UniCon3R: Unified Contact-aware 4D Human-Scene Reconstruction from Monocular Video



Tanuj Sur

Shashank Tripathi

Nikos Athanasiou

Ha Linh Nguyen

Kai Xu

Michael J. Black

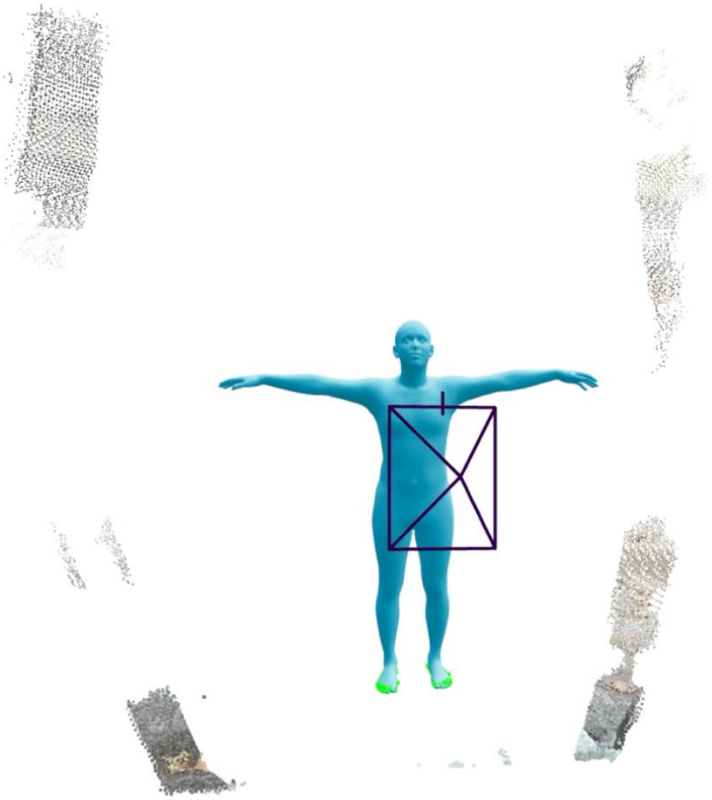
Angela Yao

Accepted to NeurIPS 2026

Problem Statement

Input video





Output

- Metric 3D scene points
- Camera pose
- Human body mesh placed in that scene
- Human-scene contact (green).

Goal

Enable plausible 4D human-scene reconstruction from streaming videos.

*** Generated by UniCon3R (Ours)**

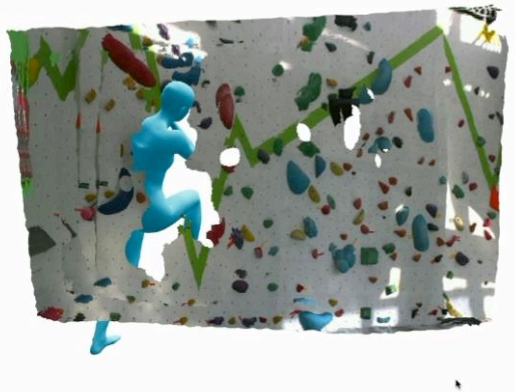
Physically Implausible Reconstruction

Input video





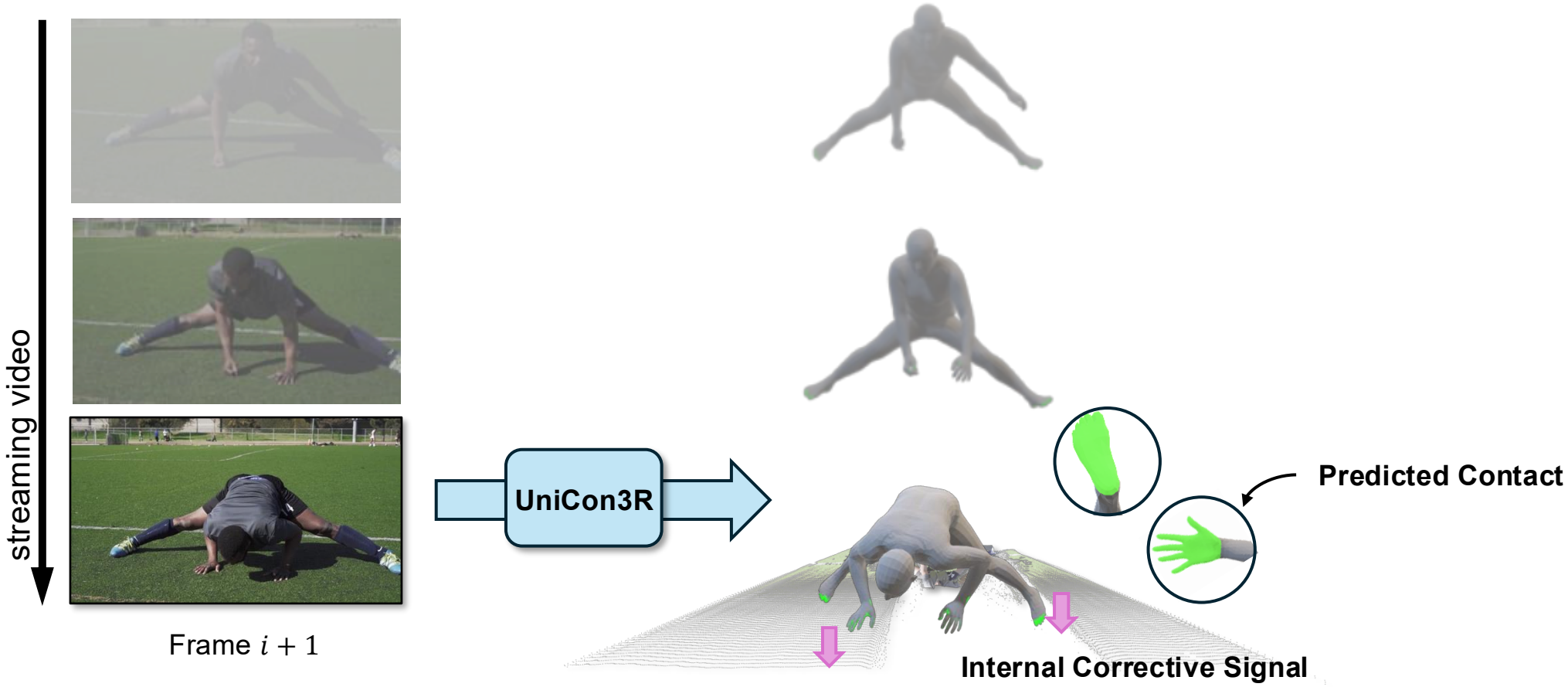
4D reconstruction





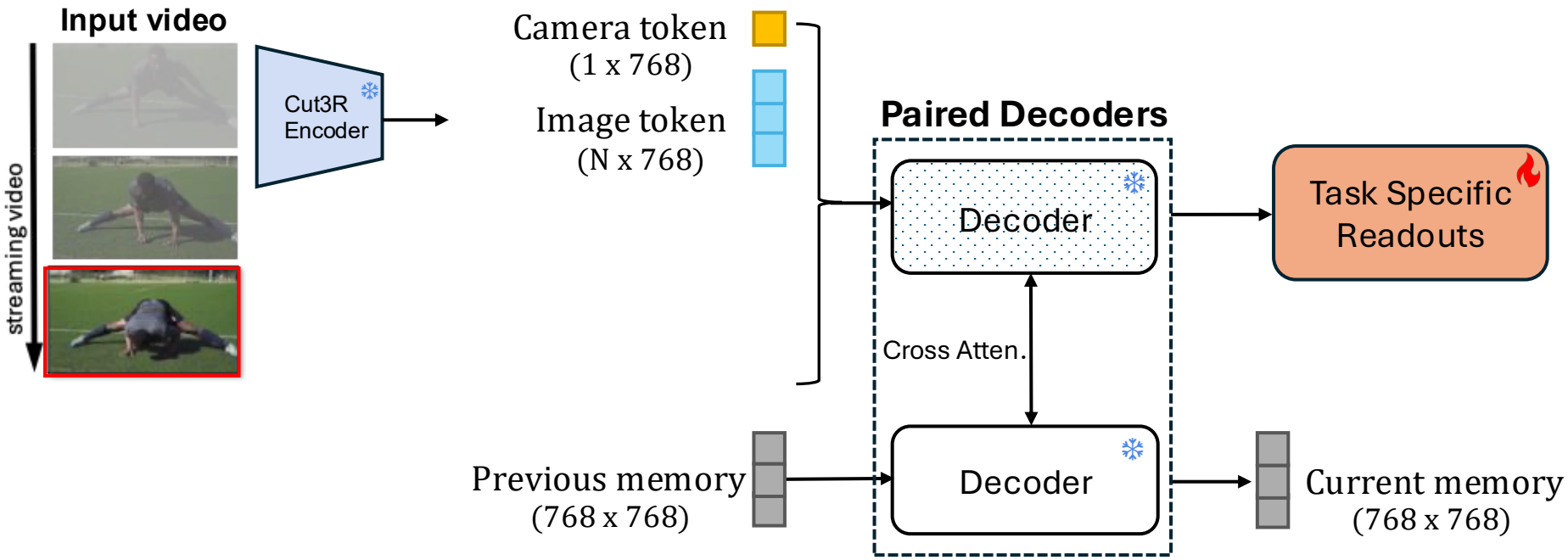
*Generated by SOTA Feed-Forward 4D Human-Scene Reconstruction Method[2].

Contact as an Internal Corrective Cue



Preliminaries: Recurrent Decoding with a Persistent Memory

1. CUT3R[1] – DUSt3R[2]'s paired decoders, with the second image replaced by a persistent memory: reads a persistent memory S_{t-1}, decodes the current frame, writes S_t.

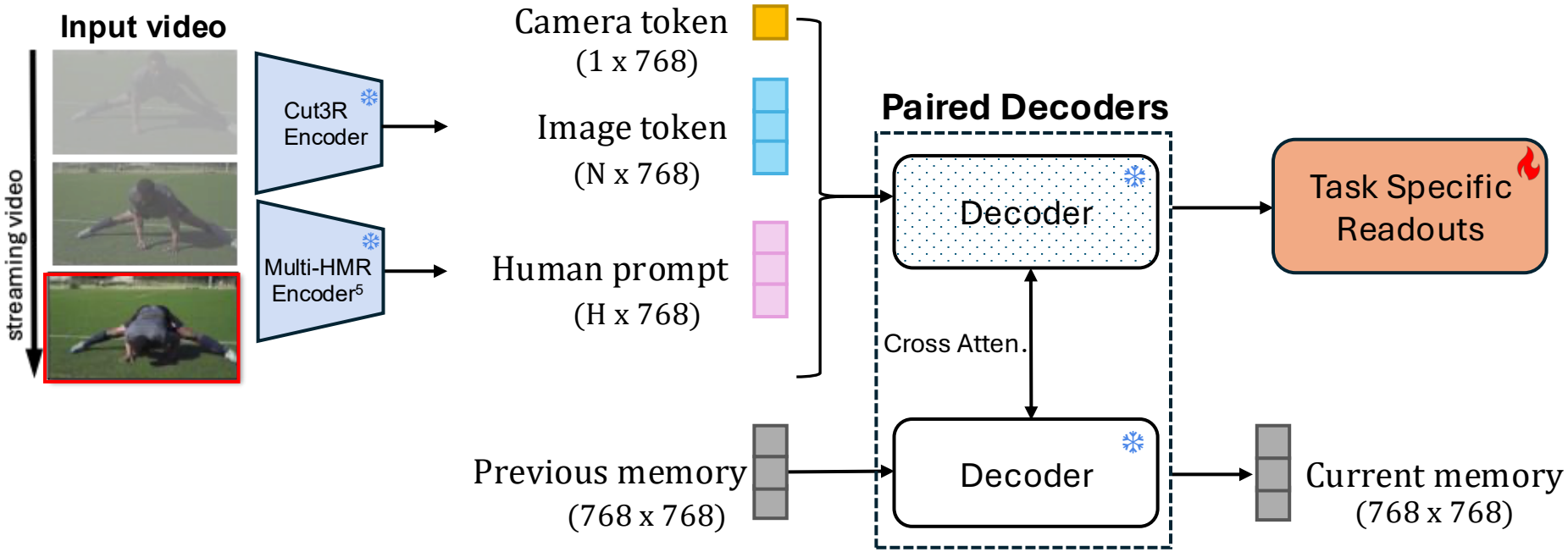


N =[512·h / (16·max(h, w))] · [512·w / (16·max(h, w))]

Preliminaries: Recurrent Decoding with a Persistent Memory

1. CUT3R[1] – DUSt3R[2]'s paired decoders, with the second image replaced by a persistent memory: reads a persistent memory S_{t-1}, decodes the current frame, writes S_t.

2. Human3R[3] – Leverages Visual Prompt Tuning[4] (VPT) i.e. append one learned token per person to the input list.



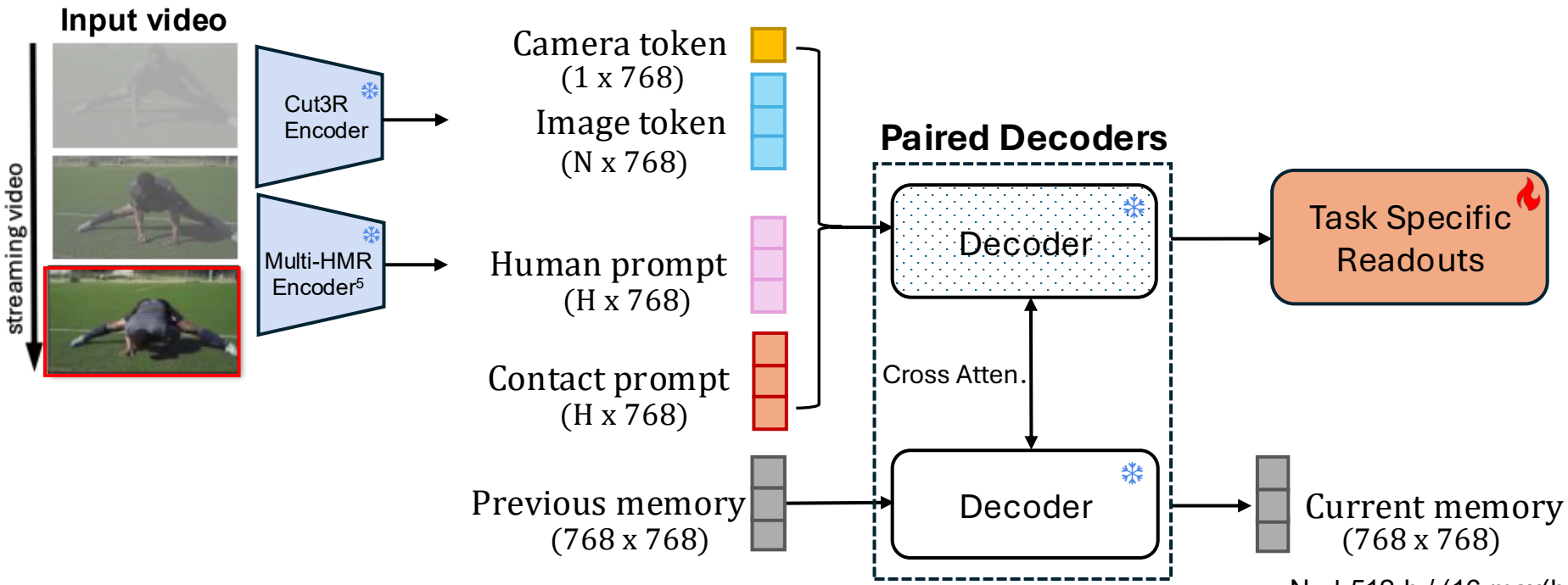
N =⌊ 512·h / (16·max(h, w)) ⌋ · ⌊ 512·w / (16·max(h, w)) ⌋
H = # humans in each frame

Preliminaries: Recurrent Decoding with a Persistent Memory

1. CUT3R[1] – DUSt3R[2]'s paired decoders, with the second image replaced by a persistent memory: reads a persistent memory S_{t-1}, decodes the current frame, writes S_t.

2. Human3R[3] – Leverages Visual Prompt Tuning[4] (VPT) i.e. append one learned token per person to the input list.

3. **UniCon3R (Ours)** – Design and append special *contact prompt* per person in addition to previous prompts.



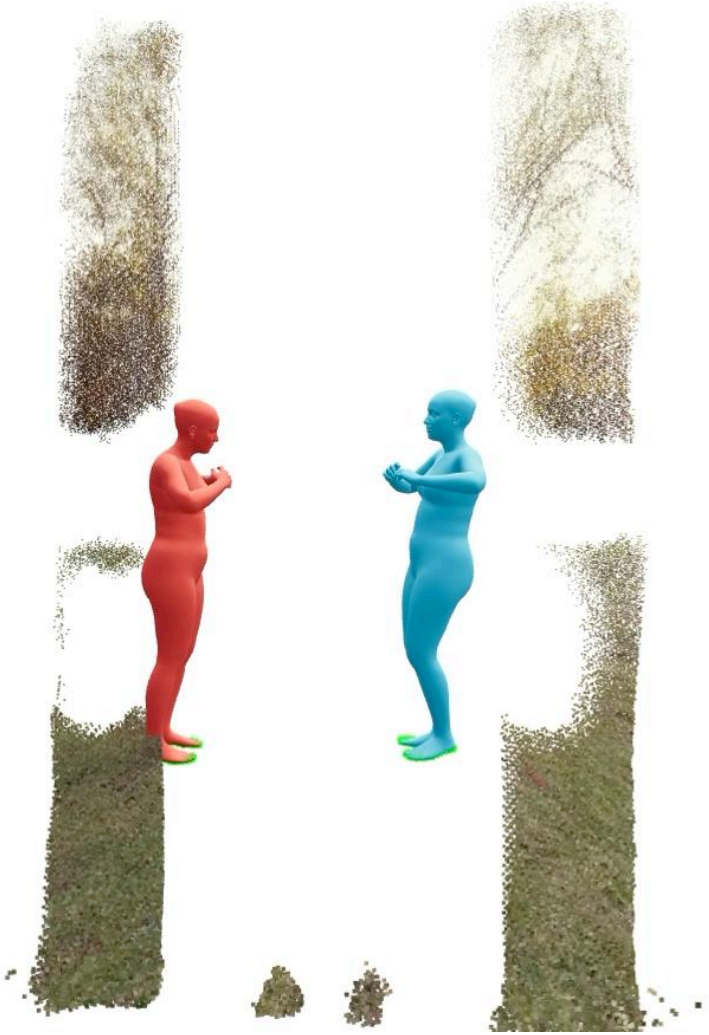
N =⌊ 512·h / (16·max(h, w)) ⌋ · ⌊ 512·w / (16·max(h, w)) ⌋
H = # humans in each frame

UniCon3R Improves Joint Human-Scene Reconstruction

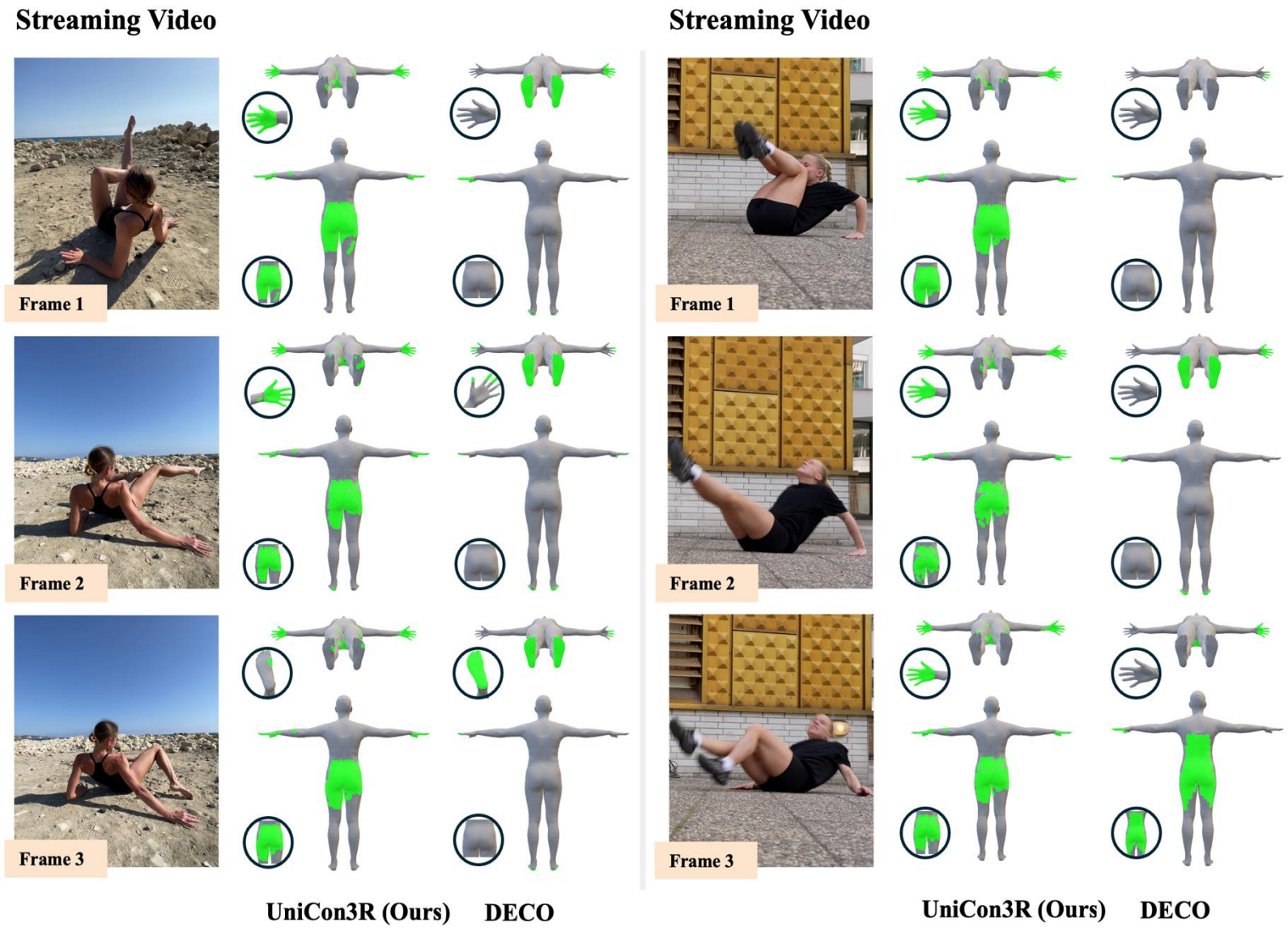
Input video



4D reconstruction



UniCon3R improves Binary Contact Prediction



Limitations

1. **Contact is a binary per-vertex label.** No proximity, pressure or friction, so the corrective signal says *whether* a body is supported, not *how*.

2. **Contact quality depends on scene quality.** Scene reconstruction errors propagate into contact and body–scene alignment; the pathway only runs one way

3. **Plausibility metrics assume rigid support.** Grounding is measured against the reconstructed scene as rigid geometry; deformable objects, non-rigid scenes and fine-grained object interaction are not modelled.

4. **Temporal momentum aligns people by slot index, not identity.** Exact for one person, a simplification in crowds.

5. **Interaction is human-scene only, not human-object-scene.** Contact grounds the body on floors and furniture, but a manipulated object is neither reconstructed nor fed back to the body.