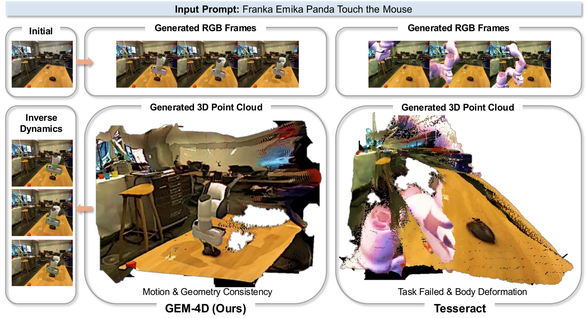
GEM-4D: Geometry-Enhanced Video World Models for Robot Manipulation

Kaichen Zhou[1,2,*] Yuzhen Chen[1,*] Fangneng Zhan[2] Hang Hua[4] Grace Chen[1] Xinhai Chang[1]
Ao Qu[2]
Yilun Du[1] Zhuang Liu[3] Paul Pu Liang[2,†] Mengyu Wang[3,†]
[1] Harvard University [2] MIT [3] Princeton University [4] MIT-IBM Watson AI Lab

CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation



Why does geometry matter?

- Video world models can look realistic while **failing** to preserve the same physical points across time.
- For robot control, this **breaks** the link between predicted videos and executable actions: if points cannot be tracked consistently, generated rollouts cannot reliably guide manipulation.
- **GEM-4D** makes video futures actionable by distilling dense 4D correspondence supervision during training, improving real-world manipulation success **from 61% to 81%** with no extra inference cost.

Methods

Geometry-Grounded World Model Training

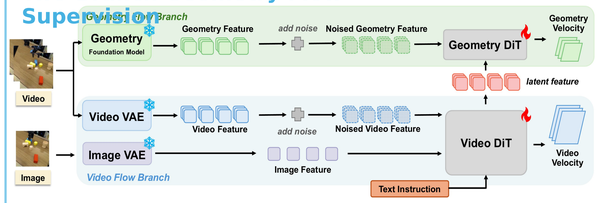
- Dense point correspondence is determined by 3D geometry, not just RGB appearance.

$$\mathbf{p}_{t+1} \sim \mathbf{K}\left[\mathbf{R}_{t\to t+1}\ \mathbf{D}(\mathbf{p}_t)\ \mathbf{K}^{-1}\ \mathbf{p}_t + \mathbf{T}_{t\to t+1} + \Delta\mathbf{X}_t\right].$$

Pixel in next frame · Intrinsic matrix · Scene flow / object motion · Camera translation · Camera rotation · Depth at $\mathbf{p}_t$ · Pixel in current frame · Inverse intrinsic matrix

Inter-frame correspondence is governed by geometry: depth, camera motion, and scene flow determine where each physical point moves. Thus, robot-ready video prediction requires dense geometric consistency, not just visual realism.

Two-Branch Geometry Flow Supervision



GEM-4D training. A Video DiT learns video velocity for RGB generation, while a training-only Geometry DiT predicts geometry velocity from intermediate video features. This joint training enforces geometry-consistent rollouts. At inference, the geometry branch is removed, leaving efficient RGB-only generation.

- **RGB Video Flow Branch:** Video DiT predicts video velocity from noisy to clean latents, learning RGB rollout generation but mainly supervising appearance.
- **Geometry Flow Branch:** A training-only Geometry DiT predicts geometry velocity from intermediate Video DiT features, using frozen 4D geometry features as supervision.

$$\nabla_\theta \mathcal{L} = \nabla_\theta \mathcal{L}_{\mathrm{FM}}^{\mathrm{vid}} + \alpha \cdot \frac{\partial \mathcal{L}_{\mathrm{FM}}^{\mathrm{geo}}}{\partial \mathbf{m}_t} \cdot \frac{\partial \mathbf{m}_t}{\partial \theta}$$

Total gradient · Geometry-induced gradient · Appearance supervision · Balancing weight

- **Joint Objective:** RGB flow and geometry flow are jointly optimized, grounding the backbone in both visual appearance and geometric correspondence.

Adaptive Inverse Dynamics for Action Extraction

- **3D Scene Grounding + Confidence-Gated Tracking:** The generated

$$\hat{\mathcal{M}}_{\mathrm{ee}}^{t} = \begin{cases} \text{re-anchor tracker at } t, & \text{if } s_t < \tau, \\ \text{Qwen3.5-VL}(I_t, c), & \text{if } \Delta s_t < -\delta, \\ \mathcal{M}_{\mathrm{ee}}^{t}, & \text{otherwise.} \end{cases}$$

Generated rollouts are lifted into 3D using depth and camera information, then task-relevant points are tracked across frames.

A dual confidence gate filtersunreliable tracks using bothtracker confidence andmotion/geometry consistencybefore action extraction.

- **Pose Fallback + Grasp Insertion + Action Synthesis:** When 6D pose recovery is unreliable, the system switches to a stable fallback, inserts grasp poses at key interaction moments, and interpolates waypoints into a smooth executable trajectory.

$$\mathbf{T}_{\mathrm{grasp}}^{*} = \arg\min_{\mathbf{T}_{\mathrm{grasp}}^{(i)}} \left(\lambda_t \left\|\mathbf{t}_{\mathrm{grasp}}^{(i)} - \mathbf{t}_{\mathrm{ref}}\right\|_2 + \lambda_R\, d_{\mathrm{geo}}\left(\mathbf{R}_{\mathrm{grasp}}^{(i)}, \mathbf{R}_{\mathrm{ref}}\right)\right)$$

Experiment

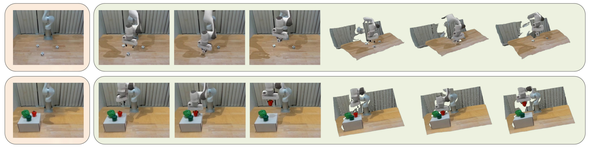
GEM-4D Improves Video and Geometry Prediction

Setting	Method	RGB FVD ↓	RGB SSIM ↑	RGB PSNR ↑	Depth AbsRel ↓	Depth δ_1 ↑	Depth δ_2 ↑	Points Chamfer ↓	Tracking $\delta^{\mathrm{vis}}_{avg}$ ↑
R	CogVideoX [64]	35.56	75.91	20.18	22.33	68.32	83.17	0.2670	66.22
	Wan 2.2-14B [48]	33.43	76.24	20.70	21.39	71.18	84.35	0.2349	68.18
	TesserAct [69]	33.28	75.66	20.08	22.07	66.80	82.60	0.2630	67.14
	Geometry-Forcing [58]	33.17	76.12	20.53	21.96	69.74	83.83	0.2443	67.97
	GEM-4D	**31.82**	**82.05**	**21.11**	**20.13**	**78.19**	**88.21**	**0.2001**	**71.23**
S	CogVideoX [64]	40.21	75.51	20.03	15.41	70.99	92.90	0.2913	58.32
	Wan 2.2-14B [48]	49.20	73.01	19.87	17.81	67.07	90.16	0.1762	61.99
	TesserAct [69]	41.97	76.72	19.71	16.02	69.26	93.03	0.1813	61.15
	Geometry-Forcing [58]	34.06	77.92	19.48	15.34	68.96	92.80	0.1488	60.84
	GEM-4D	**27.94**	**80.27**	**23.36**	**14.11**	**74.13**	**95.01**	**0.0702**	**68.18**

From Generated Videos to Successful Robot Actions

Setting	Task	CogVideoX	Tesseract	Ours
R	AUTOLab	49	58	75
	CLVR	64	65	83
	RAIL	39	59	87
S	LiftNumberedBlock	-	21	78
	PutRubbishInBin	-	0	75
	ReachTarget	-	2	82
	LampOn	-	36	67
	PickUpCup	-	49	81
	SlideBlockToTarget	-	18	80
	SolvePuzzle	-	33	63

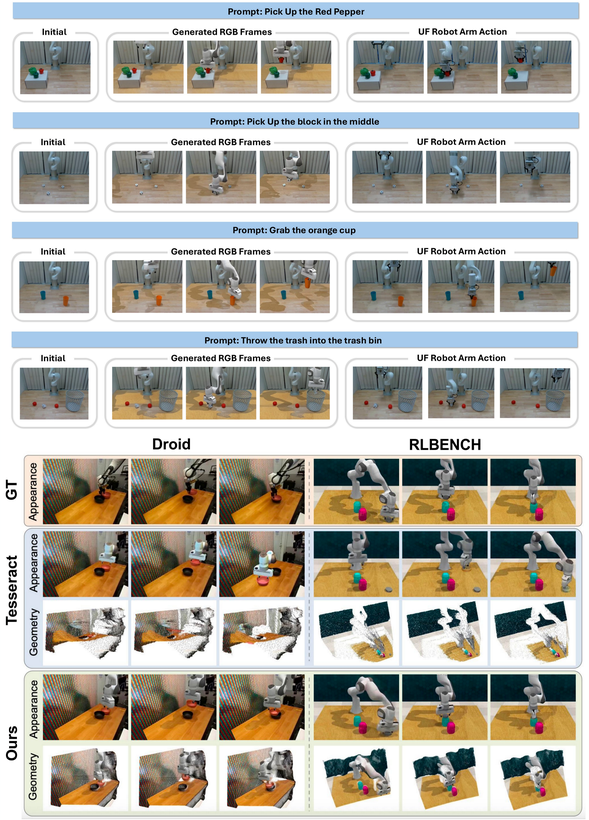
- On **real-world Droid tasks**, success improves from ~61% to ~82%.
- On **synthetic RLBench execution**, success improves from ~23% to ~75%.



Geometry Supervision Matters Actions

Domain	Method	RGB (Appearance) FVD ↓	RGB SSIM ↑	RGB PSNR ↑	Depth (Geometry) AbsRel ↓	Depth δ_1 ↑	Depth δ_2 ↑	P (Geometry) Chamfer ↓
Real	Wan 2.2-14B	33.43	76.24	20.70	21.39	71.18	84.35	0.2349
	GEM-4D(Dep)	32.91	78.58	20.75	20.89	74.60	86.67	0.2229
	GEM-4D(VGGT)	33.68	75.89	20.64	21.73	71.03	83.80	0.2370
	GEM-4D	**31.82**	**82.05**	**21.11**	**20.13**	**78.19**	**88.21**	**0.2001**

Result



Conclusion

GEM-4D shows that grounding video world models in dense 4D geometry turns visually plausible futures into physically trackable, actionable rollouts for robot manipulation.