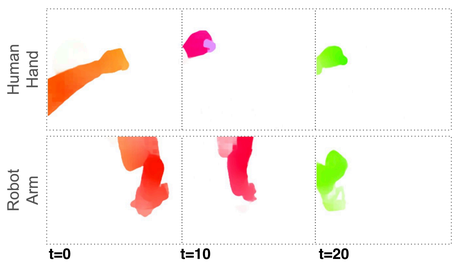
Future Optical Flow Prediction Improves Robot Control & Video Generation

Kanchana Ranasinghe, Honglu Zhou, Yu Fang, Luyu Yang, Le Xue, Ran Xu, Caiming Xiong, Silvio Savarese, Michael S Ryoo, Juan Carlos Niebles

Code + Model + Demo

Optical Flow as Motion Representation



Optical Flow is highly embodiment agnostic

See *common colors* for common movements

Colors represent motion direction of each pixel

Color intensity represents motion magnitude

Problem: Learn from Internet-Scale Video Data

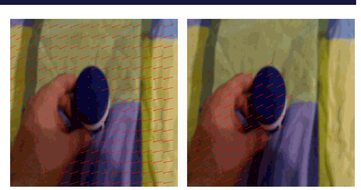
- Robot Teleop Data is finite and expensive
- Internet Data is almost infinite and free

BUT Internet Data is noisy

A) Random Camera Motions

B) Long Static Regions

C) Caption Misalignment



Camera Motion

Caption Misalignment

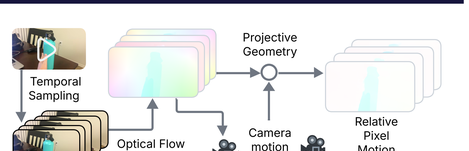
We explore

- ❖ Relative Optical Flow Calculation
- ❖ Motion Aware Frame Sampling
- ❖ Leveraging VLM-Diffusion Backbone

Relative Optical Flow Calculation

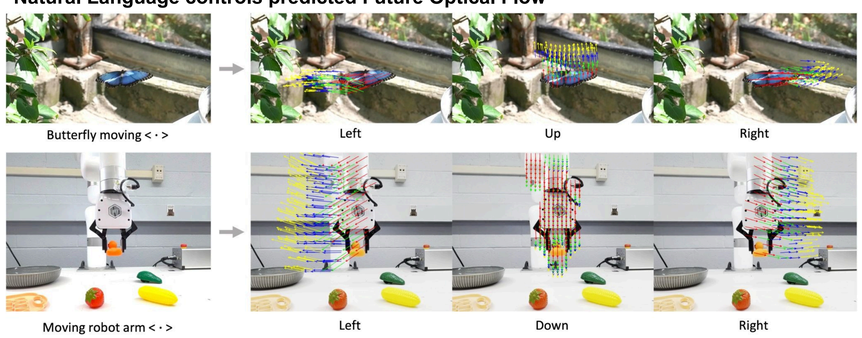
Disentangle object motions from noisy camera motion using,

- learned visual features
- homography techniques
- projective geometry

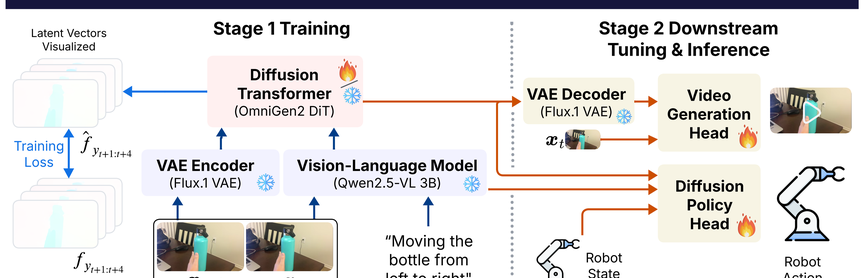


Language-Driven Future Optical Flow Prediction

Natural Language controls predicted Future Optical Flow



FOFPred Joint VLM-Diffusion Architecture



Key Design Choices

Pretrain Dataset	Train Steps	Avg Len
DROID	2000	4.04
SSv2	2000	4.39 (+0.35)

Human Videos provide value even at same quantity

Intuition: Better video diversity

Method	Avg Len
No Video Pretraining	3.91 (-0.48)
Static Frame Targets	4.28 (-0.11)
Raw Motion Targets	3.89 (-0.50)
Disentangled Motion Targets	4.39

Disentangling Motion Targets is extremely important

Method	AL ↑
Ours	4.39
No DiT temporal adaptation	4.05 (-0.34)
Replace VLM w/ Sentence-T5	3.95 (-0.44)
No web-video pretraining	3.91 (-0.48)
No image-edit pretraining	4.14 (-0.25)
Stable Diffusion-XL (SD-XL)	4.01 (-0.38)
Stable-Video Diffusion (SVD)	4.10 (-0.29)
CogVideoX	4.31 (-0.08)

Evaluations on Robot Control and Video Generation

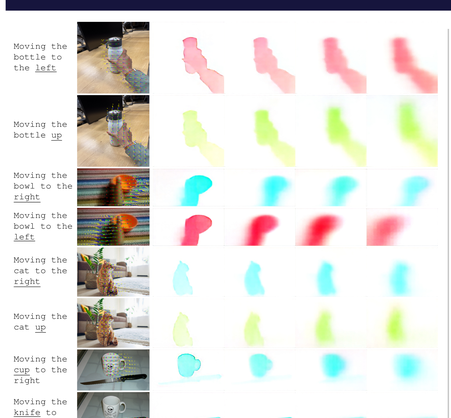
Method	Training Data	i^{th} Task Success Rate ↑ 1	2	3	4	5	Avg. Len ↑
RT-1	100% ABC	53.3	22.2	9.4	3.8	1.3	0.90
Diffusion Policy	100% ABC	40.2	12.3	2.6	0.8	0.0	0.56
Robo-Flamingo	100% ABC	82.4	61.9	46.6	33.1	23.5	2.47
Uni-Pi	100% ABC	56.0	16.0	8.0	8.0	4.0	0.92
MDT	100% ABC	63.1	42.9	24.7	15.1	9.1	1.55
Susie	100% ABC	87.0	69.0	49.0	38.0	26.0	2.69
GR-1	100% ABC	85.4	71.2	59.6	49.7	40.1	3.06
Vidman	100% ABC	91.5	76.4	68.2	59.2	46.7	3.42
RoboUniview	100% ABC	94.2	84.2	73.4	62.2	50.7	3.65
LTM	100% ABC	97.1	82.4	72.8	67.2	60.6	3.81
DreamVLA	100% ABC	98.2	94.6	89.5	83.4	78.1	4.44
VPP	100% ABC	96.5	90.9	86.6	82.0	76.9	4.33
FOFPred	100% ABC	98.8 (+2.3)	95.0 (+4.1)	90.4 (+3.8)	84.6 (+2.6)	78.7 (+1.8)	4.48 (+0.15)
GR-1	10% ABC	67.2	37.1	19.8	10.8	06.9	1.41
VPP	10% ABC	87.8	74.6	63.2	54.0	45.3	3.25
FOFPred	10% ABC	90.4 (+2.6)	77.4 (+2.8)	65.9 (+2.7)	62.8 (+8.8)	46.9 (+1.6)	3.43 (+0.18)

Method	Handover Block	Handover Mic	Pick Diverse Bottles	Pick Dual Bottles	Place Dual Shoes	Average
RDT	45	90	2	42	4	36.6
ACT	42	85	7	31	9	34.8
DP	10	53	6	24	8	20.2
DP3	70	100	52	60	13	59.0
π_0	45	98	27	57	15	48.4
VPP	54	80	60	63	52	61.8
FOFPred (ours)	61 (+7)	87 (+5)	67 (+7)	68 (+5)	60 (+8)	68.6 (+6.8)

Method	SSIM ↑	PSNR ↑	LPIPS ↓	FVD↓	KVD↓	MF ↑
Seer	41.8	10.71	58.8	287.46	81.31	—
Dynamicrafter	—	—	—	204.11	31.81	—
CosHand-I	61.5	16.87	31.3	91.18	19.24	0.432
CosHand-A	53.1	14.92	40.8	90.30	13.68	0.570
InterDyn	66.4	18.60	26.0	19.27	1.99	0.633
InterDyn-R	68.0	19.04	25.2	22.22	2.09	0.641
CogVideoX	67.2	21.51	30.3	78.47	12.46	0.594
FOFPred (ours)	68.4 (+1.2)	22.26 (+0.75)	28.5 (+1.8)	75.39 (+3.08)	11.38 (+1.08)	0.662 (+0.068)

- ❖ Clear improvements on CALVIN Robotic benchmark (left)
- ❖ Improvements on bi-manual RoboTwin benchmark (top-right)
- ❖ Video Generation improves over CogVideoX baseline (bottom-right)

Optical Flow & Video Generation Visualization



Key Insights

1. Learning grounded language-motion correspondences from noisy video-caption data.
2. Utilizing *optical flow* as generalizable motion representation.
3. Learned motion knowledge transfers to language conditioned robotic and video generation tasks.