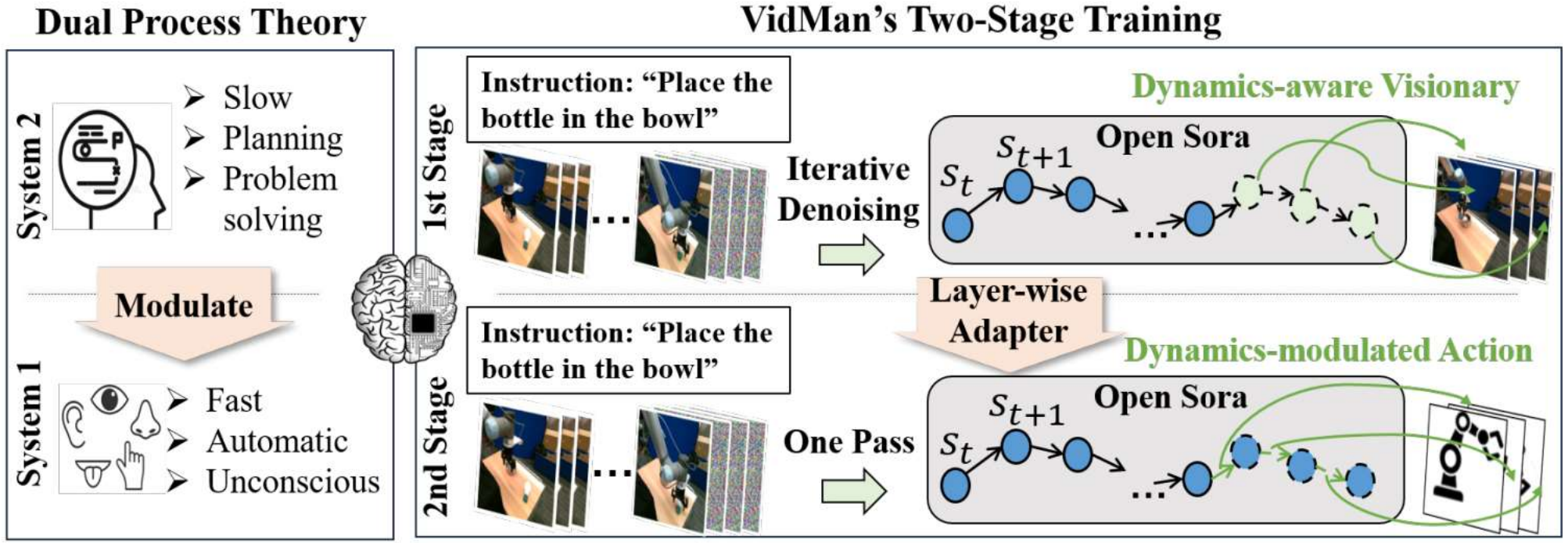
VidMan: Exploiting Intrinsic Dynamics from Video Diffusion Model for Effective Robot Manipulation

Youpeng Wen*, Junfan Lin*, Yi Zhu, Jianhua Han,
Hang Xu, Shen Zhao, Xiaodan Liang

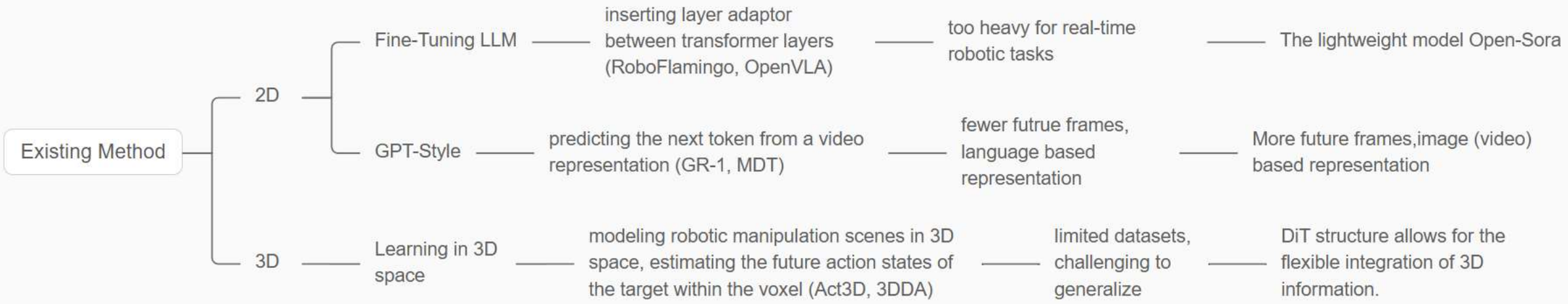
VidMan: Exploiting Intrinsic Dynamics from Video Diffusion Model for Effective Robot Manipulation



Using the state-of-the-art video diffusion architecture Open-Sora, we achieved video prediction **pretraining on a large-scale embodied dataset (OXE)**. It was then fine-tuned on **downstream instruction-action datasets**, enabling applications in visual instruction action prediction, future action prediction, and other downstream tasks.

VidMan: Exploiting Intrinsic Dynamics from Video Diffusion Model for Effective Robot Manipulation

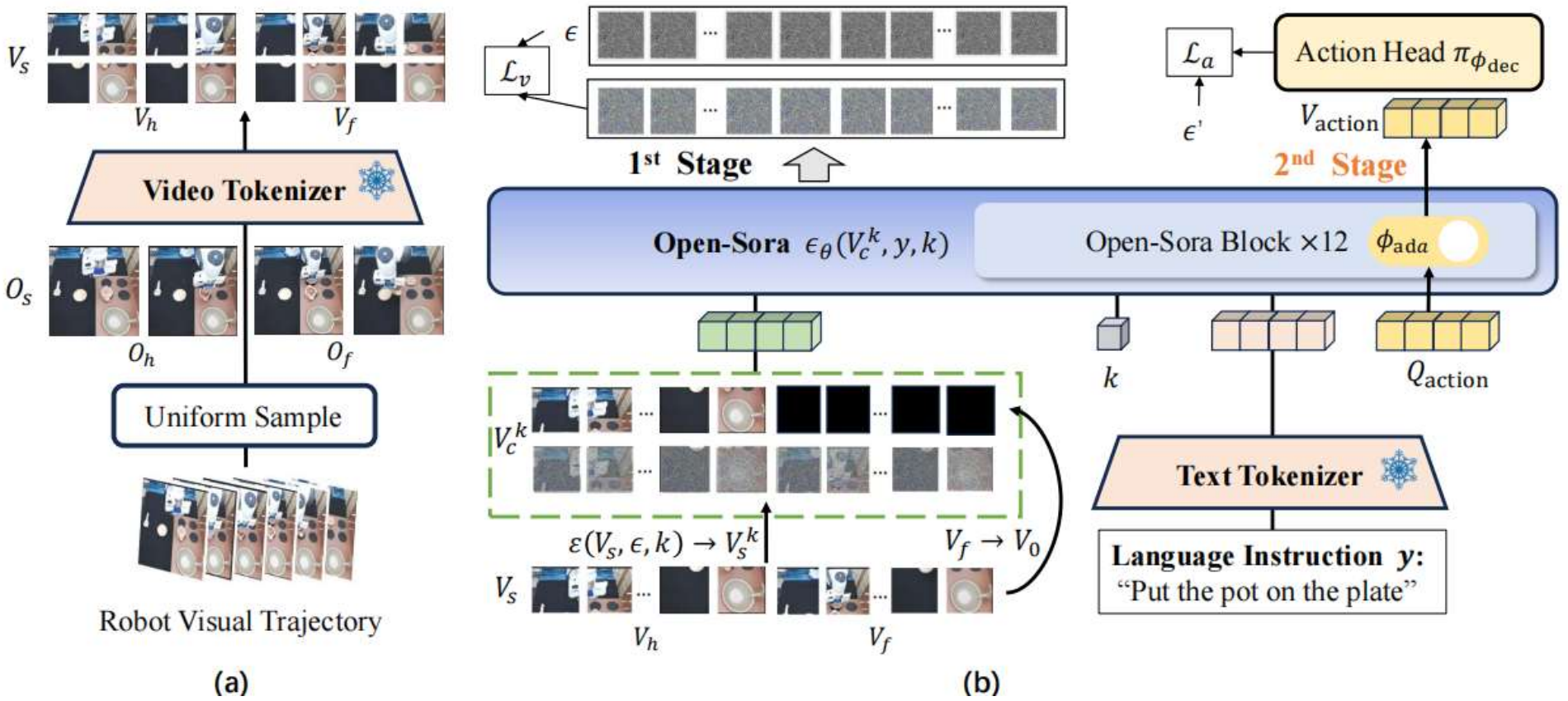
Background and Motivation



We summarized useful structures and key designs from previous methods, such as layer-wise adaptors in LLMs, more future frames, and scalable datasets.

VidMan: Exploiting Intrinsic Dynamics from Video Diffusion Model for Effective Robot Manipulation

Overall Architecture



VidMan: Exploiting Intrinsic Dynamics from Video Diffusion Model for Effective Robot Manipulation

NEURAL INFORMATION PROCESSING SYSTEMS

Results: Performance on CALVIN Benchmark

Table 1: **Zero-shot long-horizon evaluation on CALVIN.** *All* denotes that the model is trained on the entire dataset, including visual data without language annotations, while *Lang* refers to training on only the language-labeled data.. Our method outperforms the hierarchical 2D policies (MCIL [31], HULC [32] and SuSIE [33]) and large-scale 2D transformer-based policies (RT-1 [47] RoboFlamingo [26] and GR-1 [9]), while also remaining competitive compared to 3D-based policies (3D Diffusion Policy [34] and 3D Diffuser Actor [35]).

Method	Training Data	Tasks completed in a row					
		1	2	3	4	5	Avg. Len.
3D Diffusion Policy [34]	Lang	28.7	2.7	0	0	0	0.31
MCIL [31]	All	30.4	1.3	0.2	0	0	0.31
HULC [32]	All	41.8	16.5	5.7	1.9	1.1	0.67
RT-1 [47]	Lang	53.3	22.2	9.4	3.8	1.3	0.9
RoboFlamingo [26]	Lang	82.4	61.9	46.6	33.1	23.5	2.48
SuSIE [33]	All	87	69	49	38	26	2.69
GR-1 [9]	Lang	85.4	71.2	59.6	49.7	40.1	3.06
3D Diffuser Actor [35]	Lang	93.8	80.3	66.2	53.3	41.2	3.35
VidMan (Ours)	Lang	91.5	76.4	68.2	59.2	46.7	3.42

VidMan: Exploiting Intrinsic Dynamics from Video Diffusion Model for Effective Robot Manipulation

Results: Performance on OXE small scale dataset

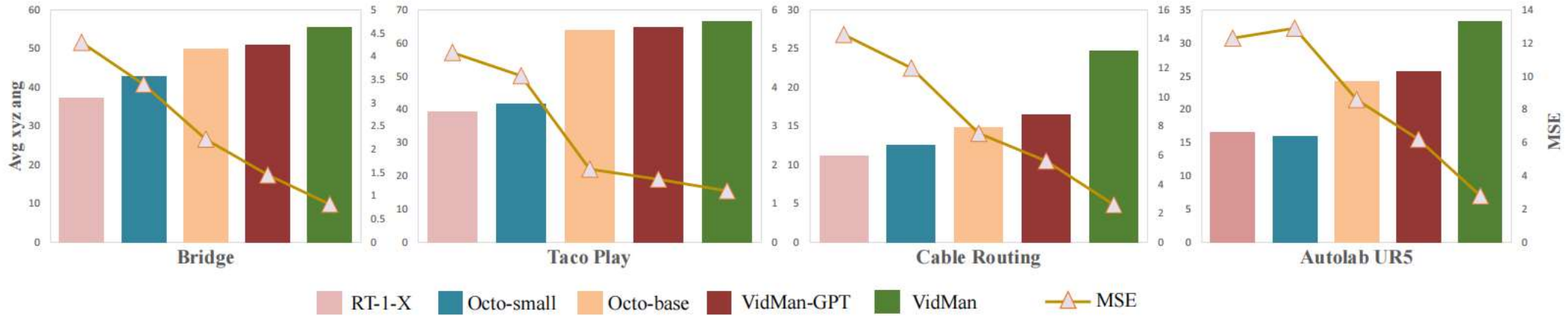


Figure 3: **Offline Performance.** The average accuracy (Avg xyz ang) of xyz accuracy and angle accuracy and MSE correspond to the left and right y-axes of the graph respectively. All models were trained on OXE and validated on offline performance across four datasets. VidMan outperformed Octo-base [7] by 5.6% on Bridge, 2.6% on Taco Play, 9.9% on Cable Routing, and 9.0% on Autolab UR5. Additionally, Our method also shows improvements over the VidMan-GPT approach.

VidMan: Exploiting Intrinsic Dynamics from Video Diffusion Model for Effective Robot Manipulation

NEURAL INFORMATION PROCESSING SYSTEMS

Video Prediction

Move red pepper to above green towel

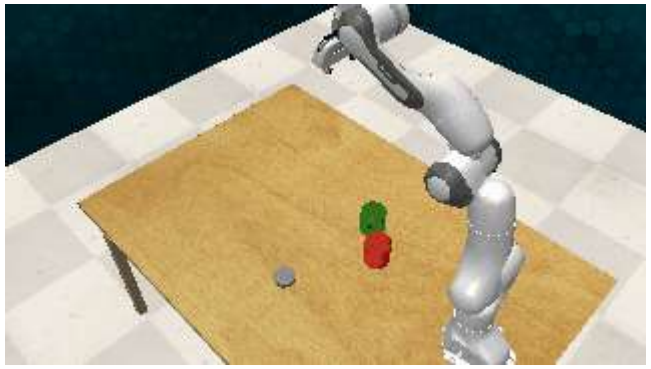
Fold the cloth from top left to bottom right

Put carrot in pot cardboard fence

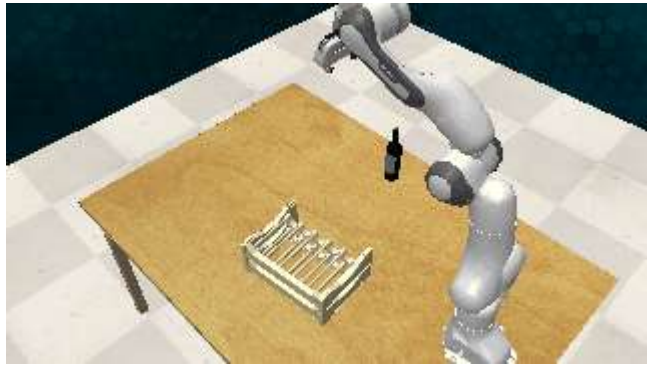
Open the silver pot

Unfold the cloth from top left to bottom right

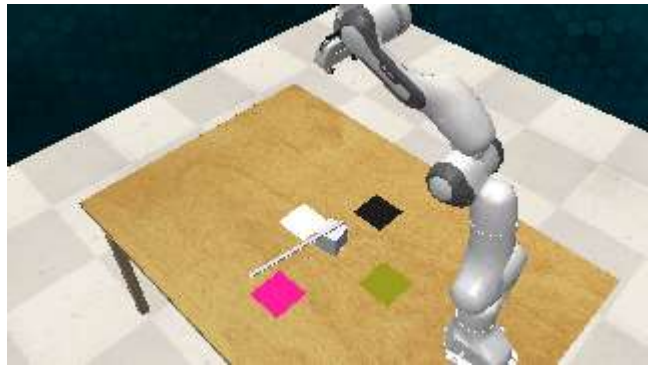
Simulation



Close jar



Place wine at rack location

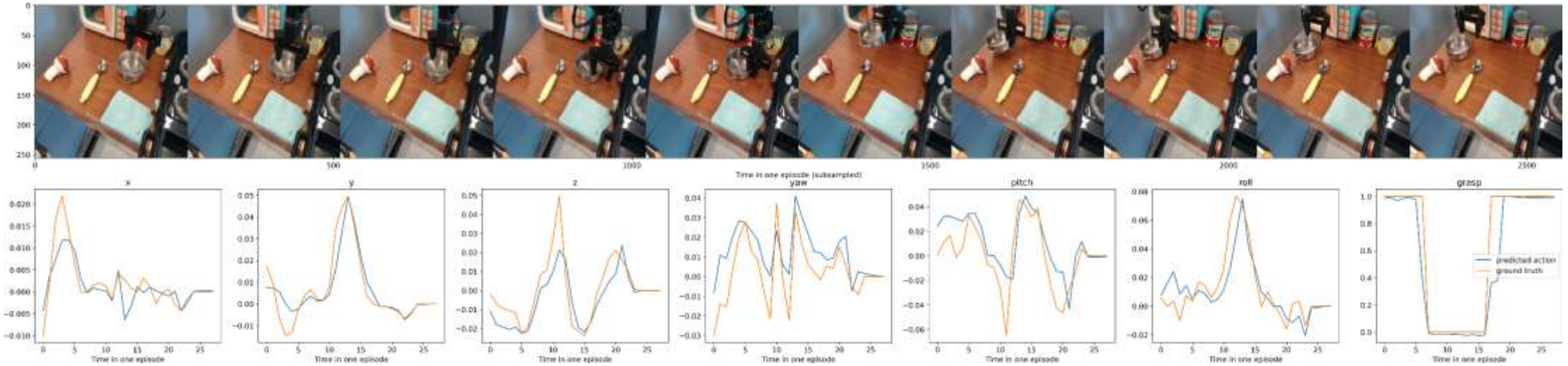


Reach and drag

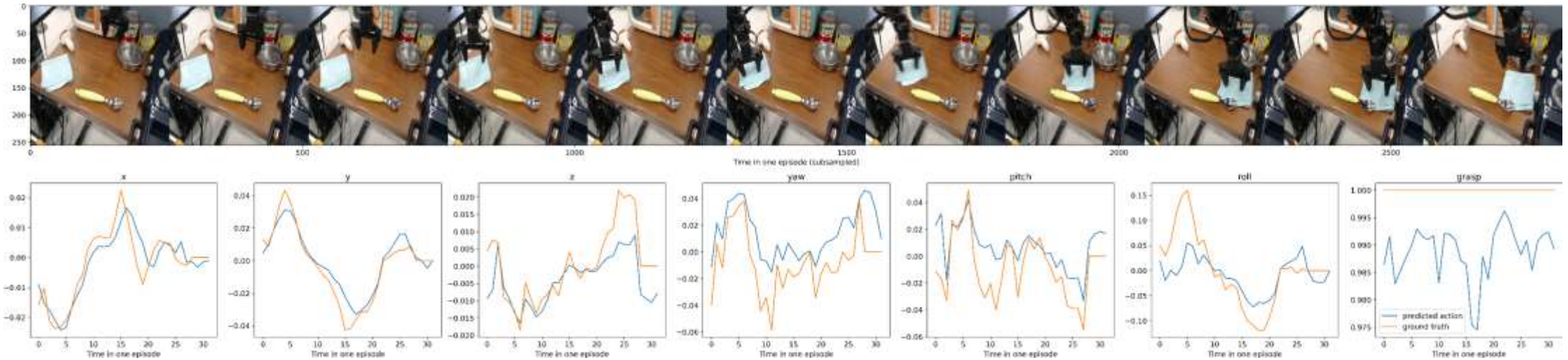
VidMan: Exploiting Intrinsic Dynamics from Video Diffusion Model for Effective Robot Manipulation

NEURAL INFORMATION PROCESSING SYSTEMS

Action Prediction



Move the pan to the front of the microwave.



Move the rag to the other side of the ice cream.


SUN YAT-SEN UNIVERSITY


JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Thank you!