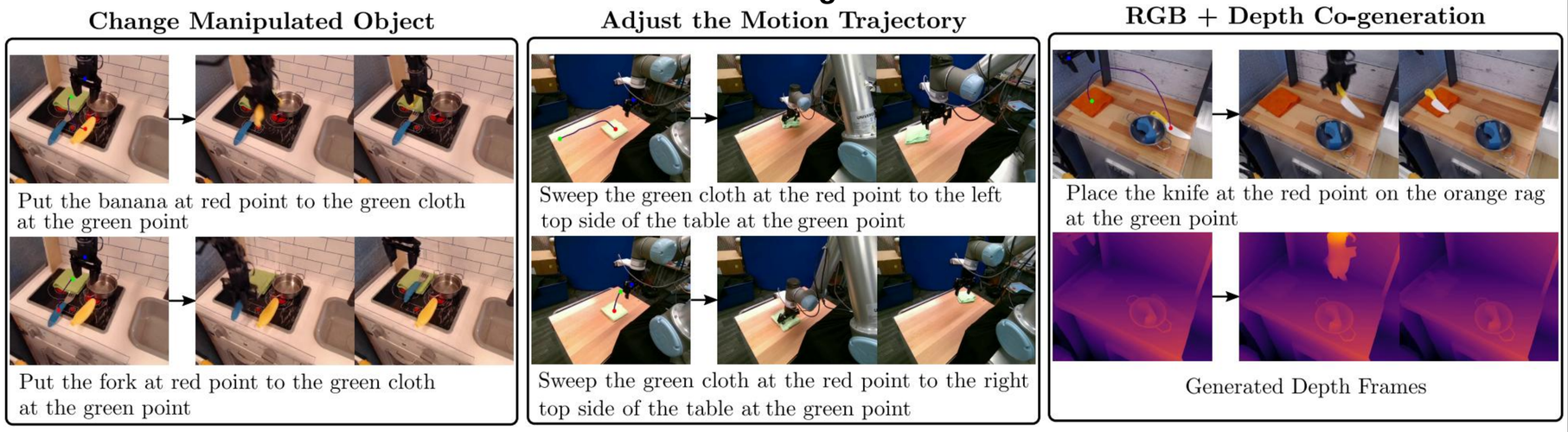
DRAW2ACT: Turning Depth-Encoded Trajectories into Robotic Demonstration Videos

Yang Bai, Liudi Yang, George Eskandar, Fengyi Shen
Mohammad Altillawi, Ziyuan Liu, Gitta Kutyniok

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
HUAWEI
Munich Center for Machine Learning
universität freiburg
Technische Universität München

ICRA IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION 2026 VIENNA

Challenge



We propose **DRAW2ACT** to jointly generate spatially aligned RGB and depth videos, leveraging cross-modality attention mechanisms and depth supervision to enhance the Spatio-Temporal consistency, and a multimodal policy model conditioned on the generated RGB and depth sequences to regress the robot's actions.

Contributions

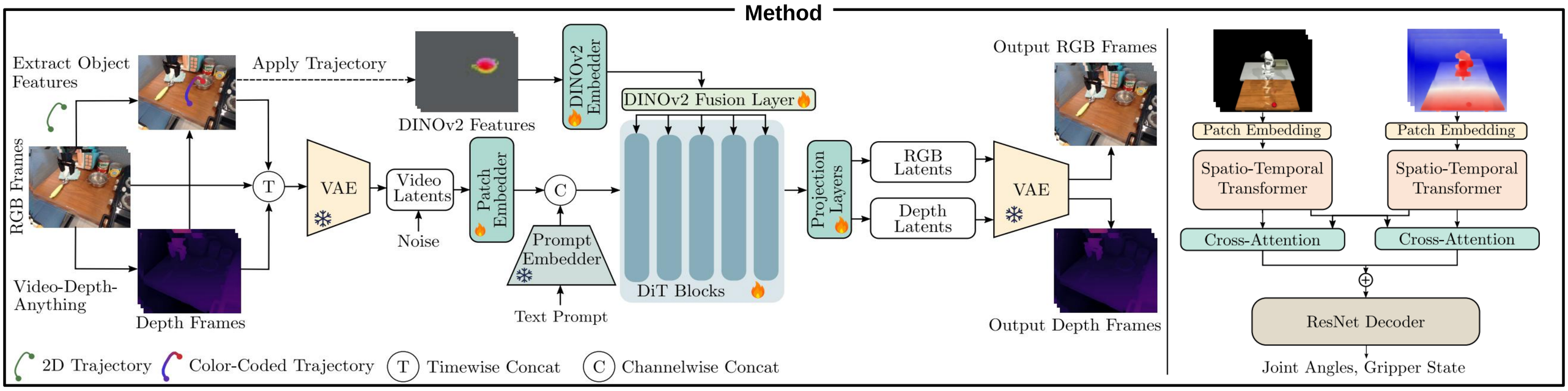
1.Trajectory-Conditioned Model
- Integration of **3D trajectories**, **semantic features**, and **coordinate-enhanced text** within a diffusion transformer.

2. Object-Centric Feature Control
- First-frame object feature extraction and **trajectory-based propagation** for shape, semantics, and motion guidance.

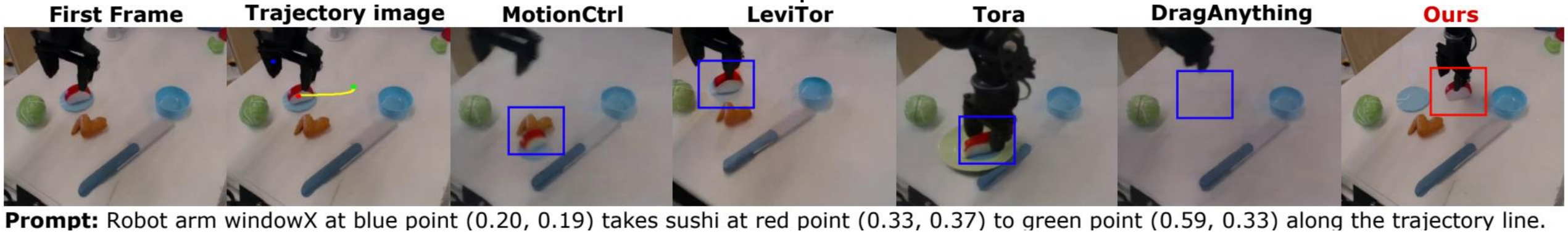
3. Multimodal RGB–Depth Generation
- Joint RGB and depth video synthesis with a **multimodal policy** for robotic manipulation learning.

4. Performance Gains
- **Higher video quality** and **task success rates** compared to prior trajectory-conditioned methods.

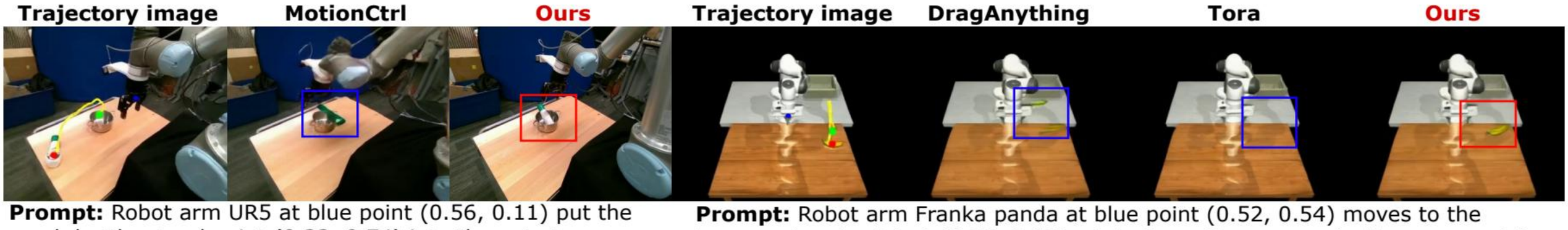
Method



Video Comparison



Prompt: Robot arm windowX at blue point (0.20, 0.19) takes sushi at red point (0.33, 0.37) to green point (0.59, 0.33) along the trajectory line.



Prompt: Robot arm UR5 at blue point (0.56, 0.11) put the ranch bottle at red point (0.23, 0.74) into the pot at green point (0.43, 0.51) along the trajectory line.

Prompt: Robot arm Franka panda at blue point (0.52, 0.54) moves to the banana at red point at (0.72, 0.67), picks it up, then moves to the green point and releases it. at (0.72, 0.60) along the trajectory line.

Qualitative Comparison

Dataset	Method	Vbench Evaluation Mot.Cons.↑	Bg.Cons.↑	Subj.Cons.↑	Tem.Fli.↑	Trajectory Deviation Object Traj. Error ↓	Task Evaluation Success Rate(%)↑
Bridge V2	LeviTor	0.9712	0.9289	0.9272	0.9817	46.52	N/A
	Tora	0.9875	0.9507	0.9346	0.9821	35.67	N/A
	MotionCtrl	0.9792	0.9471	0.9317	0.9811	38.24	N/A
	DragAnything	0.9810	0.9442	0.9289	0.9832	37.11	N/A
	Ours	**0.9891**	**0.9512**	**0.9383**	**0.9849**	**25.30**	N/A
Berkeley UR5	LeviTor	0.9803	0.9436	0.9325	0.9761	47.29	N/A
	Tora	0.9818	0.9502	0.9410	**0.9833**	35.73	N/A
	MotionCtrl	0.9844	0.9472	0.9391	0.9761	37.91	N/A
	DragAnything	0.9827	0.9488	0.9402	0.9761	39.76	N/A
	Ours	**0.9845**	**0.9509**	**0.9417**	**0.9833**	**22.37**	N/A
Simulator	LeviTor	0.9711	0.9394	0.9381	0.9819	33.52	0.0
	Tora	0.9844	0.9441	0.9452	0.9803	35.44	36.8
	MotionCtrl	0.9832	0.9387	0.9392	0.9813	29.83	29.6
	DragAnything	0.9811	0.9424	0.9433	0.9811	30.27	31.2
	Ours	**0.9865**	**0.9473**	**0.9495**	**0.9821**	**19.88**	**65.2**