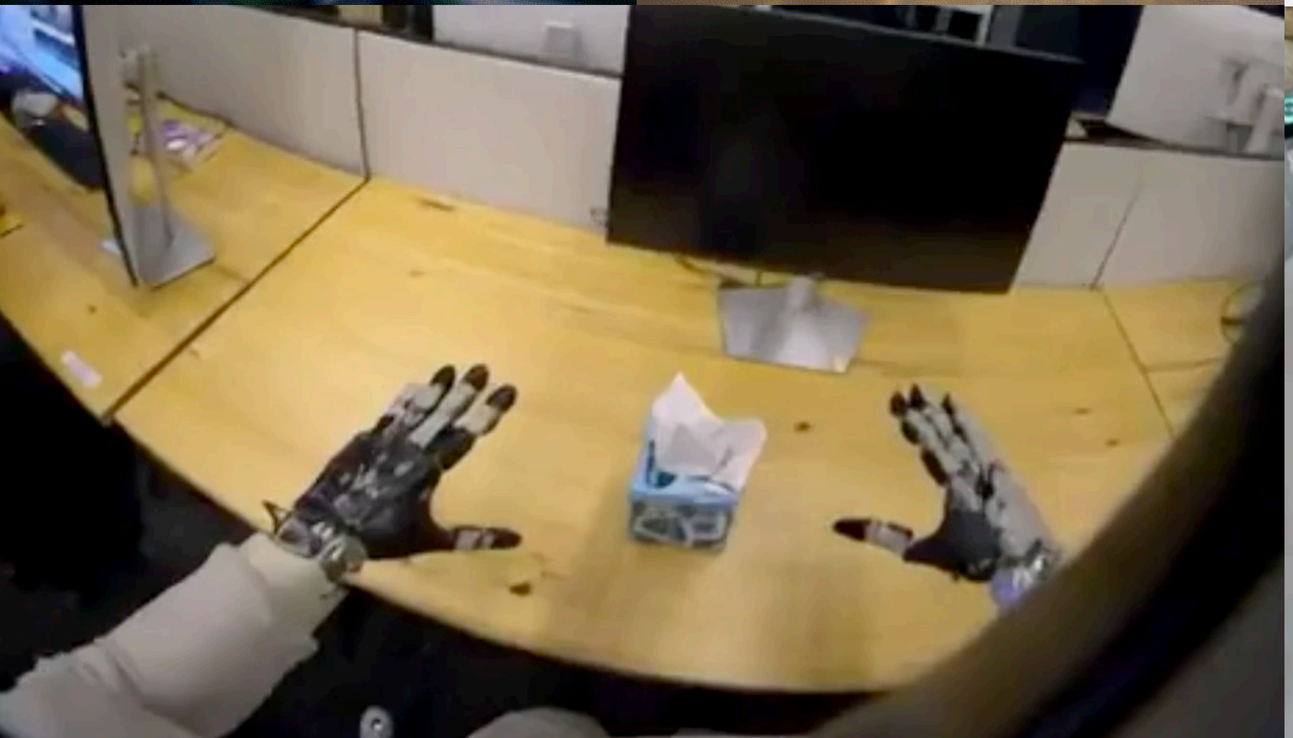
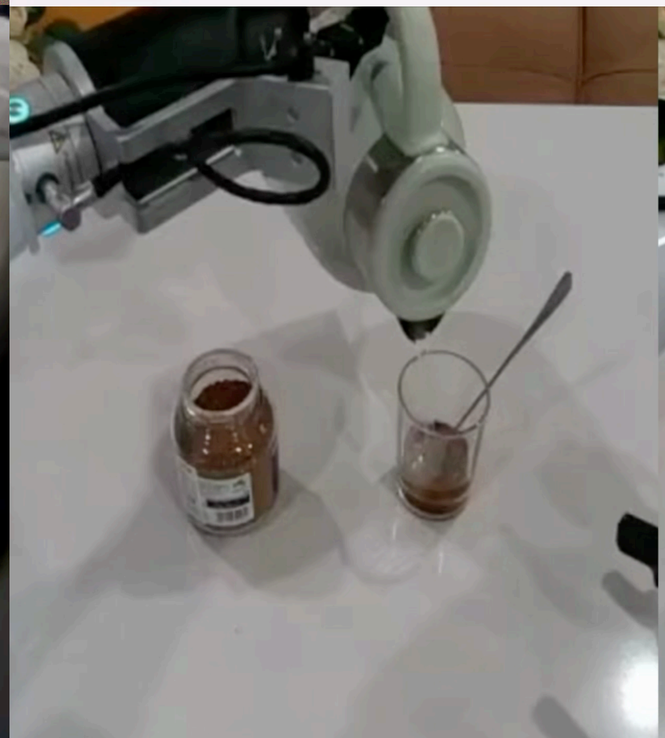
Geometry-aware 4D Video Generation for Robot Manipulation

Zeyi Liu[1], Shuang Li[1], Eric Cousineau[2], Siyuan Feng[2],
Benjamin Burchfiel[2], Shuran Song[1]

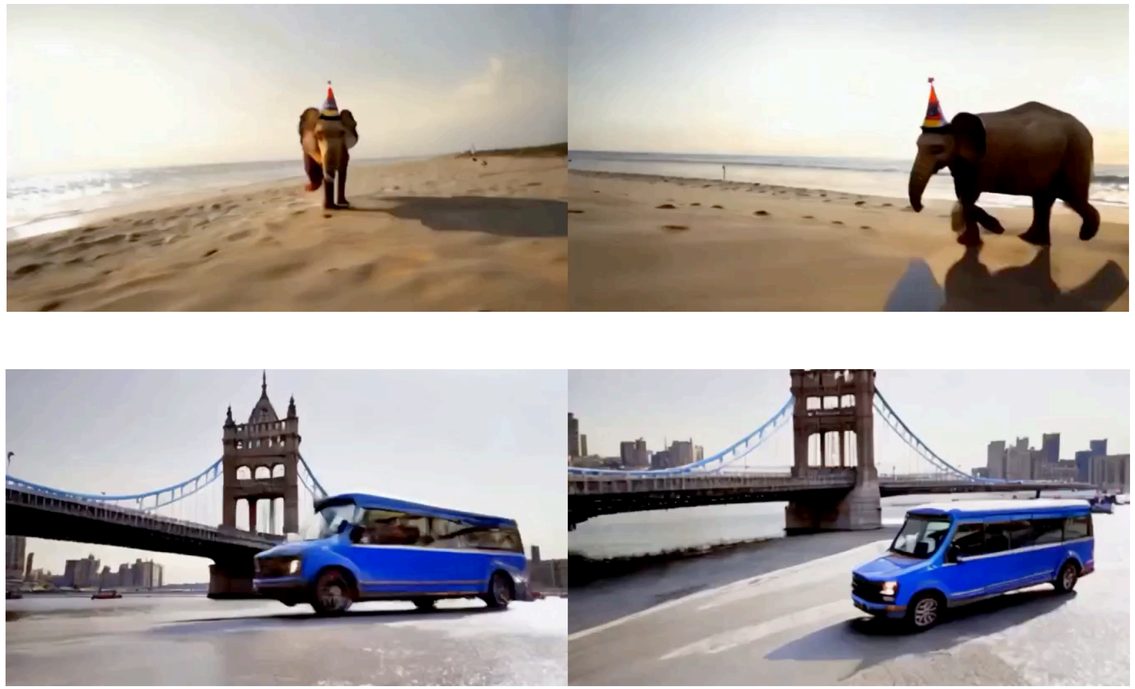
[1]Stanford University, [2]Toyota Research Institute


COITIZEN
COUNTET

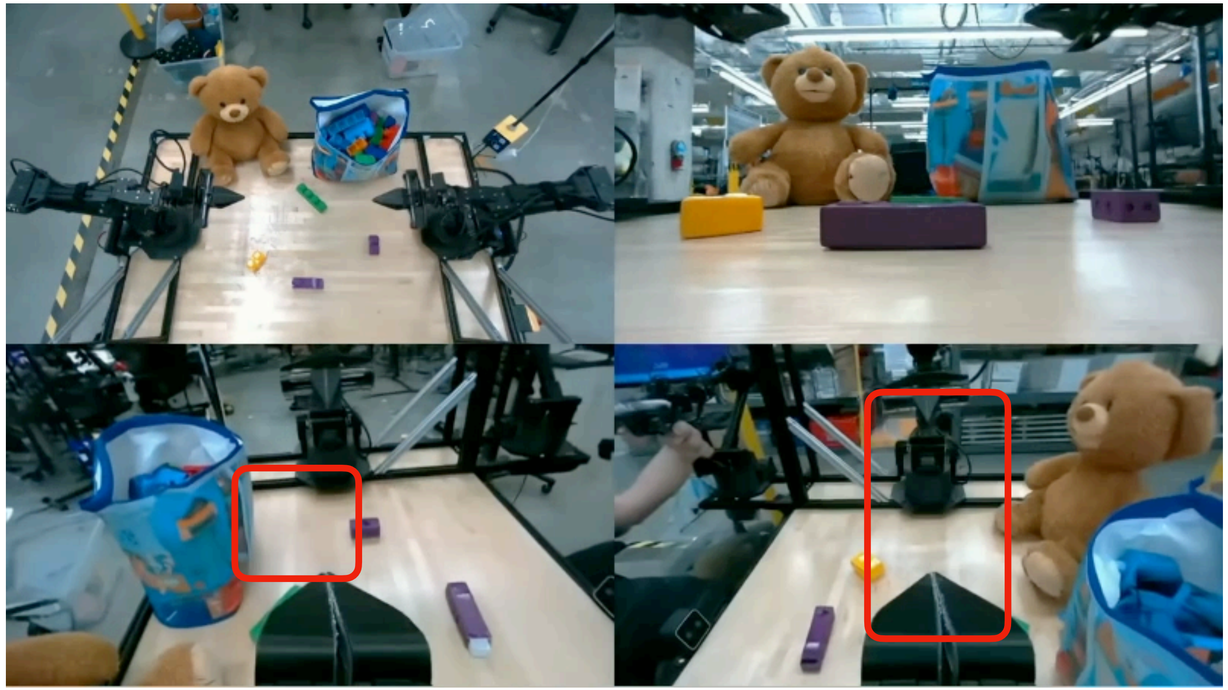




Challenge: view consistency

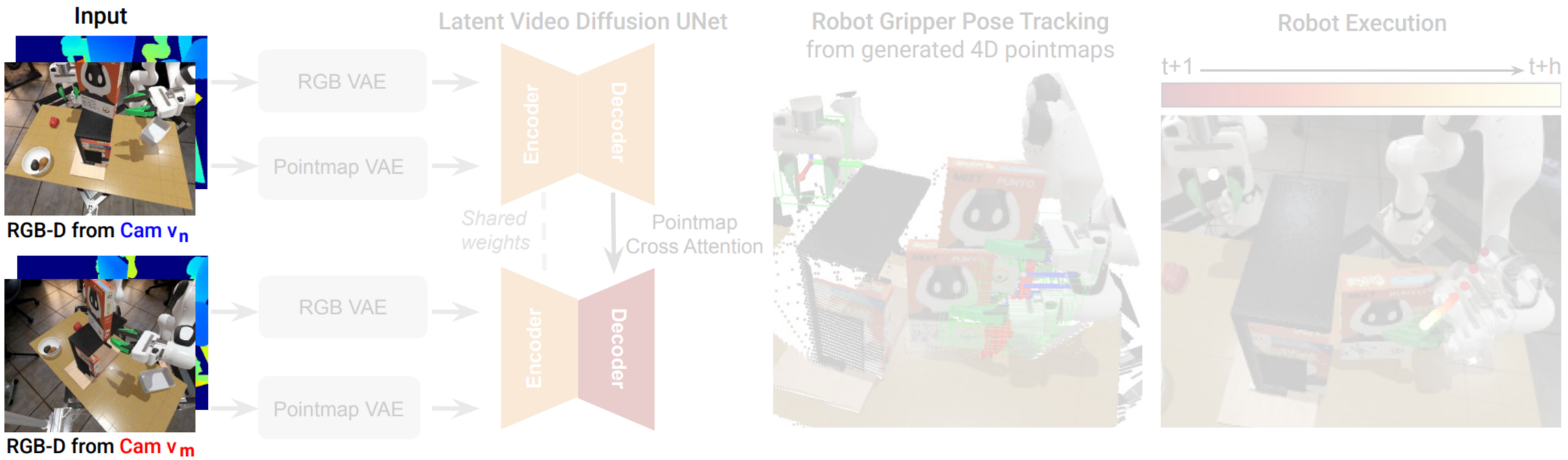


[1] Bai et al., SynCamMaster, 2024



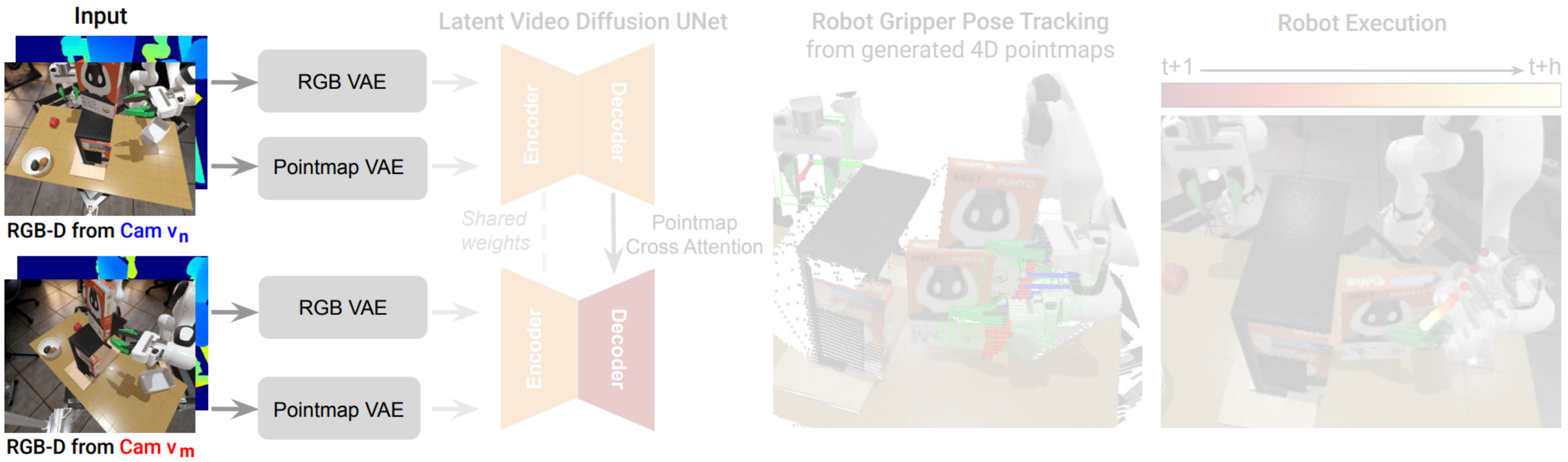
[2] Team, Gemini Robotics, et al., 2025

Method



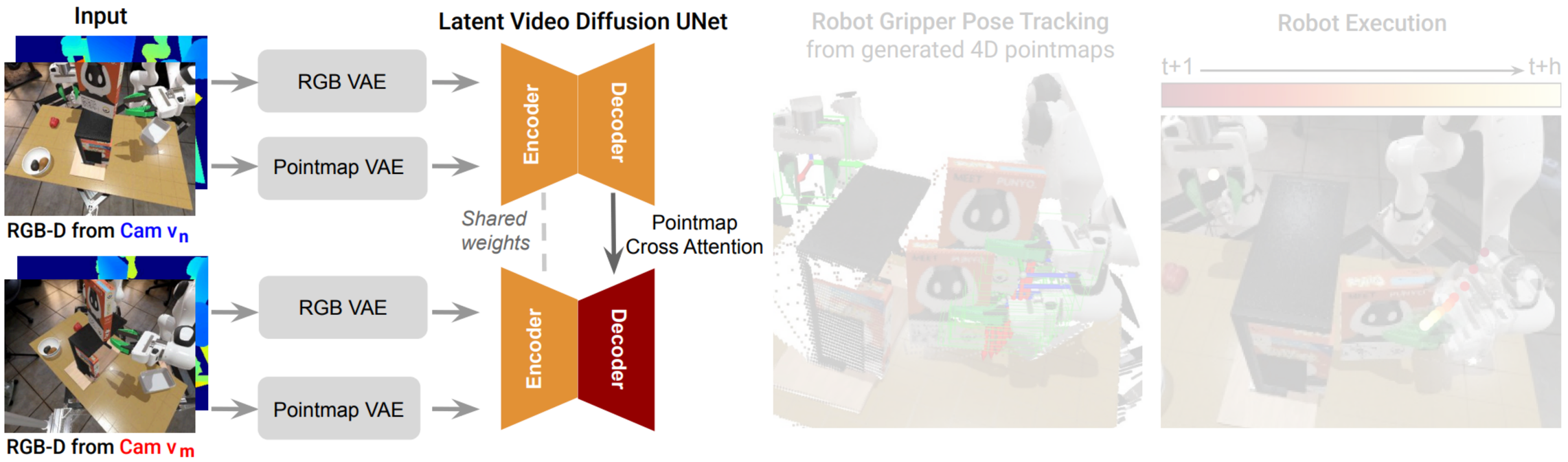
Input: paired RGB-D videos from two randomly sampled camera views

Method



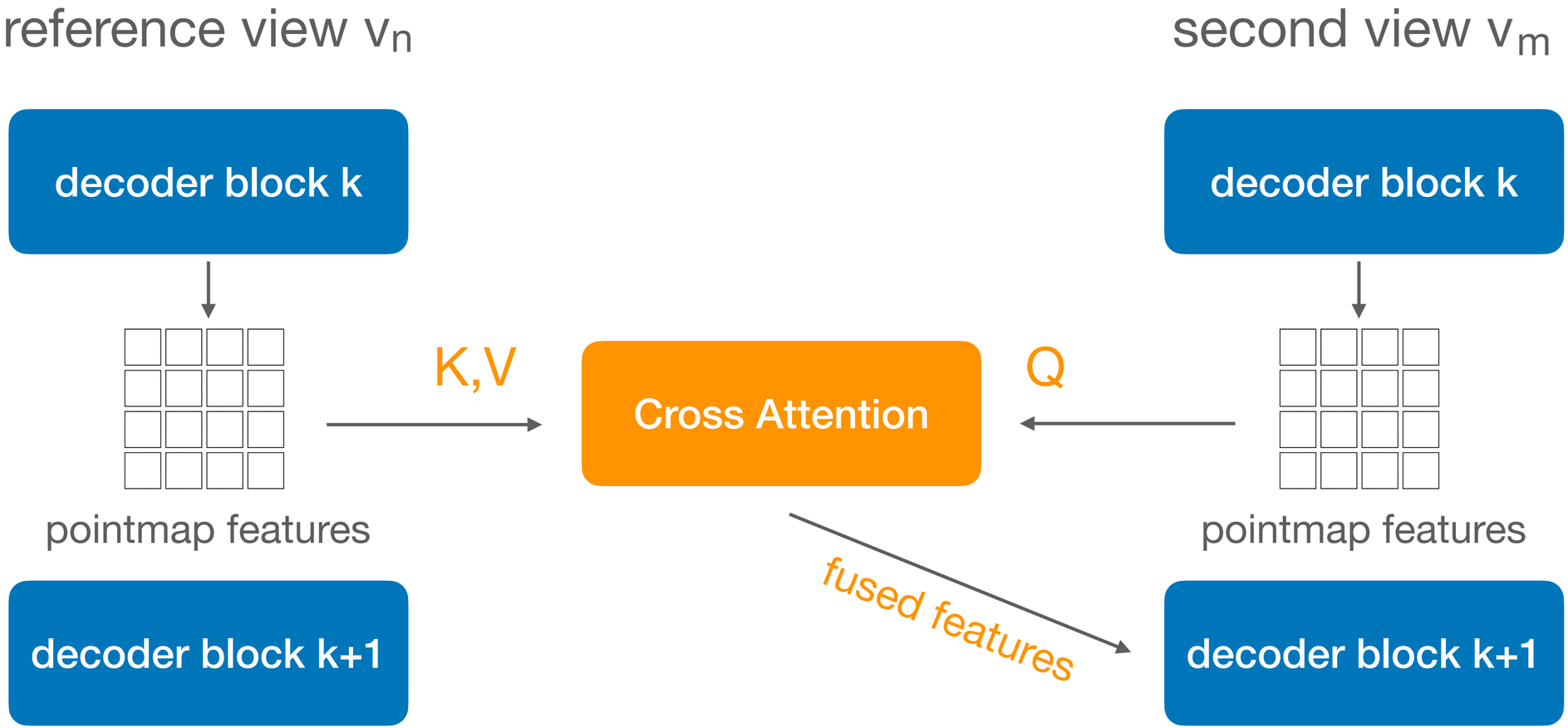
Encode: $z = \text{concat}\left(\mathcal{E}_{\text{rgb}}(x_{\text{rgb}}),\ \mathcal{E}_{\text{pointmap}}(x_{\text{pointmap}})\right)$

Method



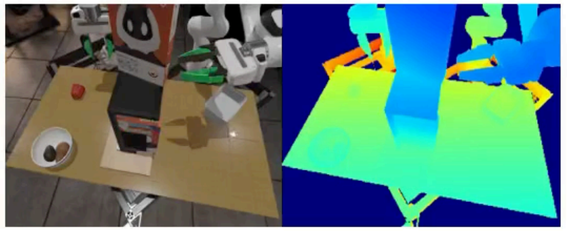
Prediction: separate U-Net decoders with cross-attention over pointmaps

Multi-View Cross-Attention for 3D consistency

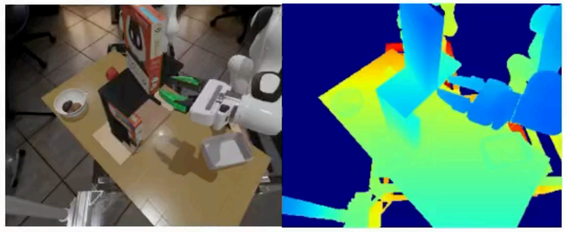


Geometry-consistent Supervision

RGB-D observations in Cam v_n



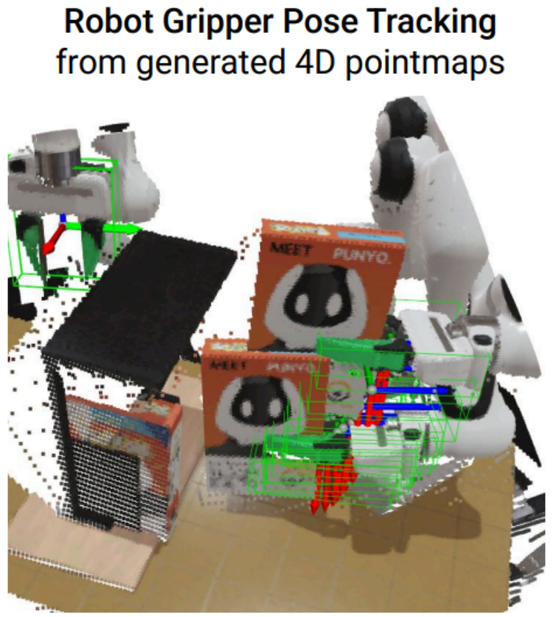
RGB-D observations in Cam v_m

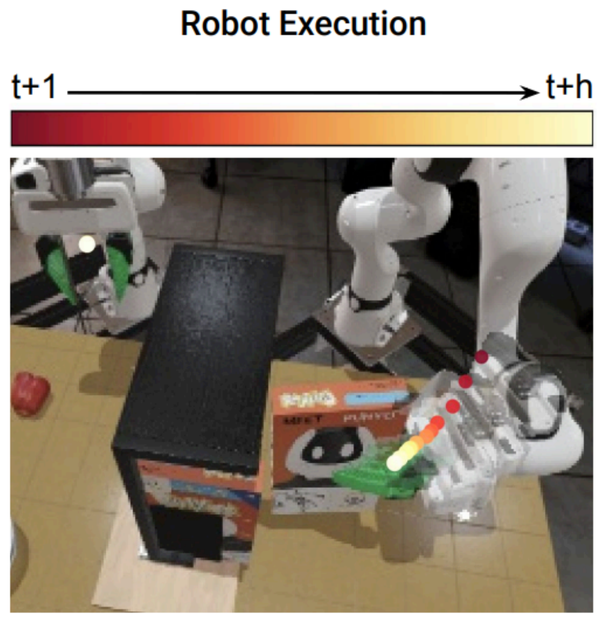


Novel Views

$$\mathcal{L} = \sum_{t'=t+1}^{t+h} \left[\underbrace{\mathcal{L}^n_{\text{diff}}(t') + \mathcal{L}^m_{\text{diff}}(t')}_{\text{RGB loss}} + \lambda \cdot \underbrace{\mathcal{L}_{\text{3D-diff}}(t')}_{\text{pointmap loss}} \right]$$

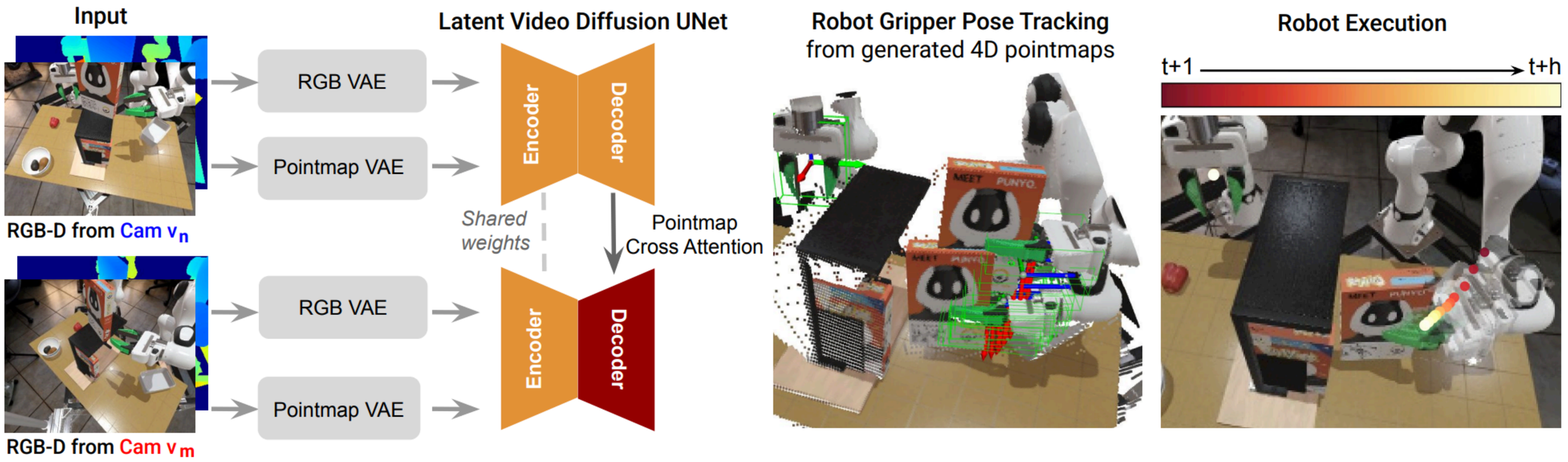
Method





Application: pose tracking from predicted multi-view RGB-D videos

Method



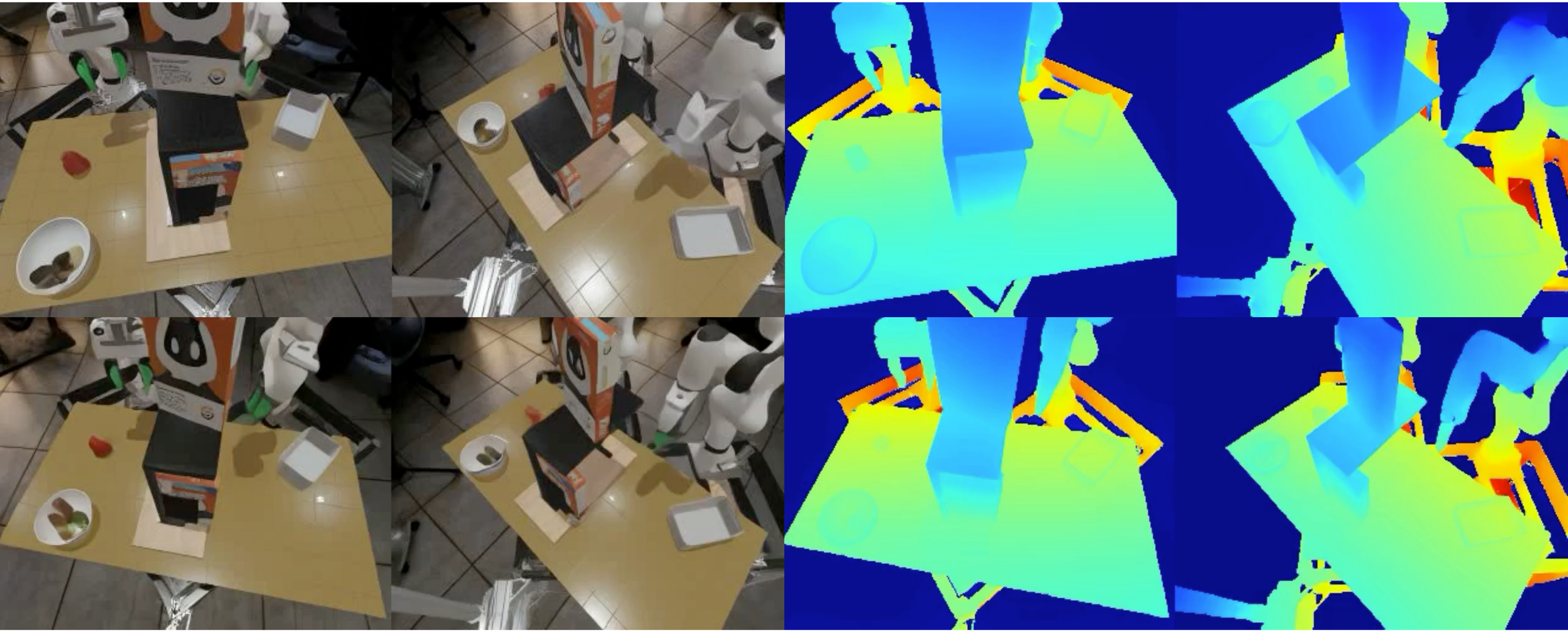
NO need for camera pose input during inference time

Video Generation Results

Our method generates **high-fidelity** RGB and depth predictions with strong **cross-view consistency**.

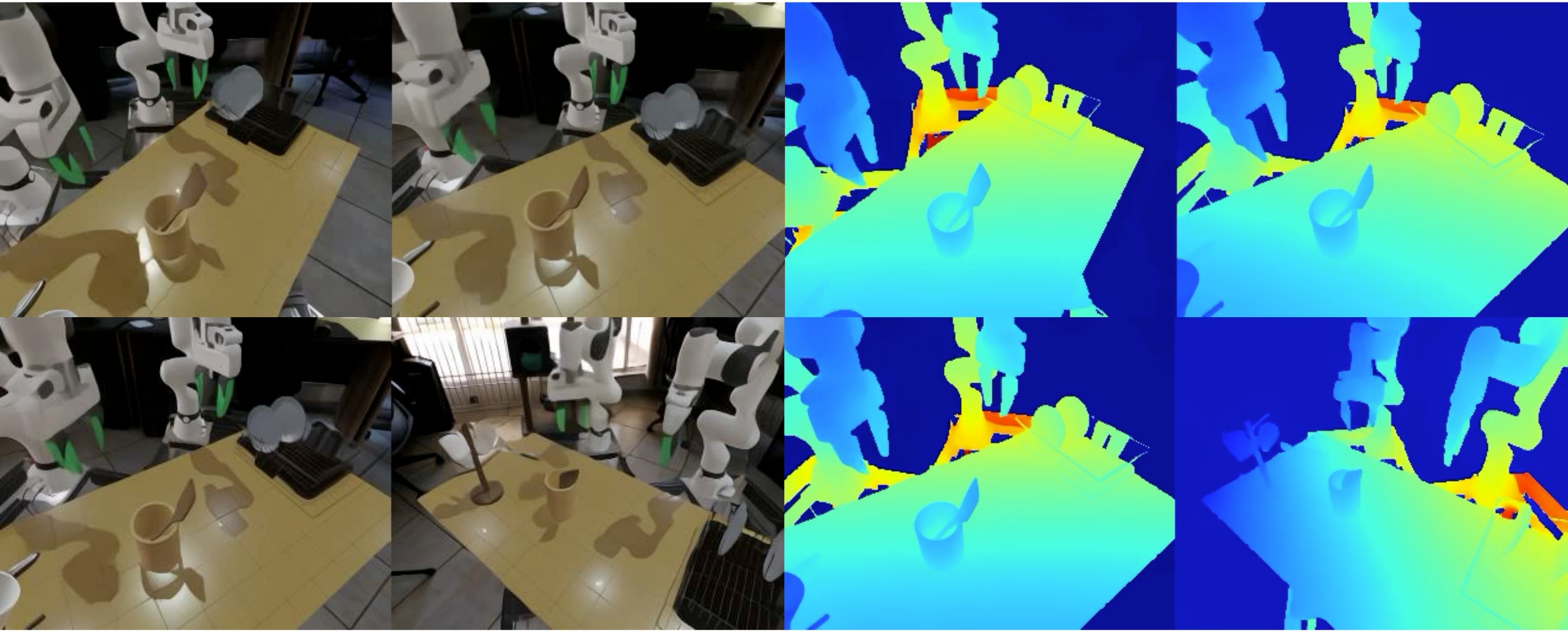
Task 1 🥣: Put Cereal Box Under Shelf

RGB view 1 | RGB view 2 | Depth view 1 | Depth view 2



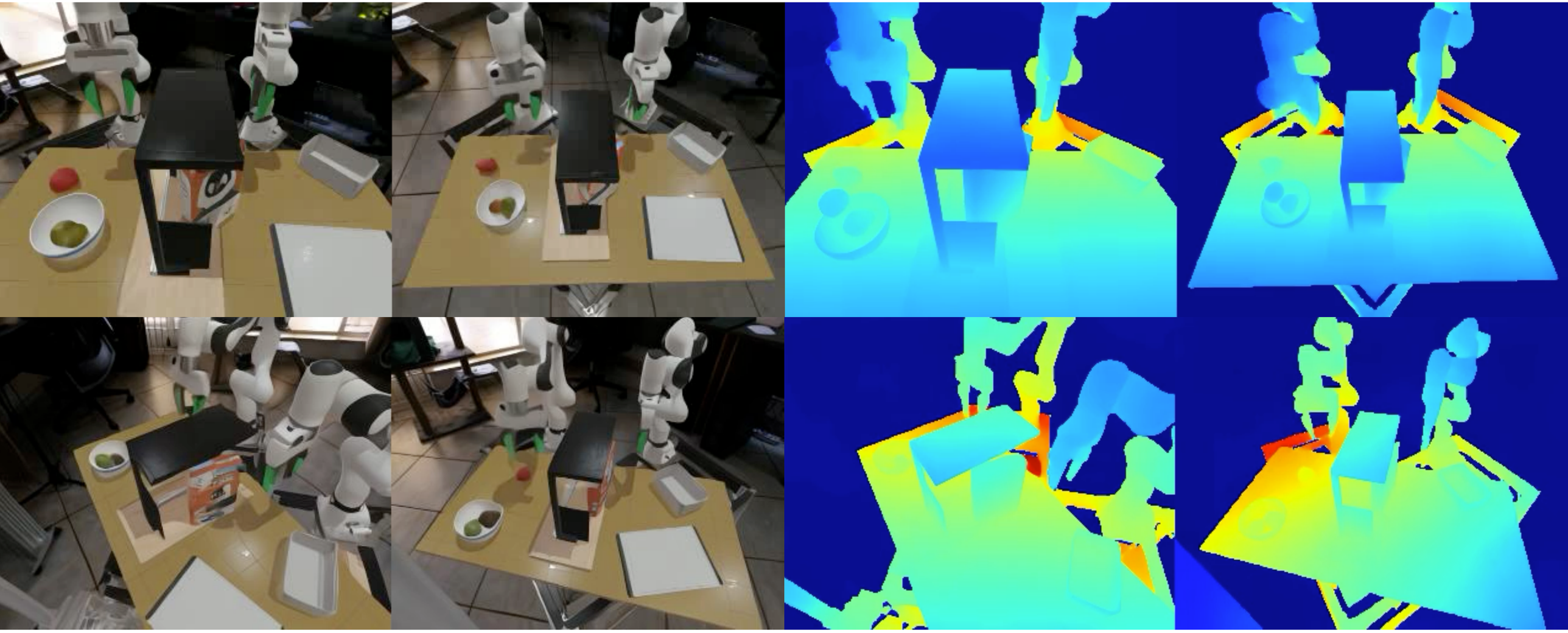
Task 2 👨‍🍳: Put Spatula On Table

RGB view 1 | RGB view 2 | Depth view 1 | Depth view 2



Task 3 🍎: Place Apple From Bowl Into Bin

RGB view 1 | RGB view 2 | Depth view 1 | Depth view 2



Comparison with Baselines

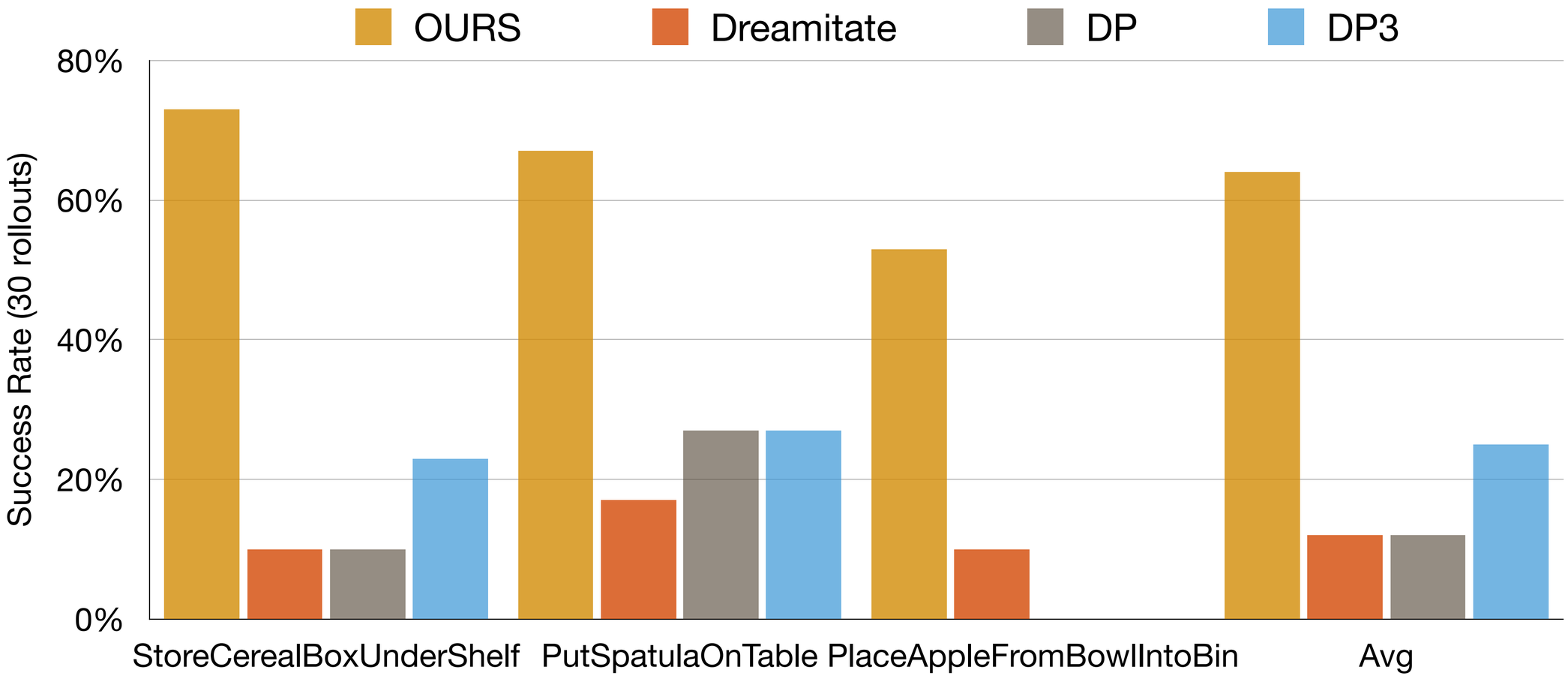
OURS w/o cross attn

SVD w/ cross attn

SVD

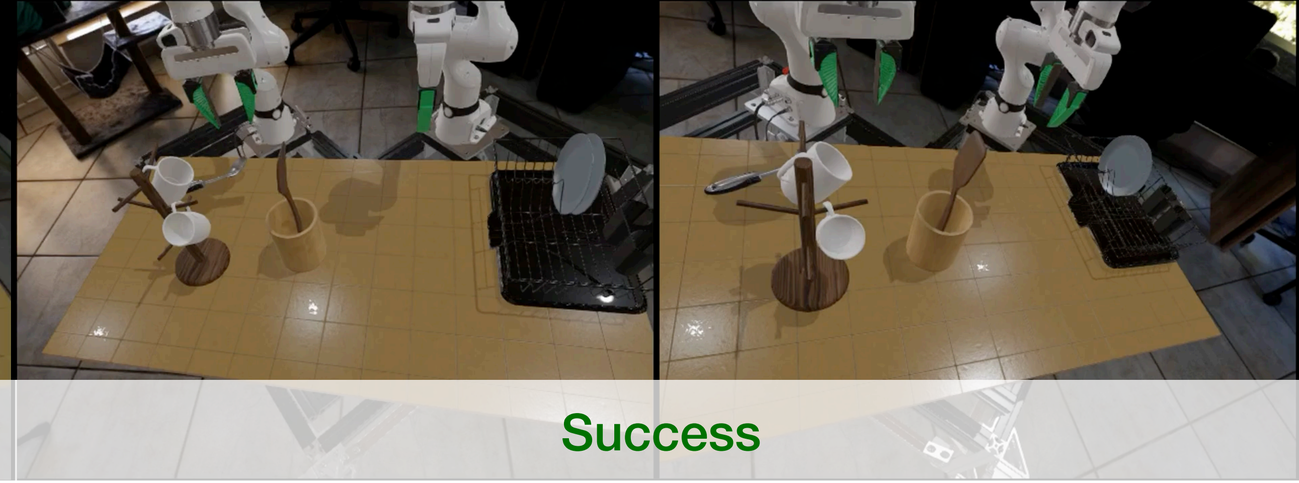
Policy Results

policy performance in unseen camera views



OURS

Success



Success

Real World Fine-Tuning

Fine-tuning on real-world dataset enables high-fidelity multi-view RGB-D generation for real-world manipulation tasks.

Future Directions

- Leverage RGB-based depth estimation to enable scalable, high-quality data curation in real-world settings without requirement on depth cameras.

- Explore faster generative backbones, such as flow matching and autoregressive transformers to enable more reactive robot policies.

- Extend the model to egocentric video generation (e.g., wrist-mounted cameras), which are common in robotic applications.

Stanford University

TOYOTA RESEARCH INSTITUTE

Geometry-aware 4D Video Generation for Robot Manipulation

robot4dgen.github.io

POSTER SESSION 6 Sat, Apr 25th, 2026 3:15 PM - 5:45 PM