Predicting the Future in 3D

Roozbeh Mottaghi

Skild AI

University of Washington

Jun 3, 2026
Workshop on Multimodal Spatial Intelligence

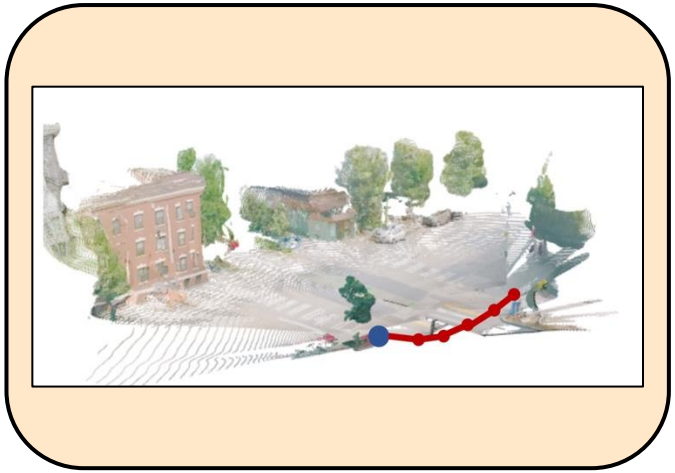
Why future prediction?

As we move towards physical AI, we need a model of how the world works to act effectively within it.

- Rolling out actions before execution leads to successful planning.
- A model that understands the world dynamics is much more sample efficient.
- Safety is important in interactive tasks. Sometimes actions can lead to irreversible damages.

Why 3D?

- 2D pixel reasoning is inherently ambiguous.
- There is a latent 3D structure that 2D foundation models leverage.
- 3D representation leads to more systematic generalization and efficient learning.



ODIN

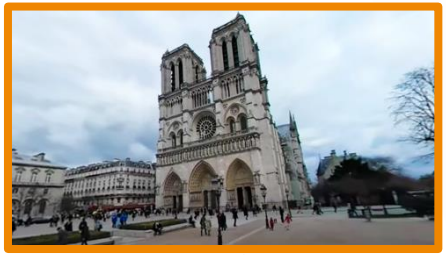
3D prediction for navigation

ObjectForesight

3D prediction for manipulation

Long-Range Novel View Synthesis

Starting input view



Synthesized view

Translate: Forward 30m, Right 5m
Rotate: 20° left, 5° Upward

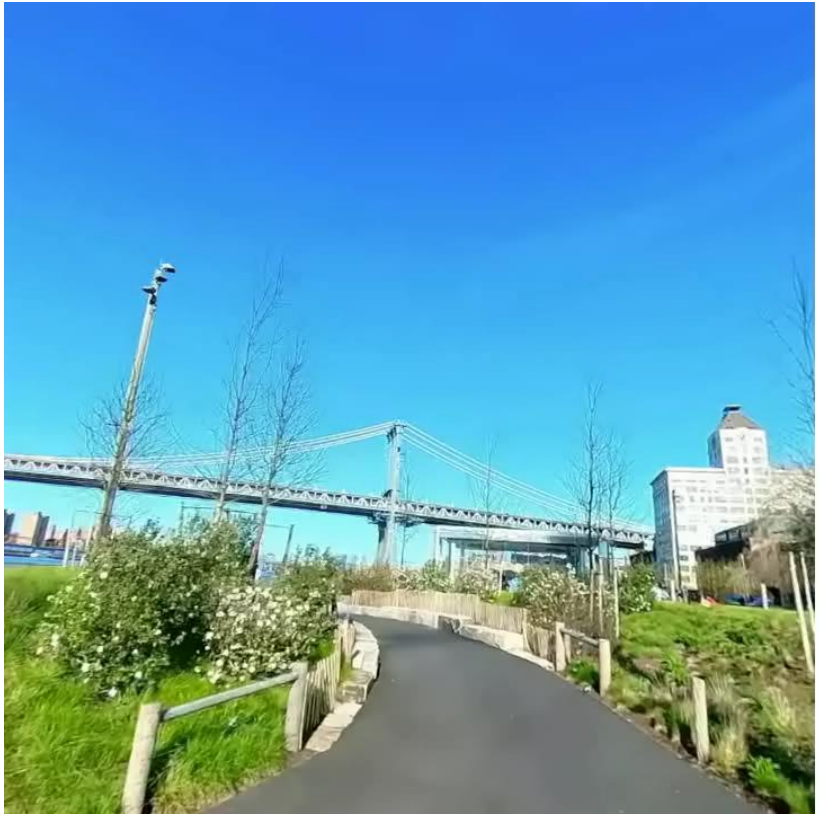
Synthesized view

Translate: Back 20m, Right 15m
Rotate: 15° left

Synthesized view

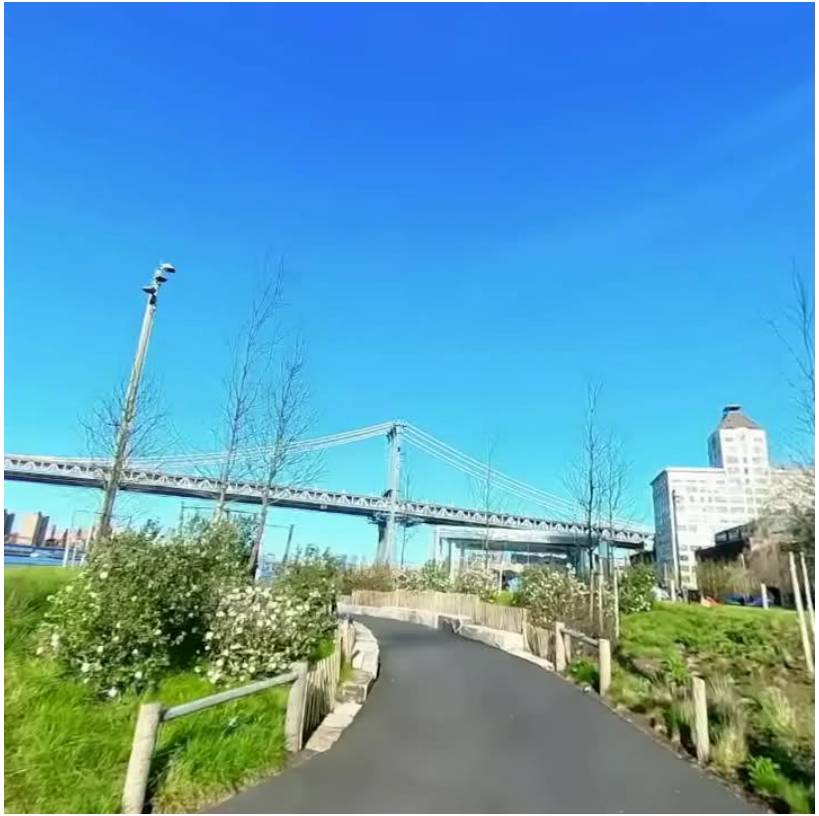
Translate: Right 20m
Rotate: 5° left

Video as 3D Data



Challenges of Learning from Video

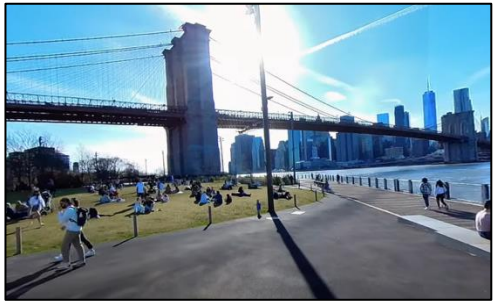
- Diverse multi-view data is sparse in each video
- Difficult to find corresponding frames for diverse multi-view data
- Traditional SfM methods require many frames and are expensive



We usually don’t get the views we want

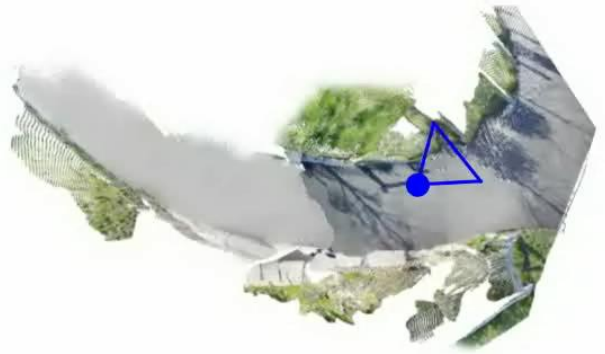


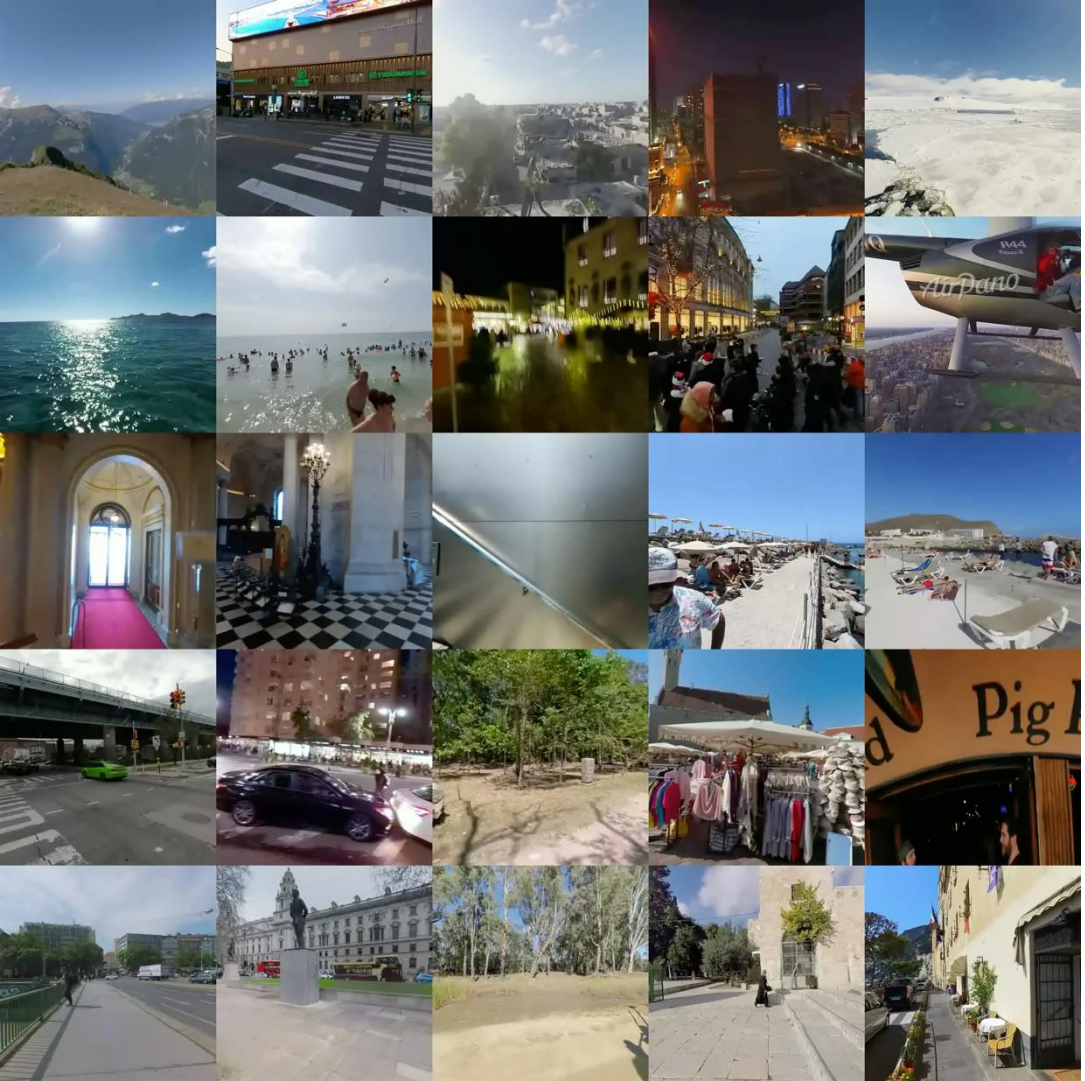




Exploring the World with 360˚ Video

Bird's-Eye View



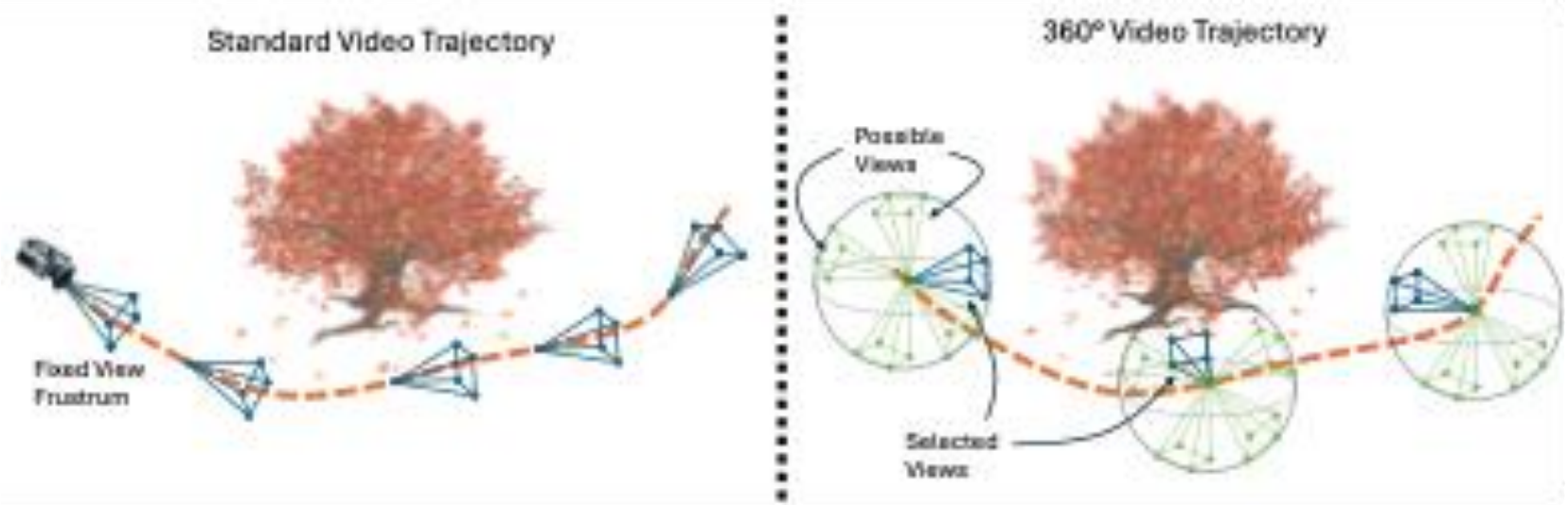


360-1M

Collected 1 million 360° Videos spanning 15 diverse categories

Wallingford et al. (NeurIPS 2024)

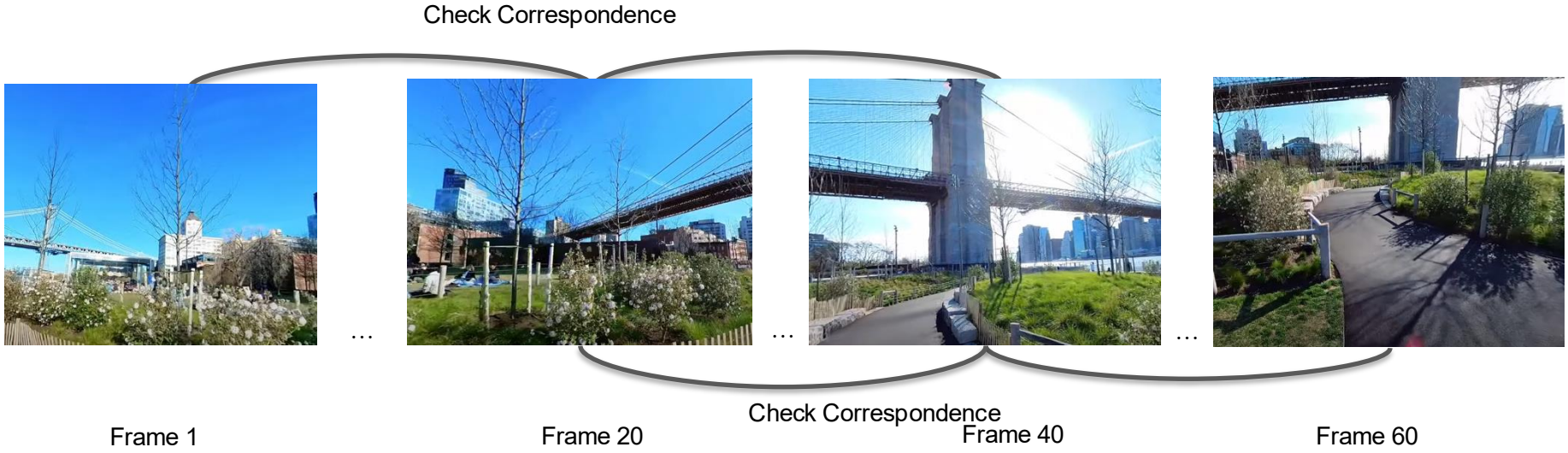
360˚ Video to Multi-View Data



Difficult to find corresponding views

Can freely rotate the camera view to find overlapping content

Correspondence Search



- Check for corresponding frames within a sliding window (40 frames in this example)
- $O(K^2)$ where K is the size of the sliding window
- Searching over entire video for correspondence is intractable

Searching with 3D Correspondence

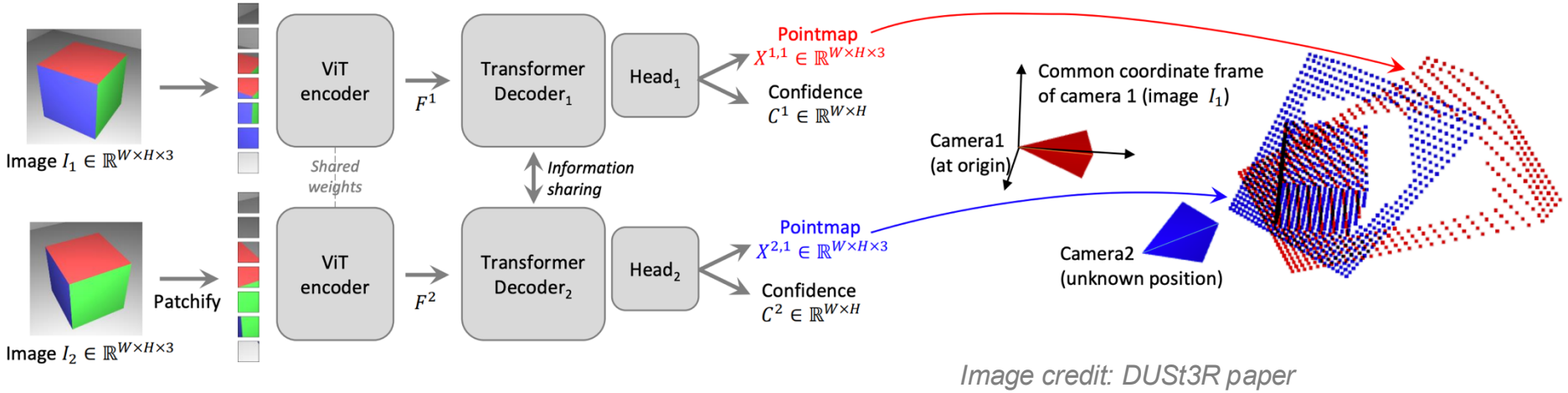


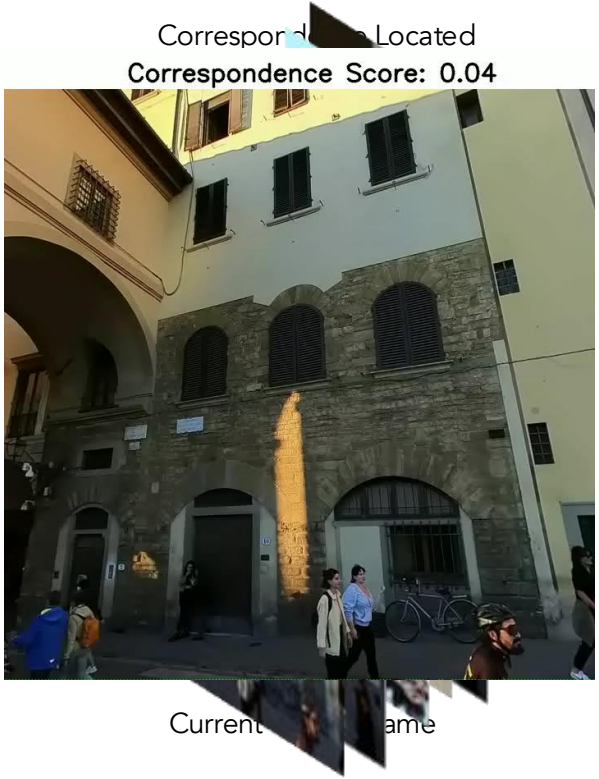
Image credit: DUSt3R paper

$$\text{Confidence Score} = \frac{1}{HW} \sum_i \sum_j C_{ij}$$

We perform gradient descent with respect to camera viewing angle to maximize the confidence score and overlapping content between frames.

Correspondence Search

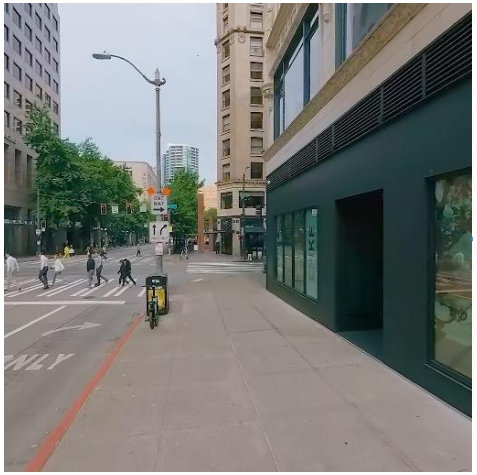
Initial Frame

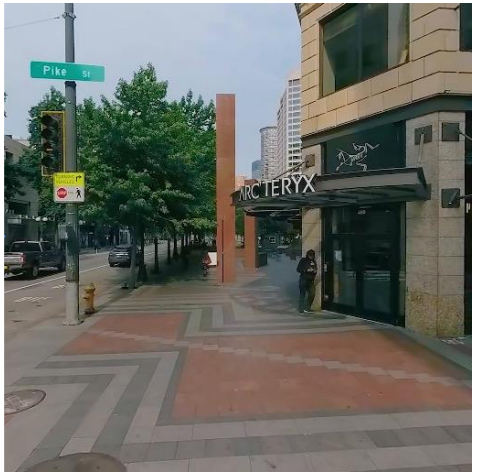


Current Frame

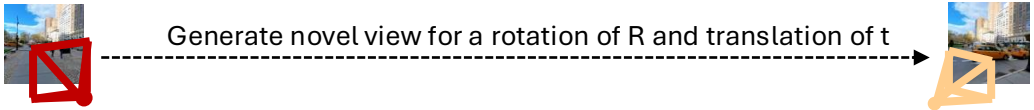
Query Frames

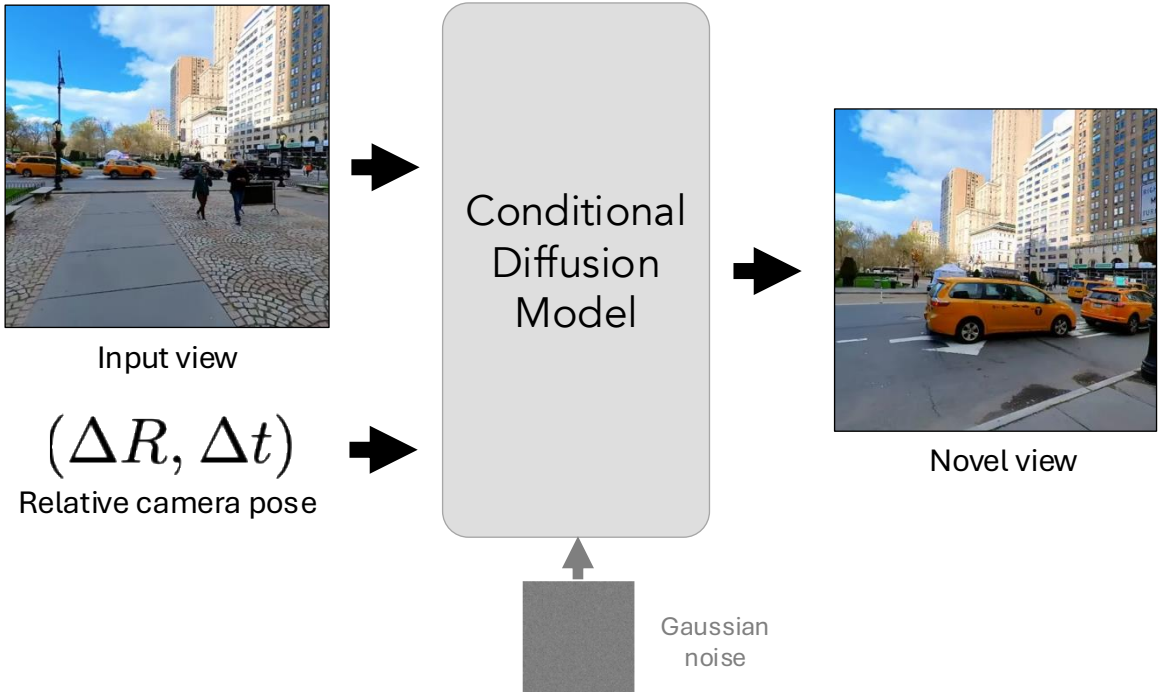
Frame Correspondence Example





ODIN





Wallingford et al. (NeurIPS 2024)

Learning from Dynamic Scenes

- Novel view synthesis objective assumes static scenes.
- Most videos have dynamic elements.
- Leads to unstable training and object flickering.

Learning from Dynamic Scenes

$$\mathcal{L} = \|(\epsilon - \epsilon_\theta(z_t, t, f_\theta(x, R, t))) \cdot M\|_2^2$$

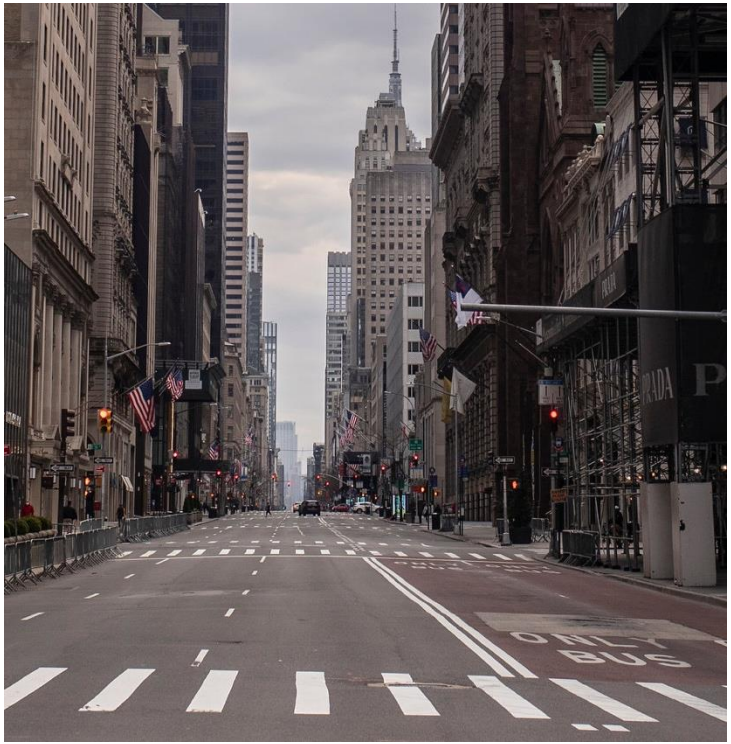
$$\mathcal{L}_{\text{auxiliary}} = -\lambda \sum_{i,j} M_{ij}$$

M is the temporal masking predicted by the decoder during training to mask the loss

- Network predicts regions of the scene where objects may move
- Mask regions of movement from the loss function
- Regularize size of the mask as auxiliary loss

Large-Scale Real-World Scene Generation

Input Image



Generated Trajectory

Generating Long-Range Fly-Through

Input View

Generated Fly-Through

Generating Indoor Scenes

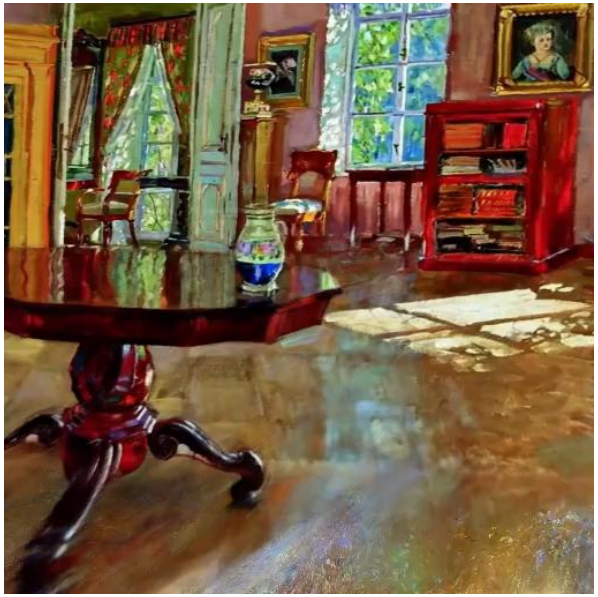
Input Image

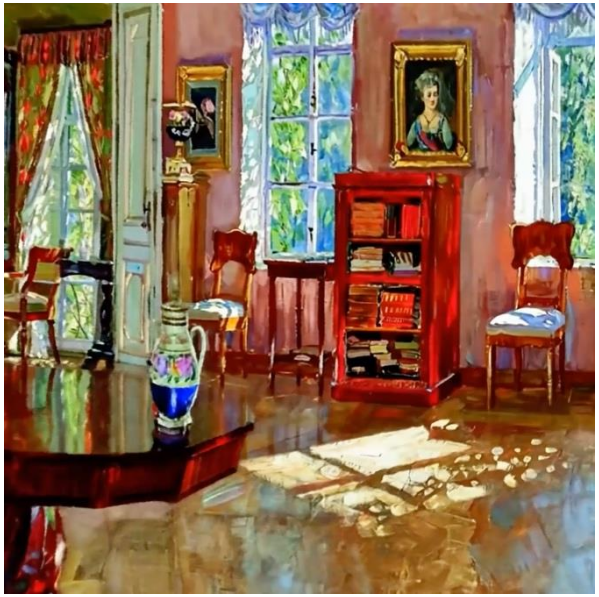
Generated Fly-Throughs

Out-of-Domain Generalization

Input Image

Generated Fly-Throughs

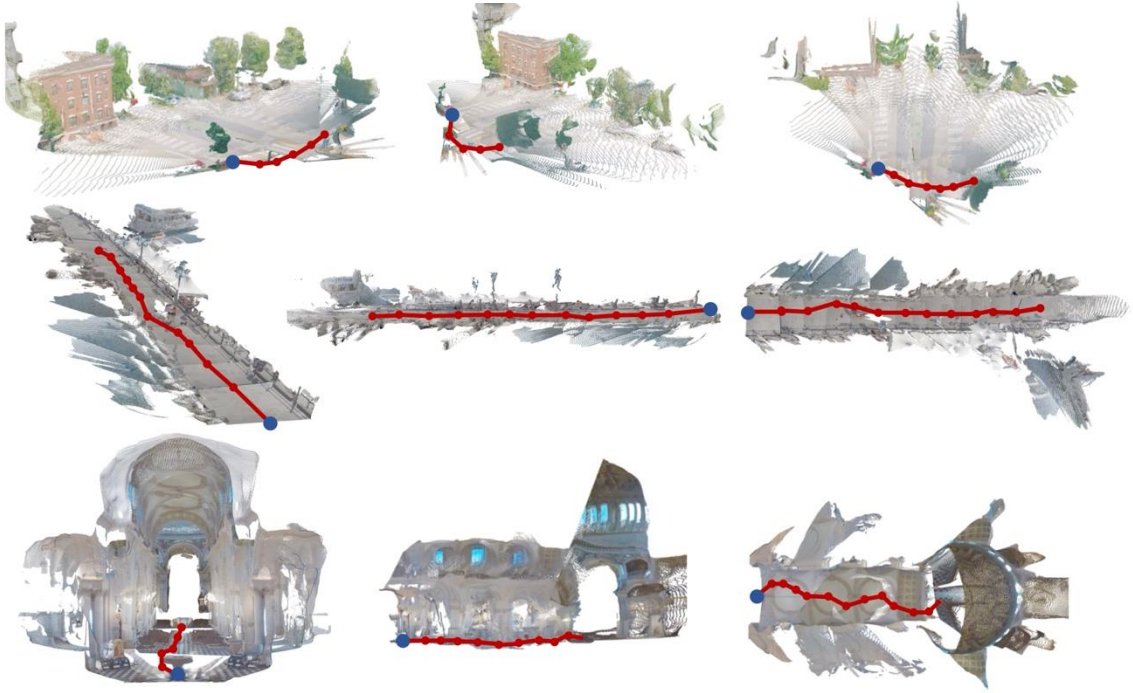




Estimating Scene Geometry from an Image

Single Input Images

Reconstructed Scene Geometry



Given a single image and a trajectory, imagine the geometry and appearance of the scene

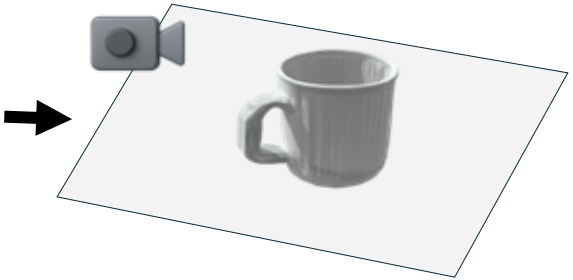
ODIN

3D prediction for navigation

ObjectForesight

3D prediction for manipulation

Input video



Estimate the current 6D pose

*Predict the **future** 6D pose*

Input (context)

$\mathcal{I}_{1:C}$ images

$$\mathbf{P}_{1:C} = [\mathbf{p}_1, \ldots, \mathbf{p}_C], \ \mathbf{p}_t \in \mathrm{SE}(3)$$

Context size — Object pose

Output (future prediction)

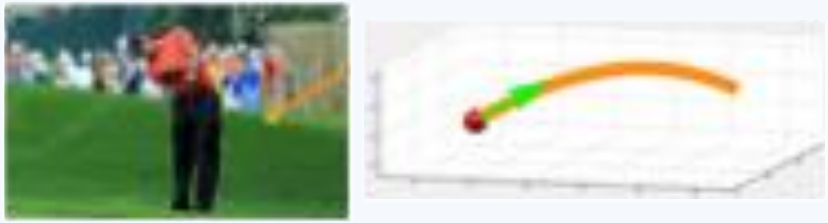
$$\mathbf{P}_{\text{future}} = [\mathbf{p}_{t_a}, \ldots, \mathbf{p}_{t_a+H-1}], \quad t_a = C+1.$$

H steps into the future

ObjectForesight, R. Soraki, H. Bhardhwaj, A. Farhadi, R. Mottaghi, 2026.

Examples of early work in future prediction

Newtonian Image Understanding, 2015



Predicting the future motion of the object from a single image.

"What happens if"…, 2016

Predicting counterfactuals about future motion in 3D.

Current “in-the-wild” indoor 3D datasets with dynamic interactions are limited

Small scale

Most datasets contain a small number of scenes or interactions.

Limited diversity

Many datasets are captured in controlled environment that do not reflect the diversity of the real-world settings.

Datasets such as HOT3D (Banerjee et al., 2024) are good, but not sufficient

We create a new in-the-wild dataset with more than 2 million clips of object interaction with 3D annotations based on EpicKitchens.

Action segment filtering

Select annotated single-activity segments and discard clips longer than 10 seconds

Action segment filtering

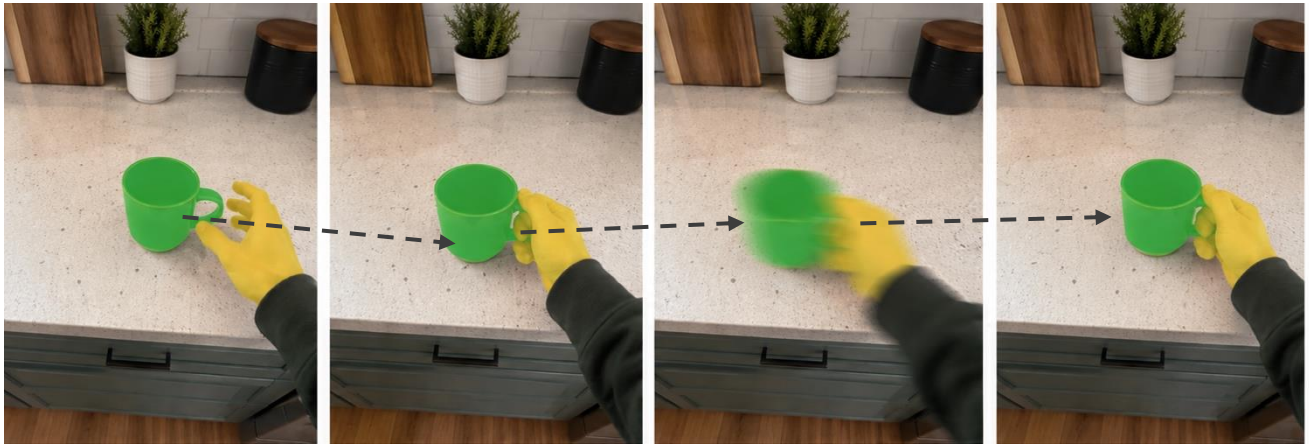
Hand–object discovery

Using EgoHOS (Zhang et al., 2022), frames without hands or without any object hypotheses are removed

Action segment filtering

Hand–object discovery

Robust object masks



Using SAM2 (Ravi et al., 2024), we robustly track the object across frames.

Action segment filtering

Hand–object discovery

Robust object masks

VLM gating

Using a VLM to: (1) discard objects that don’t move, (2) remove blurry images or occluded objects.

Action segment filtering

Hand–object discovery

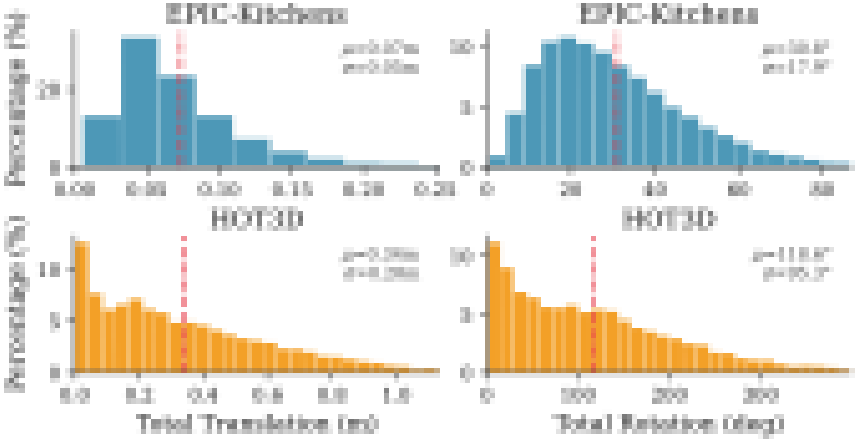
Robust object masks

VLM gating

Object 3D reconstruction

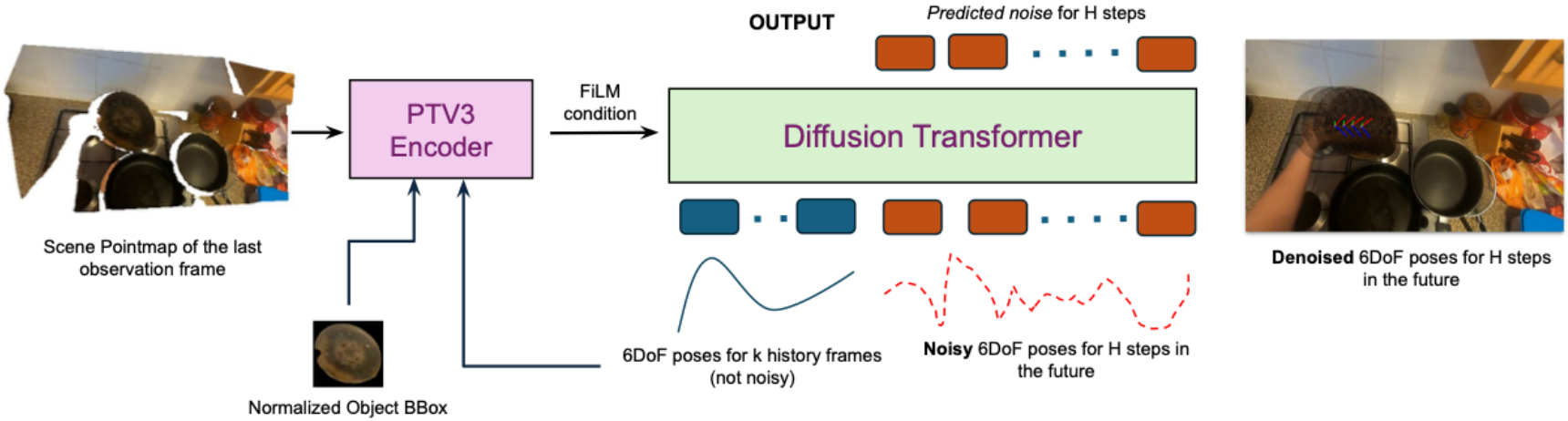
Using Trellis (Xiang et al., 2025) to create the 3D mesh from clean views.

2+ million clips with 6D annotations



Step Name	Number
Action Segments	76,885 vids
Selected Vids (hands, $\leq$ 10s)	72,046 vids
SAM2 Tracks	229,102 tracks
Filtered Tracks	112,057 tracks
TRELLIS Models	71,296 models
Objects with Pose Tracks	39,174 tracks
Pre-Filter Trajectories	3,065,568 trajs
Post-Filter Trajectories	2,073,109 trajs

Model architecture



Losses

Standard diffusion objective (v-parameterization)

$$\mathcal{L}_v = \mathbb{E}\left[w(\tau)\left\|\mathbf{v}_\theta(\bar{\mathbf{Y}}_\tau, \tau) - \mathbf{v}_\tau\right\|_2^2\right]$$

- Denoising objective for diffusion model
- $\mathbf{v}_\theta$: predicted target
- $\mathbf{v}_\tau$: groundtruth target
- $w(\tau)$: timestep-dependent weight

Direct supervision of future frames

$$\mathcal{L}_{\text{aux}} = \mathbb{E}\left[\bar{\alpha}_\tau\left(\lambda_R\,\bar{d}_{\text{geo}} + \lambda_{\text{trans}}\,\bar{e}_{\text{trans}}\right)\right]$$

- Auxiliary loss on predicted future poses
- $\bar{d}_{\text{geo}}$: geodesic rotation error
- $\bar{e}_{\text{trans}}$: translation error

Encouraging smooth trajectories

$$\mathcal{L}_{\text{vel}} = \overline{\|\Delta\mathbf{t}_k - \Delta\hat{\mathbf{t}}_k\|_2^2 + \mathrm{d}_{\text{geo}}(\Delta\mathbf{R}_k, \Delta\hat{\mathbf{R}}_k)^2}$$

- Penalizes inconsistent velocity
- $\Delta\mathbf{t}_k, \Delta\mathbf{R}_k$: target linear / angular velocities
- $\Delta\hat{\mathbf{t}}_k, \Delta\hat{\mathbf{R}}_k$: predicted velocities

$$\mathcal{L}_{\text{acc}} = \overline{\|\Delta^2\mathbf{t}_k - \Delta^2\hat{\mathbf{t}}_k\|_2^2 + \mathrm{d}_{\text{geo}}(\Delta^2\mathbf{R}_k, \Delta^2\hat{\mathbf{R}}_k)^2}$$

- Penalizes inconsistent accelerations
- $\Delta^2\mathbf{t}_k, \Delta^2\mathbf{R}_k$: target linear / angular acc.
- $\Delta^2\hat{\mathbf{t}}_k, \Delta^2\hat{\mathbf{R}}_k$: predicted accelerations

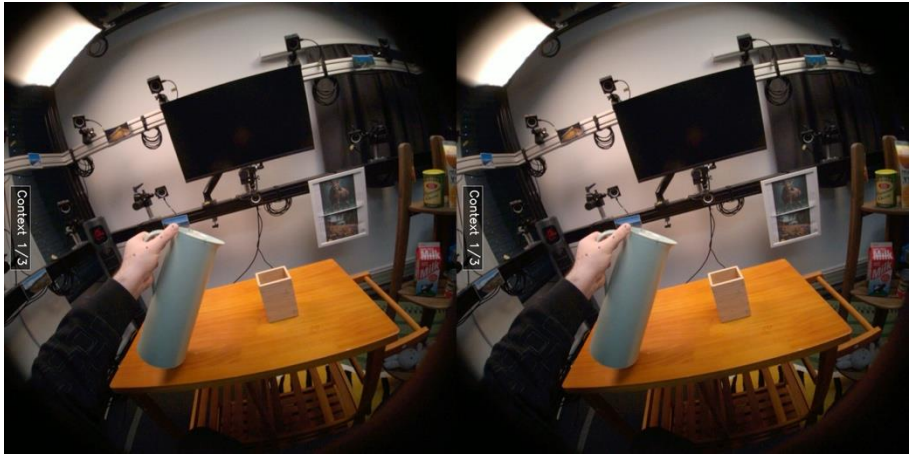
Preventing degenerate small depth generations

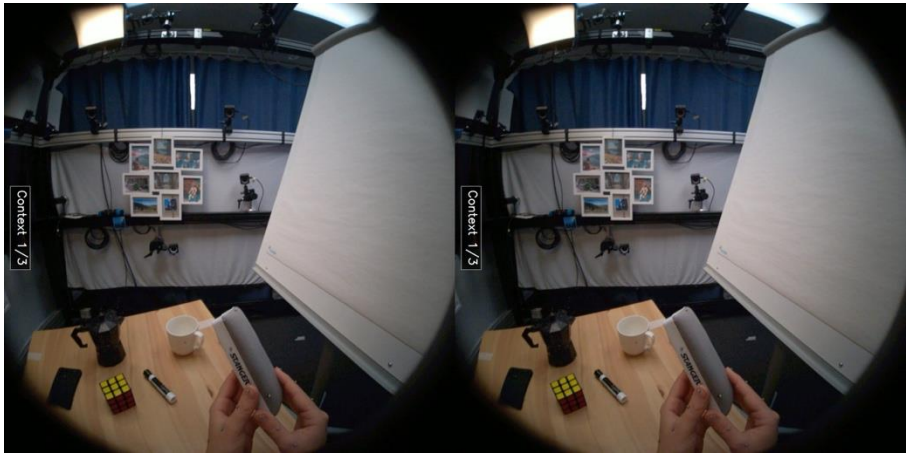
$$\mathcal{L}_{z_{\min}} = 0.01\,\mathrm{ReLU}(z_{\min} - \hat{z}_t)$$

- Encourages the predicted depth $\hat{z}_t$ to be above minimum $z_{\min}$
- Prevents collapse to unrealistically small depth

Total loss

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{v}} + \mathcal{L}_{\text{aux}} + \mathcal{L}_{z_{\min}} + \lambda_{\text{vel}}\,\mathcal{L}_{\text{vel}} + \lambda_{\text{acc}}\,\mathcal{L}_{\text{acc}}$$





Left: groundtruth | Right: prediction

Dataset	Method	ADE ↓	FDE ↓	DES ↓	ARE ↓	FRE ↓	RES ↓
Epic-Kitchens	Constant Velocity	0.027	0.053	0.007	2.47°	5.60°	0.80°
	ObjectForesight-AR	0.067	0.074	**0.002**	9.48°	12.58°	0.93°
	ObjectForesight-DiT	**0.016**	**0.029**	0.004	**2.30°**	**4.82°**	**0.66°**
	vs. Video Generation						
	Luma AI Ray3	0.084	0.149	0.020	12.86°	20.90°	2.62°
	ObjectForesight-DiT	**0.029**	**0.059**	**0.008**	**7.29°**	**13.98°**	**1.77°**
HOT3D-Clips	Constant Velocity	0.136	0.280	0.040	38.70°	68.53°	9.85°
	ObjectForesight-AR	0.055	0.082	0.007	9.80°	14.95°	1.55°
	ObjectForesight-DiT	**0.021**	**0.026**	**0.003**	**8.92°**	**12.58°**	**1.16°**

ADE: average displacement error across all timesteps

FDE: displacement error at the final timestep

DES: displacement error slope

ARE: average rotation error

FRE: final rotation error

RES: rotation error slope

Lower is better for all metrics.

Conclusion

Future prediction and 3D world modeling are foundational capabilities for Physical AI.

Current foundation models remain heavily biased toward 2D perception. While repurposing 2D models for 3D tasks is possible, it is not ideal.

The next generation of models should unify scene-level and object-centric prediction.

Thank you !