Better Slots, Better Worlds:

Representation Quality & Robustness in Object-Centric World Models

Shukrullo Nazirjonov[1], Sai Prasanna[1,2,†], Anna Manasyan[1,2,†], Georg Martius[1,2]

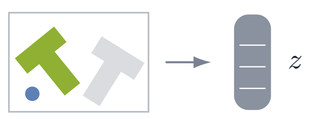
[1]University of Tübingen [2]Max Planck Institute for Intelligent Systems [†]Equal contribution

Paper Code

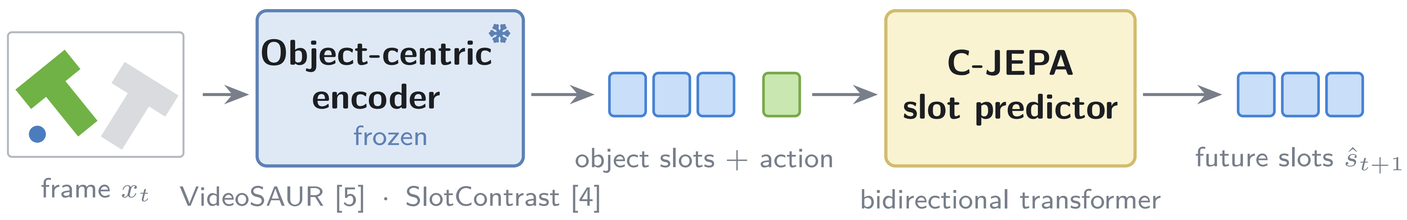
Summary

- **Contributions:** three controlled research questions about what makes object-centric world models work for visual planning.
- **Q1** Planning improves with object binding quality and plateaus at high quality.
- **Q2** With well-bound slots, neither masking nor robot-state input improves planning in our evaluations.
- **Q3** Frozen pretrained features are a key driver of robustness to visual shifts.

One Scene, Different Representations



Experimental Context: C-JEPA Framework [3]

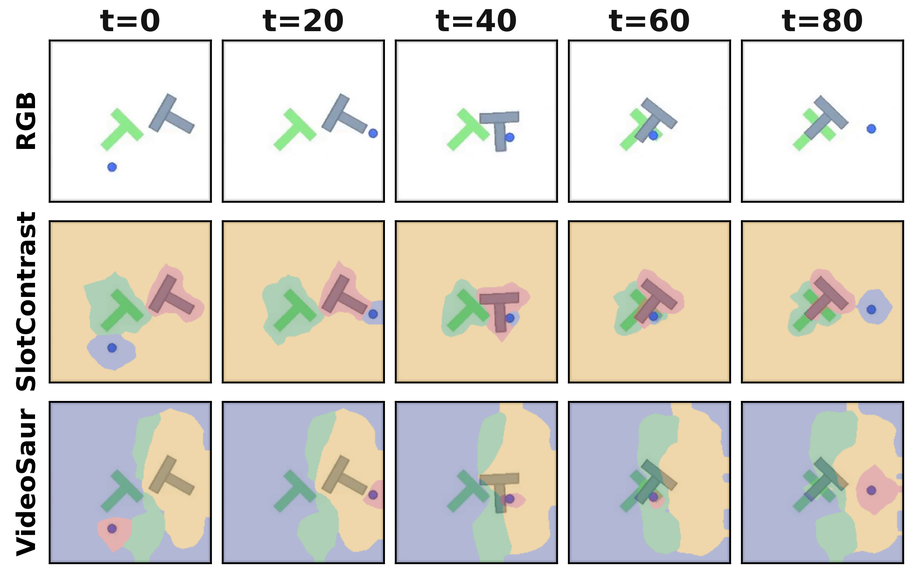


We use C-JEPA [3] as a **fixed testbed**: a frozen encoder maps each frame to object slots, and an action-conditioned transformer predicts future slots. We vary only **encoder quality**, **slot masking**, and the **robot-state input**; predictor architecture and planner stay fixed.

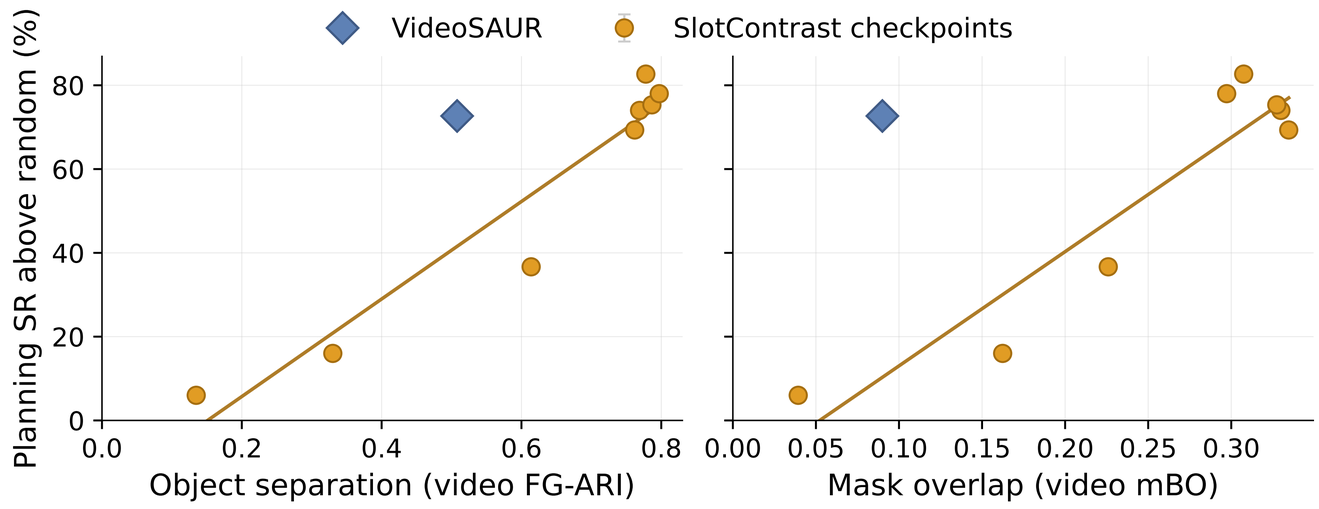
Evaluation: MPC + CEM planning **Environments:** 2D PushT · 3D OGBench-Cube **Baselines:** DINO-WM [1] (frozen DINOv2 patches) · LeWM [2] (end-to-end ViT) · C-JEPA [3] recipe — same planner budget for all

Q1 · Does slot quality drive planning success?

SlotContrast [4] produces **temporally consistent object slots**. Across its PushT checkpoints, object separation and mask overlap track planning until slot quality plateaus; predictor and planner are **fixed**.



The relation is weaker on OGBench-Cube, where large objects can dominate mask scores even when the small, task-relevant objects are poorly represented.



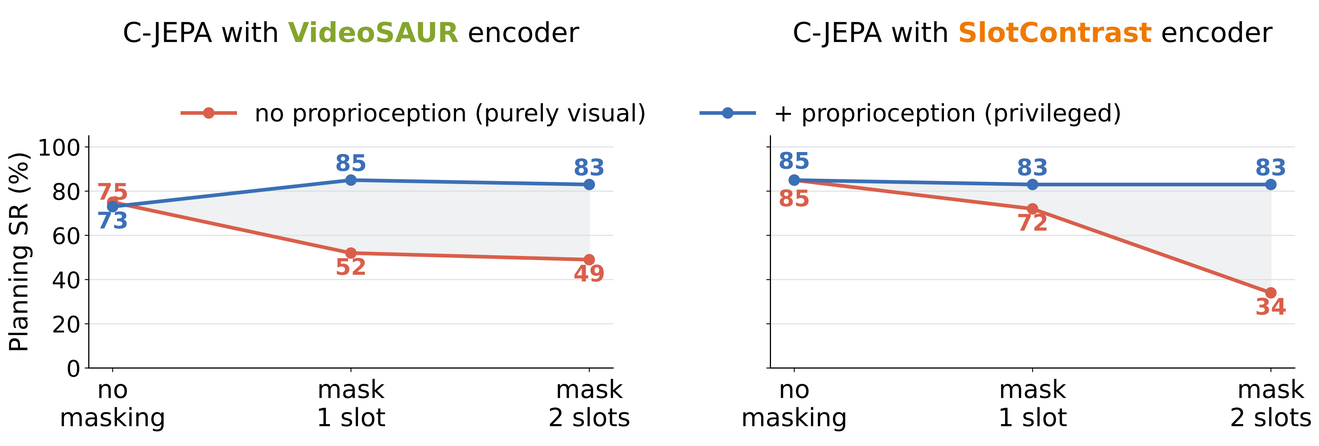
Takeaway 1: *Better-bound slots yield better planning, with performance plateauing at high slot quality.*

Q2 · Are auxiliary objectives & privileged inputs needed?

C-JEPA [3] partitions slots into **masked** and **unmasked** subsets and predicts the masked slots from the rest; it also supplies a robot-state (**proprioception**) token. We sweep 0–2 masked slots, with and without that input, using SlotContrast [4] and the weaker VideoSAUR encoder [5].

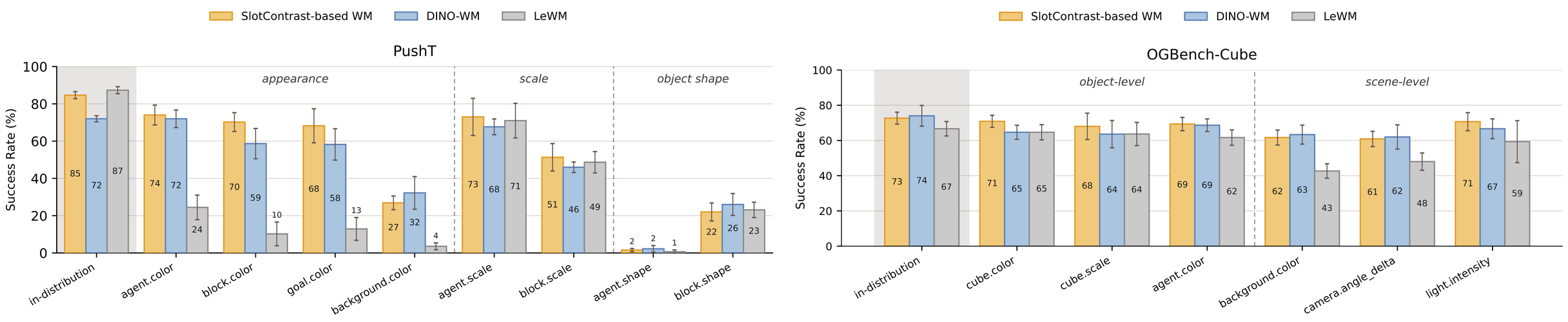
Across this sweep, masking lowers **purely visual** planning for both encoders. It helps the weaker encoder only when paired with robot state (73.3 → 85.3), indicating a **proprioceptive shortcut** (shaded gap).

Takeaway 2: *With well-bound object slots, neither masking nor robot-state input improves planning in our evaluations.*



Q3 · What drives robust planning under distribution shift?

Under **unseen visual shifts**, SlotContrast-based WM [4] degrades least and DINO-WM [1] stays close, while end-to-end LeWM [2] degrades sharply. The shared frozen pretrained features likely contribute substantially to this robustness. All models fail when object shape changes because new geometry changes **contact dynamics**.



Takeaway 3: *In our evaluations, world models planning over frozen pretrained features (object-centric or not) retain performance under appearance and frame-level shifts.*

References

[1] G. Zhou, H. Pan, Y. LeCun, L. Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. ICML 2025.
[2] L. Maes et al. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels. arXiv:2603.19312, 2026.
[3] H. Nam et al. Causal-JEPA: Learning world models through object-level latent interventions. arXiv:2602.11389, 2026.
[4] A. Manasyan, M. Seitzer, F. Radovic, G. Martius, A. Zadaianchuk. Temporally consistent object-centric learning by contrasting slots. CVPR 2025.
[5] A. Zadaianchuk, M. Seitzer, G. Martius. Object-centric learning for real-world videos by predicting temporal feature similarities. NeurIPS 2023.
[6] F. Locatello et al. Object-centric learning with slot attention. NeurIPS 2020.
[7] Z. Wu, N. Dvornik, K. Greff, T. Kipf, A. Garg. SlotFormer: Unsupervised visual dynamics simulation with object-centric models. ICLR 2023.