Northeastern University
Electrical and Computer Engineering Department
Augmented Cognition Laboratory (ACLab)
ostadabbas.sites.northeastern.edu

Physical AI Research (PAIR) Initiative

PanoWorld: Geometry-Consistent Panoramic Video World Modeling

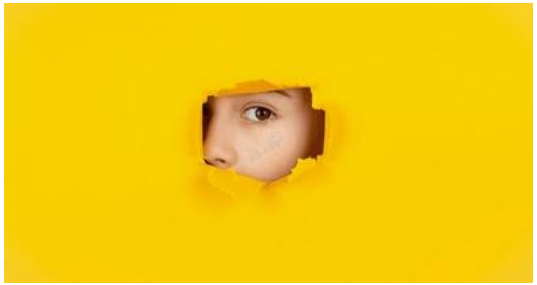
Presenter: Le Jiang
Electrical and Computer Engineering
Northeastern University

The Problem in One Sentence

A perspective camera sees only 10% of the world

Field-of-View Constraint

- Standard video world models (Cosmos, SVD, CogVideo, Lumiere) all operate on perspective frustums — at any moment they observe only $\Omega_t \subset \mathbb{S}^2$ a narrow cone of the viewing sphere.
- A typical 90° FOV camera captures **~10%** of the surrounding scene; the remaining **~90%** of the world is unmodelled.



Information-Poor!

Why this matters?

Everything breaks at the edge of the frame

- 🏟️ Application: Reconstruction & VR/AR need a closed full sphere; perspective generators can never close it.
- 🤸 Long-trajectory tracking: when motion diverges, one view cannot follow every direction at once.
- 🤖 Multi-agent scenes: objects outside the frustum are lost; multi-camera rigs only add fragile sync.

Tracking View



Static View



Panoramic View



Related Work: Limitations & Where PanoWorld Fits

Others miss one axis; we cover all three

Marble



✅ 360° view, ❌ Dynamic, ✅ Geometric Grounding

Nvidia Cosmos



❌ 360° view, ✅ Dynamic, ✅ Geometric Grounding

Panoramic Generators

360DVD | Imagine360 | ARGUS



✅ 360° view, ✅ Dynamic, ❌ Geometric Grounding

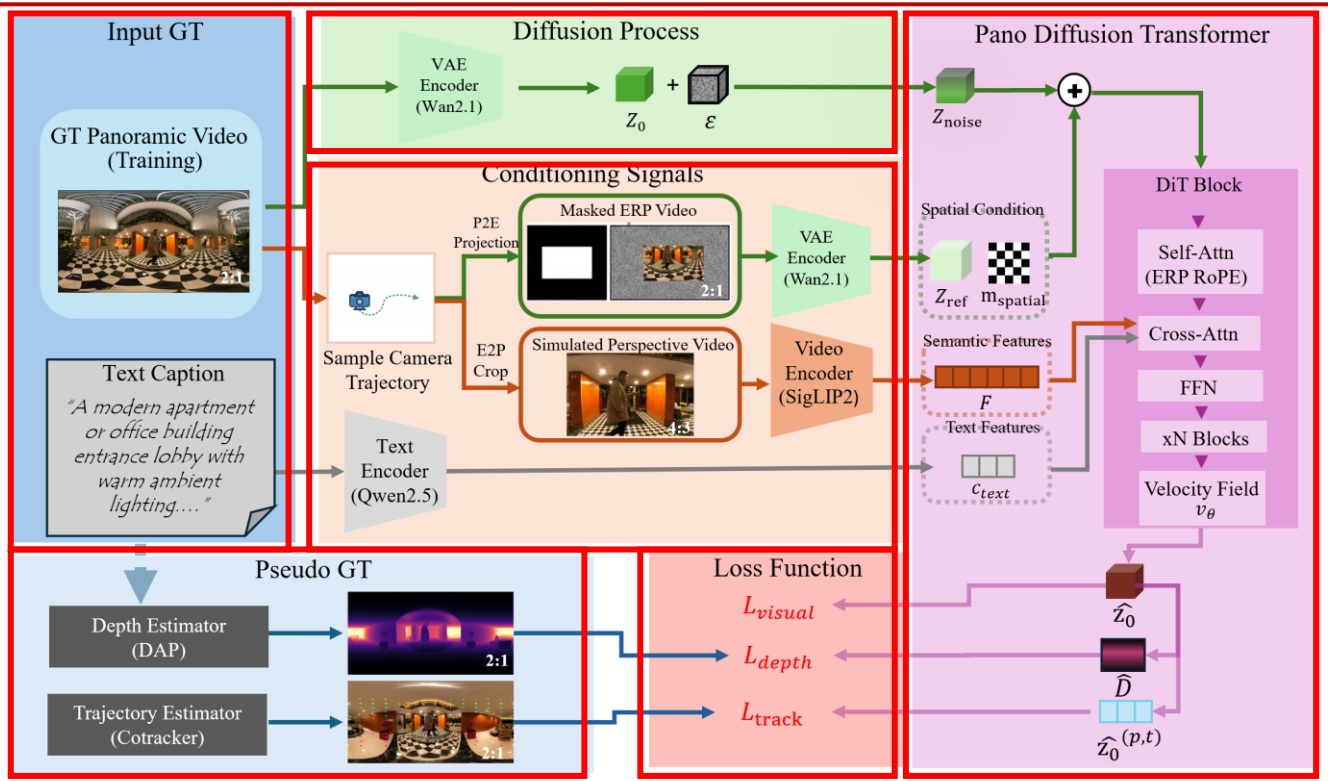
PanoWorld



✅ 360° view, ✅ Dynamic, ✅ Geometric Grounding

PanoWorld Model Structure

Panoramic diffusion with geometry-aware losses



Equirectangular Projection (ERP)

Teaching a transformer to speak the sphere

Why standard Transformers break on ERP?

Equirectangular projection maps longitude $\lambda \in [-\pi, \pi]$and latitude $\phi \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ onto a non-uniform 2D image. Pole rows are **stretched by** $\frac{1}{\cos\phi}$, so a linear positional embedding implicitly treats unequal angular spacings as equal, and a pixel-uniform loss over-represents polar content.

Modification	Linear (broken)	Our adaptation
Latitude-aware RoPE position	$\text{pos}_h = h$ — equal pixel ≡ equal angle	$\text{pos}_h = \frac{H-1}{2}\left(\sin\left(\frac{\pi h}{H-1} - \frac{\pi}{2}\right) + 1\right)$ — compresses near poles; longitude stays linear because ERP preserves uniform λ spacing
Spherical area weighting $w_h = \cos\phi\ (h)$	Equal pixel weight	Pixel-space losses ($\mathcal{L}_{\text{depth}}$,$\mathcal{L}_{\text{track}}$ (weighted by $\cos\phi \approx$ solid angle $d\Omega$
Circular shift augmentation	$\lambda = -\pi$ and $\lambda = \pi$ treated as maximally distant	Random horizontal roll exposes all boundary alignments → seamless wrap-around at inference

Geometry-Aware Training Objectives

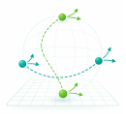
Depth and trajectory consistency, nearly for free



L_{depth} Depth Consistency

$$h_d(\hat{z_0}) = \hat{D}, \quad \hat{D} \in [0,1]$$
$$L_{\text{depth}} = c(\sigma) \cdot \text{Trim}_{98\%}\left(w_h \cdot \left|\hat{D} - D^{gt}\right| + L_{\text{edge}}\right),$$
$$w_h = \cos\phi(\text{h})$$

- L1 against DAP pseudo-GT (both pre-normalized to [0,1])
- Edge-preserving gradient term sharpens object boundaries
- Adds only a tiny depth head to the DiT graph



L_{track} Trajectory Consistency

$$\mathbf{X}_{p,t}^{\text{pred}} = \mathbf{R}_t^{\top}\left(\pi_{\mathbb{S}^2}\left(\mathbf{u}_{p,t}, \hat{D}(\mathbf{u}_{p,t})\right) - \mathbf{t}_t\right),$$
$$\xi_{p,t} = [X,\ \alpha\Delta X,\ \beta\Delta^2 X] \qquad (\alpha = 0.5,\ \beta = 0.25),$$
$$L_{\text{track}} = c(\sigma)\sum_{p,t} w_{p,t}\left\|\xi_{p,t}^{\text{pred}} - \xi_{p,t}^{gt}\right\|_1 \quad w_{p,t} = v_{p,t}\cos\phi_{p,t}$$

- Lift 2D track to the world frame via predicted depth
- Motion descriptor (position + velocity + acceleration):
- CoTracker3 dense grid + visibility mask

Both act on the clean latent $\hat{z}_0 = z_\sigma - \sigma \cdot v_\theta$,

faded in by confidence $c(\sigma)$ · area-weighted by $\cos\phi$

Qualitative Comparison

Riverside

Input (perspective image)



A leafy riverside setting with churning green water in the foreground and an old stone bridge or wall along the front. People in casual summer clothes stand on a gravel bank and wooden railings; a person in dark clothing appears in the surf. Bright daylight filters through tall green trees.



PanoWorld



OmniRoam



Argus



Imagine360



Follow-Your-Canvas



DVD360

Quantitative Comparison

Top two on every metric; best on all in Stage 2

Method	Visual Realism			Caption	Geometric Self-Consistency		
	FVD↓	FAED↓	FID↓	CLIP-T↑	3D-Smooth↓	Depth-σ ↓	Tr-Life↑
Stage 1: single perspective image + caption							
360DVD [Wang et al., 2024]	100.7	94.6	191.8	0.262	0.763	0.065	0.881
Imagine360 [Tan et al., 2025]	67.7	96.6	142.9	0.253	0.364	0.099	0.594
ARGUS [Luo et al., 2025]	65.0	**82.7**	152.5	0.212	**0.100**	0.053	**0.991**
Follow-Your-Canvas [Chen et al., 2025a]	82.2	108.8	182.3	**0.279**	0.463	0.140	0.600
OmniRoam [Liu et al., 2026]	98.8	131.2	178.4	0.132	1.063	0.280	0.200
PanoWorld(Ours)	**56.1**	83.4	**136.4**	0.264	0.101	**0.033**	0.977
Stage 2: single ERP first frame + caption (frame-0 anchored)							
OmniRoam [Liu et al., 2026]	39.7	53.6	69.7	0.233	1.197	0.180	0.608
PanoWorld(Ours)	**28.5**	**41.5**	**55.4**	**0.253**	**0.511**	**0.060**	**0.776**

Controllability of PanoWorld

One photo, six prompts, full control



Base



Cherry Blossom



Autumn



Crowded



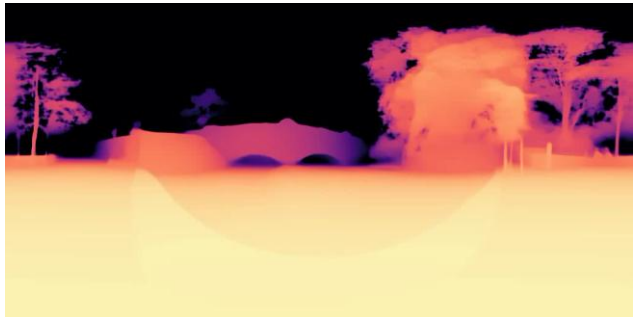
Camera Move



Swimmers

Controllability of PanoWorld

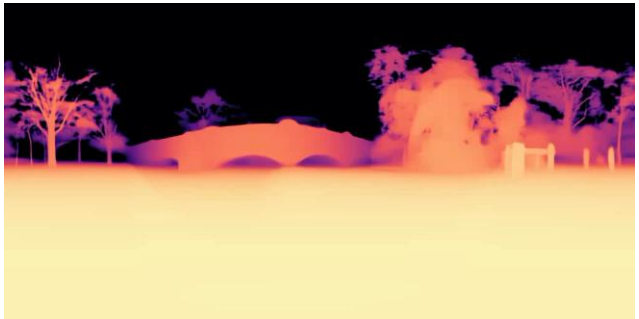
Appearance changes, geometry stays



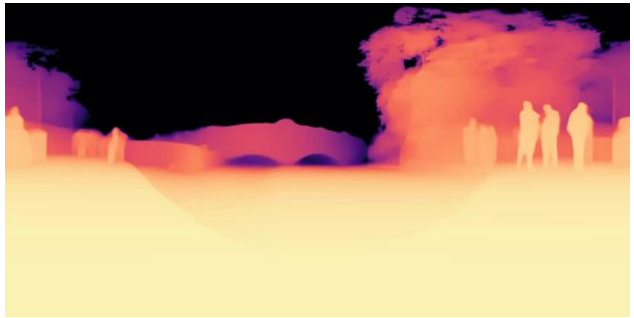
Base



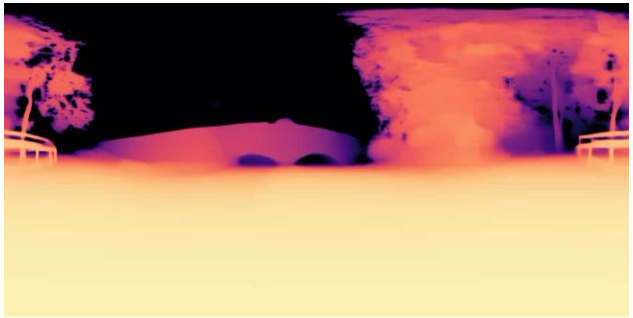
Cherry Blossom



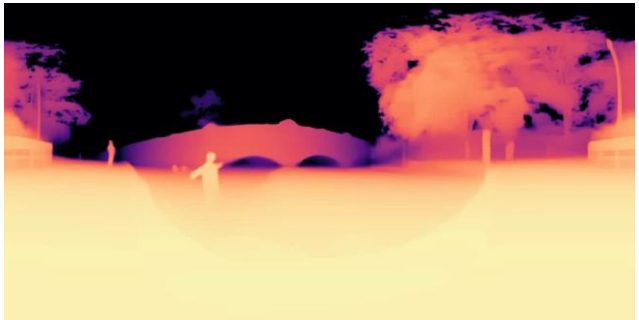
Autumn



Crowded



Camera Move



Swimmers

From One Image to an Explorable 3D Scene

Photo → panorama → 3D → free viewpoint



(a) Input Perspective Image

(b) Panoramic Video Generation

(c) 3D Reconstruction

(d) Novel View