Target-Aware Video Diffusion Models

ICLR 2026

Taeksoo Kim

Hanbyul Joo

Motivation

- Remarkable developments on video diffusion models
- OpenAI Sora: “Video generation models as world simulators”

“A stylish woman walks down a Tokyo street
filled with warm glowing neon and animated city signage.”

OpenAI. Sora. 2024

Motivation

- Controlling outputs via text is challenging → Image-to-video generation
- Semantically aligned, but undesired outputs

Input image

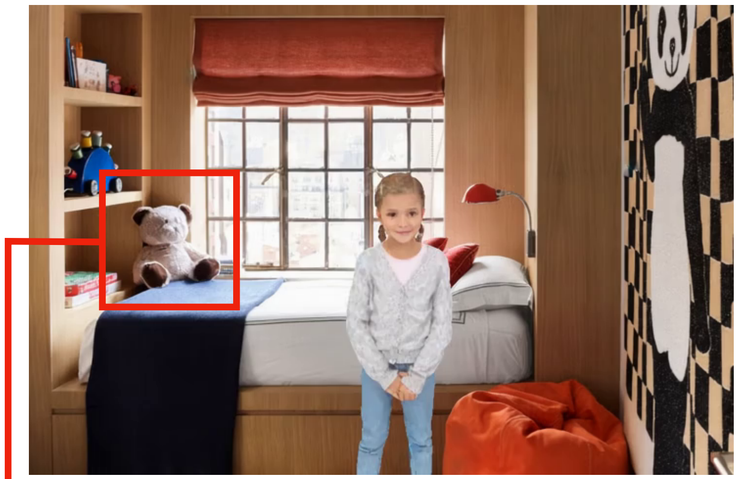
Image-to-video
Diffusion Model

“The girl turns and picks up the teddy bear resting on the bed.”

Input text prompt

Motivation

- Controlling outputs via text is challenging → Image-to-video generation
- Semantically aligned, but undesired outputs



Input image

“The girl turns and picks up the teddy bear resting on the bed.”

Input text prompt

Image-to-video Diffusion Model

Kling 1.6 (Closed-sourced)

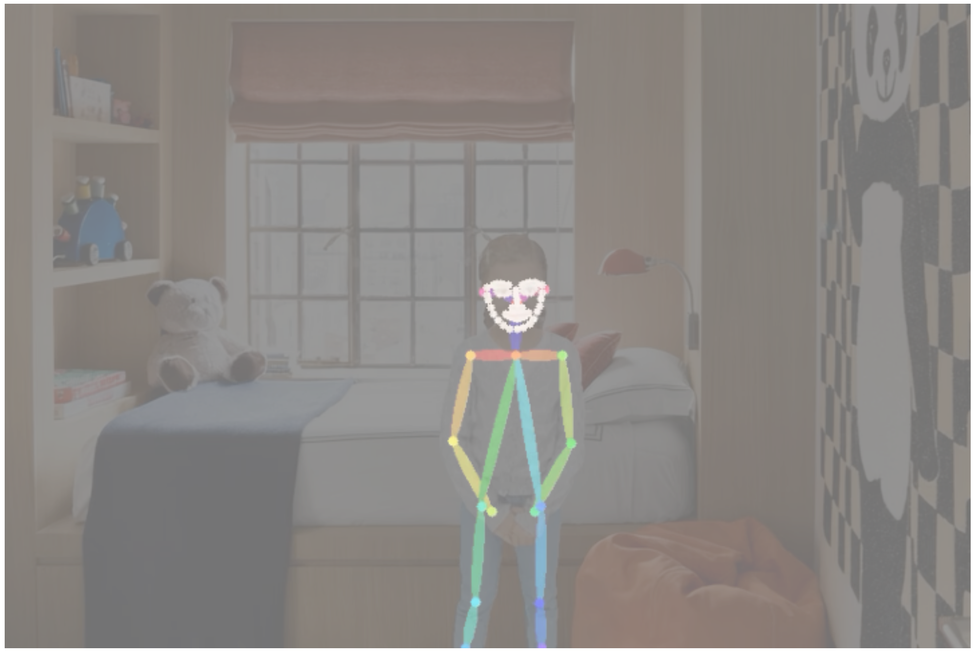
CogVideoX (Open-sourced)

Kuaishou. Kling. 2025

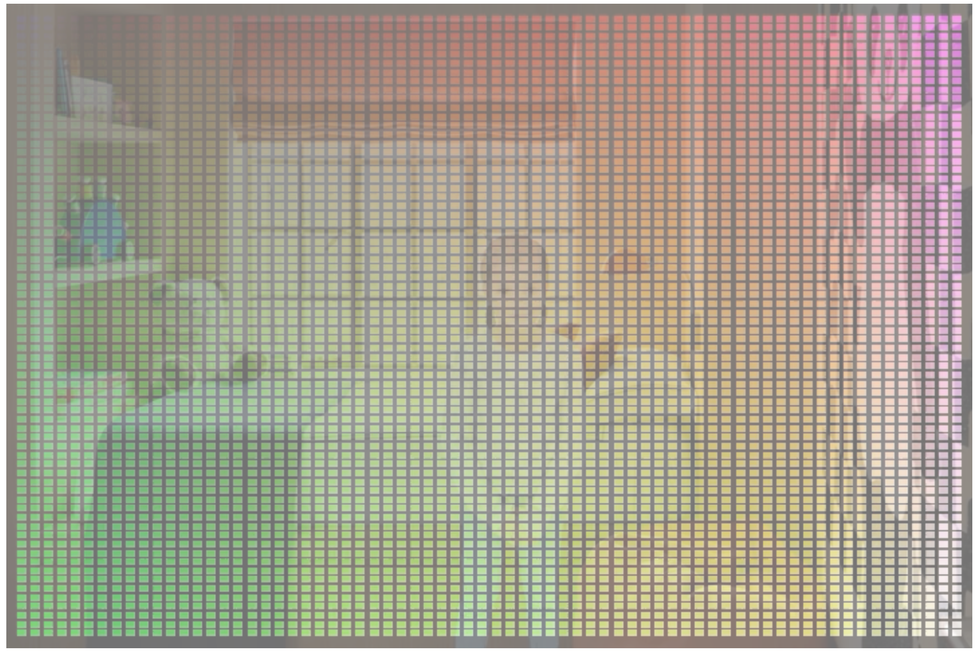
Yang et al. CogVideoX: Text-to-Video Diffusion Models with an Expert Transformer. ICLR 2025

Motivation

- Some use dense structural or motion cues for control → Usually inaccessible
- Especially challenging for motions with complex interactions



OpenPose / DWPose conditions



3D tracking conditions

Cao et al. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. TPAMI 2019
Xiao et al. SpatialTracker: Tracking Any 2D Pixels in 3D Space. CVPR 2024
Gu et al. Diffusion as Shader: 3D-aware Video Diffusion for Versatile Video Generation Control. arXiv 2025

Target-Aware Video Diffusion Models

Accurate interactions between the actor and the target

- Use segmentation mask to specify the target

Input image and target mask

Target-Aware
VDM

Output video

“The girl turns and picks up the
[TGT] teddy bear resting on the bed.”

Input text prompt

Target-Aware Video Diffusion Models

Accurate interactions between the actor and the target

- Use segmentation mask to specify the target

Input image and target mask

Target-Aware VDM

Output video

“The man turns around and lifts the [TGT] vase filled with vibrant hydrangeas.”

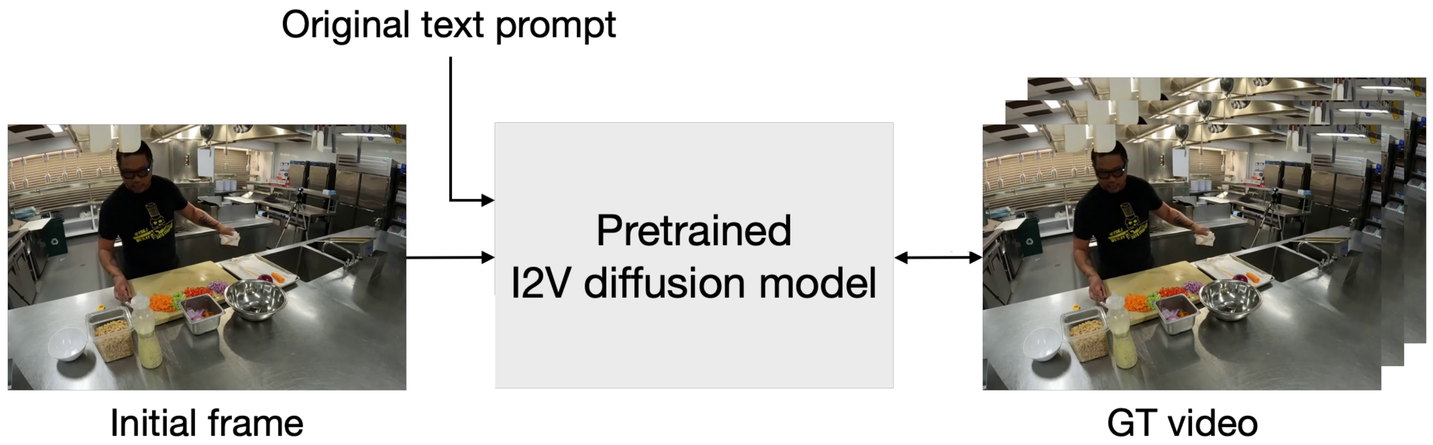
Input text prompt

Method

Method

Target-awareness via cross-attention loss

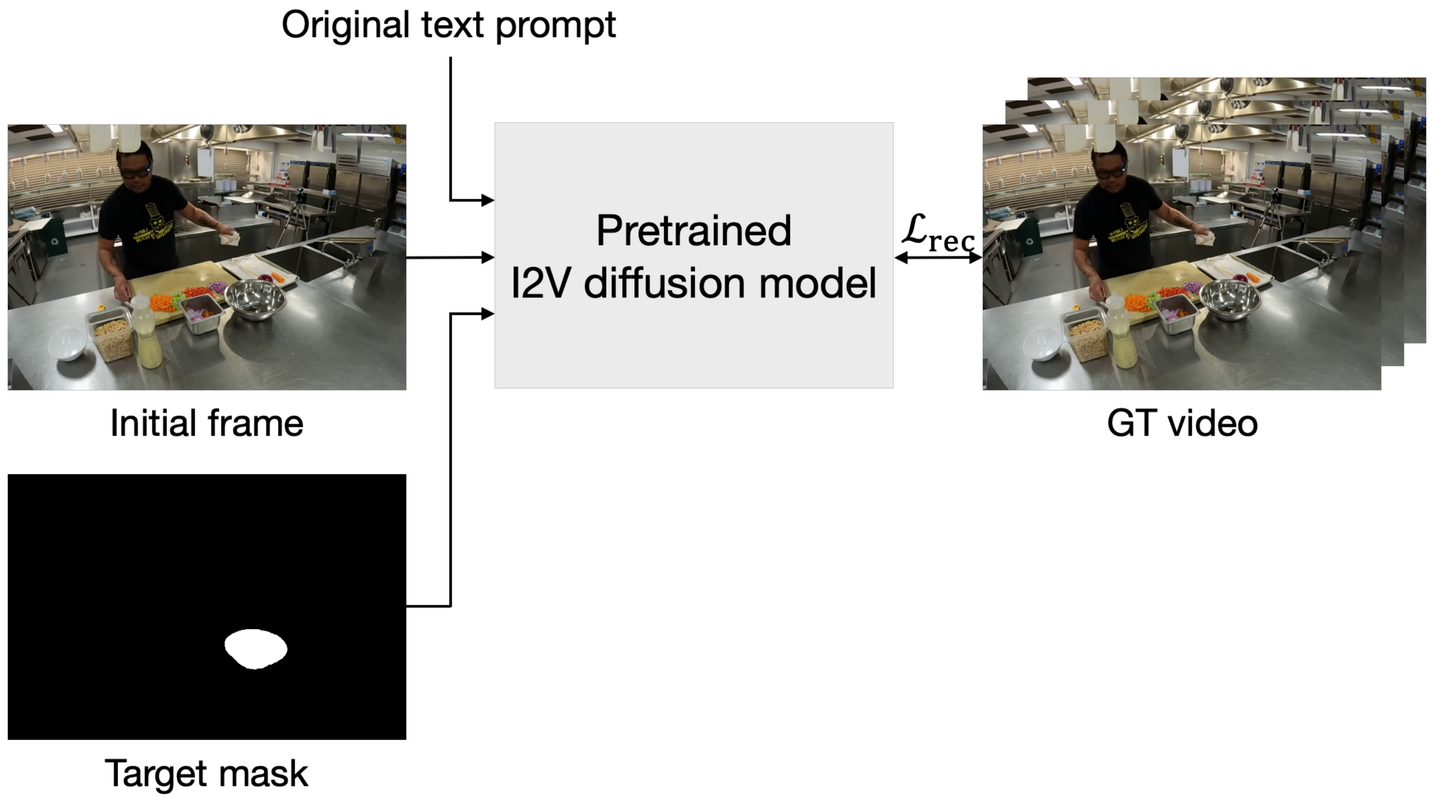
- Base image-to-video pipeline



Method

Target-awareness via cross-attention loss

- Extending the base model to take an additional mask as input

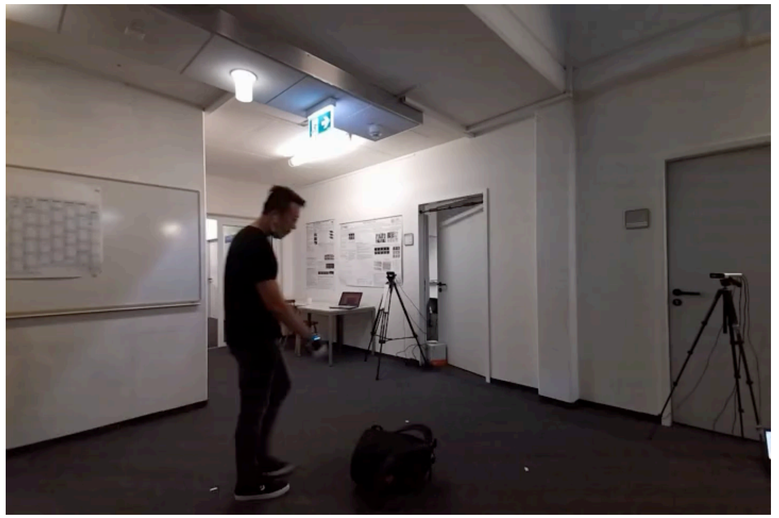


Method

Dataset

- Curate 1290 videos from BEHAVE and Ego-Exo4D datasets
- Criteria:
    1. Initial frame with the actor and the target present, but not yet interacting
    2. Subsequent frames must capture the actor engaging with the target

Training data from EgoExo4D



Training data from BEHAVE

Grauman et al. Ego-Exo4D: Understanding Skilled Human Activity from First. CVPR 2024

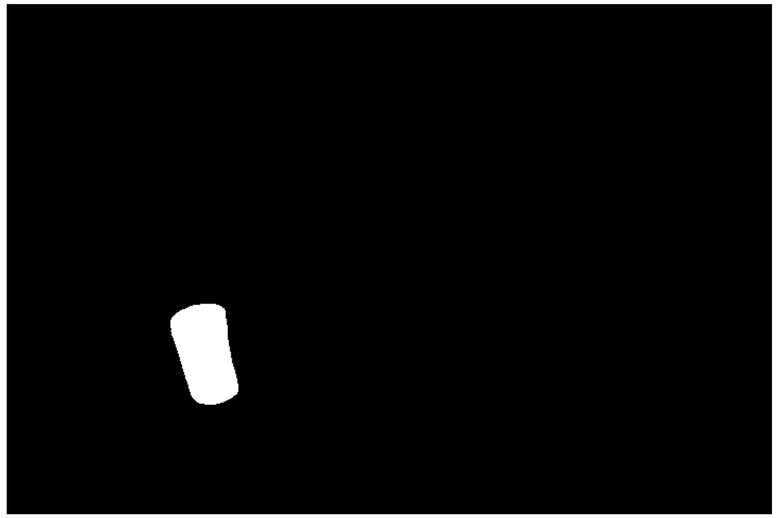
Bhatnagar et al. Dataset and Method for Tracking Human Object Interactions. CVPR 2022

Method

Dataset

- Segmentation mask for the target in the first frame with SAM and manual selection
- General video descriptions with CogVLM2

Training data from EgoExo4D



Target mask in the first frame

“A woman in a striped top and glasses is seen in a cluttered kitchen surrounded by cooking ingredients and utensils. She pours a white liquid from a can into a bowl, with a sandwich and various condiments on the counter ~ with a camera on a tripod set up, indicating she might be recording her cooking process.”

Text prompt

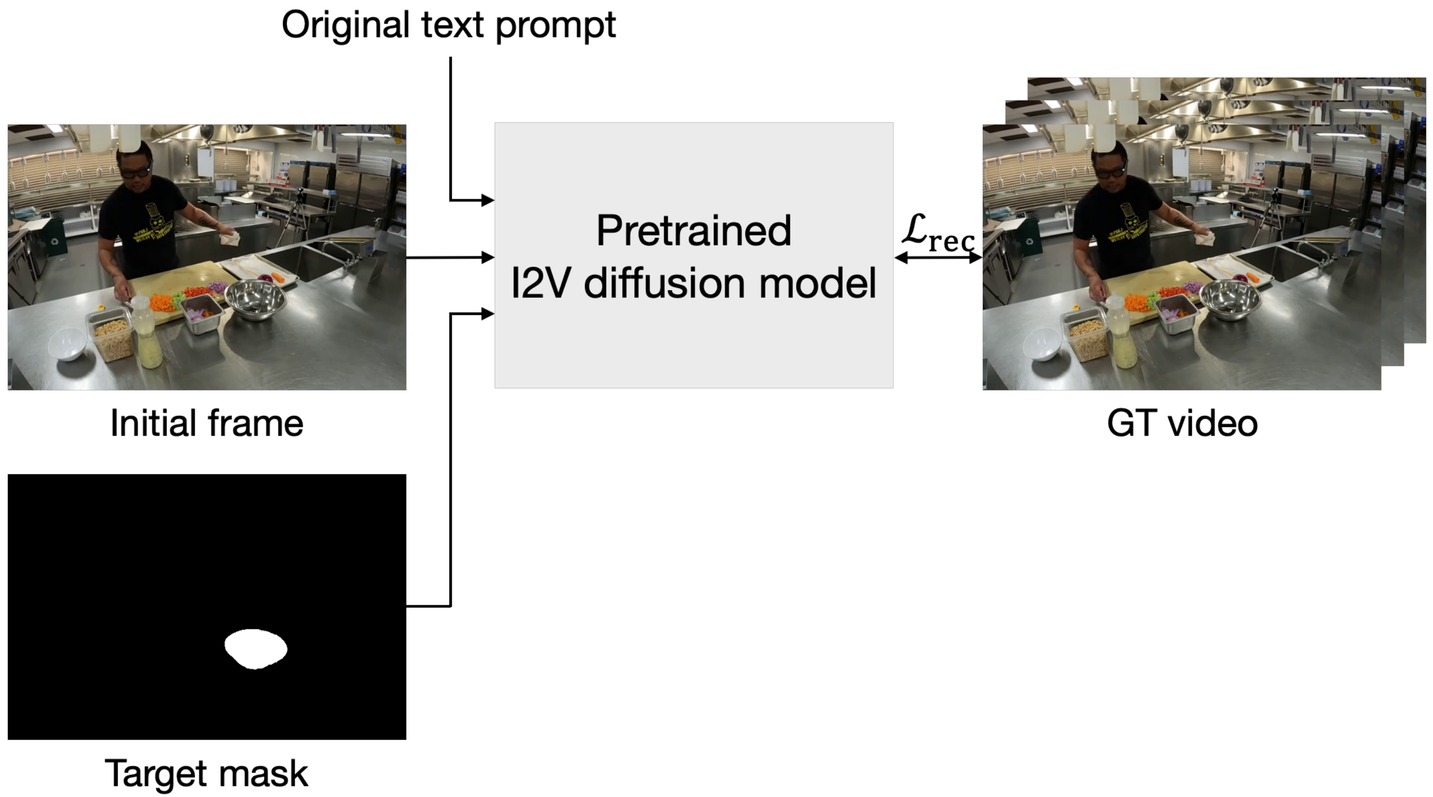
Kirillov et al. Segment Anything. ICCV 2023

Yang et al. CogVideoX: Text-to-Video Diffusion Models with an Expert Transformer. ICLR 2025

Method

Target-awareness via cross-attention loss

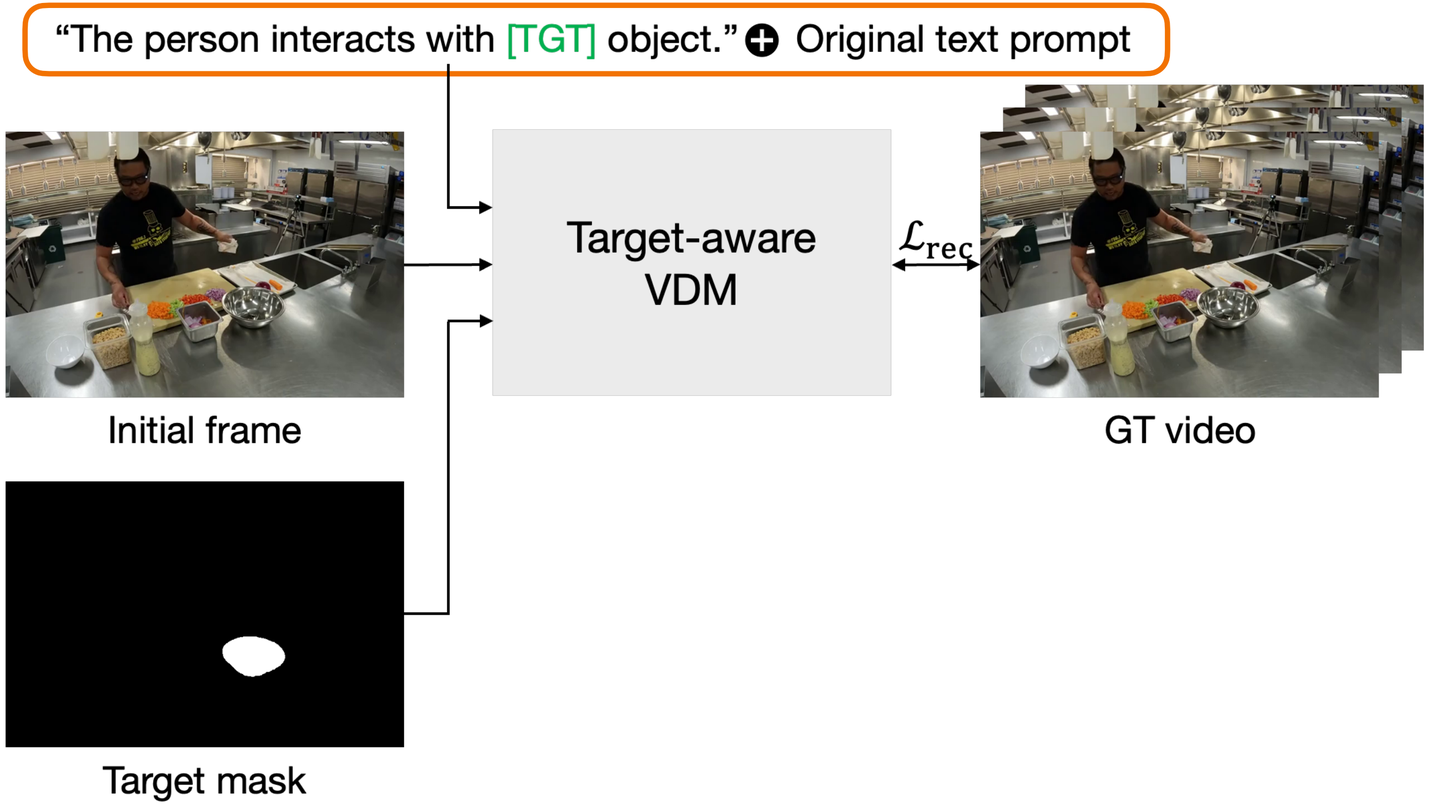
- Simple finetuning with the mask input doesn't help!



Method

Target-awareness via cross-attention loss

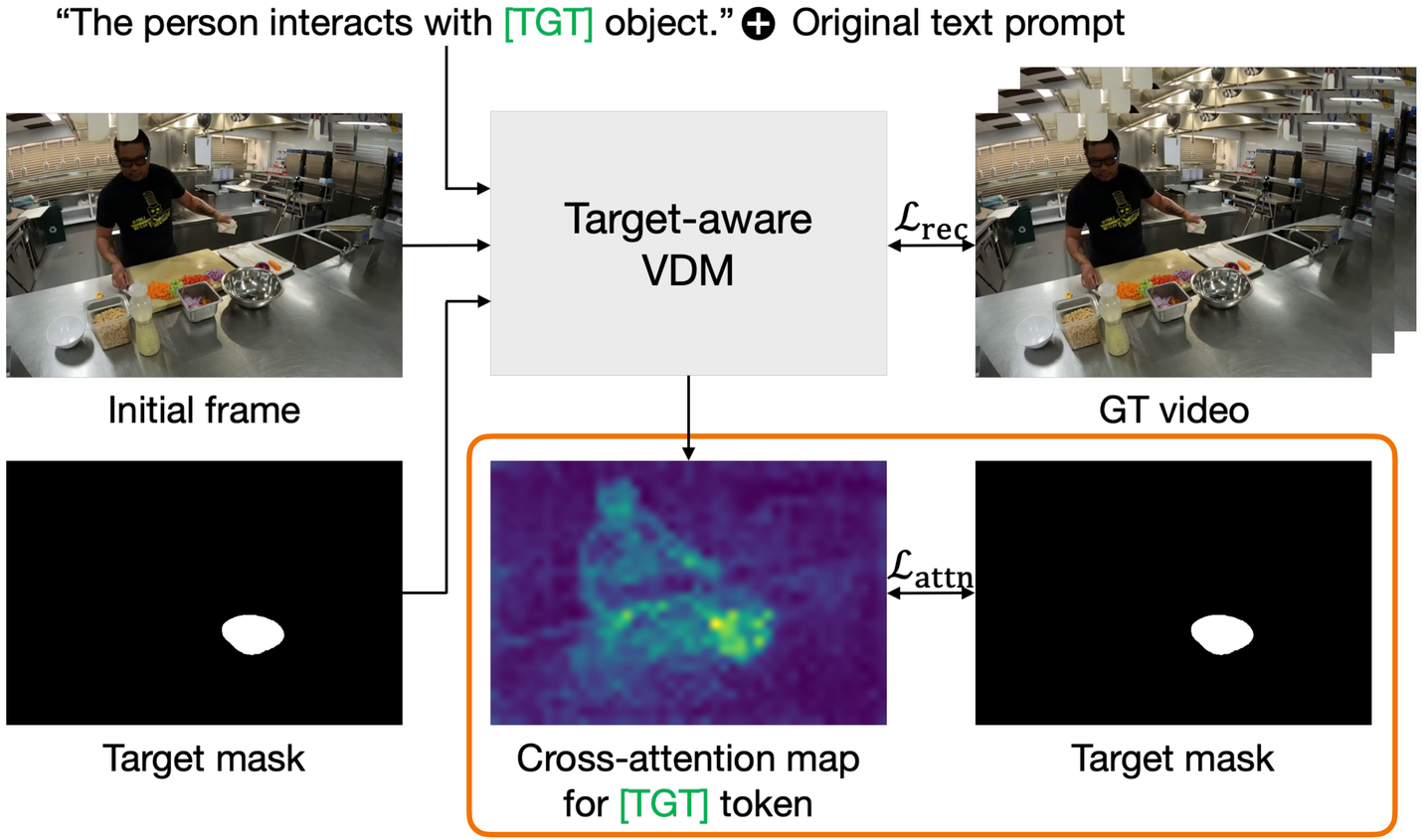
- Prepend special token “[TGT]” to encode spatial information of target
- Align cross-attention map of [TGT] with target mask (for the first frame of the video)



Method

Target-awareness via cross-attention loss

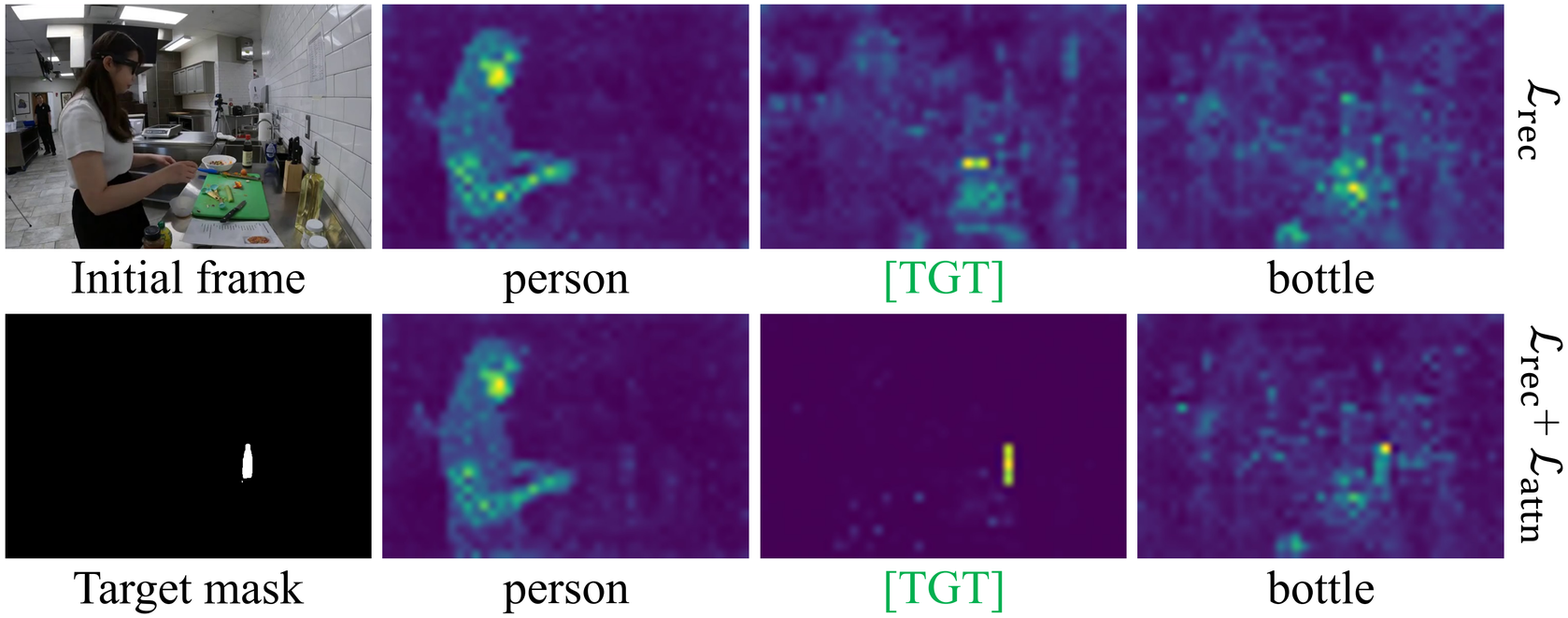
- Prepend special token “[TGT]” to encode spatial information of target
- Align cross-attention map of [TGT] with target mask (for the first frame of the video)



Method

Target-awareness via cross-attention loss

- Our cross-attention loss effectively guides the [TGT] token to focus on the target region

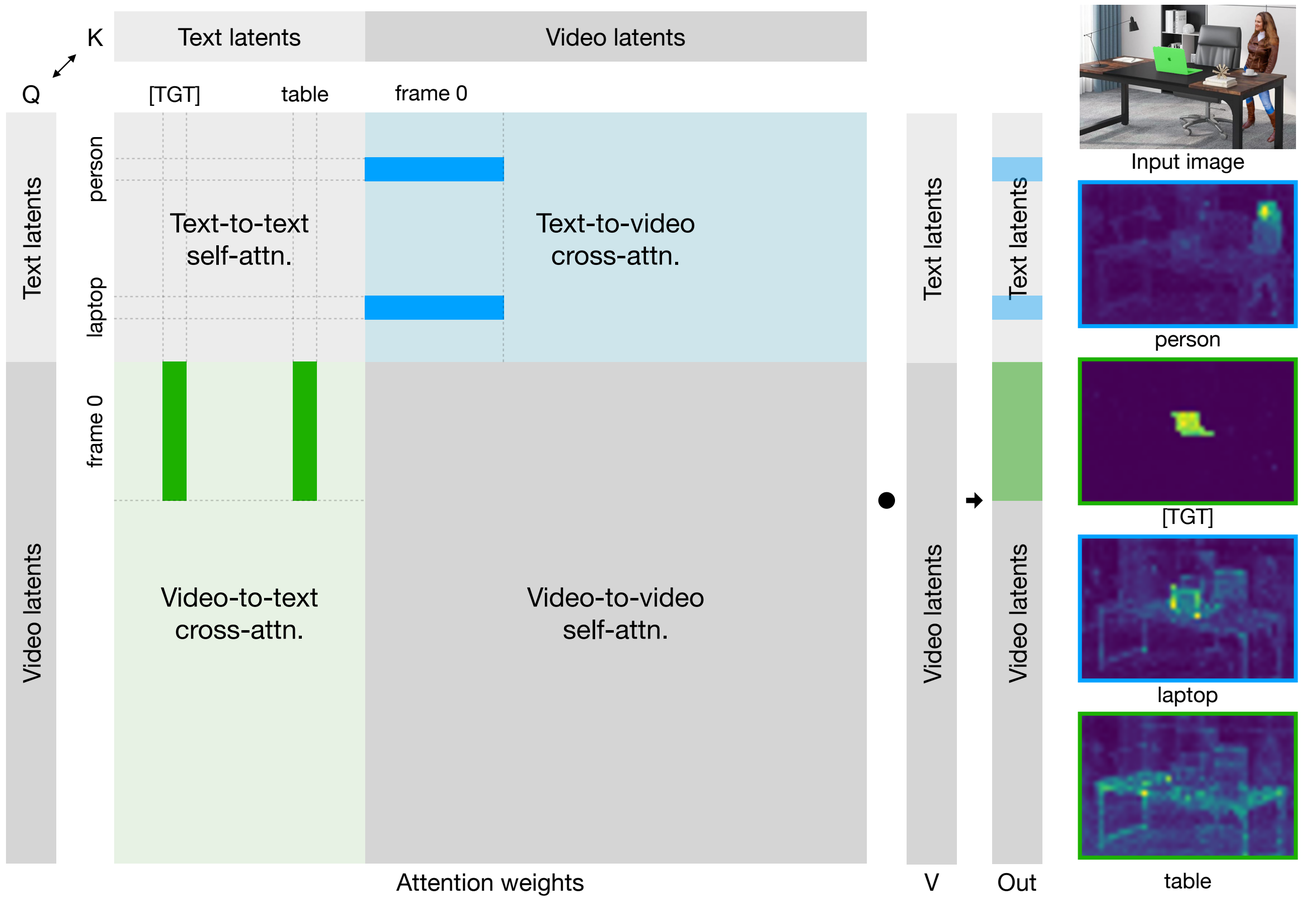


"The person picks up the [TGT] plastic bottle with a red cap."

Method

Selective cross-attention loss

- Multi-modal diffusion transformers have two cross-attention regions and both encode semantics
- We apply loss to video-to-text cross-attention regions since they directly impact the video latents

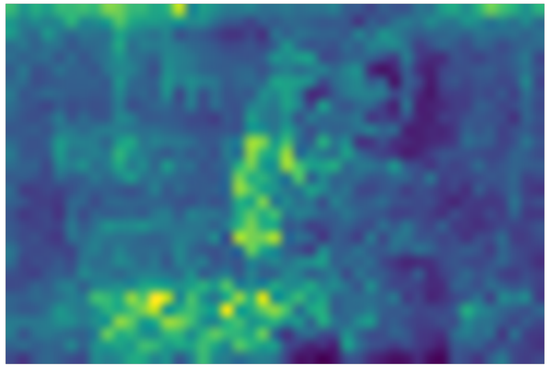


Method

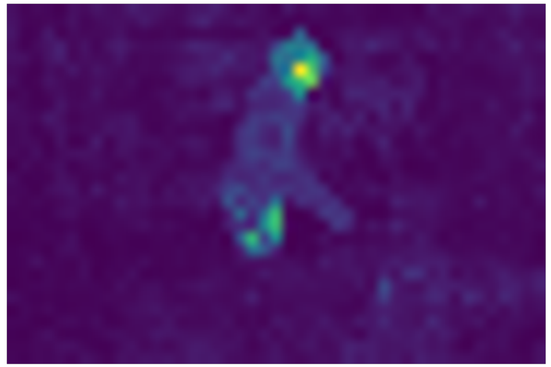
Selective cross-attention loss

- Certain transformer blocks capture richer semantic details than others → apply loss to those blocks
- Generate 100 videos and calculate L2 error between mask and attention map of each block

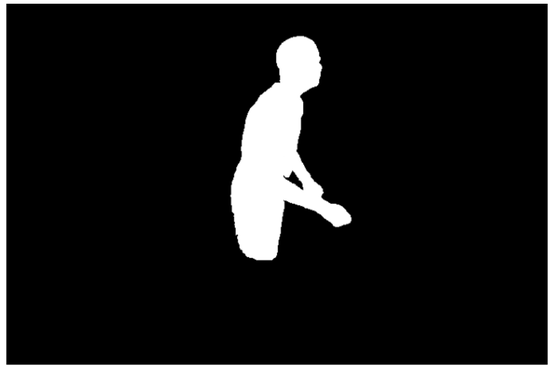
Input image



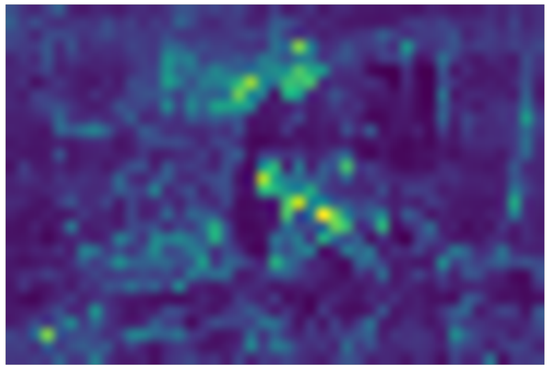
1st block



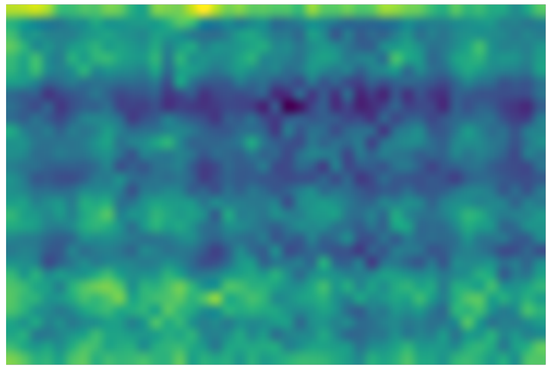
12th block



Segmentation mask



25th block

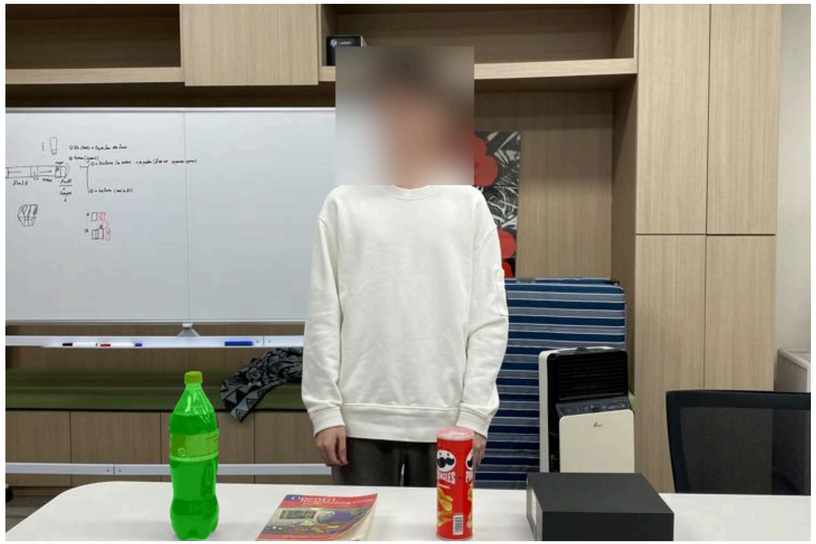


36th block

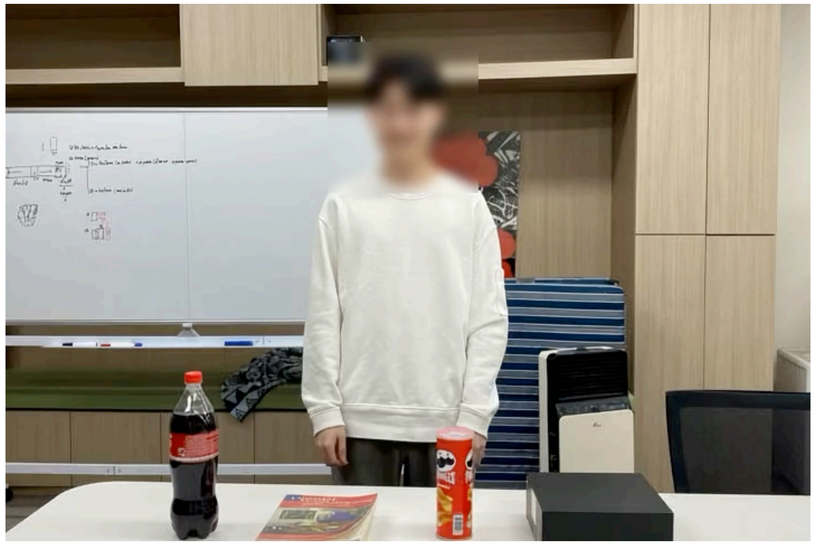
Results

Results

Accurate interactions with the target



Input image and target mask



Vanilla CogVideoX

Ours

“The man lifts the [TGT] bottle of coke from the table and takes a slow sip.”

Yang et al. CogVideoX: Text-to-Video Diffusion Models with an Expert Transformer. ICLR 2025

Results

Accurate interactions with the target

Input image and target mask

Vanilla CogVideoX

Ours

“The woman steps closer to the couch and gently picks up the [TGT] blue cushion with both hands.”

Yang et al. CogVideoX: Text-to-Video Diffusion Models with an Expert Transformer. ICLR 2025

Results

Outdoor scenes

“The woman picks up the [TGT] acorn.”

Results

Outdoor scenes

“The woman picks up the [TGT] acorn.”

Results

Outdoor scenes

“The woman picks up the [TGT] acorn.”

Results

Outdoor scenes

“The woman picks up the [TGT] acorn.”

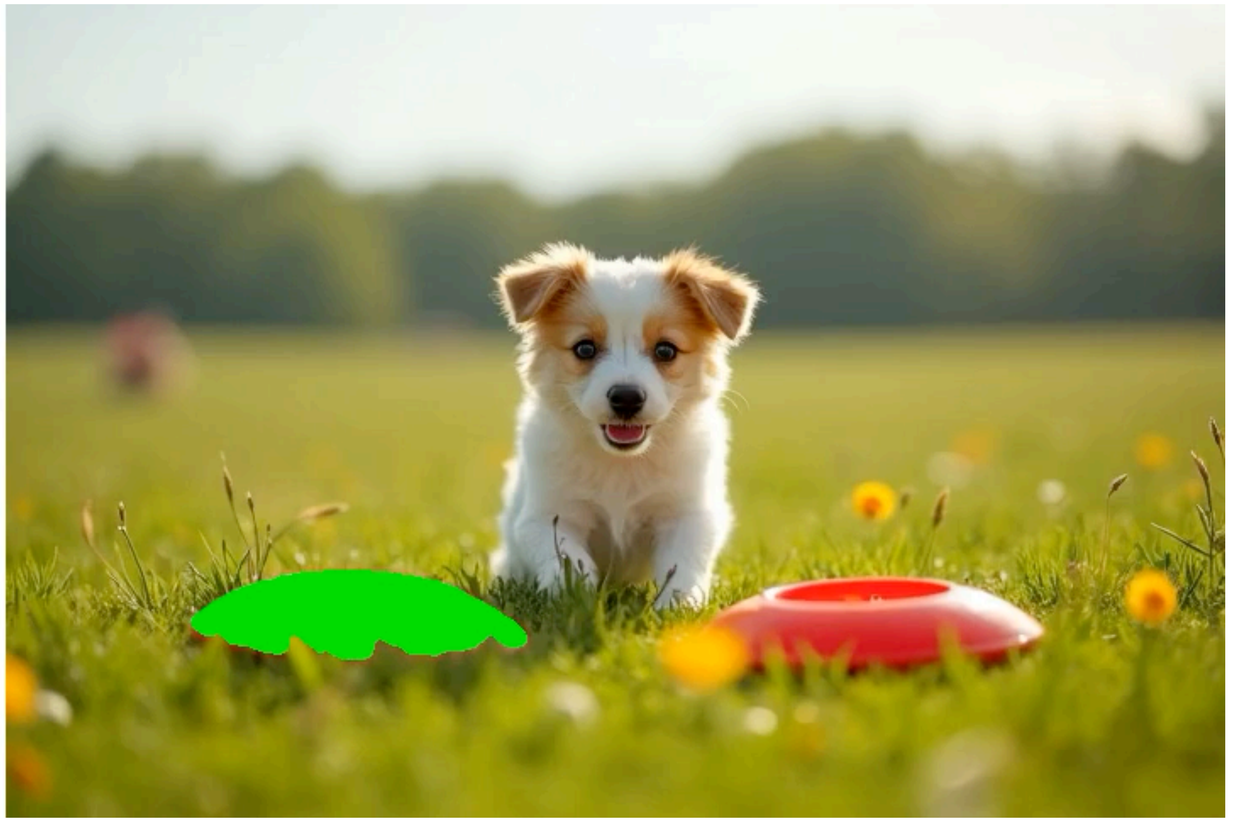
Results

Non-human interactions

“The rabbit turns its head towards the [TGT] carrot and takes a bite of it.”

Results

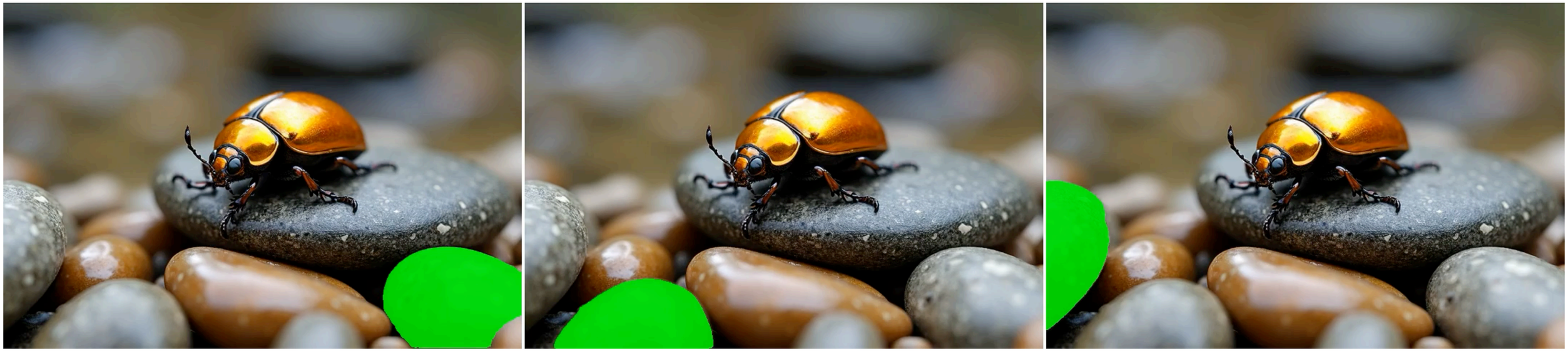
Non-human interactions



“The dog bites onto the [TGT] frisbee and lifts it off the ground.”

Results

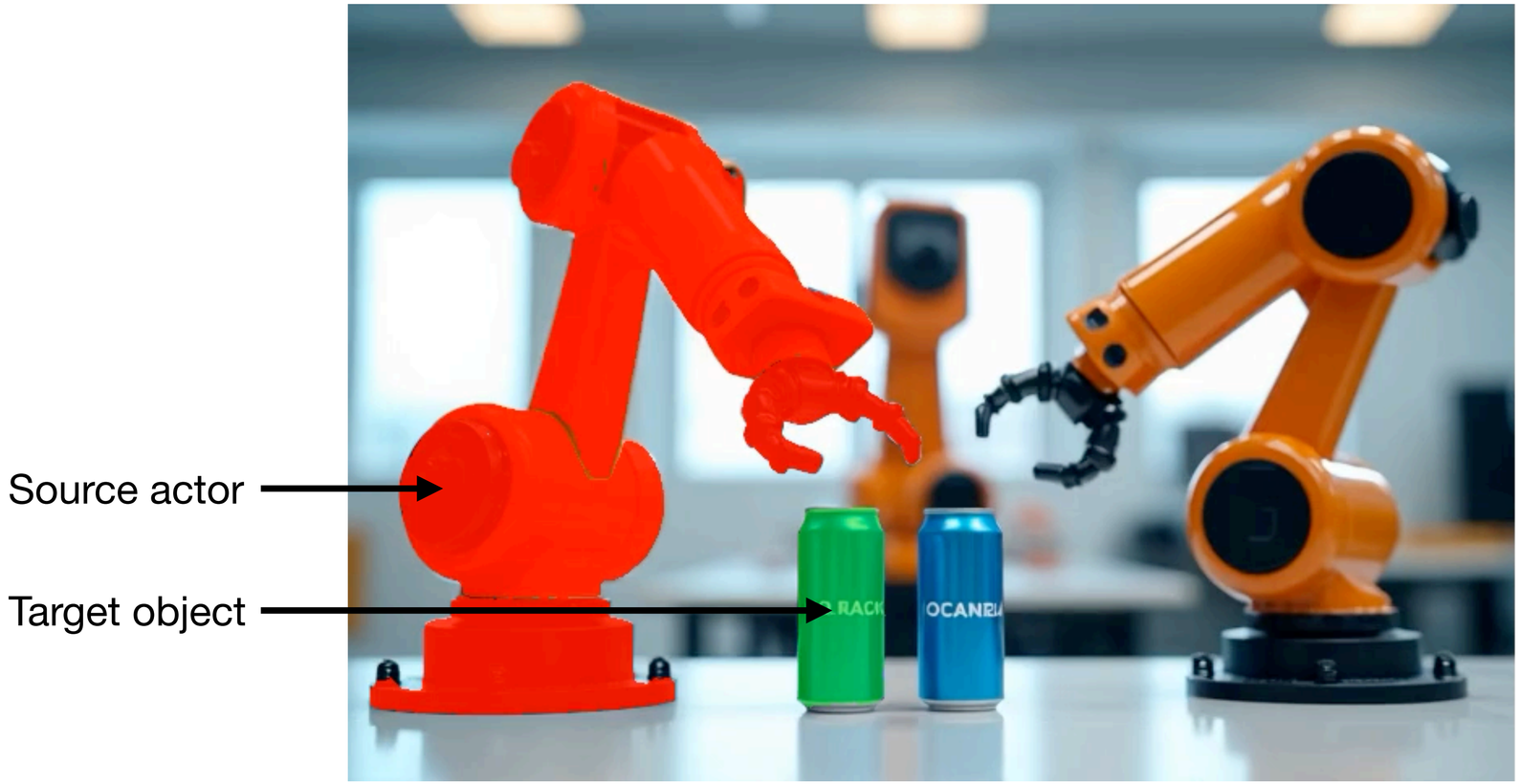
Scenes with multiple objects of the same type



“The beetle moves towards the [TGT] pebble and leans on it.”

Results

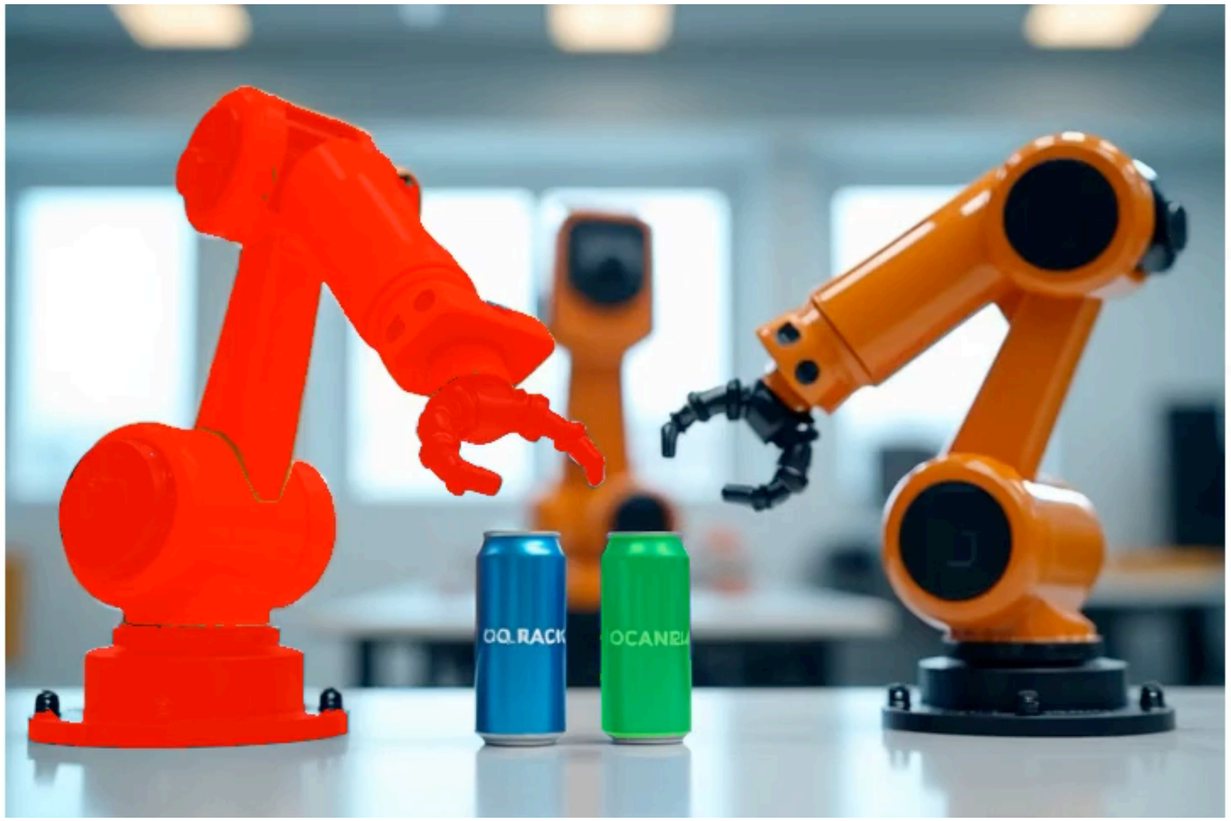
Specifying both the actor and the target



“The [SRC] robotic arm picks up the [TGT] blue can with its robot hand.”

Results

Specifying both the actor and the target



“The [SRC] robotic arm picks up the [TGT] blue can with its robot hand.”

Results

Specifying both the actor and the target

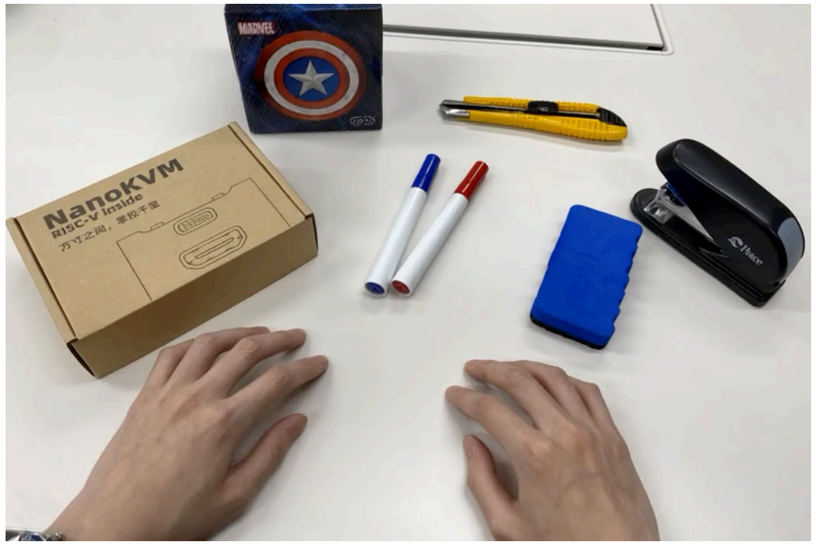
“The [SRC] robotic arm picks up the [TGT] blue can with its robot hand.”

Results

Target-Aware Egocentric Video Generation

Input image and target mask

CogVideoX-Ego



CogVideoX-Ego + Ours

Yang et al. CogVideoX: Text-to-Video Diffusion Models with an Expert Transformer. ICLR 2025

Results

Comparison with drag-based methods

Input image and target mask

Vanilla CogVideoX

Ours

“The man walks towards the [TGT] chair by the windows and sits on it.”

DragDiffusion

Shi et al. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. CVPR 2024

Yang et al. CogVideoX: Text-to-Video Diffusion Models with an Expert Transformer. ICLR 2025

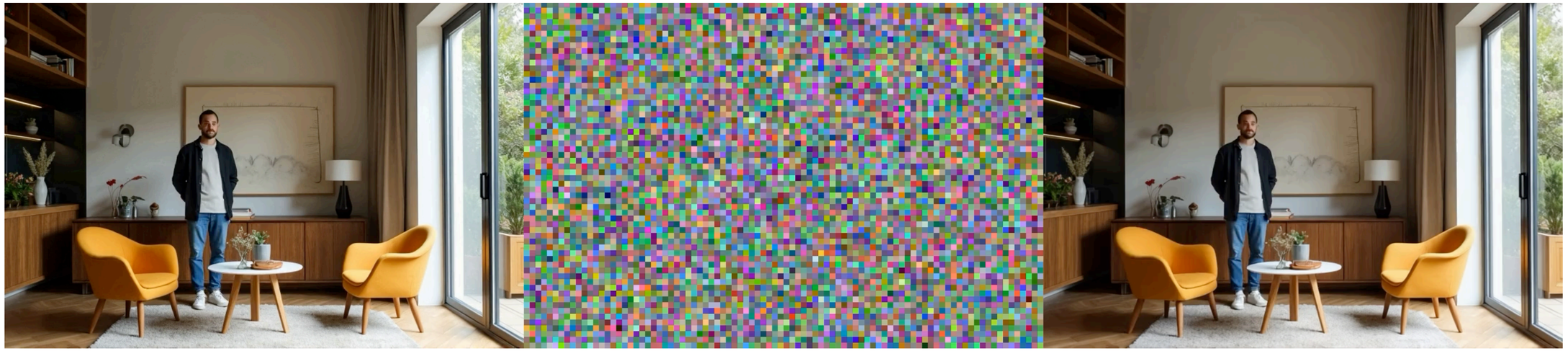
Results

Comparison with drag-based methods

Input image and target mask

Vanilla CogVideoX

Ours



Go-with-the-Flow

Burgert et al. Go-with-the-Flow: Motion-Controllable Video Diffusion Models Using Real-Time Warped Noise. CVPR 2025

Yang et al. CogVideoX: Text-to-Video Diffusion Models with an Expert Transformer. ICLR 2025

Results

Accurate interactions in complex scenes

Input image and target mask

Vanilla CogVideoX

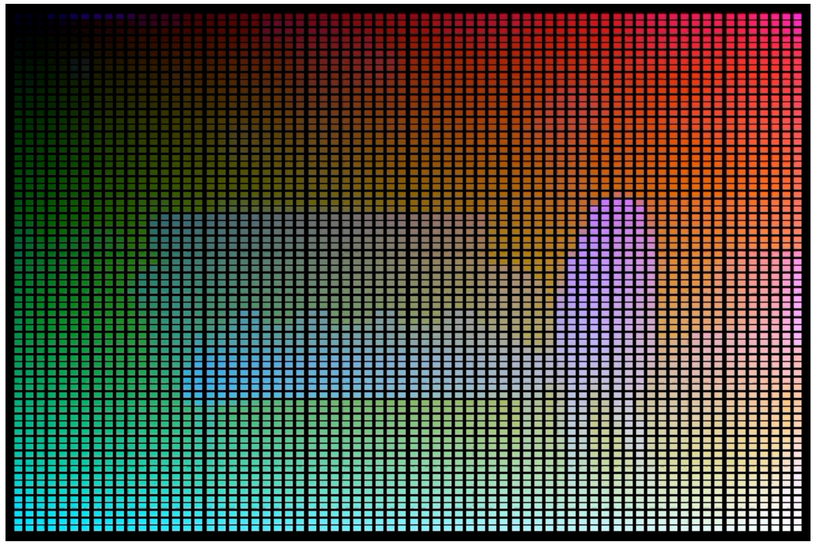
Ours

Yang et al. CogVideoX: Text-to-Video Diffusion Models with an Expert Transformer. ICLR 2025

Results

Providing motion source for controllable video generation pipeline

Our output



3D tracking condition

Motion transfer result

Gu et al. Diffusion as Shader: 3D-aware Video Diffusion for Versatile Video Generation Control. Siggraph 2025

Applications

Zero-Shot 3D HOI Motion Synthesis

Physics-based imitation learning

Input image and target

Zero-Shot 3D HOI Motion Synthesis

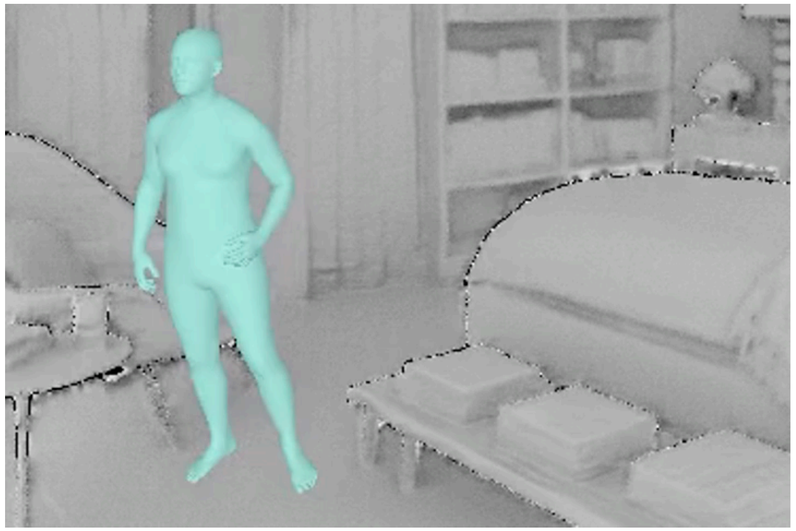
Physics-based imitation learning

Our output

Zero-Shot 3D HOI Motion Synthesis

Physics-based imitation learning

Our output



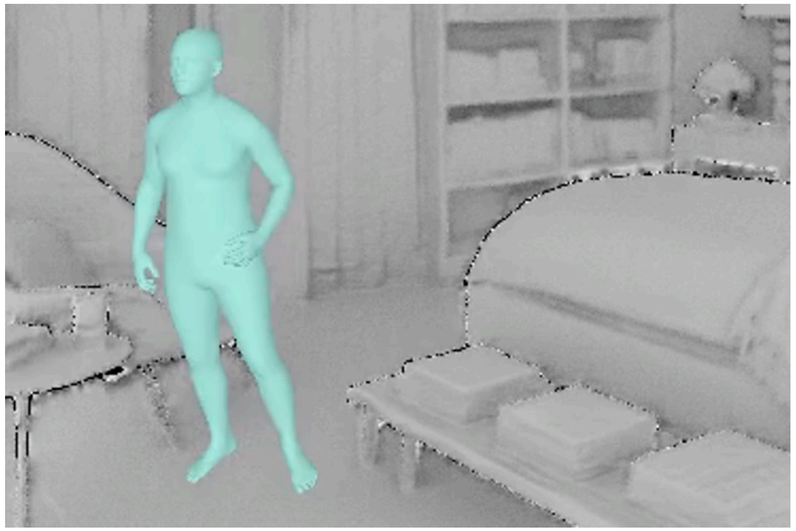
Estimated 3D human poses

Shen et al. World-grounded human motion recovery via gravity-view coordinates. Siggraph Asia 2024

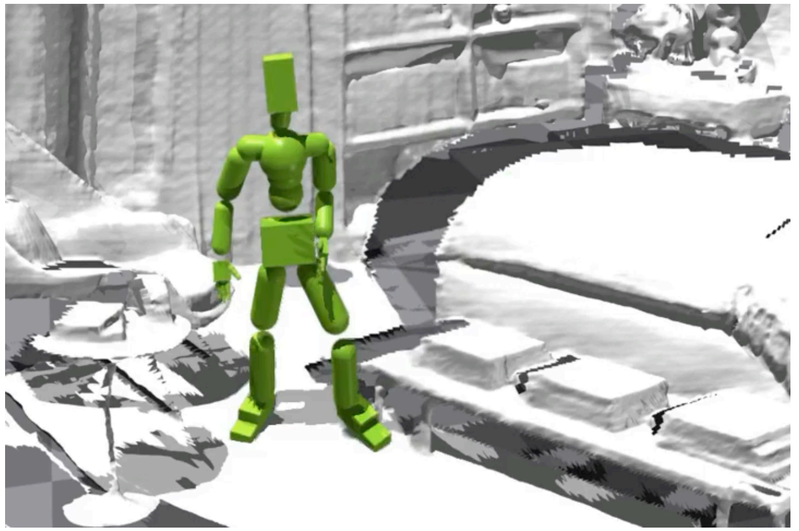
Zero-Shot 3D HOI Motion Synthesis

Physics-based imitation learning

Our output



Estimated 3D human poses



Physics-based imitation learning output

Shen et al. World-grounded human motion recovery via gravity-view coordinates. Siggraph Asia 2024

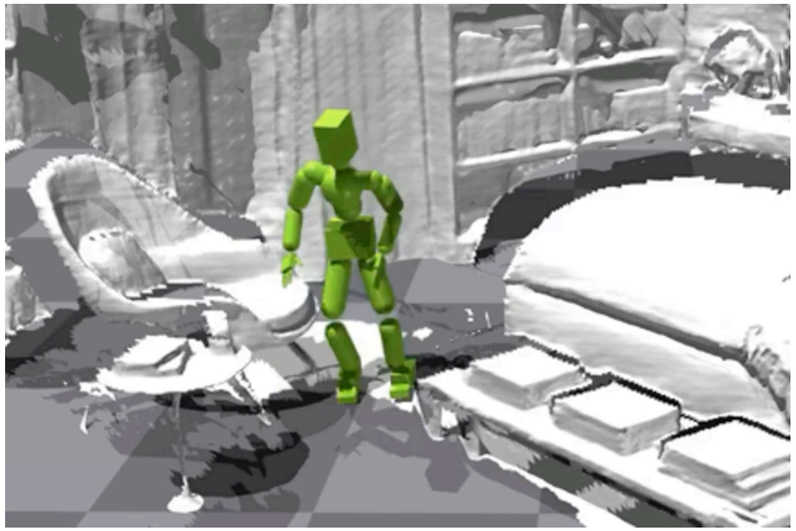
Wang et al. Physhoi: Physics-based imitation of dynamic human-object interaction. arXiv 2023

Zero-Shot 3D HOI Motion Synthesis

Physics-based imitation learning

Our output

Estimated 3D human poses



Physics-based imitation learning output

Shen et al. World-grounded human motion recovery via gravity-view coordinates. Siggraph Asia 2024

Wang et al. Physhoi: Physics-based imitation of dynamic human-object interaction. arXiv 2023

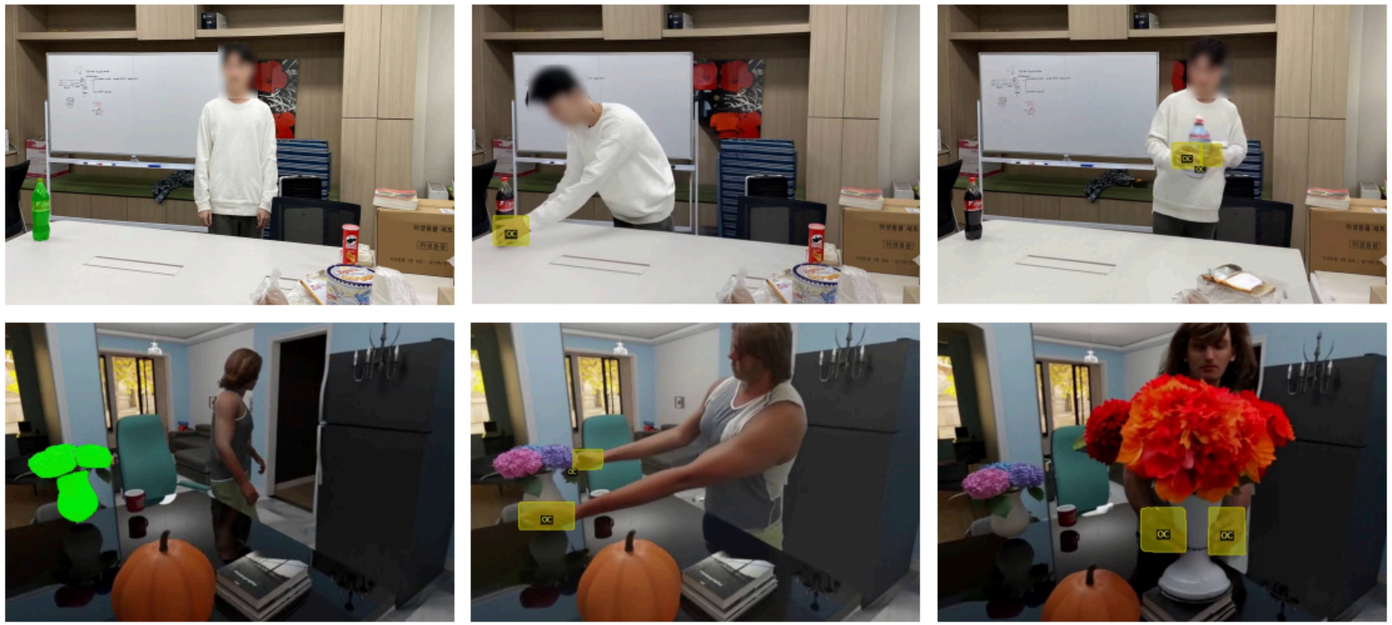
Evaluations & Ablations

Evaluation

Targeting Quality & Video Quality

- Contact Score: rate of videos with accurate interactions over all generated videos

	Targeting Quality			Video Quality						
	Contact Score↑	Hum. Eval.↑	User Pref.↑	SC↑	BC↑	DD↑	MS↑	AQ↑	IQ↑	Avg.↑
CogVideoX	0.560	0.456	28.4%	0.893	0.898	0.883	0.988	0.502	0.694	0.810
CogVideoX w.data	0.638	0.596	36.2%	0.914	0.907	0.907	0.990	0.492	0.653	0.810
Attn. Mod.	0.546	0.508	22.2%	0.872	0.889	0.786	0.986	0.499	0.687	0.786
Ours	**0.878**	**0.892**	(100%-above)	0.933	0.919	0.899	0.937	0.496	0.656	0.807



Input image Success Failure

Ablation

Cross-Attention Loss

- CA loss on selective blocks

	Contact Score↑	Video Quality↑
Random	0.819	0.807
Equally-Spaced	0.816	0.800
Ours	**0.878**	0.807

- CA loss on selective regions

	Contact Score↑	Video Quality↑
T2V Cross-Attn.	0.740	0.806
Both Cross-Attn.	0.860	0.810
Ours (V2T Cross-Attn.)	**0.878**	0.807

- CA loss weight

	Contact Score↑	Video Quality↑
$\lambda_{attn} = 0.0$	0.647	0.815
$\lambda_{attn} = 0.05$	0.727	0.811
$\lambda_{attn} = 0.1$	0.878	0.807
$\lambda_{attn} = 0.25$	**0.890**	0.806
$\lambda_{attn} = 0.5$	0.888	0.807
$\lambda_{attn} = 1.0$	0.888	0.804

	T
	Contact Score↑
CogVideoX	0.560
CogVideoX w.data	0.638
Attn. Mod.	0.546
Ours	**0.878**

Ablation

Masks

- Shape of masks

	Contact Score↑	Video Quality↑
original	**0.896**	0.812
circular-15	0.838	0.813
circular-30	0.888	0.809

- Automatic mask acquisition

	Contact Score↑	Video Quality↑
Original	0.896	0.812
Automatic	0.864	0.810

- Quality of masks

	Contact Score↑	Video Quality↑
original	0.896	0.812
dilate-3	0.884	0.808
dilate-5	0.872	0.815
erode-3	**0.904**	0.815
erode-5	0.880	0.813

Limitations

Limitations

Targeting Multiple Objects

- Model struggles when a mask spans multiple objects

Initial frame

“The woman picks up both [TGT] cups with each hand.”

“The person reaches out and picks up both [TGT] objects with each hand.”

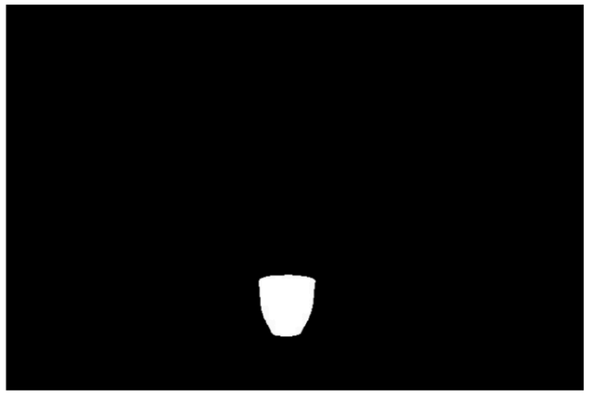
Limitations

Targeting Multiple Objects

- Model struggles to switch targets within a video

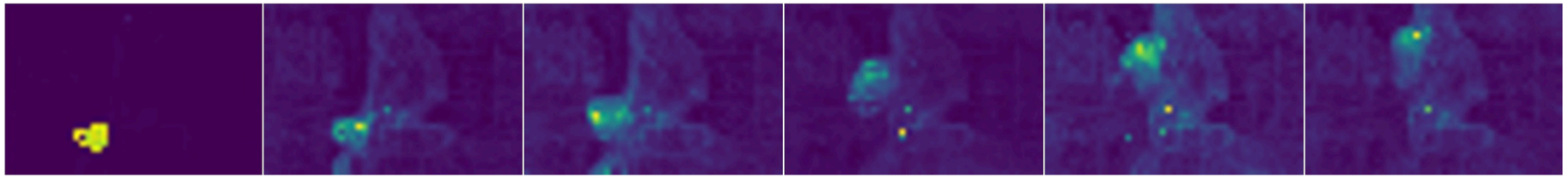
Initial frame

First frame mask



Middle frame mask

Generation output



[TGT] attention visualization