SViMo: Synchronized Diffusion for Video and Motion Generation in Hand-object Interaction Scenarios

Lingwei Dang [1*], Ruizhi Shao [2*], Hongwen Zhang [3], Wei MIN [4], Yebin Liu [2], Qingyao Wu [†1]

[1] South China University of Technology, [2] Tsinghua University, [3] Beijing Normal University, [4] Shadow AI

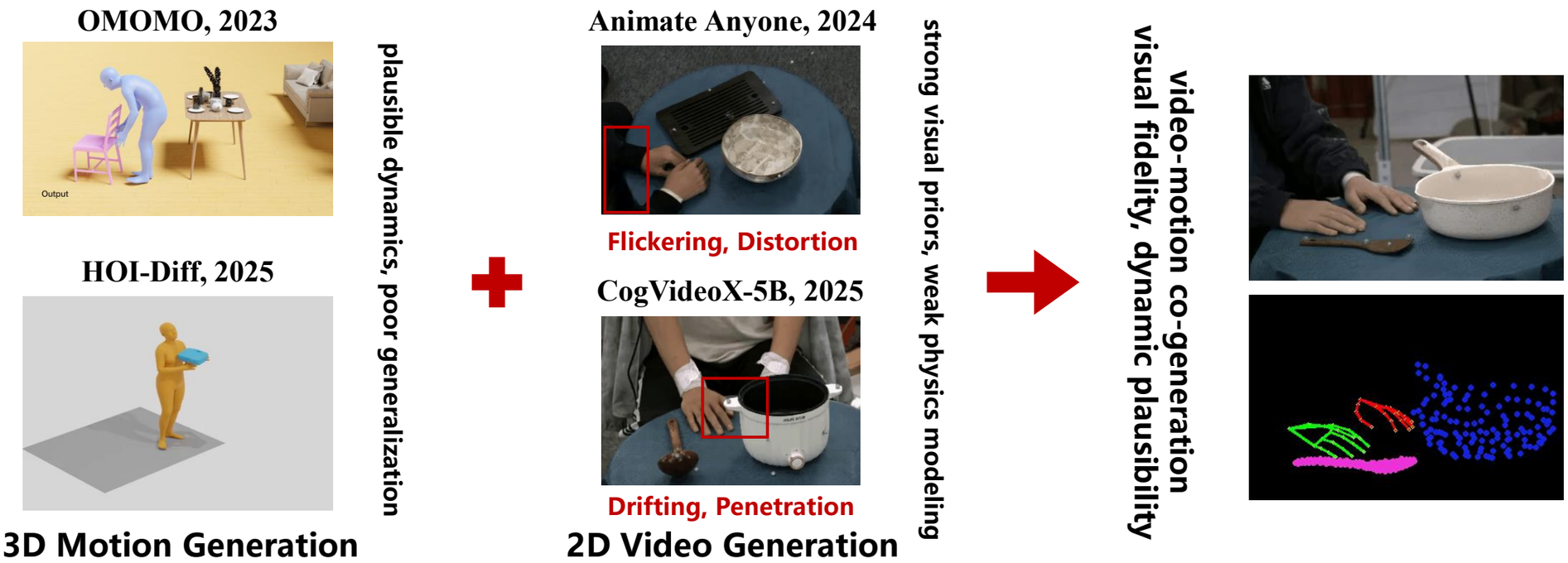
NeurIPS 2025 Spotlight

[*] Equal contributions.

[†] Corresponding Author. Email: qyw@scut.edu.cn.

1. Background & Motivation

- 3D motion generation models produce plausible dynamics but suffer from limited data and **poor generalization**
- 2D Video generation models possess rich visual priors yet **lack awareness of physics**.

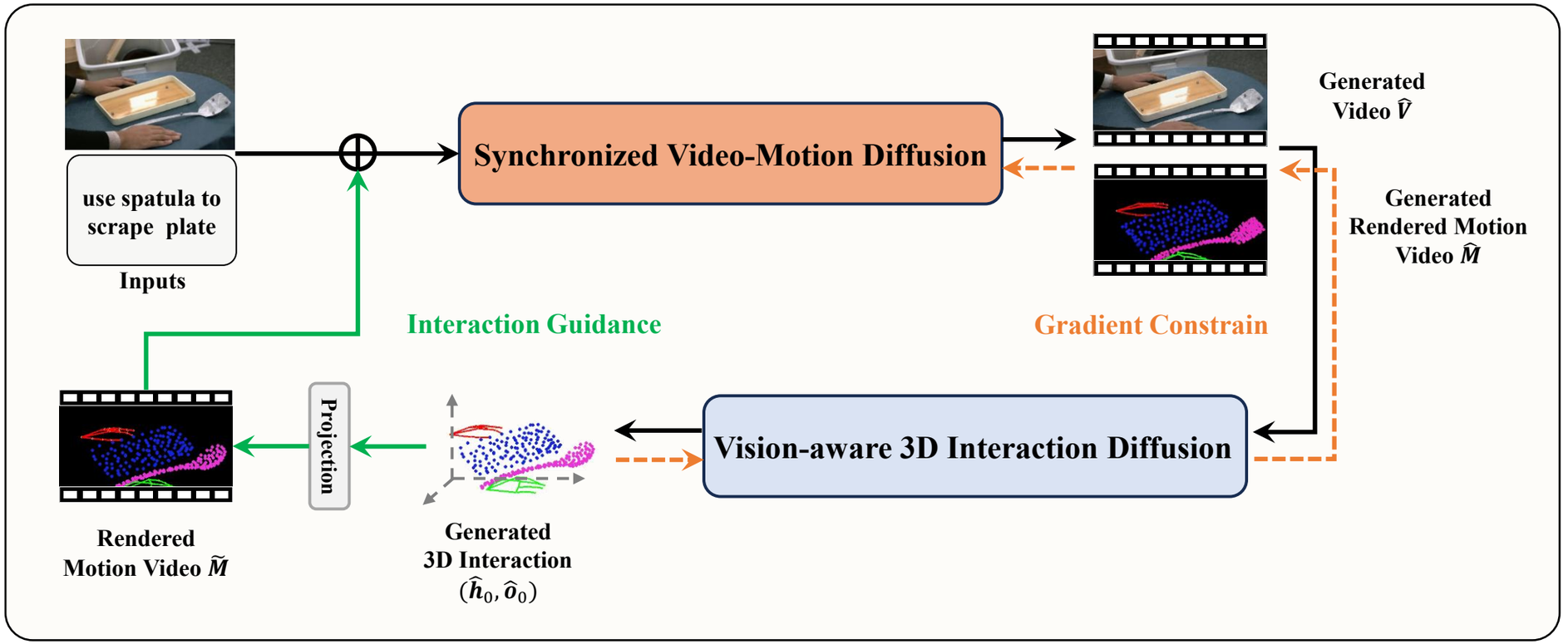


Object Motion Guided Human Motion Synthesis, SIGGRAPH Asia 2023; HOI-Diff: Text-Driven Synthesis of 3D Human-Object Interactions using Diffusion Models, CVPR 2025 Workshop; Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation, CVPR 2024; Cogvideox: Text-to-video diffusion models with an expert transformer, ICLR 2025.

Visual appearance and motion dynamics share the same physical laws. We propose to unify visual priors and kinematic constraints through synchronized video-motion co-generation.

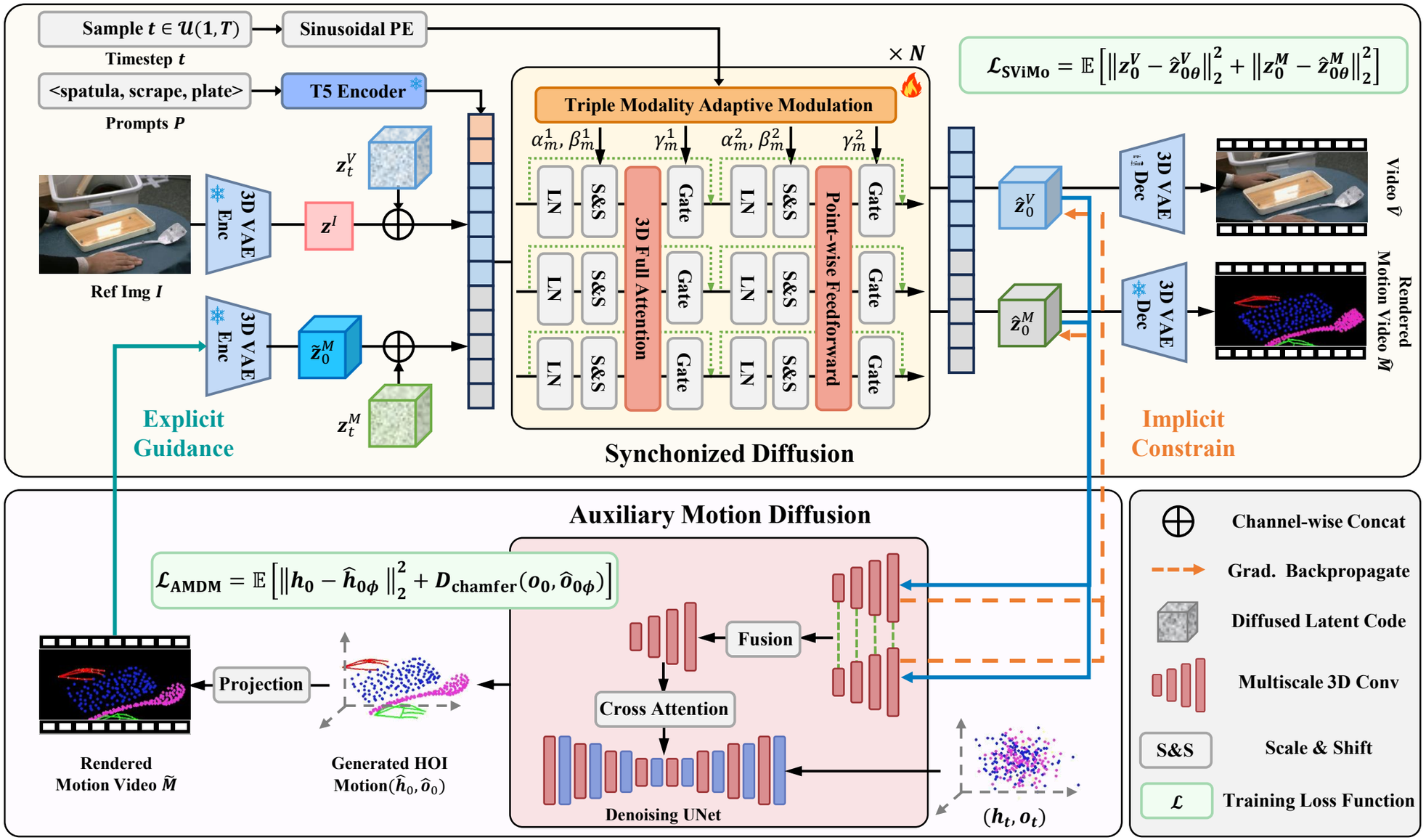
2. Method

SViMo: A synchronized diffusion model for HOI video and motion joint generation



①End-to-end video-motion synthesis; ②Visual realism and dynamic plausibility; ③Generalization ability

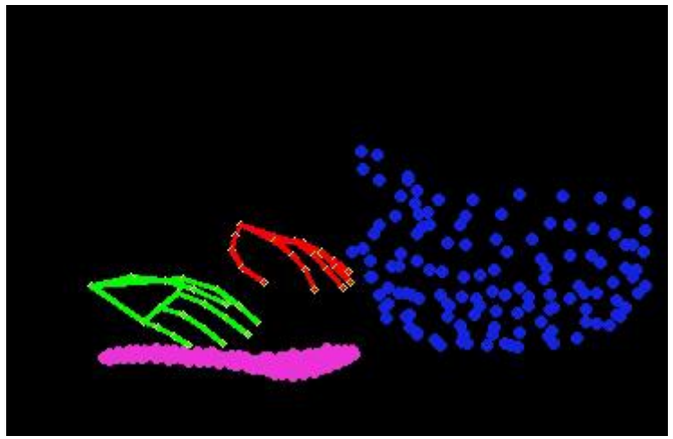
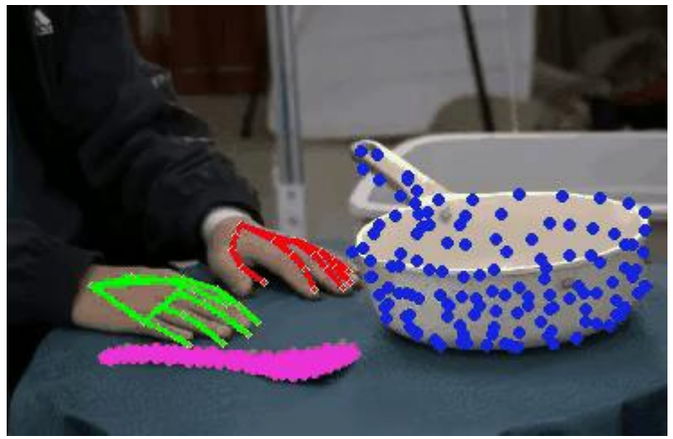
2. Method


Sample $t \in \mathcal{U}(1, T)$
Timestep t
Sinusoidal PE
<spatula, scrape, plate>
Prompts P
T5 Encoder
Ref Img I
3D VAE Enc
z_t^V
z^I
$\tilde{z}_0^M$
z_t^M
Explicit Guidance
$\times N$
Triple Modality Adaptive Modulation
α_m^1, β_m^1
γ_m^1
α_m^2, β_m^2
γ_m^2
LN
S&S
3D Full Attention
Gate
Point-wise Feedforward
Synchonized Diffusion
$\mathcal{L}_{\text{SViMo}} = \mathbb{E}\left[\left\|z_0^V - \hat{z}_{0\theta}^V\right\|_2^2 + \left\|z_0^M - \hat{z}_{0\theta}^M\right\|_2^2\right]$
$\hat{z}_0^V$
$\hat{z}_0^M$
3D VAE Dec
Video $\tilde{V}$
Rendered Motion Video $\tilde{M}$
Implicit Constrain
Auxiliary Motion Diffusion
$\mathcal{L}_{\text{AMDM}} = \mathbb{E}\left[\left\|h_0 - \hat{h}_{0\phi}\right\|_2^2 + D_{\text{chamfer}}(o_0, \hat{o}_{0\phi})\right]$
Fusion
Cross Attention
Denoising UNet
(h_t, o_t)
Projection
Rendered Motion Video $\tilde{M}$
Generated HOI Motion$(\hat{h}_0, \hat{o}_0)$
Channel-wise Concat
Grad. Backpropagate
Diffused Latent Code
Multiscale 3D Conv
S&S
Scale & Shift
$\mathcal{L}$
Training Loss Function

3. Demonstrations of Our Method: Case 1

NEURAL INFORMATION PROCESSING SYSTEMS

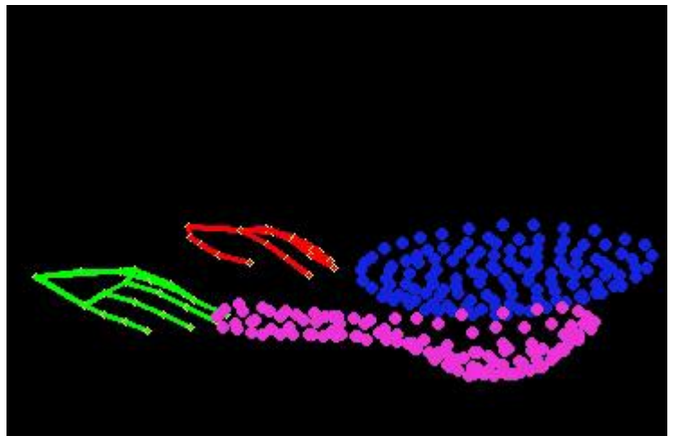
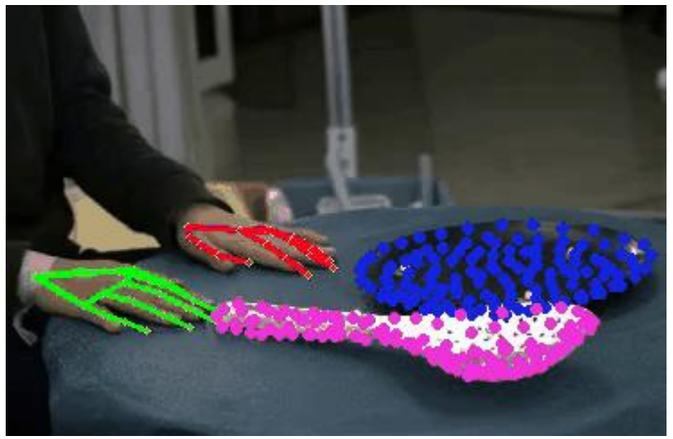
use spatula to scrape off pan

Generated Video	Generated Motion	Overlayed Results
		
Realistic Video	**Plausible Motion**	**Consistent Video and Motion**

3. Demonstrations of Our Method: Case 2

NEURAL INFORMATION PROCESSING SYSTEMS

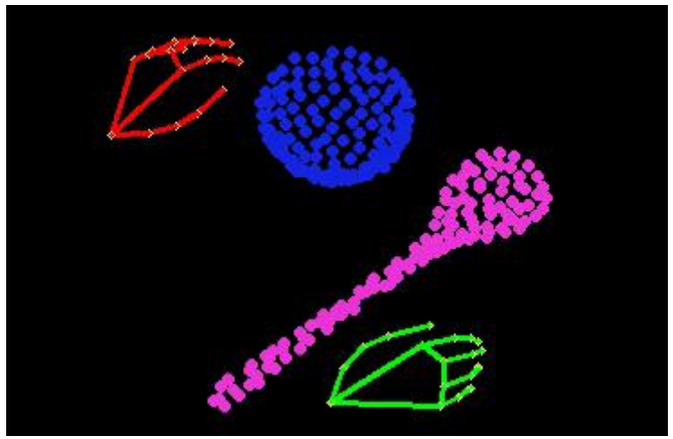
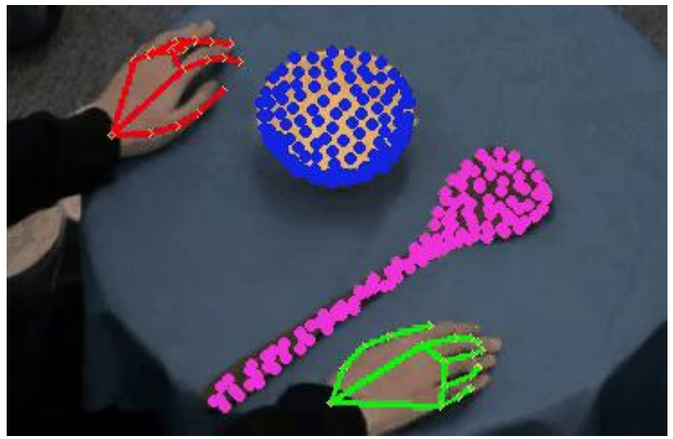
use spoon to scrape off plate

Generated Video	Generated Motion	Overlayed Results
		
Realistic Video	**Plausible Motion**	**Consistent Video and Motion**

3. Demonstrations of Our Method: Case 3

NEURAL INFORMATION PROCESSING SYSTEMS

use spatula to put out bowl

Generated Video	Generated Motion	Overlayed Results
		
Realistic Video	**Plausible Motion**	**Consistent Video and Motion**

4. Comparison of Videos: Case 1

NEURAL INFORMATION PROCESSING SYSTEMS

use bowl to put in plate

Hunyuan-13B-Zeroshot

Low-dynamic, Hallucination

Animate Anyone

Flickering, Distortion

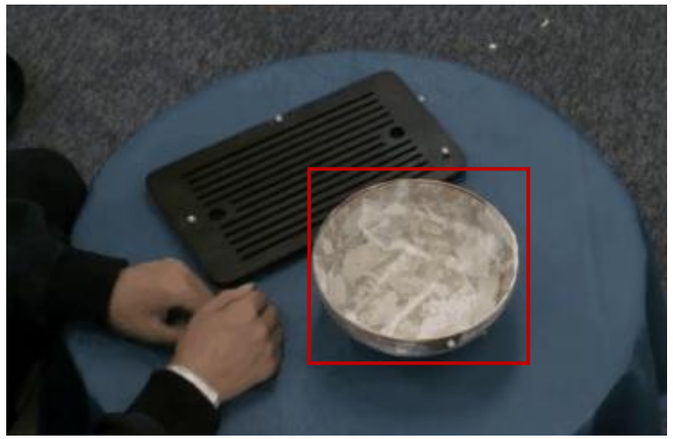
CogVideoX-5B

Implausible movements

Wan-14B-Zeroshot

Hallucination, Implausible movements

Easy Animate



Flickering, Object inconsistency

Ours

4. Comparison of Videos: Case 2

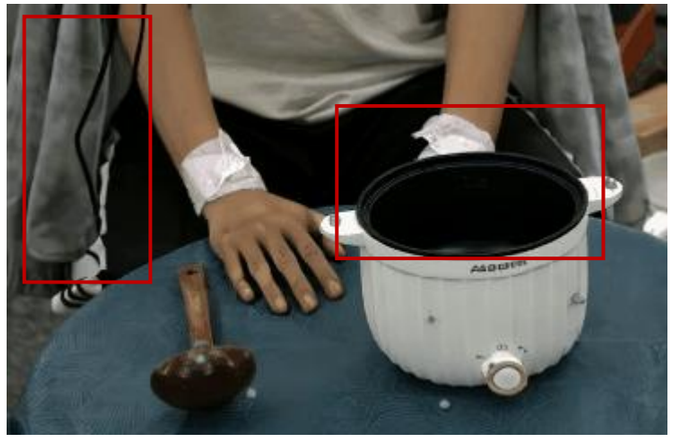
NEURAL INFORMATION PROCESSING SYSTEMS

use spoon to scrape off pan

Hunyuan-13B-Zeroshot

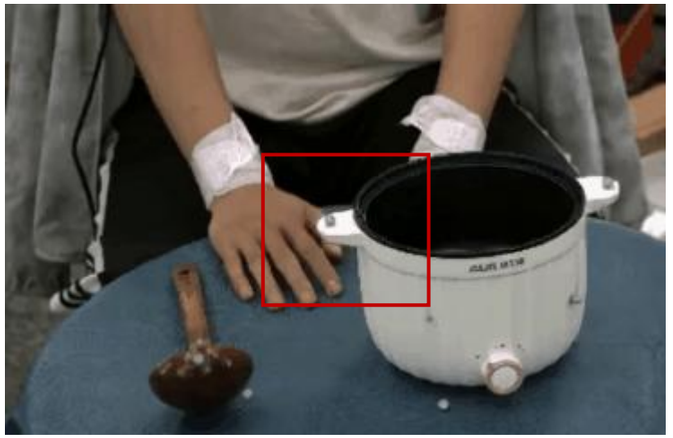
Low-dynamic

Animate Anyone



Flickering, Distortion

CogVideoX-5B



Implausible movements, Penetration

Wan-14B-Zeroshot

Hallucination

Easy Animate

Flickering, Implausible movements

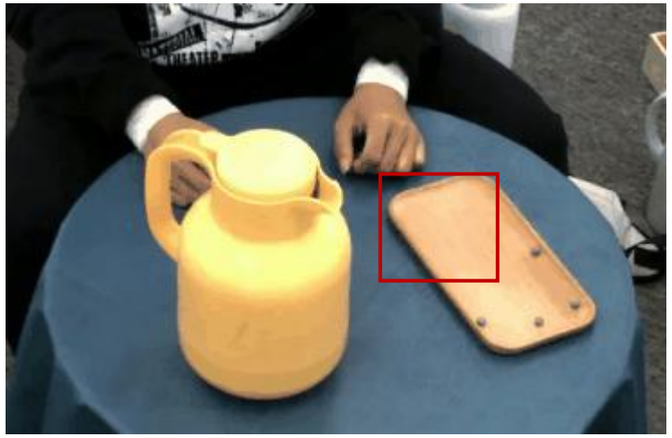
Ours

4. Comparison of Videos: Case 3

NEURAL INFORMATION PROCESSING SYSTEMS

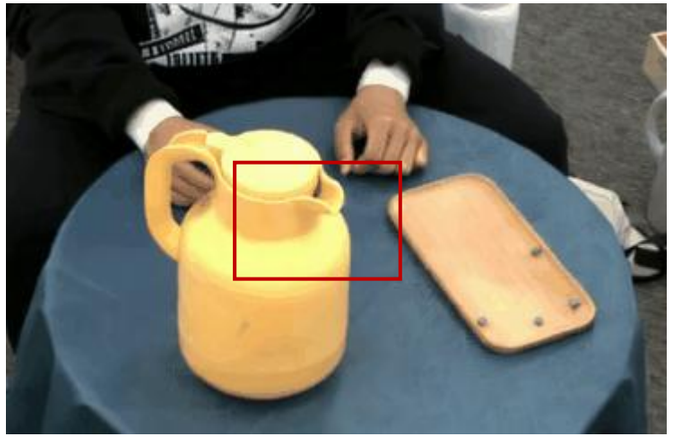
use kettle to pour in plate

Hunyuan-13B-Zeroshot



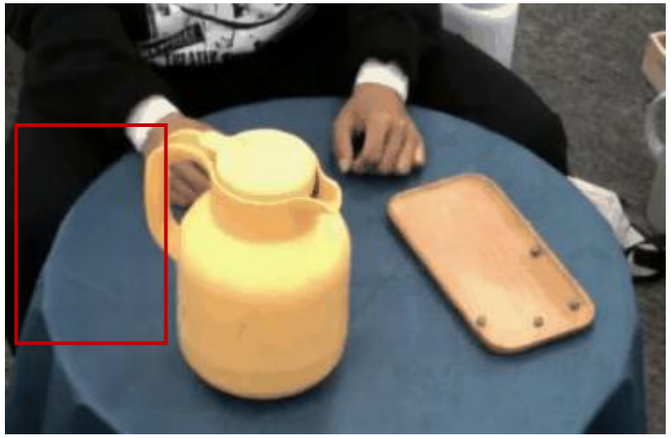
Low-dynamic, Hallucination

Animate Anyone



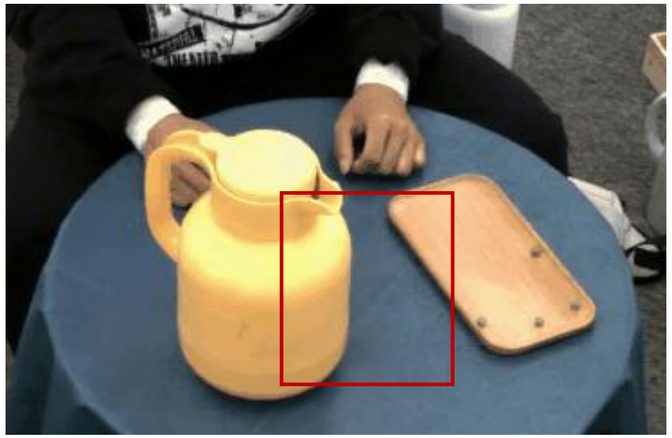
Flickering, Distortion

CogVideoX-5B



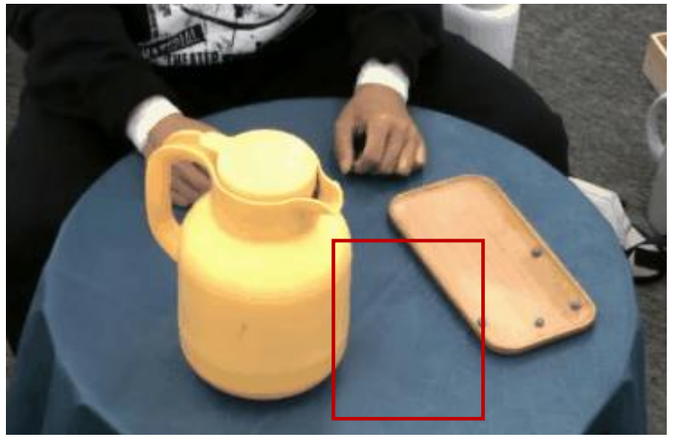
Object inconsistency

Wan-14B-Zeroshot



Hallucination, Camera shake

Easy Animate



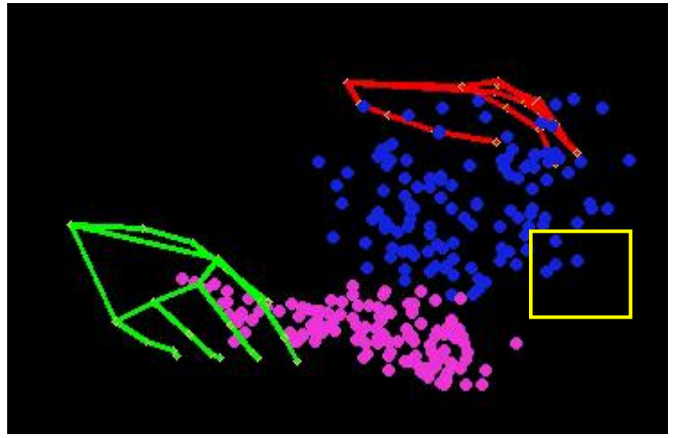
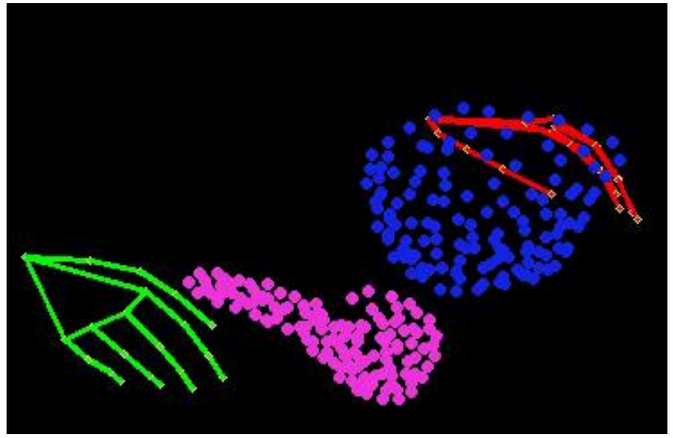
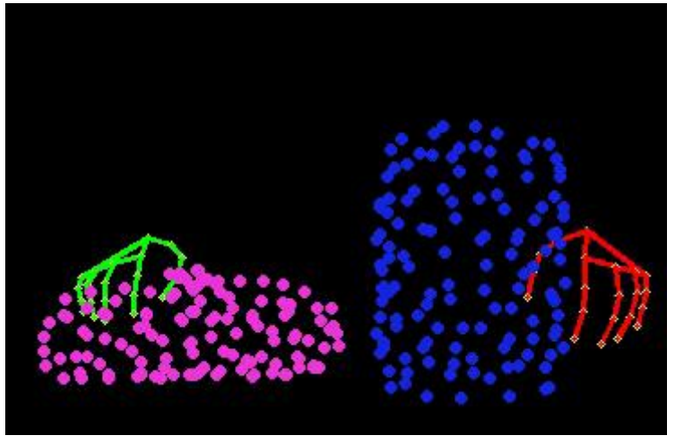
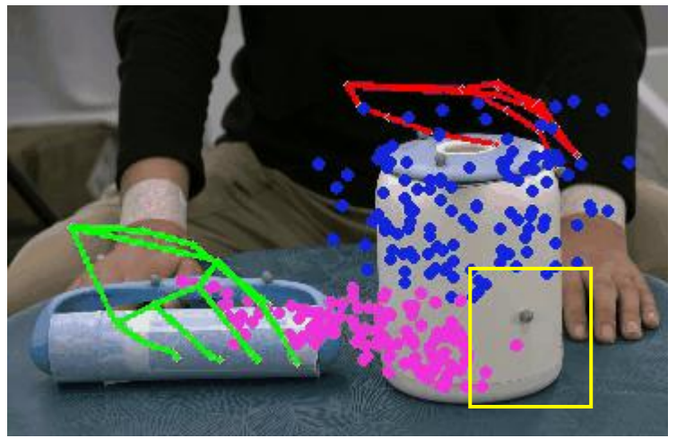
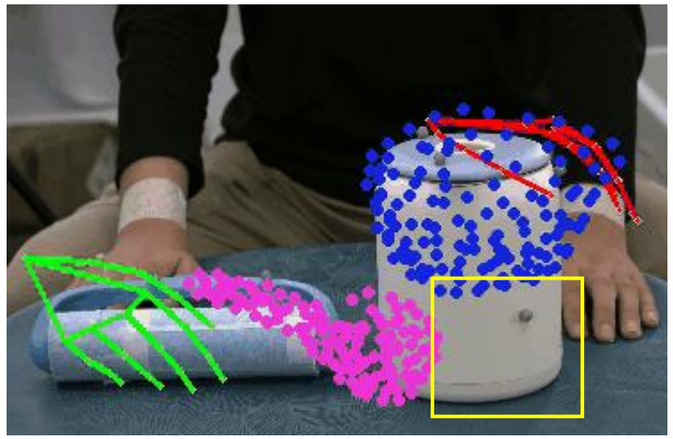
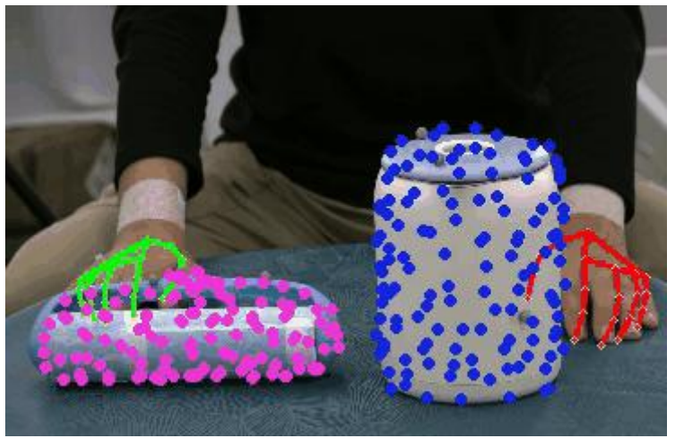
Hallucination

Ours

5. Comparison of Motions: Case 1

NEURAL INFORMATION PROCESSING SYSTEMS

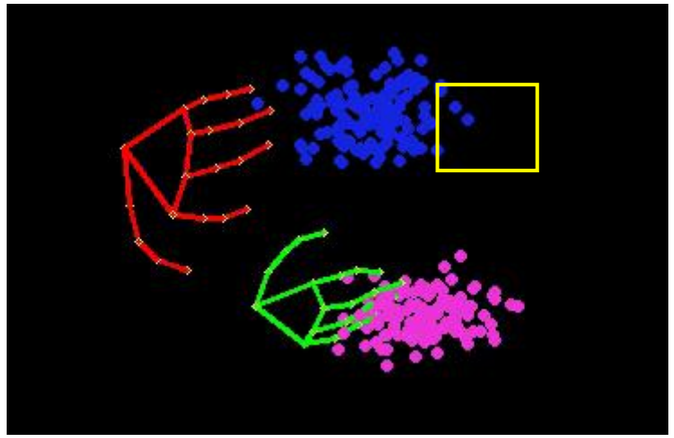
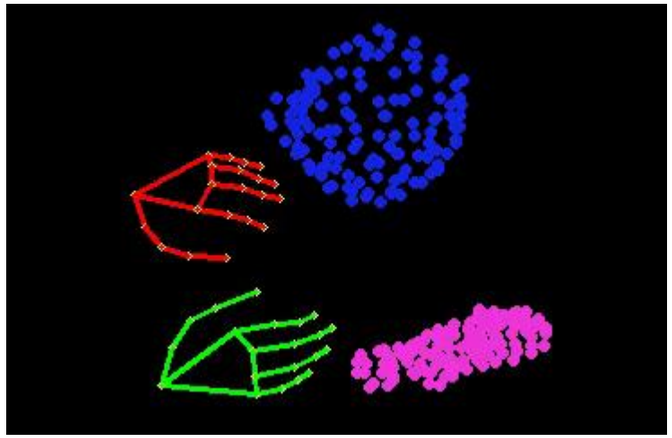
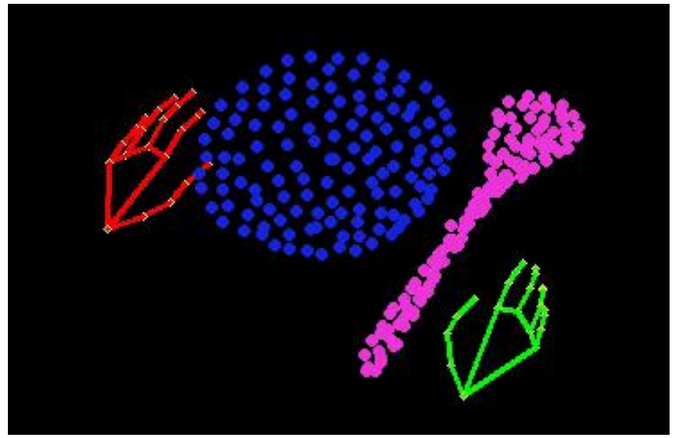
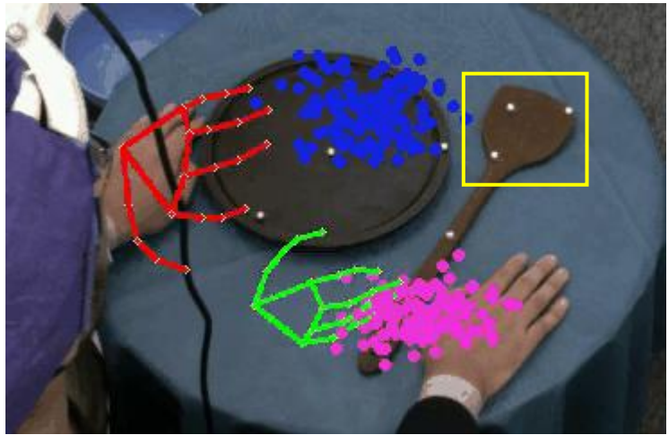
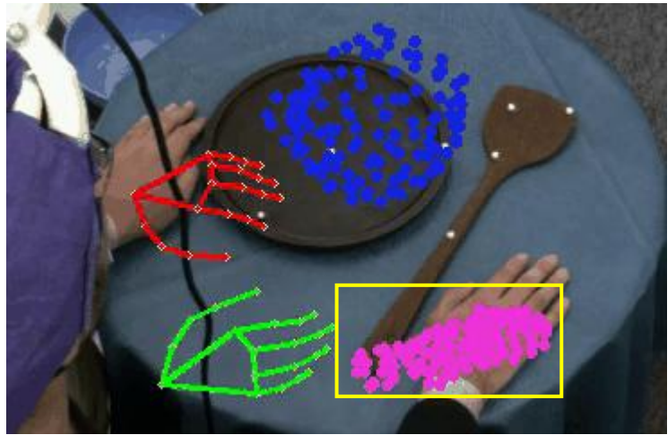
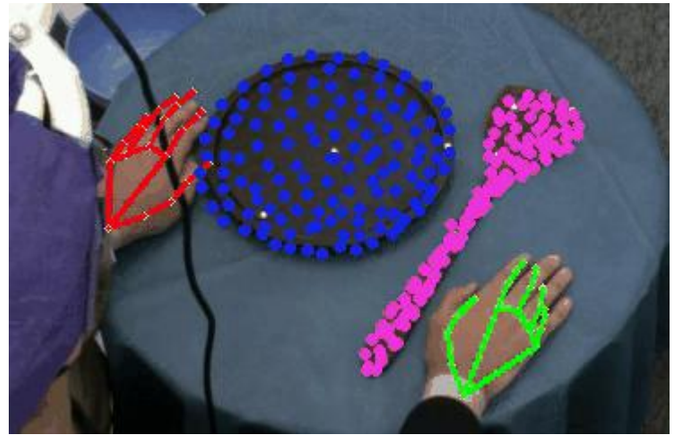
use roller to dust kettle

	MDM	EMDM	Ours
Generated Motion			
Overlayed Motion onto the Image			
	Blurred outline, Low dynamic	Image-motion inconsistency	

5. Comparison of Motions: Case 2

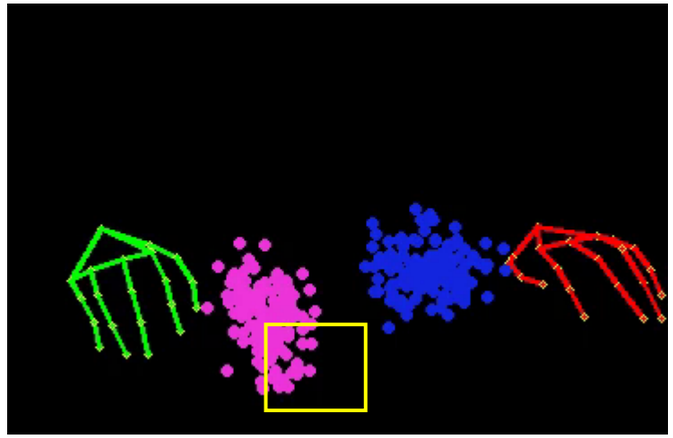
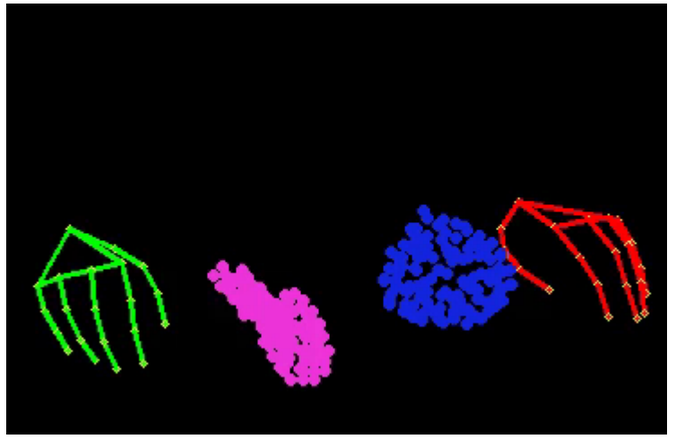
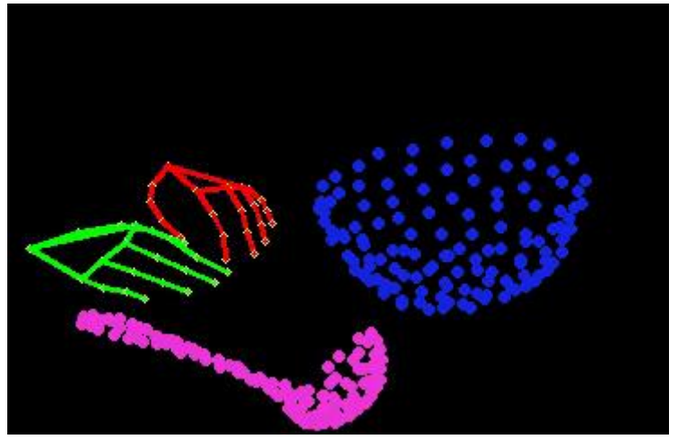
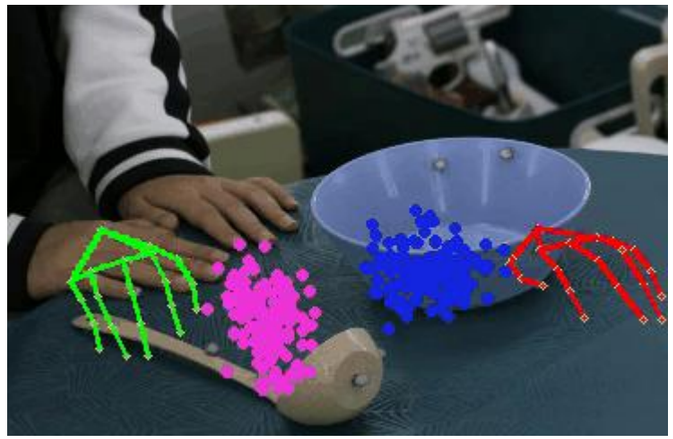
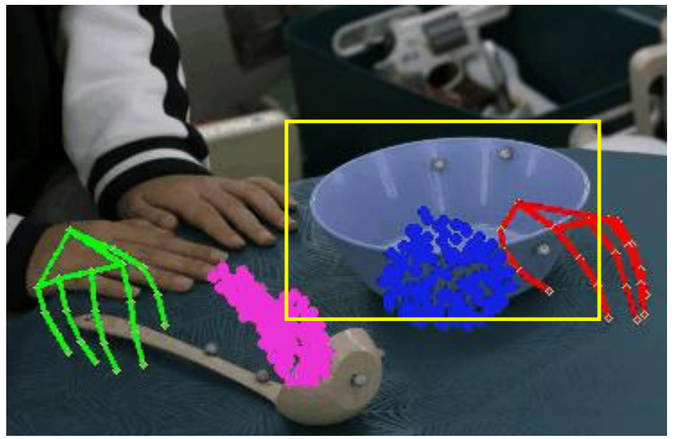
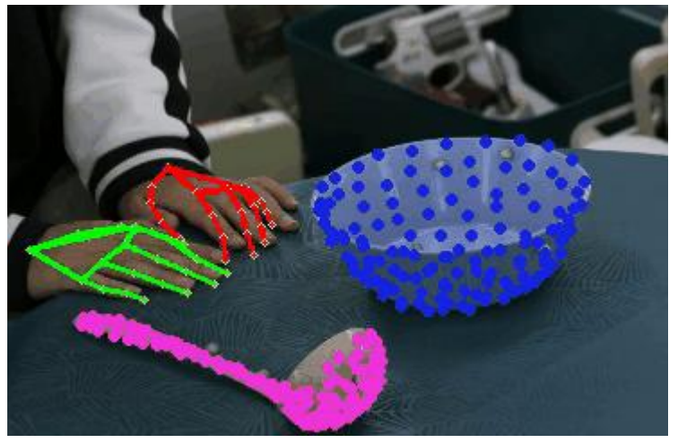
NEURAL INFORMATION PROCESSING SYSTEMS

use spatula to stir plate

	MDM	EMDM	Ours
Generated Motion			
Overlayed Motion onto the Image			
	Blurred outline, Implausible mov.	Image-motion inconsistency	

5. Comparison of Motions: Case 3

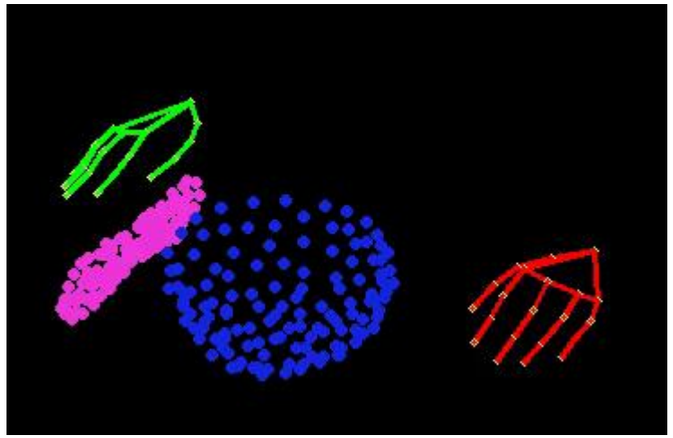
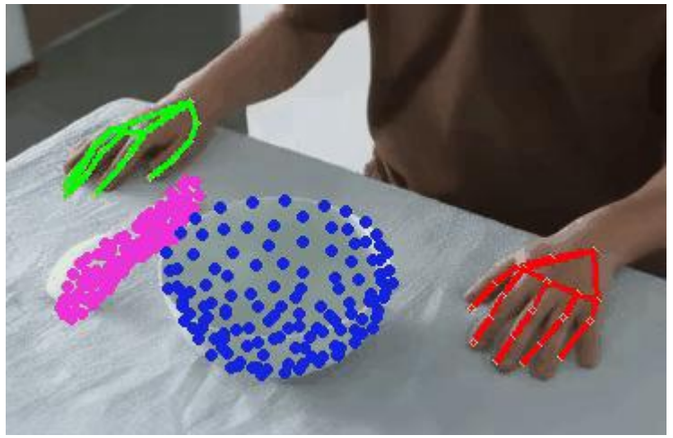
use spoon to put in bowl

	MDM	EMDM	Ours
Generated Motion			
Overlayed Motion onto the Image			
	Blurred outline, Low dynamic	Image-motion inconsistency	

6. Generalization on the Real-World Data

NEURAL INFORMATION PROCESSING SYSTEMS

use the spoon to scrape the bowl

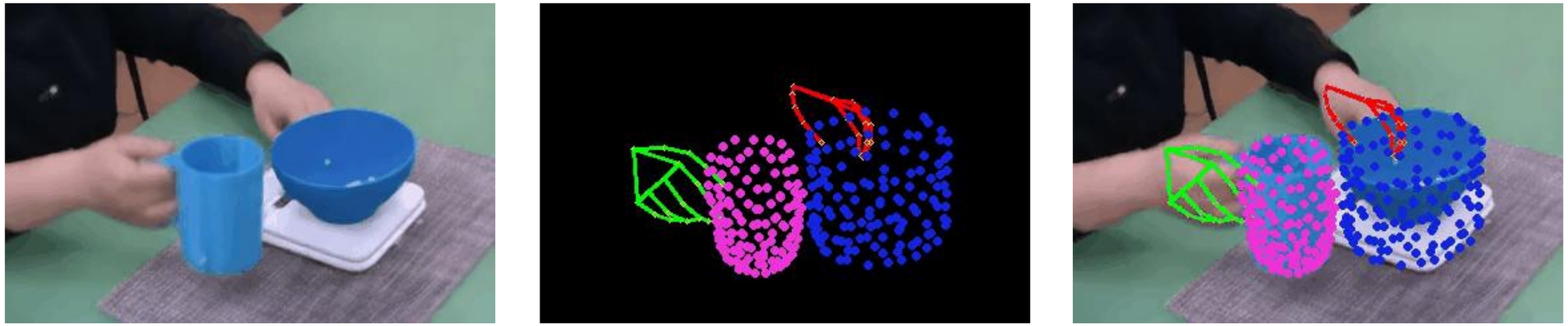
Generated Video	Generated Motion	Overlayed Results
		

6. Generalization on the OAKINK2 Data

NEURAL INFORMATION PROCESSING SYSTEMS

use the cup to pour into the bowl

Generated Video	Generated Motion	Overlayed Results



Xinyu, Zhan, et al. "Oakink2: A dataset of bimanual hands-object manipulation in complex task completion." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

7. QR Code for Our Project

https://droliven.github.io/SViMo_project/

Thank you!

Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) No.62125107 and No.62272172.