Vision and Image Processing Research Group

VIP lab

UNIVERSITY OF WATERLOO

4D World Reconstruction of Humans, Scenes, and Camera Systems

Teaching machines to see, understand, and complete the world in motion.

Jerrin Bright

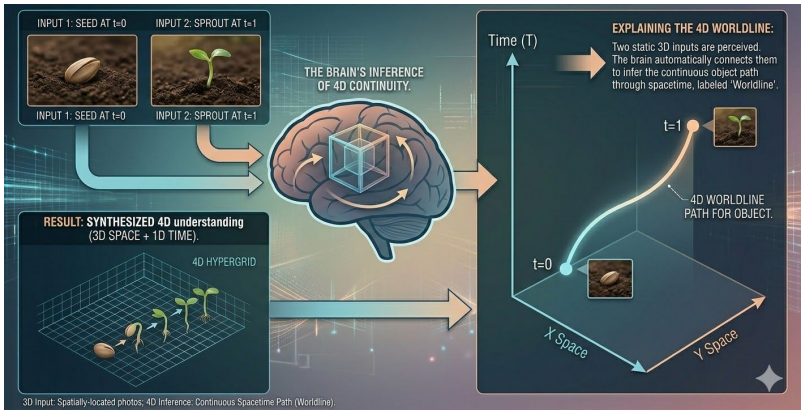
Supervisor: John Zelek

PhD Student, Systems Design Engineering
Vision and Image Processing Lab
University of Waterloo, Canada

Date: 08/04/26

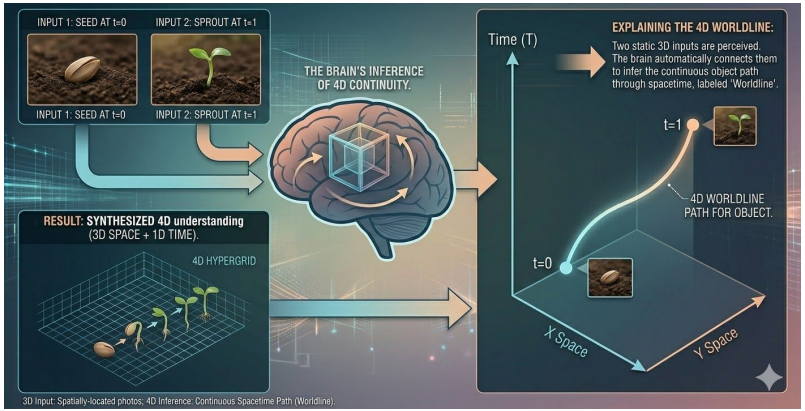
The Challenge of 4D World Understanding

- Humans understand the 4D world from just one or two images
- This remarkable ability stems from our capacity to draw on a lifetime of visual experiences.



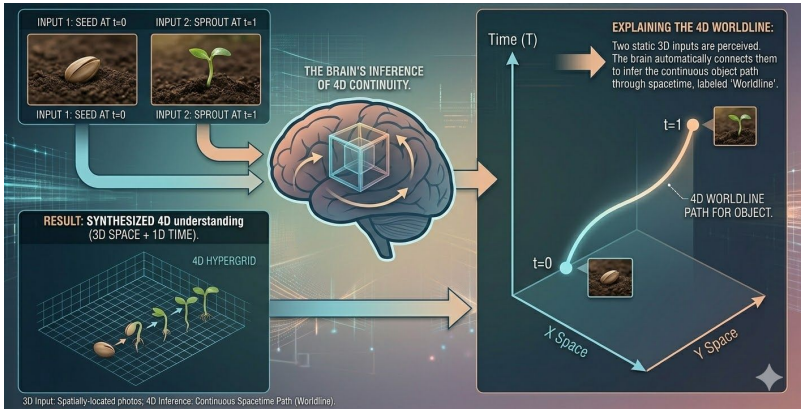
The Challenge of 4D World Understanding

- Humans understand the 4D world from just one or two images
- This remarkable ability stems from our capacity to draw on a lifetime of visual experiences.
- This suggests that human visual understanding is:
  - **Not static:** It evolves with every new observation
  - **Continuously updated:** Real-time perception adapts based on prior knowledge and context



The Challenge of 4D World Understanding

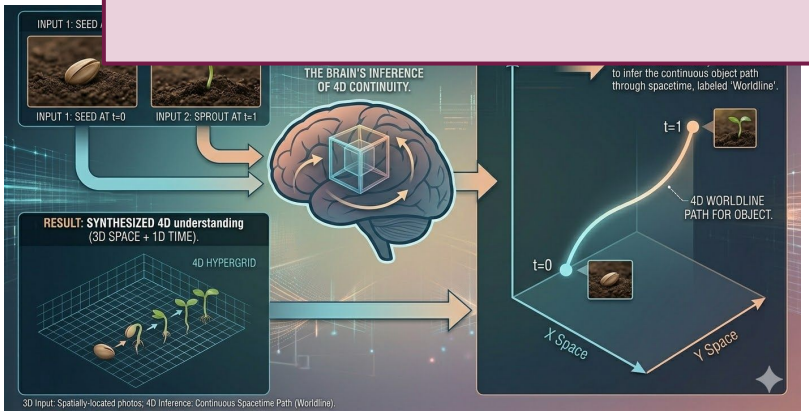
- Humans understand the 4D world from just one or two images
- This remarkable ability stems from our capacity to draw on a lifetime of visual experiences.
- This suggests that human visual understanding is:
  - **Not static:** It evolves with every new observation
  - **Continuously updated:** Real-time perception adapts based on prior knowledge and context



The Challenge of 4D World Understanding

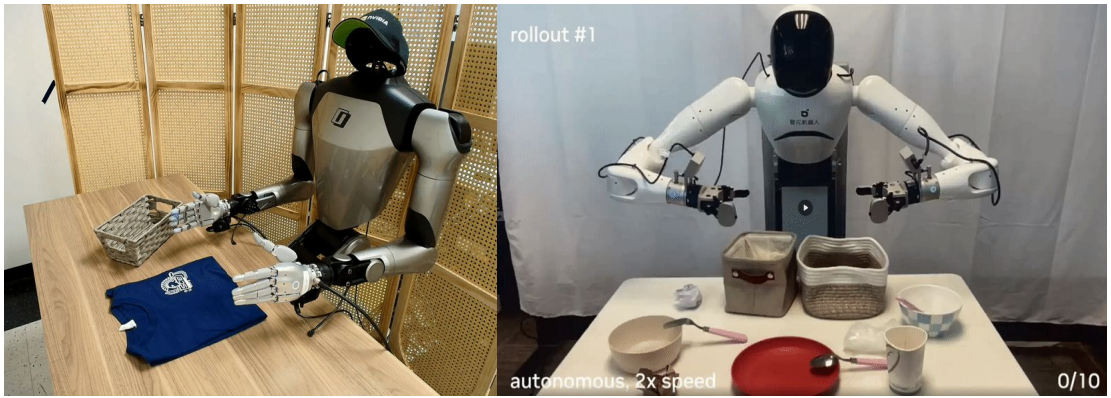
- Humans understand the 4D world from just one or two images
- This remarkable ability stems from our capacity to draw on a lifetime of visual experiences.
- This suggests that human visual understanding is:
  - **Not**
  - **Cont** …text

Can we emulate human ability to **continuously** reconstruct and update a **complete** 4D world from sparse observations?



Real-World Impact

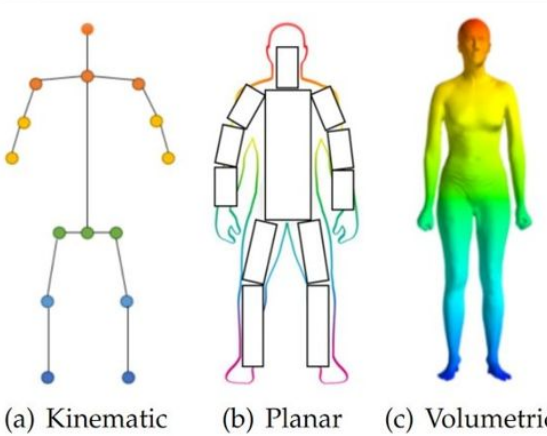
- Sports interactive experience + fan experience
- Embodied agents data layer
- AR/VR & Telepresence
- Film, VFX & Content Creation

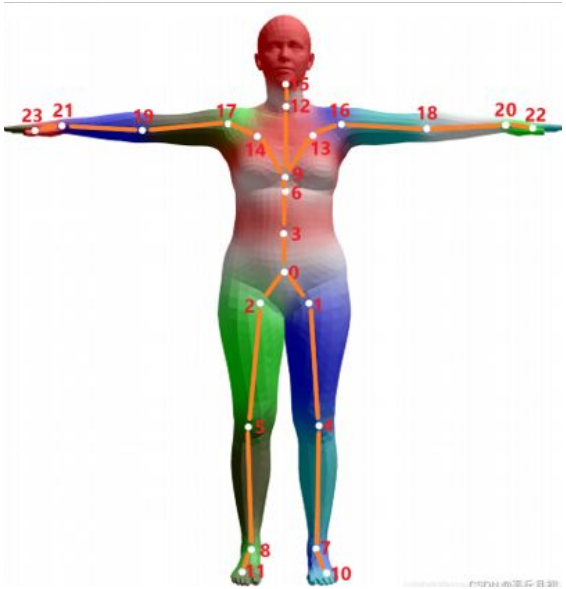


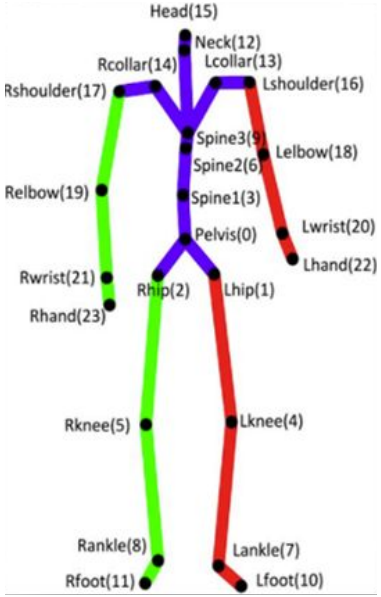
Prior Arts- Human

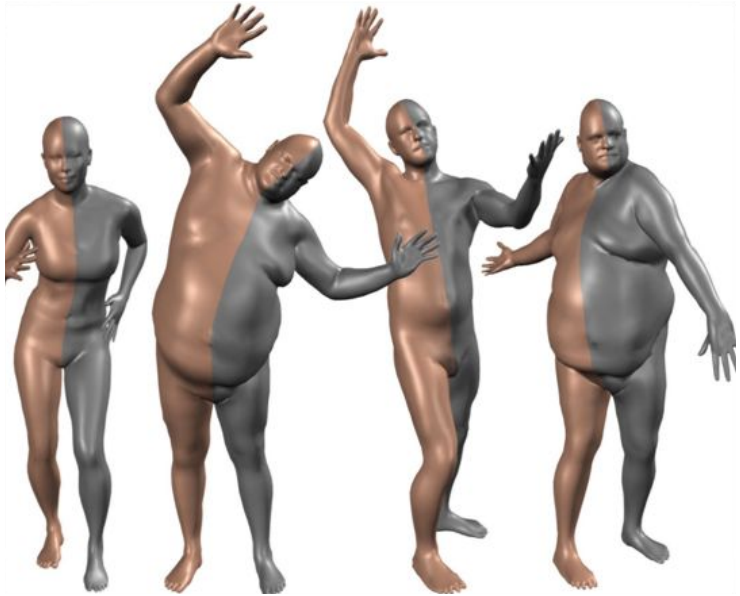
SMPL

- Skinned Multi-Person Linear model.
- 72 joint and 10 shape parameters -> 6890 vertices.

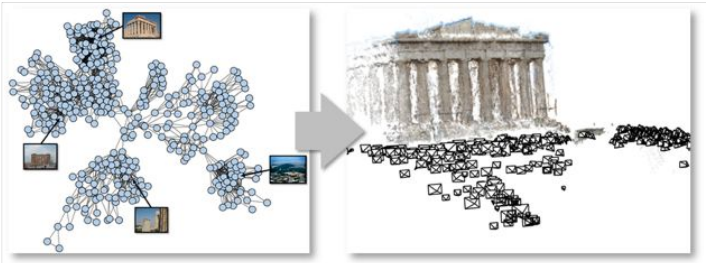




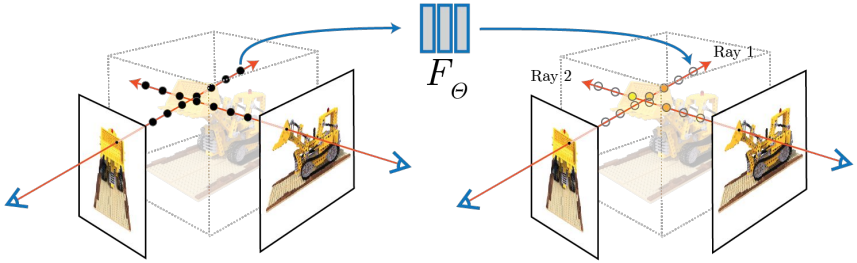




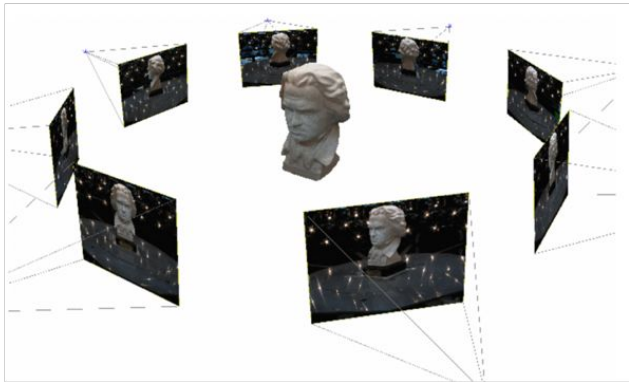
Prior Arts- Scene



Structure from Motion (SfM)



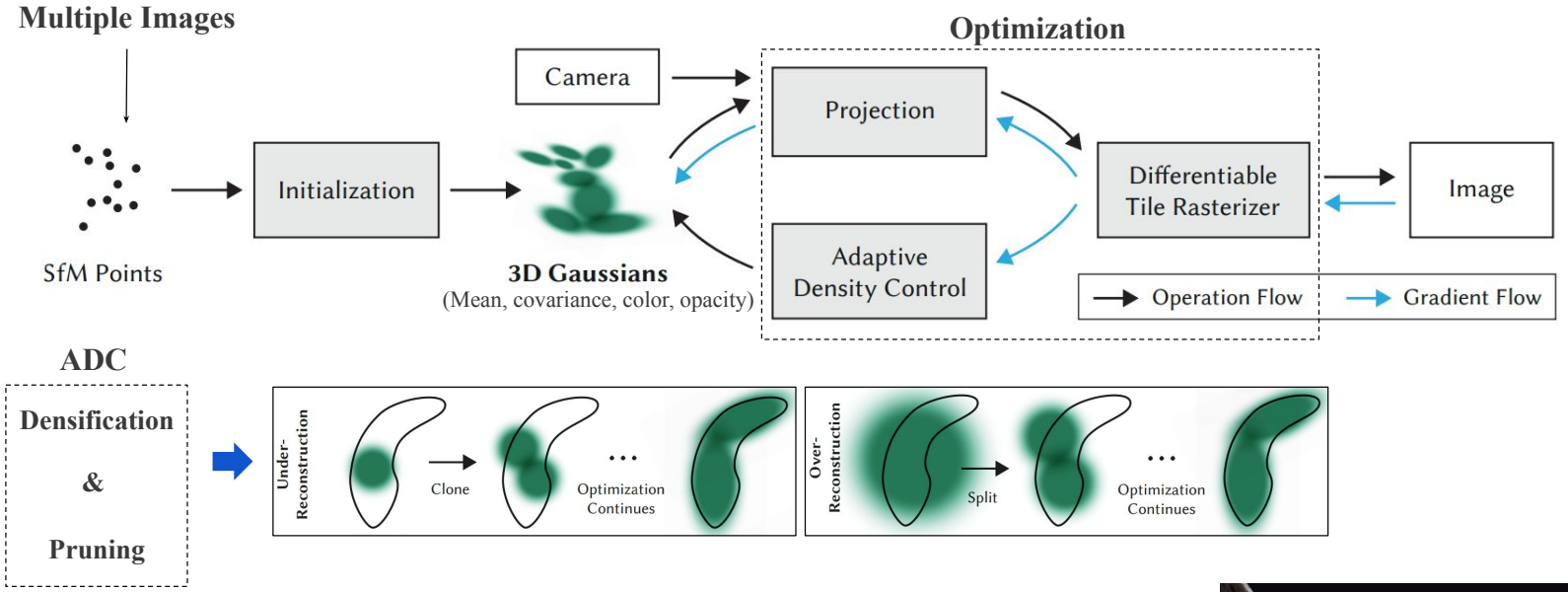
Neural Radiance Fields (NERF)



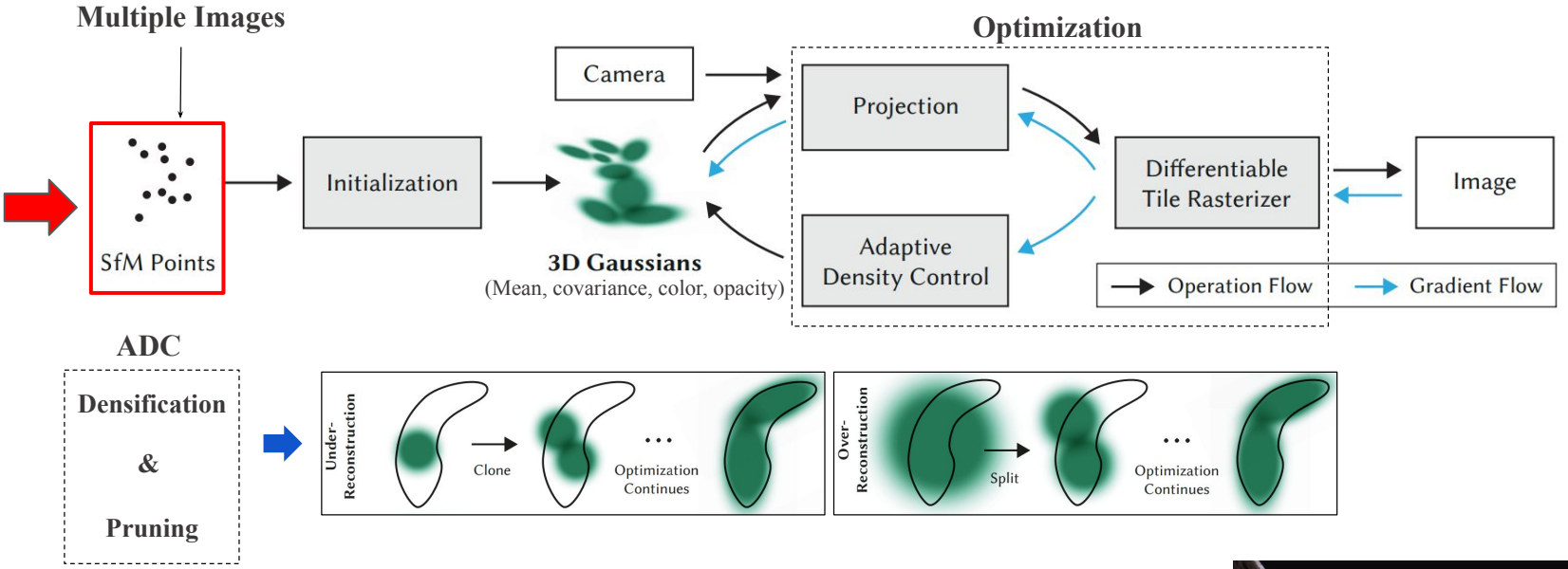
Multiview Stereo (MVS)

3D Gaussian Splatting

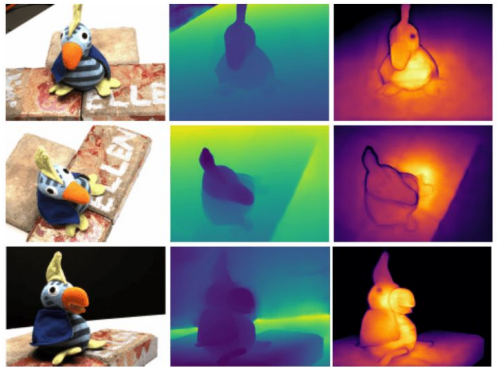
Prior Arts- Scene (3D Gaussian Splatting)

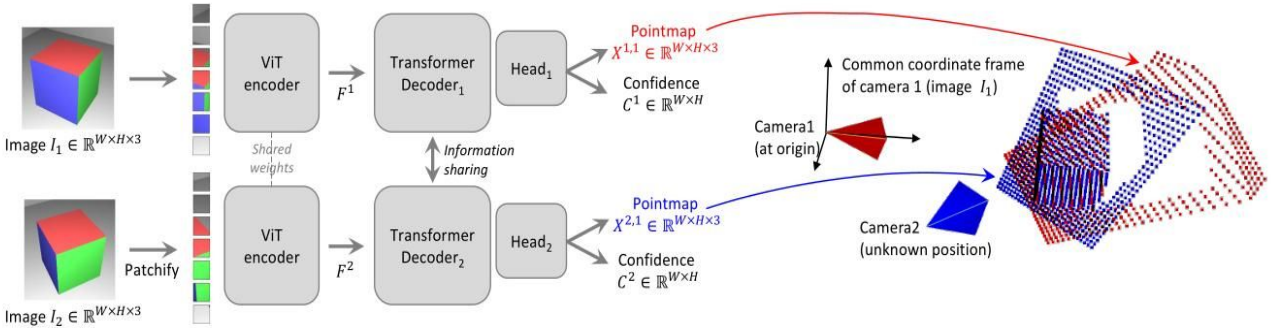


Prior Arts- Scene (3D Gaussian Splatting)



Prior Arts: Learning-based 3D Reconstruction

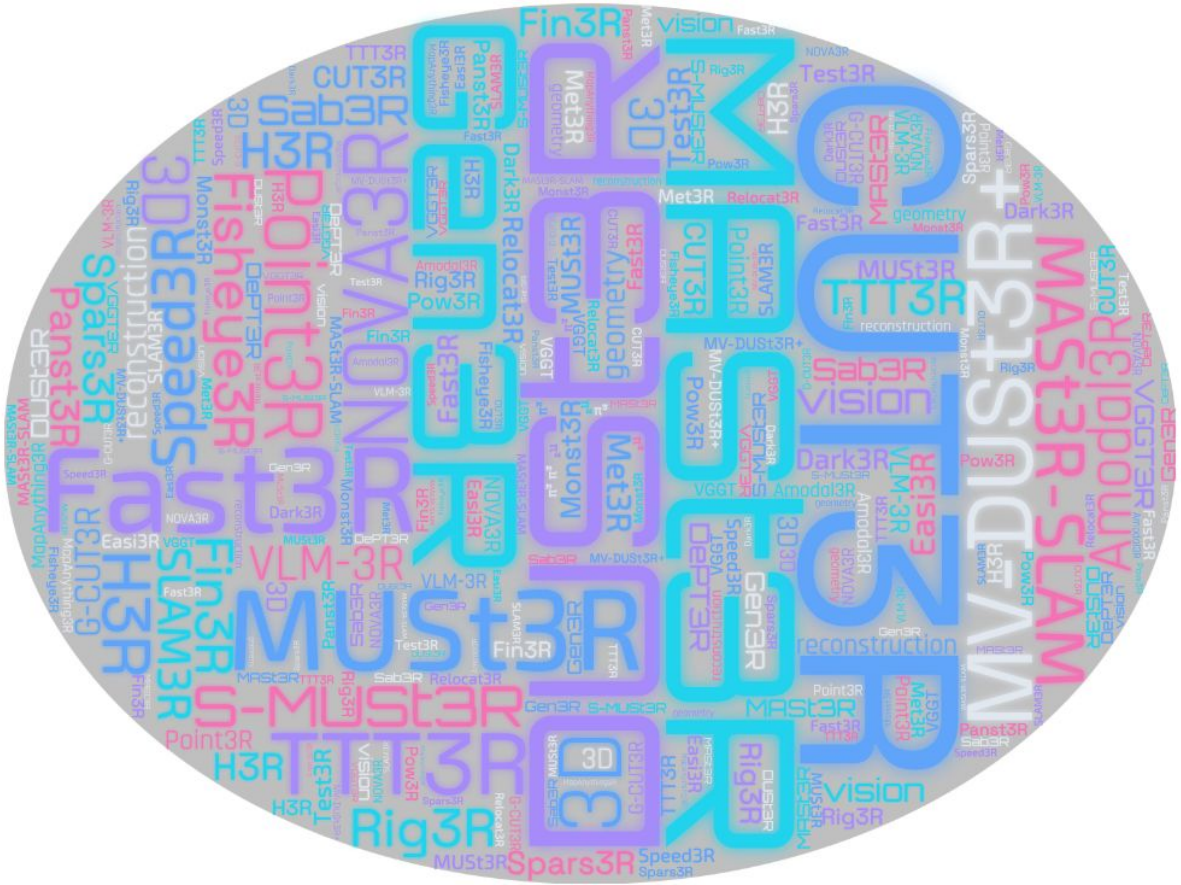




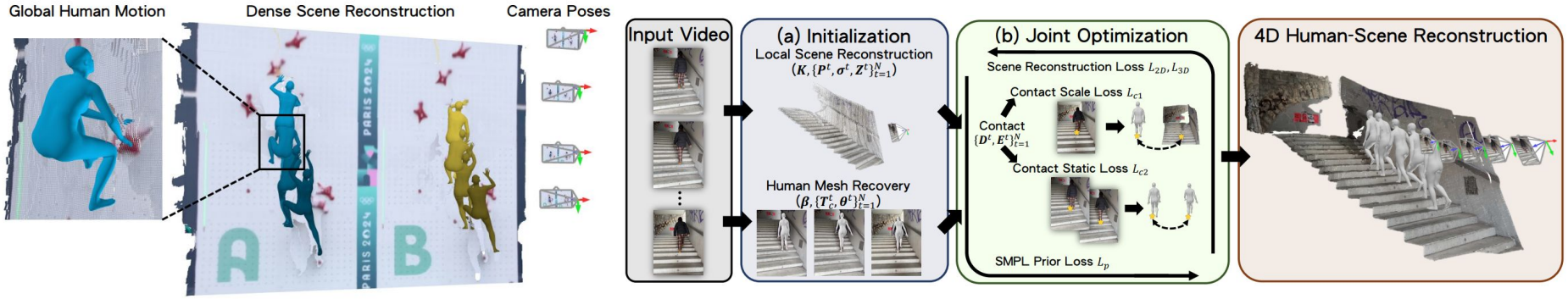
Dense Unconstrained Stereo 3D Reconstruction (Dust3R) [Wang et al. CVPR'24]

- Online, continuous update ✗
- Data-driven priors ✓

Prior Arts: Learning-based 3D Reconstruction



Prior Arts: Human-Scene Reconstruction (Opt.)

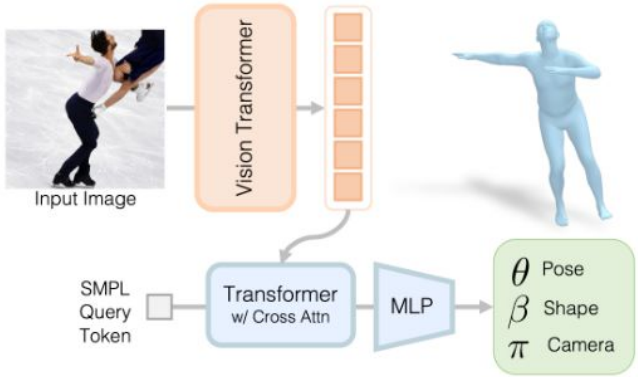


Joint Optimization for 4D Human-Scene Reconstruction in the Wild (JOSH) [ICLR'26]

- Bidirectional execution ✓
- Unified feedforward system ✗

Prior Arts: Human-Scene Reconstruction (E2E)

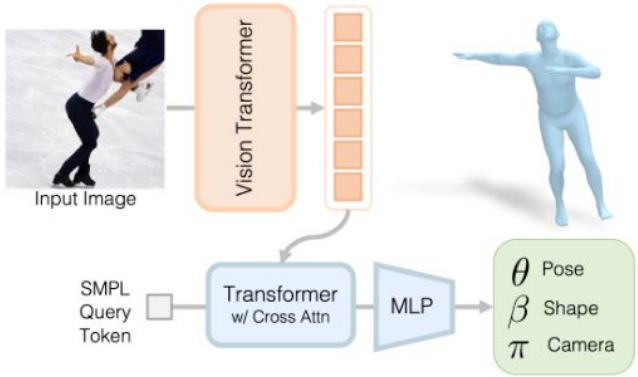
Unifying Scene and Human Reconstruction in a Feed-Forward Pass (UniSH) [CVPR'26]



Unified feedforward system ✓
Data-driven priors ✓
Bidirectional execution ✗

Prior Arts: Human-Scene Reconstruction (E2E)

Unifying Scene and Human Reconstruction in a Feed-Forward Pass (UniSH) [CVPR'26]



Unified feedforward system ✓
Data-driven priors ✓
Bidirectional execution ✗

Humans do not shape the reconstructed scene
Scenes do not constrain human motion
Contact, support, collision, affordance - not modeled

Summary: Optimization vs E2E Methods

Table 4.3: **Global human-scene reconstruction performance on EMDB-2 [24] and RICH [20].** We report world-grounded motion metrics including WA-MPJPE$_{100}$, W-MPJPE$_{100}$, and relative trajectory error (RTE). **Opt.** **Free** indicates feed-forward inference, while **Scene** denotes explicit scene reconstruction.

Method	Opt. Free	Scene	EMDB-2			RICH		
			WA-MPJPE ↓	W-MPJPE ↓	RTE(%) ↓	WA-MPJPE ↓	W-MPJPE ↓	RTE(%) ↓
TRAM [66]	✗	✗	76.4	222.4	1.4	127.8	238.0	6.0
WHAM [52]	✓	✗	135.6	334.8	6.0	108.4	190.1	4.5
GVHMR [51]	✓	✗	111.0	276.5	2.0	**78.8**	**126.3**	**2.4**
JOSH [35]	✗	✓	**68.9**	**174.7**	**1.3**	89.0	132.5	3.0
JOSH3R [35]	✗	✓	220.0	661.7	13.1	-	-	-
UniSH [31]	✓	✓	118.5	270.1	5.8	118.1	183.2	4.8

Summary: Optimization vs E2E Methods

Feature	Optimization Methods	E2E Methods
Human–Scene Coupling	Strong, explicit	Weak or indirect coupling
Use of Human–Scene Priors	Explicit modeling	Implicit or absent
Initialization Sensitivity	Sensitive to quality of initial human, scene, and camera estimates	Less sensitive due to amortized inference
Computational Cost	High; requires iterative optimization per sequence	Low; single forward pass enables real-time or near real-time inference
Robustness to Occlusion and Noise	Fragile under occlusion, fast motion, or missing contacts	More robust to noisy or incomplete observations

Joint end-to-end execution ≠ Joint reasoning

Summary: Optimization vs E2E Methods

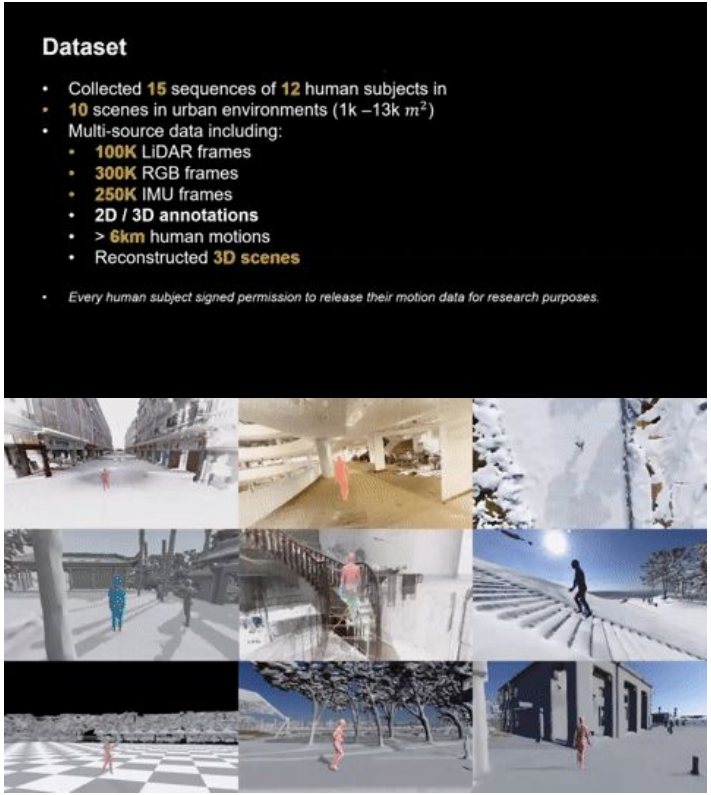
Feature	Optimization Methods	E2E Methods
Human–Scene Coupling	Strong, explicit	Weak or indirect coupling
Use of Human–Scene Priors	Explicit modeling	Implicit or absent
Initializatio		ortized
Computatio	per sequence	s enables real-time or near real-time inference
Robustness to Occlusion and Noise	Fragile under occlusion, fast motion, or missing contacts	More robust to noisy or incomplete observations

How can we achieve true bidirectional interaction when joint execution ≠ joint reasoning?

Joint end-to-end execution ≠ Joint reasoning

Existing Datasets

SLOPER4D



PROX

SLOPER4D

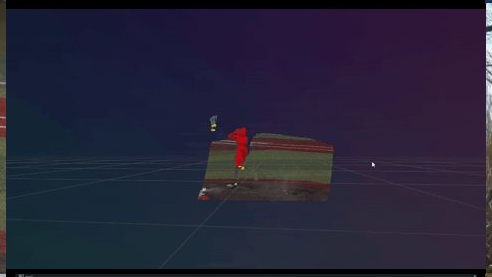
- 300K LiDAR frames at 20 Hz
- Spread across multiple outdoor scenes and subjects

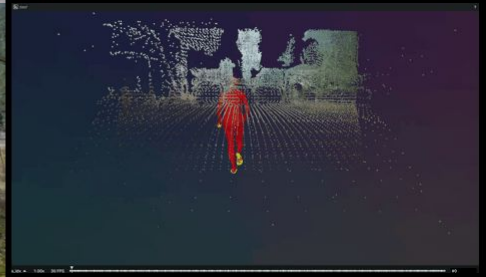
PROX

- 100K RGB-D frames at 30 FPS
- Dynamic data with human-scene interaction groundtruth



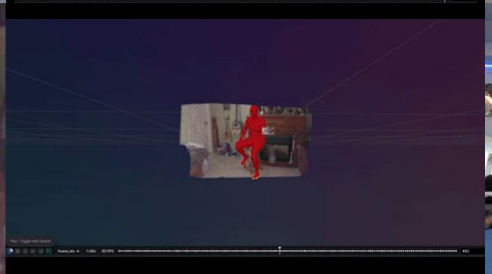


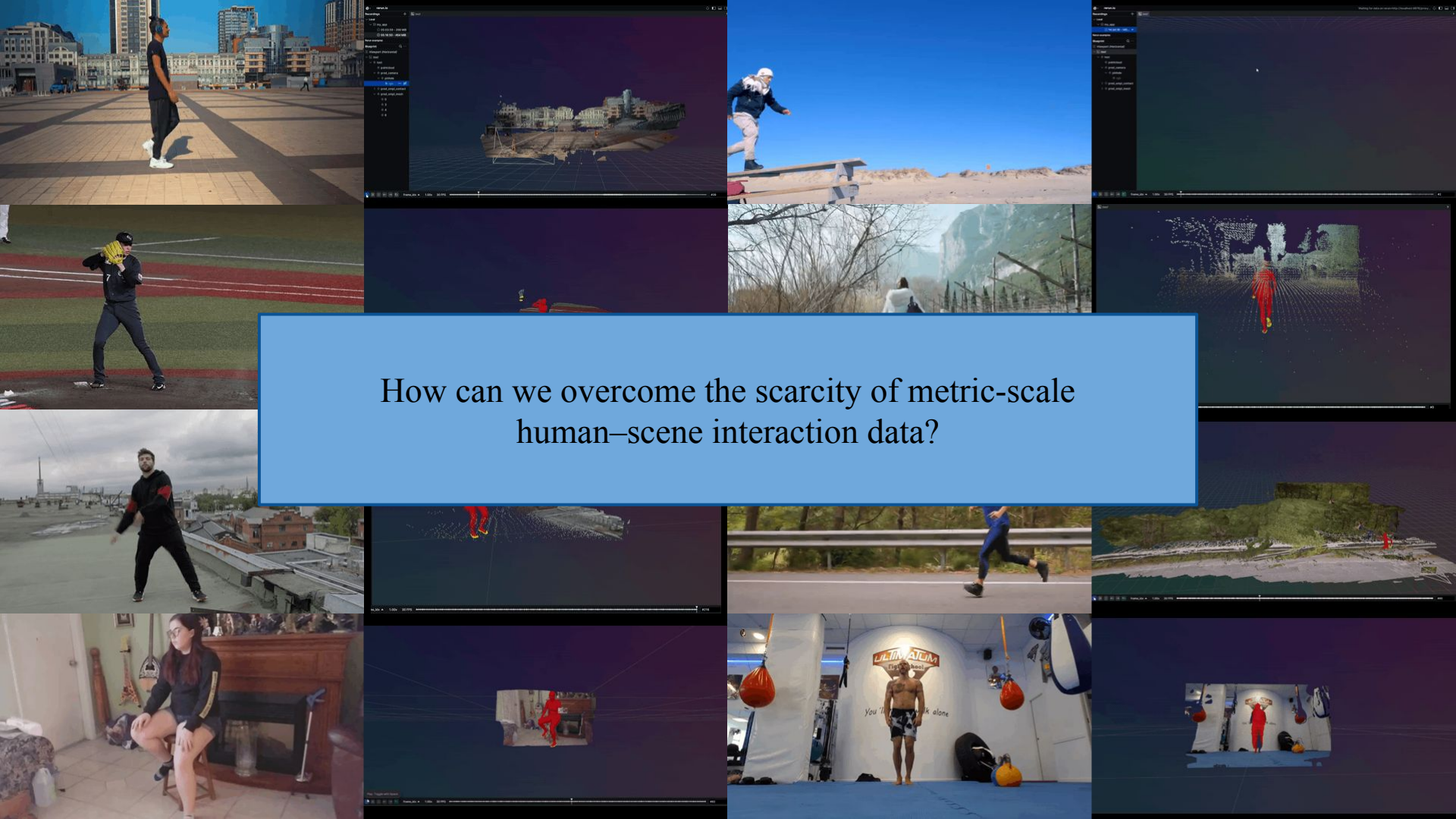








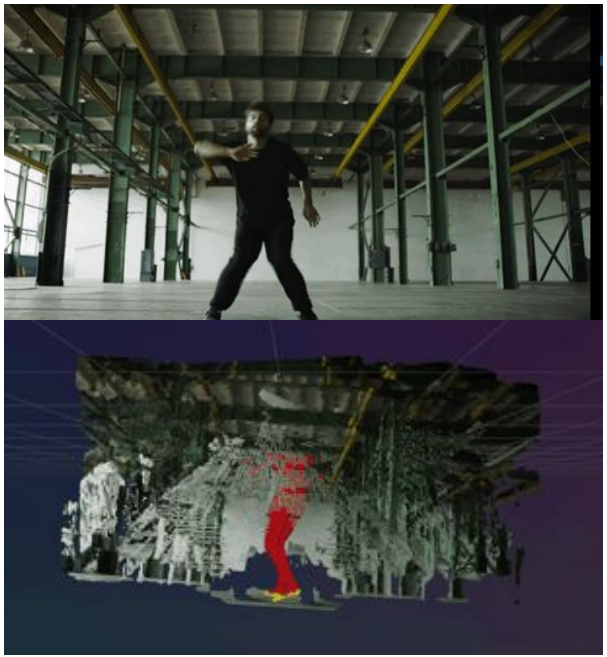
ULTIMATUM

How can we overcome the scarcity of metric-scale human–scene interaction data?

Persistent Challenges in Monocular 4D Reconstruction

- Duplicate “ghost” layers when camera drifts slightly
- Human feet can be buried to the ground, struggles with low texture
- Holes, artifacts and floaters
- No large-scale rich dataset for human+scene reconstruction



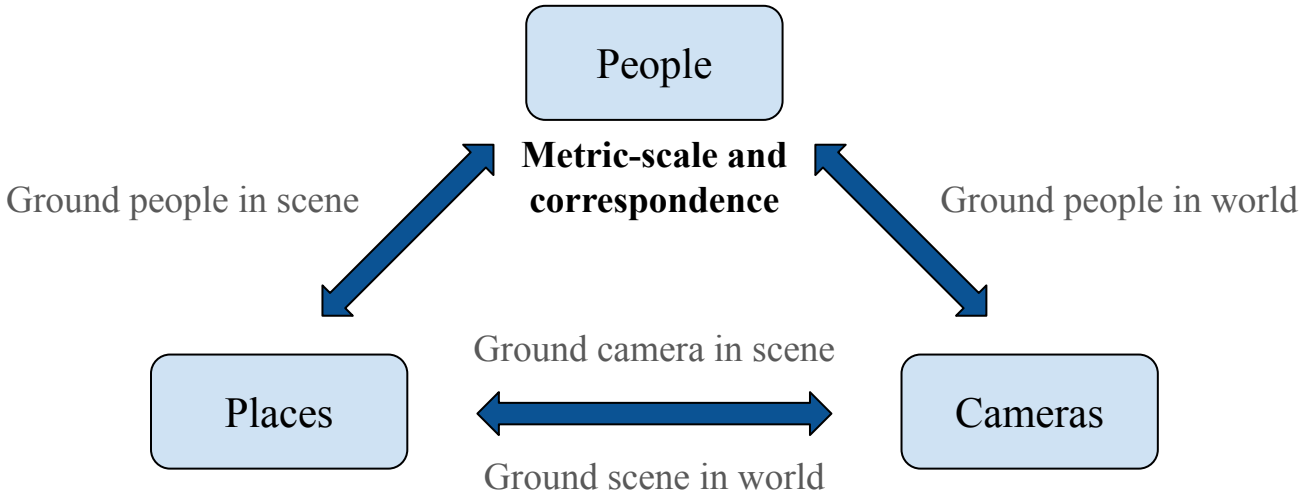
Guiding Research Objectives

Can we emulate human ability to continuously reconstruct and update a complete 4D world from sparse observations?

How can we achieve true bidirectional interaction when joint execution ≠ joint reasoning?

How can we overcome the scarcity of metric-scale human–scene interaction data?

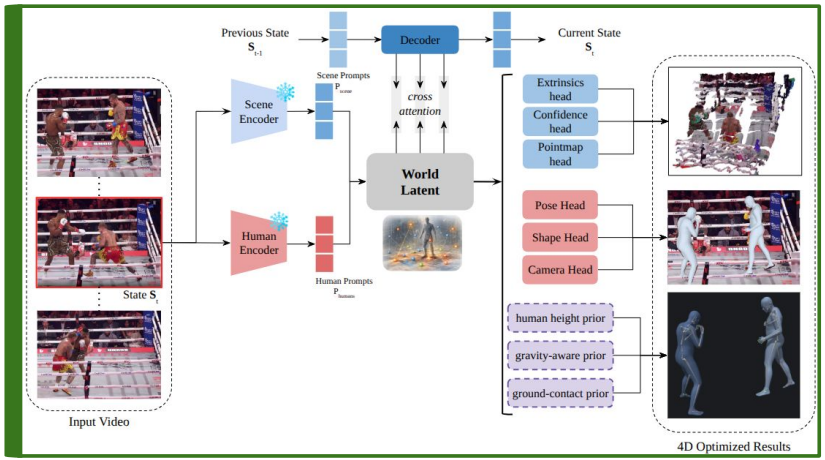
Intuition: Joint Scene, Human, and Camera Understanding



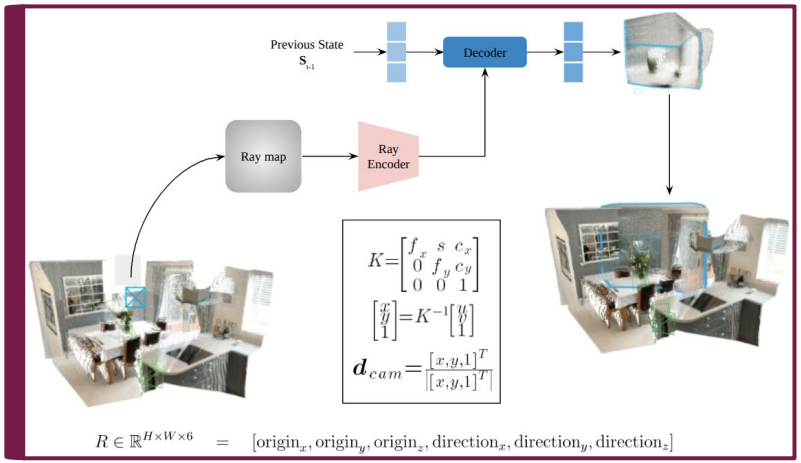
From Objectives to Action

Data Engine

Proposed Architecture

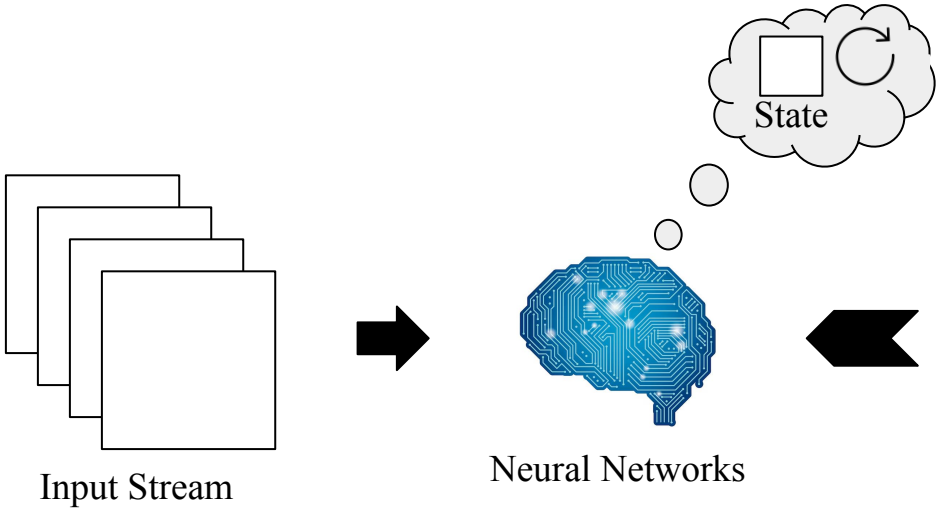


Scene Completion

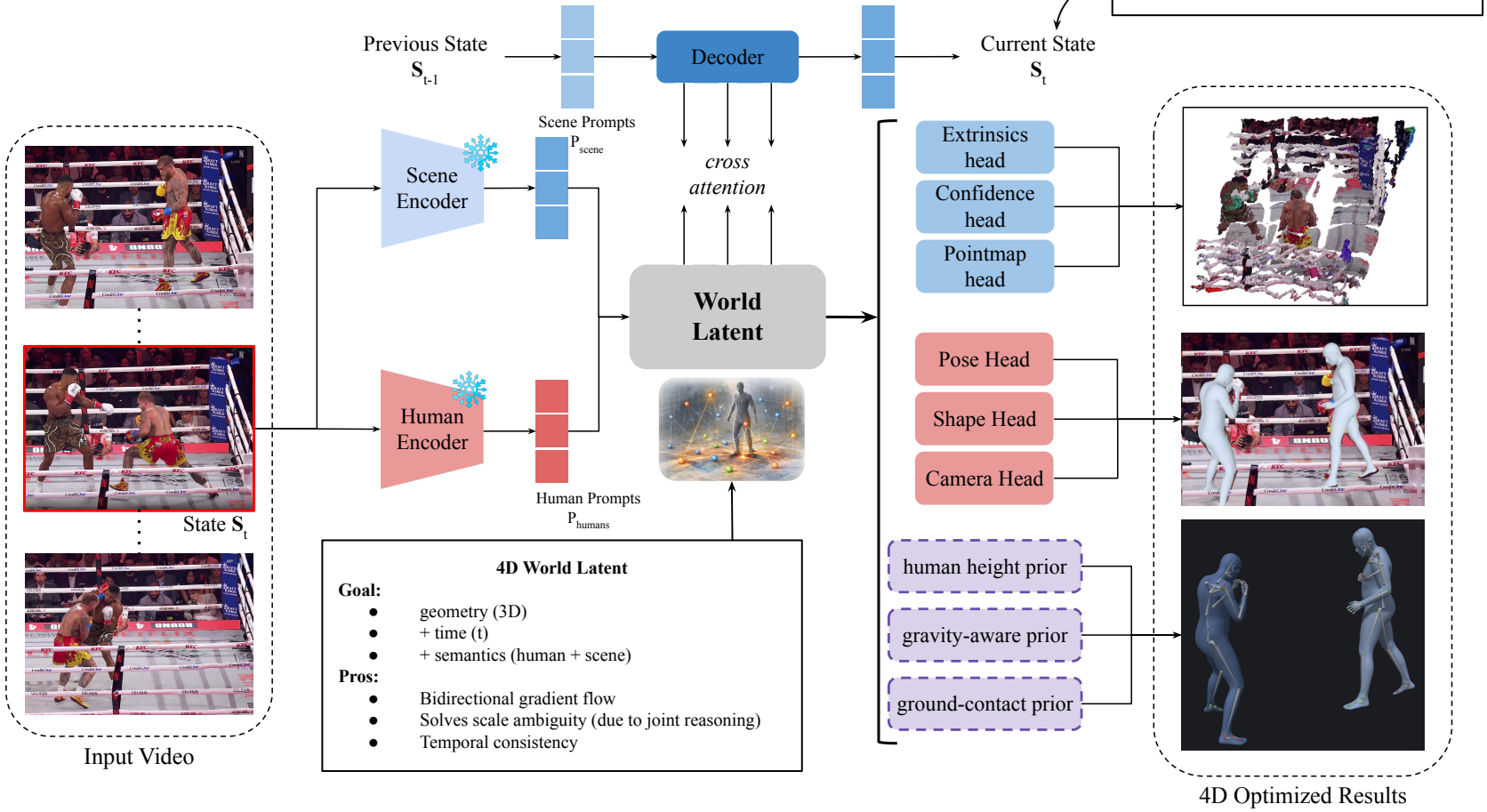


Human Perception as Inspiration

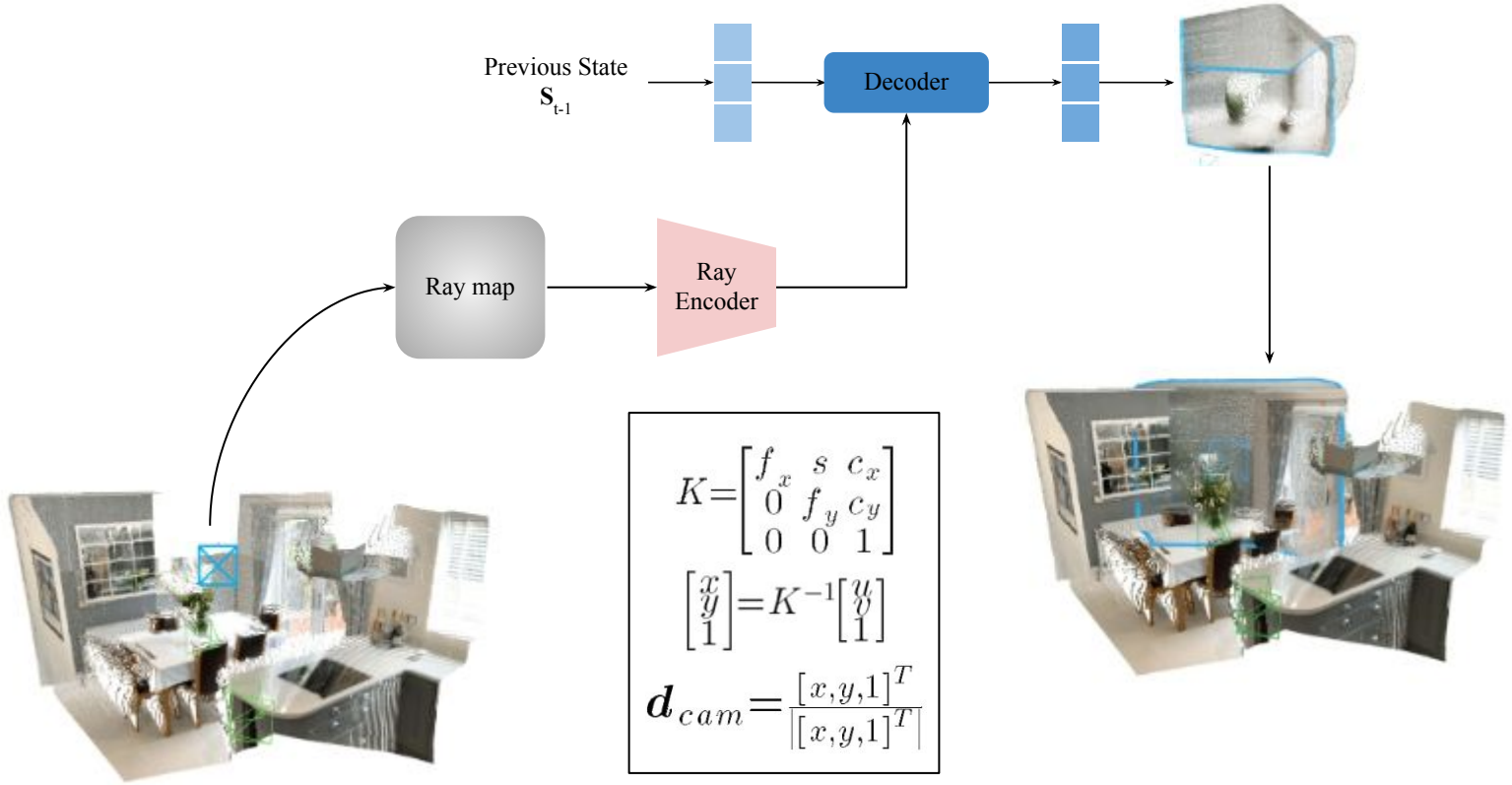
- Reconstructing 3D scenes from few observations
- Explicit inferring unseen regions beyond observation
- Continuously updating the reconstruction online with more observations



Proposed 4D World Reconstruction Framework

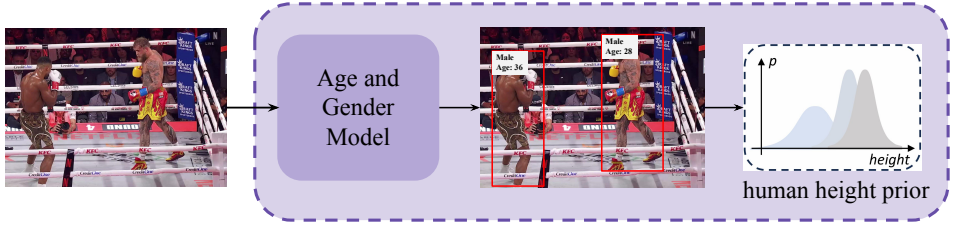


Scene Completion via Raymap Probing



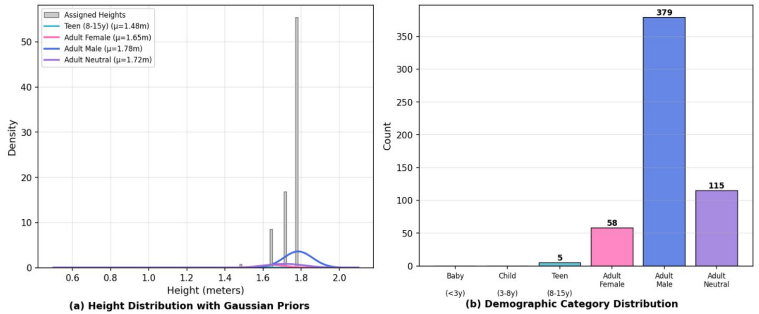
$$R \in \mathbb{R}^{H \times W \times 6} \quad = \quad [\text{origin}_x, \text{origin}_y, \text{origin}_z, \text{direction}_x, \text{direction}_y, \text{direction}_z]$$

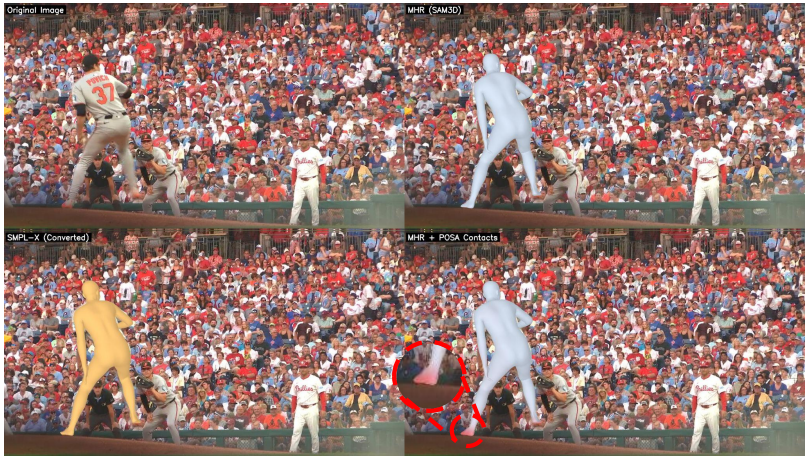
Progress to Date- Height and Contact Priors



- **Baby** ($\hat{a} < 3$): $\mu = 0.801$ m, $\sigma = 0.126$ m
- **Kid** ($3 \leq \hat{a} < 8$): $\mu = 1.122$ m, $\sigma = 0.12$ m
- **Teen** ($8 \leq \hat{a} < 15$): $\mu = 1.477$ m, $\sigma = 0.156$ m
- **Adult Female** ($\hat{a} \geq 15$, confident female): $\mu = 1.647$ m, $\sigma = 0.0707$ m
- **Adult Male** ($\hat{a} \geq 15$, confident male): $\mu = 1.784$ m, $\sigma = 0.0759$ m
- **Neutral Adult** ($\hat{a} \geq 15$, uncertain gender): $\mu = 1.715$ m, $\sigma = 0.10$ m

$$\mathcal{L}_{\text{height}} = \frac{(h_{\text{current}} - \mu)^2}{2\sigma^2}$$

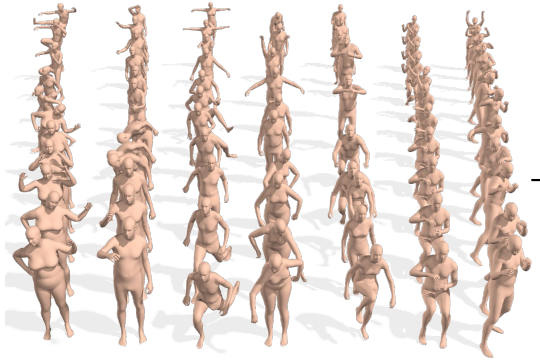




Progress to Date- Synthetic Dataset

Avatar4D Framework (Submitted to CVPRW'26)

3D Models/Assets

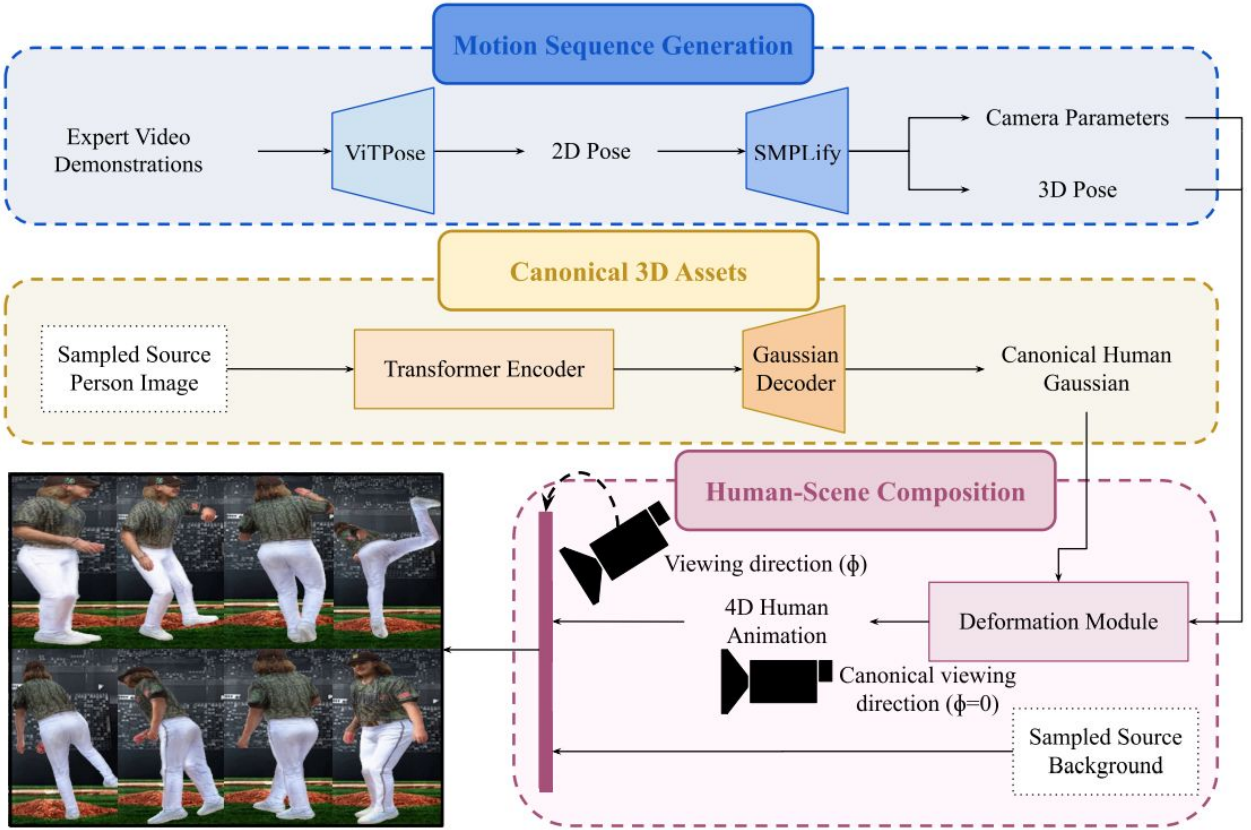


Dictionary of Motion Sequences

Deformation, Rasterization, & BG Diffusion

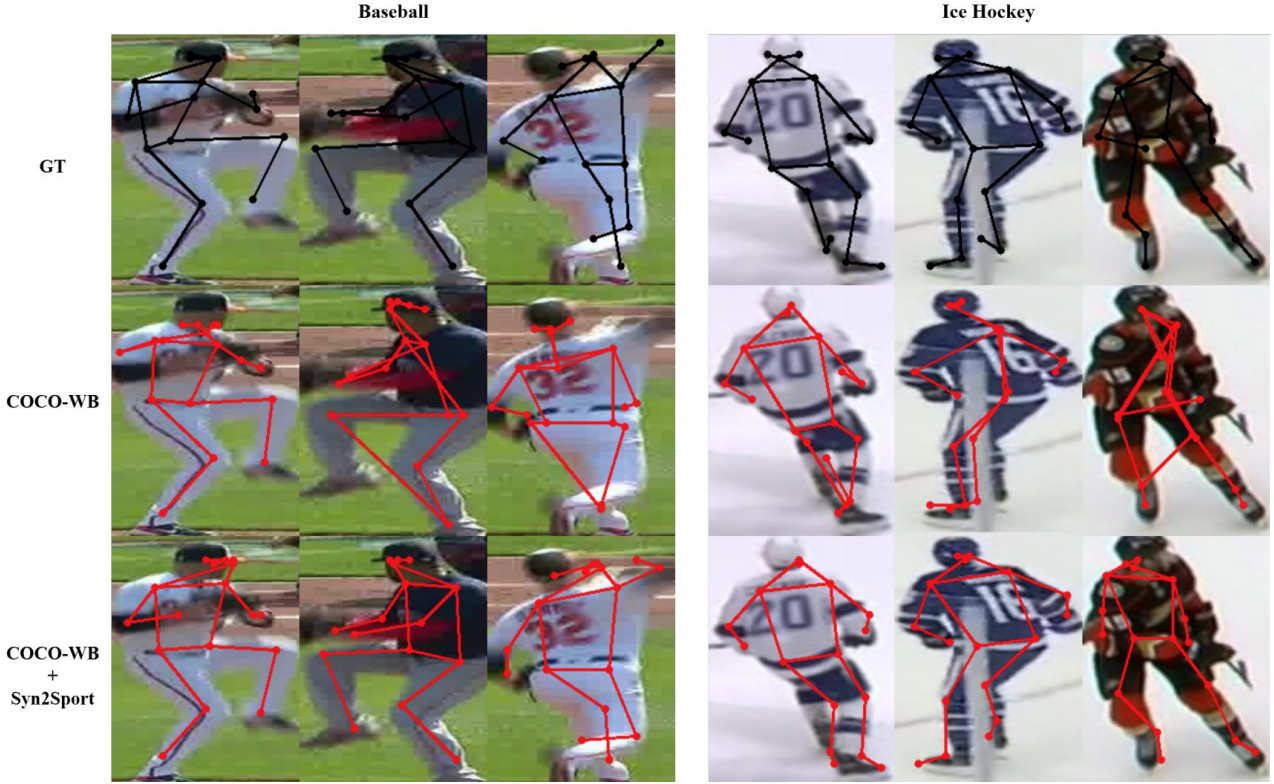
Progress to Date- Synthetic Dataset

Avatar4D Framework (Submitted to CVPRW'26)

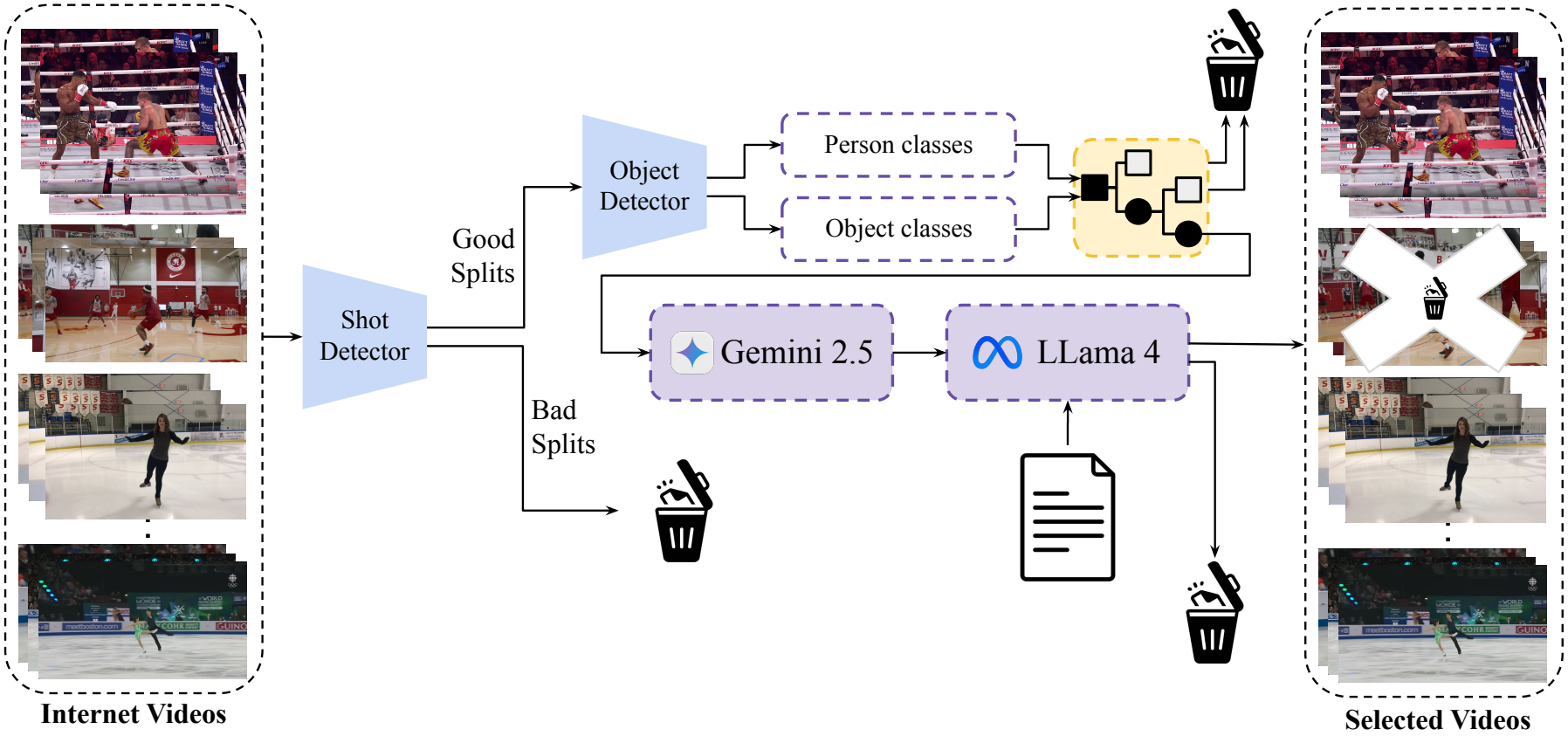


Progress to Date- Synthetic Dataset

Avatar4D Framework (Submitted to CVPRW'26)

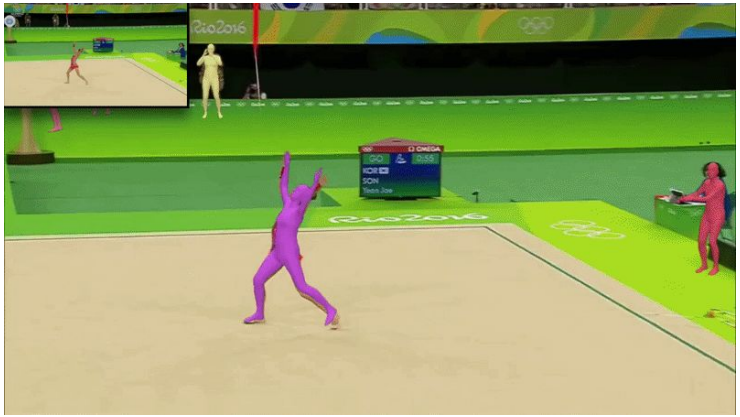


Progress to Date- Data Engine (Curation Phase)

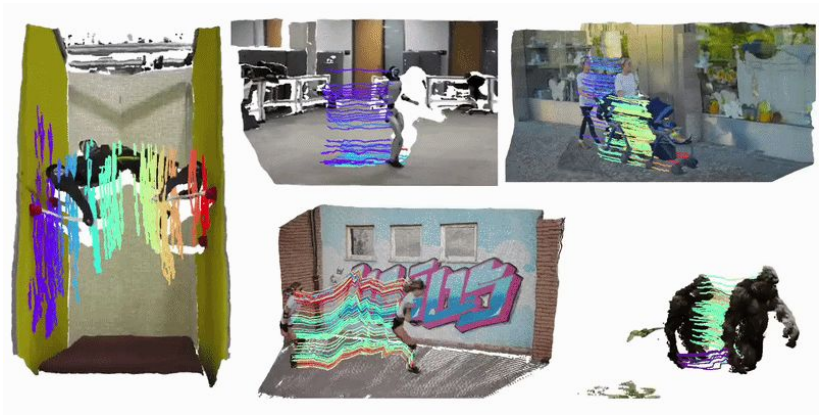



TUNETALK.COM

Progress to Date- Data Engine (Processing Phase)



Human Reconstruction



Scene Reconstruction

Joint Alignment Optimization

$$\min_{\boldsymbol{T},\, s,\, R_{root},\, \Delta\theta,\, \beta} \mathcal{L} = \underbrace{\lambda_1 \mathcal{L}contact}_{\text{ground feet}} + \underbrace{\lambda_2 \mathcal{L}penetration}_{\text{no interpenetration}} + \underbrace{\lambda_3 \mathcal{L}gravity}_{\text{upright motion}} + \underbrace{\lambda_1 \mathcal{L}reproj}_{\text{2D consistency}} + \lambda_5 \mathcal{L}height + \underbrace{\lambda_6 \mathcal{L}\beta}_{\text{beta drift regularization}} + \underbrace{\lambda_7 \mathcal{L}vel}_{\text{contact velocity}}$$

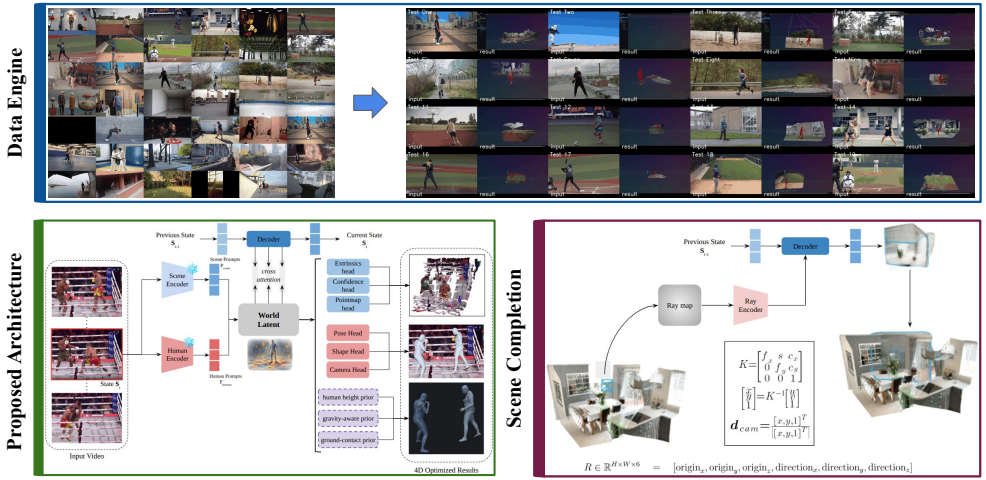
subject height anchoring

Evaluation Plan and Metrics

Objective	Quantitative Metrics	Evaluation Setup
Continuous 4D Understanding	• Hole coverage (%) • Point density (points/m²) • Trajectory drift (ATE / RTE in cm) • Raymap probe completeness	BEDLAM (metric GT) Avatar4D synthetic Long real videos (>500 frames)
Bidirectional Human–Scene Reasoning	• Contact error (cm) • Penetration volume (m³) • Scale consistency (human-scene height ratio) • Ground-contact accuracy (%)	Human3R, CUT3R, JOSH, UniSH Baseline: BEDLAM + PROX + SLOPER4D
Scalable Human–Scene Data	• Dataset scale (frames / sequences) • Diversity score (scenes / actions / lighting) • Synthetic vs. real ablation gain (%)	Avatar4D (proposed) + curated real videos Baseline: BEDLAM + PROX + SLOPER4D

Key Contributions

- First recurrent World Latent for true bidirectional human–scene reasoning in a feed-forward model
- Raymap-based scene completion for hole-free, online 4D reconstruction.
- Explicit interaction priors (height, gravity, ground-contact) injected into the latent for physical plausibility.
- Avatar4D + data engine: scalable synthetic + curated real data closing the metric-scale human–scene gap.



The Road Ahead