the synthetic MOVi-C dataset. While the baseline tends to over-segment coherent entities into multiple parts, SlotCurri maintains more consistent and compact object-level groupings across both space and time.

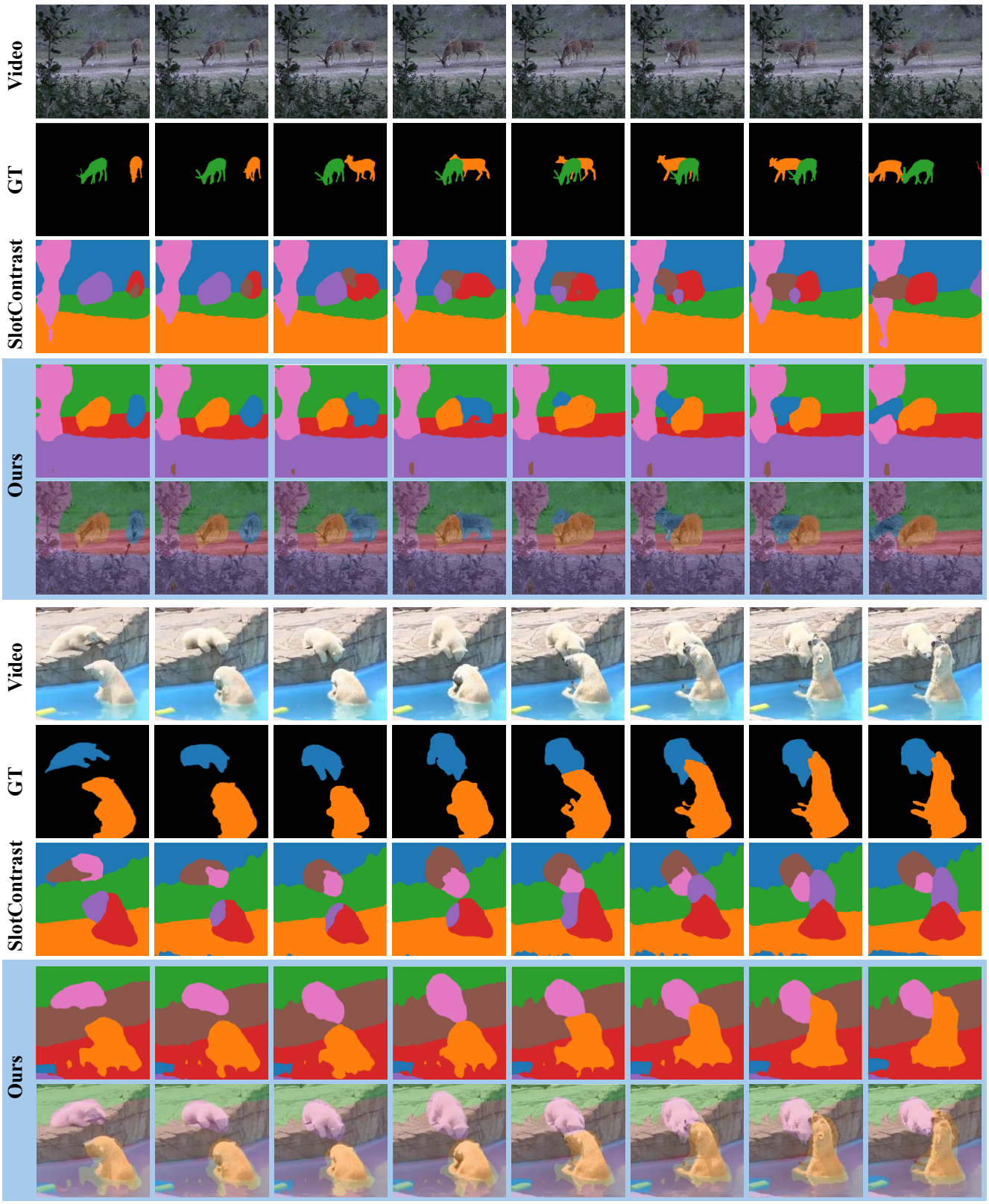

Figure A4. **Qualitative results on the YouTube-VIS dataset.**

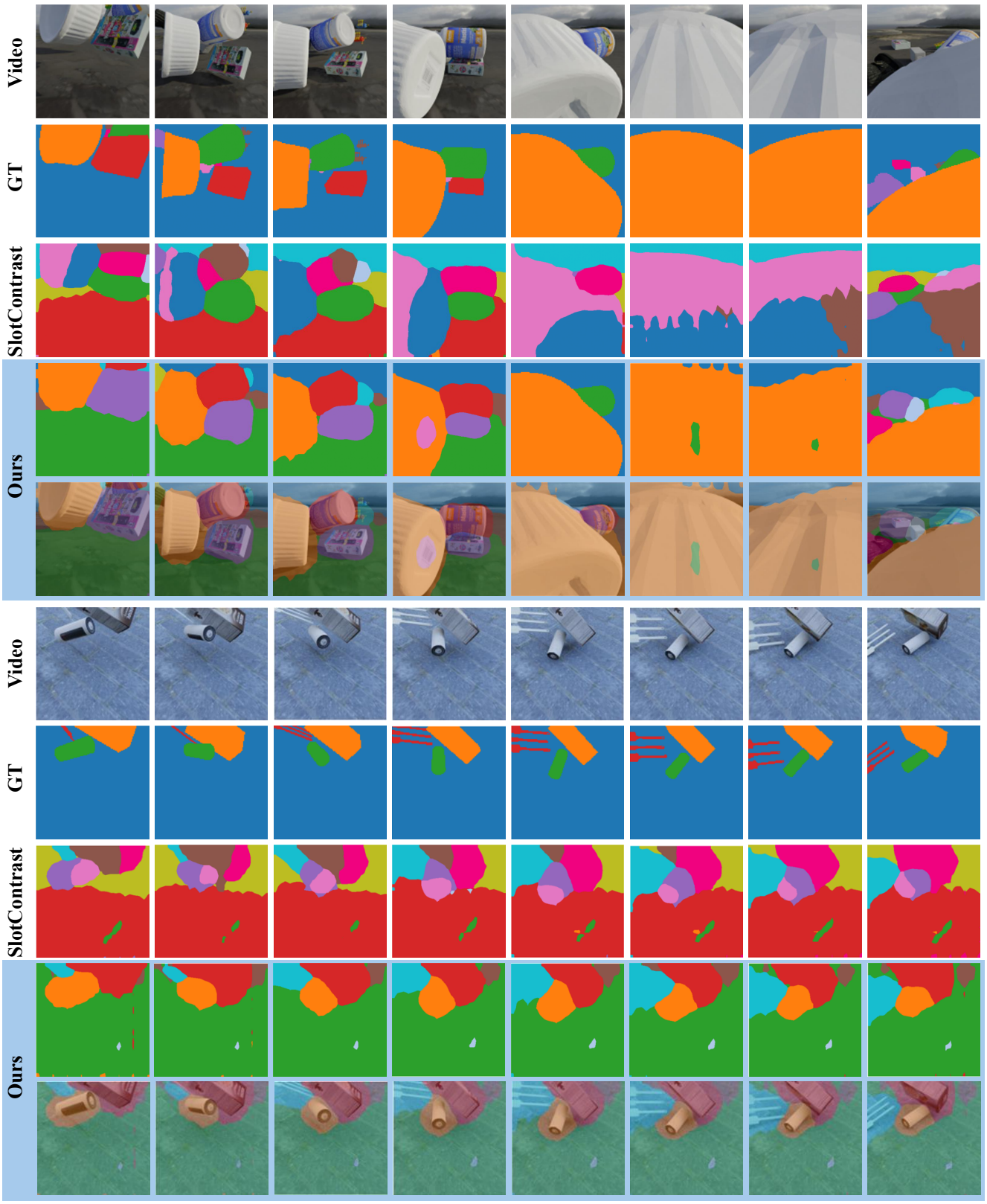

Figure A5. **Qualitative results on the MOVi-C dataset.**

References

[1] Anna Manasyan, Maximilian Seitzer, Filip Radovic, Georg Martius, and Andrii Zadaianchuk. Temporally consistent object-centric learning by contrasting slots. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5401–5411, 2025. 1, 4, 5, 6

[2] Tao Wang, Ning Xu, Kean Chen, and Weiyao Lin. End-to-end video instance segmentation via spatial-temporal graph neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10797–10806, 2021. 1

[3] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *European conference on computer vision*, pages 208–223. Springer, 2020. 1

[4] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3749–3761, 2022. 1

[5] Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised visual dynamics simulation with object-centric models. *arXiv preprint arXiv:2210.05861*, 2022. 5