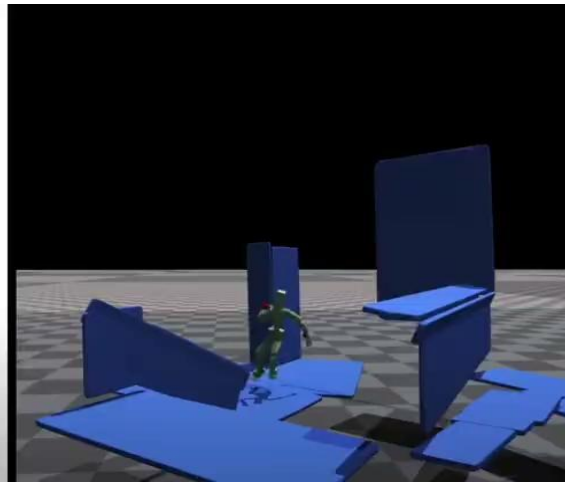
CRISP : Contact-guided Real2Sim from Monocular Video with Planar Scene Primitives

Zihan Wang* Jiashun Wang* Jeff Tan Yiwen Zhao
Jessica Hodgins Shubham Tulsiani Deva Ramanan





unposed monocular RGB video → 4D human-scene reconstruction → whole-body control policy

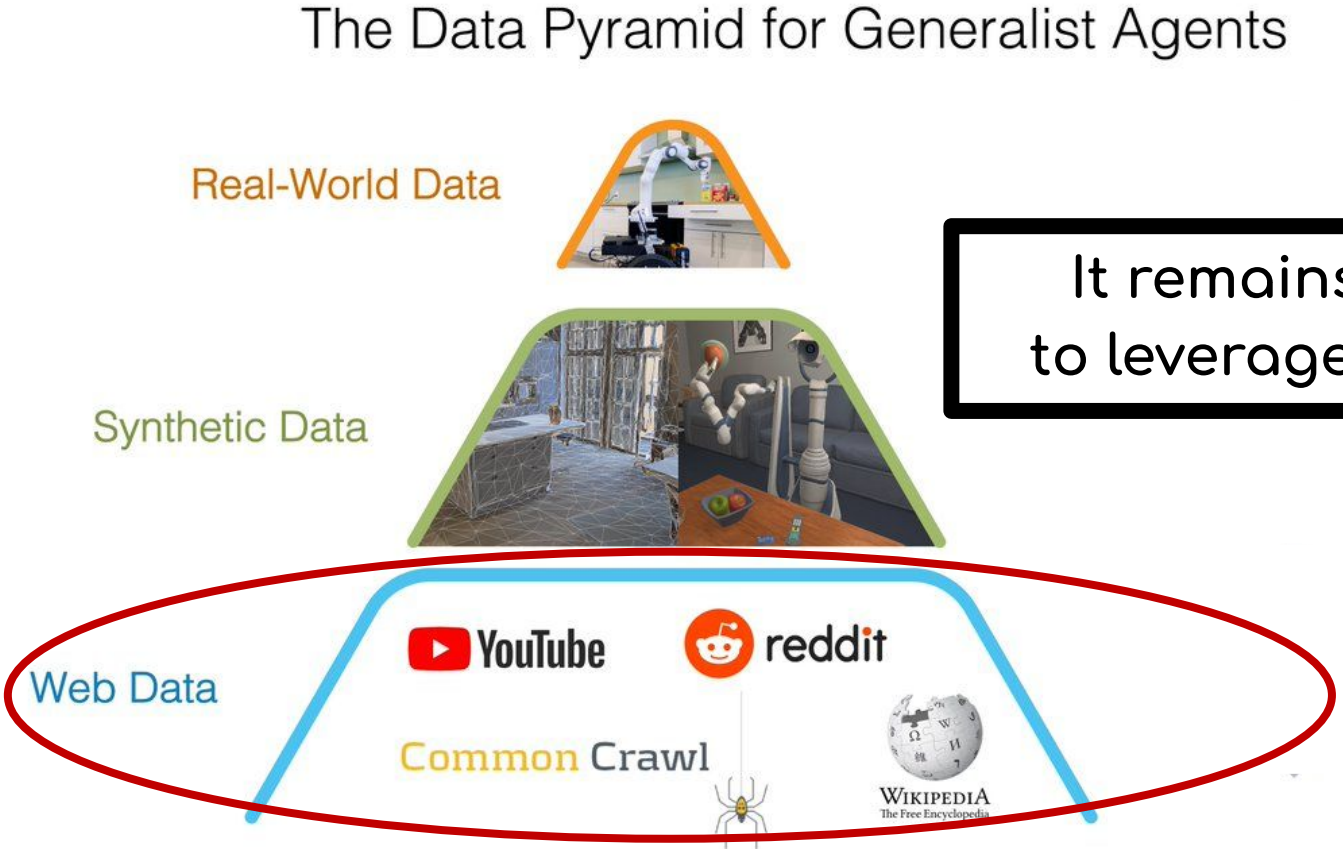
Robotics Institute, Carnegie Mellon University

Introduction

- Why humanoid ?
  - The world is built for humans → humanoids are the most general platform.
  - Leverage human data → smallest cross-embodiment gap.



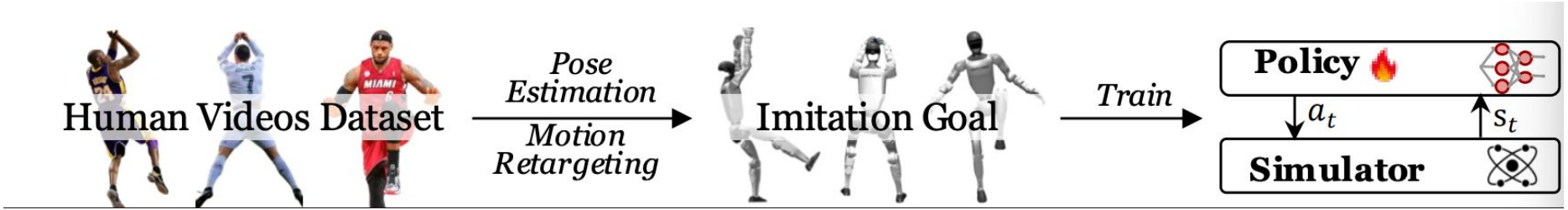

Introduction



It remains unknown how to leverage web-scale data

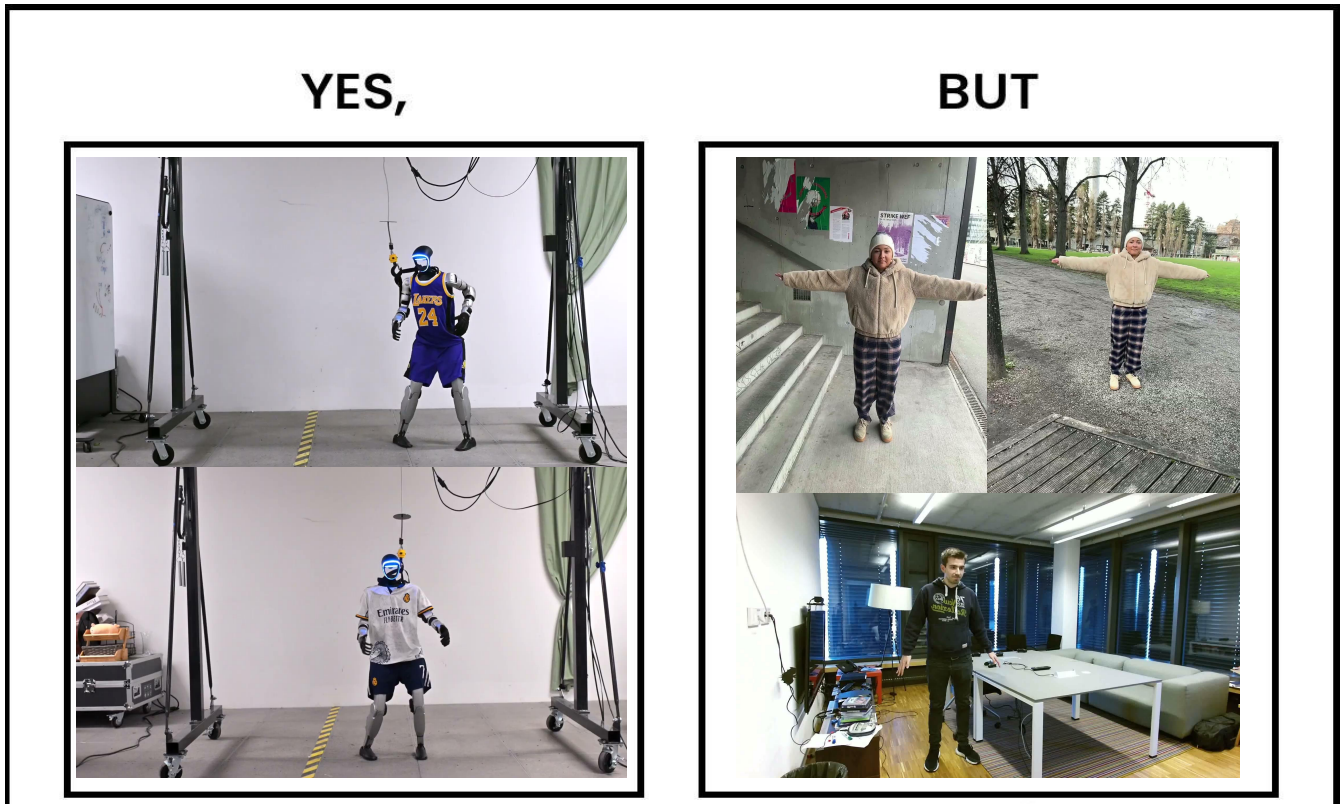
Introduction

- It is proven successful in learning from video. On flat terrain !



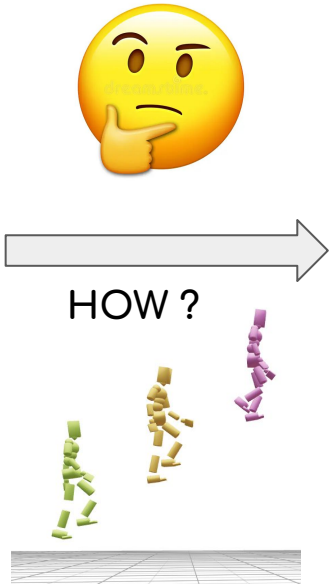
Introduction

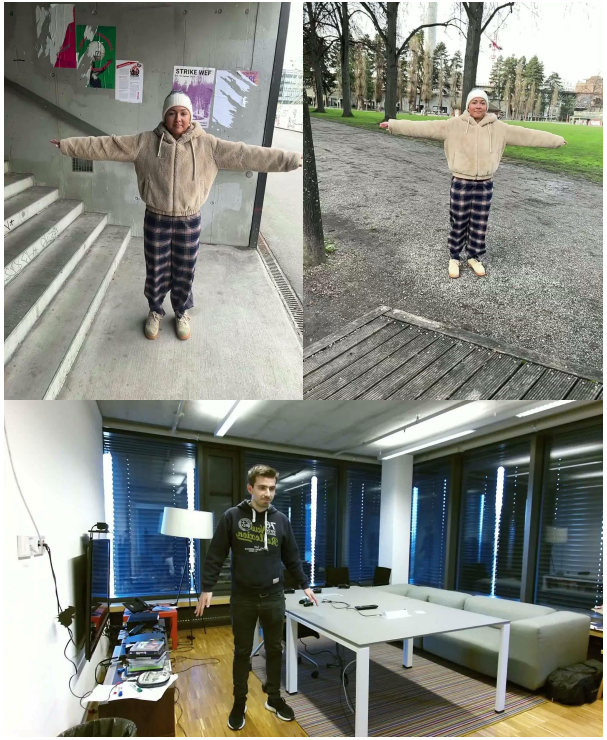
- [more than flat ground] We human constantly interact with diverse, complex terrain.



Introduction

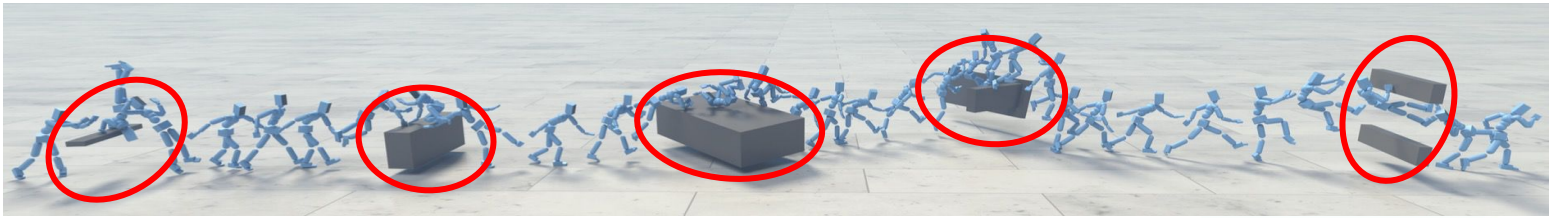
- [more than flat ground] We human constantly interact with diverse, complex terrain.

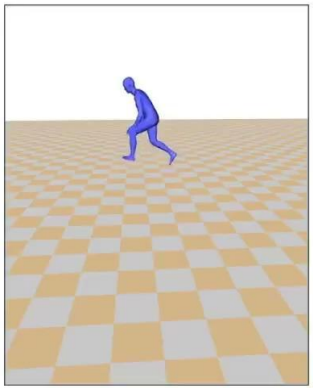




Introduction

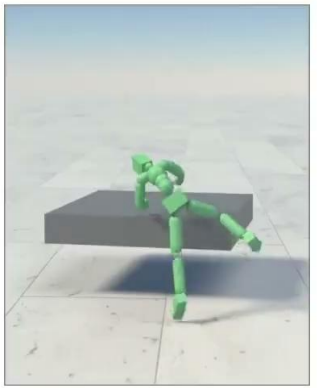
- Existing work consider geometry includes heavy hand-crafted efforts.





Pose estimation from video

Scene annotation



Simulated reference motion

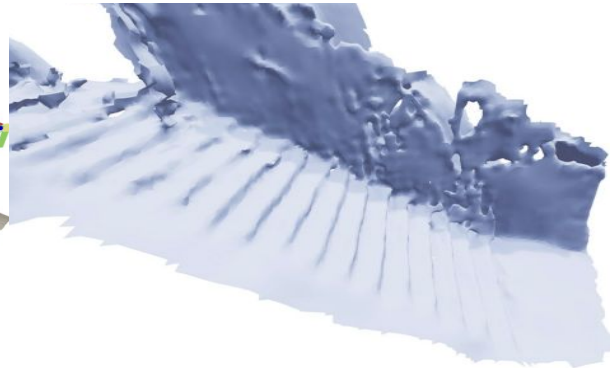
Due to the scarcity of motion capture data

Introduction

How can we bridge video → simulatable reconstruction → robot skill learning in complex, diverse environment?

Core challenge 1 – gap between vision and simulation

- Agent in simulation relies on collision mesh to provide faithful physics feedback.



Visually ✌️

Simulation🤷

- Traditional TSDF / NKSR makes visually-okish triangle mesh, but what is the cost?
  - 2.5D reconstruction – non-watertight → imprecise simulation → penetrate
  - noisy reconstruction – unstable normals → contact jitter
  - high triangle count – insufficient to simulate

- Thus, a visually-okay mesh is not necessarily a good one for simulation.

Core challenge 1 – reconstruction quality (1/3)

- Physics-based simulator requires geometry for collision, bad one ruins everything!

input video

VideoMimic

Unstable contact &
Unnatural motion.

Core challenge 1 – reconstruction quality (2/3)

- Physics-based simulator requires geometry for collision, bad one ruins everything!

input video

VideoMimic

bumpy recon hinders movement of agent.

Core challenge 1 – reconstruction quality (3/3)

- Physics-based simulator requires geometry for collision, bad one ruins everything!

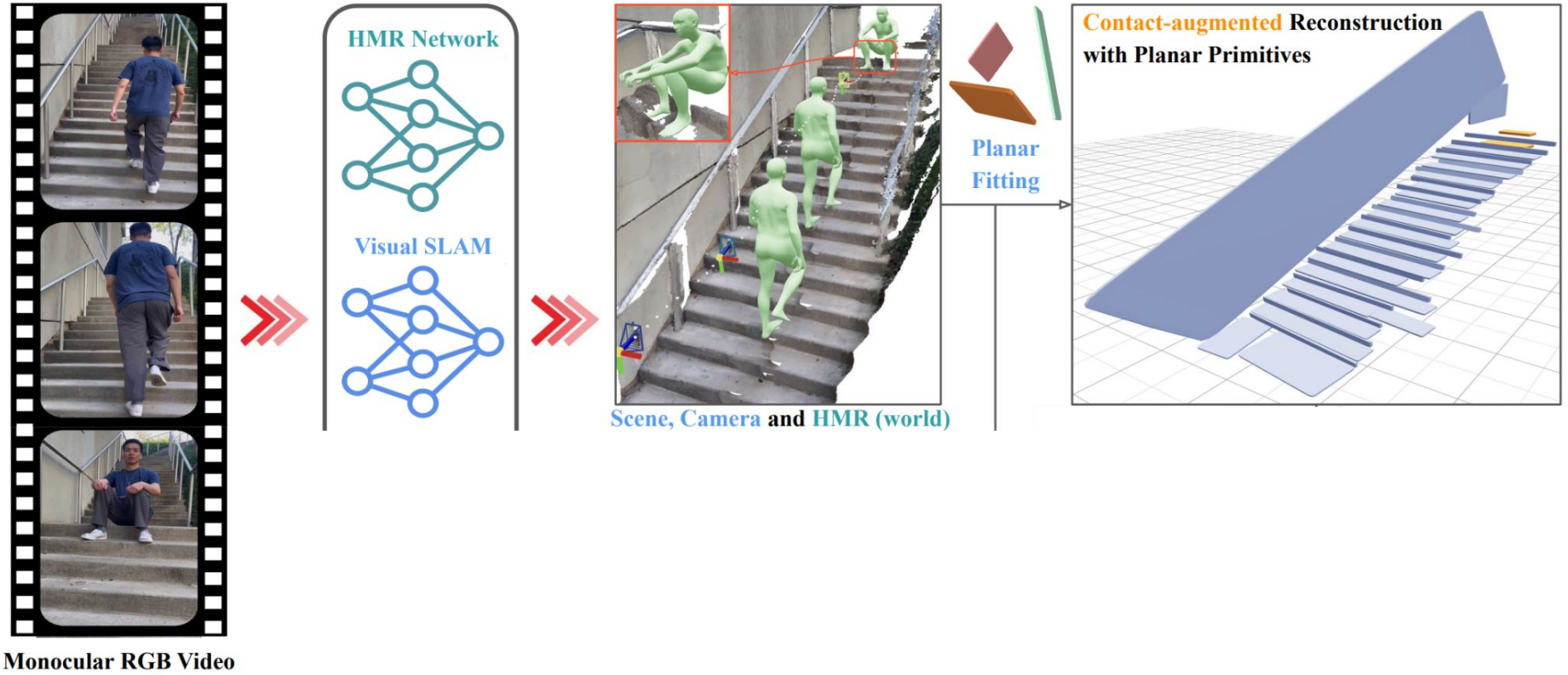
input video

VideoMimic

penetration issue & stuck by ghost layer.

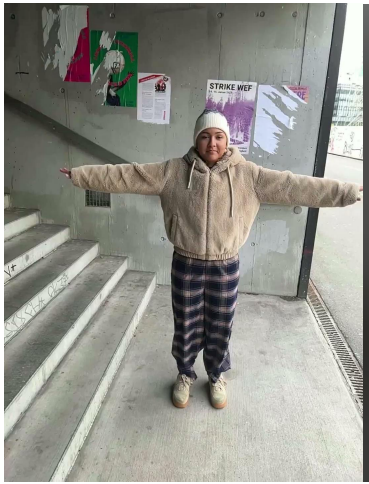
Our work is exactly about making that connection actually work in a **‘simulatable’** way!

Method Overview

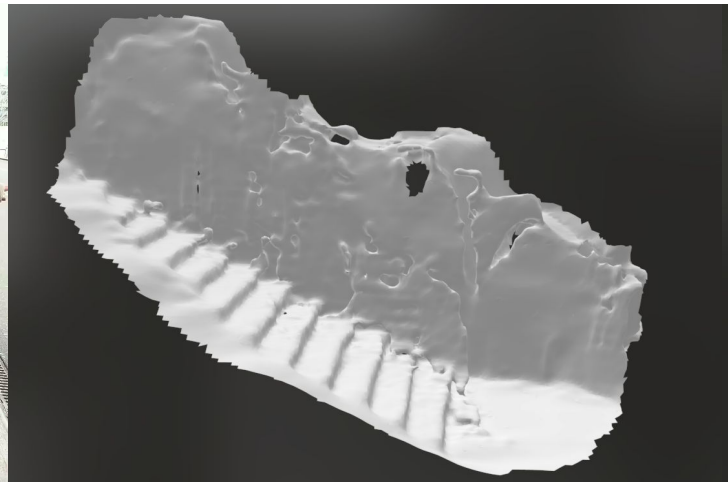


Insight 1: Build scene with simulation-ready planar primitives

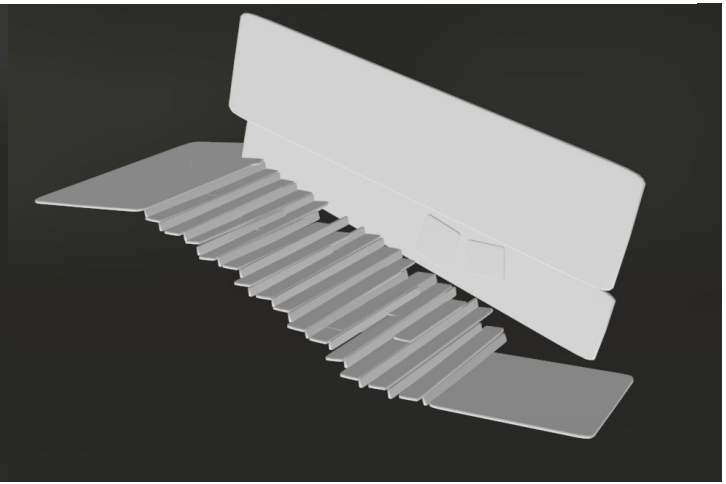
✅ enforce ‘flat prior’ → stable contact ✅ few primitives → efficient to simulate ✅ robust to noise



Input source video



Geometry of VideoMimic



Geometry of CRISP

Comparison results (1/3)

input video

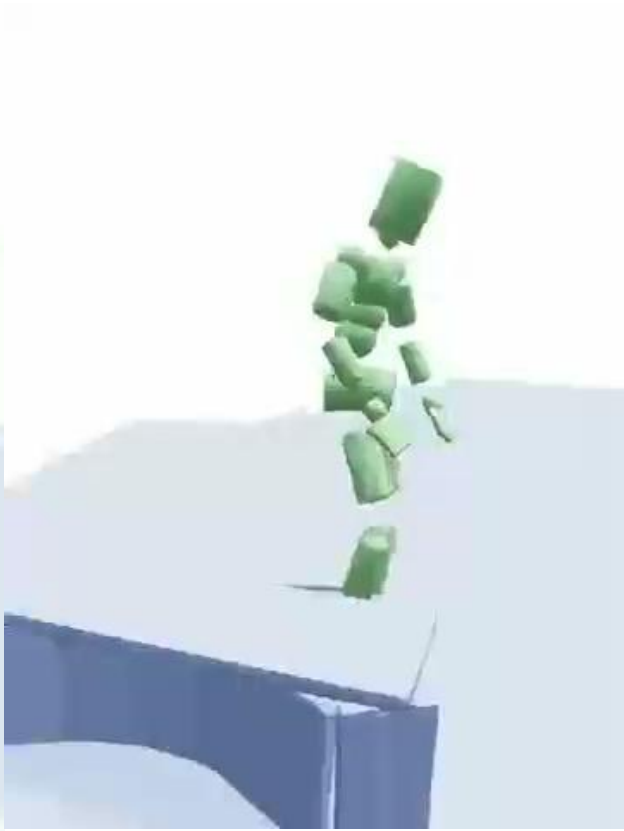
VideoMimic

CRISP

Comparison results (2/3)

input video

VideoMimic



CRISP

Comparison results (3/3)

input video

VideoMimic

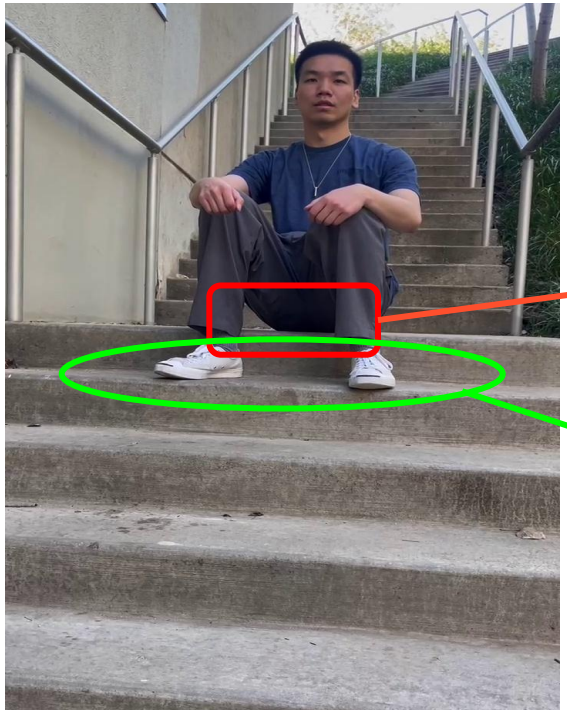
CRISP

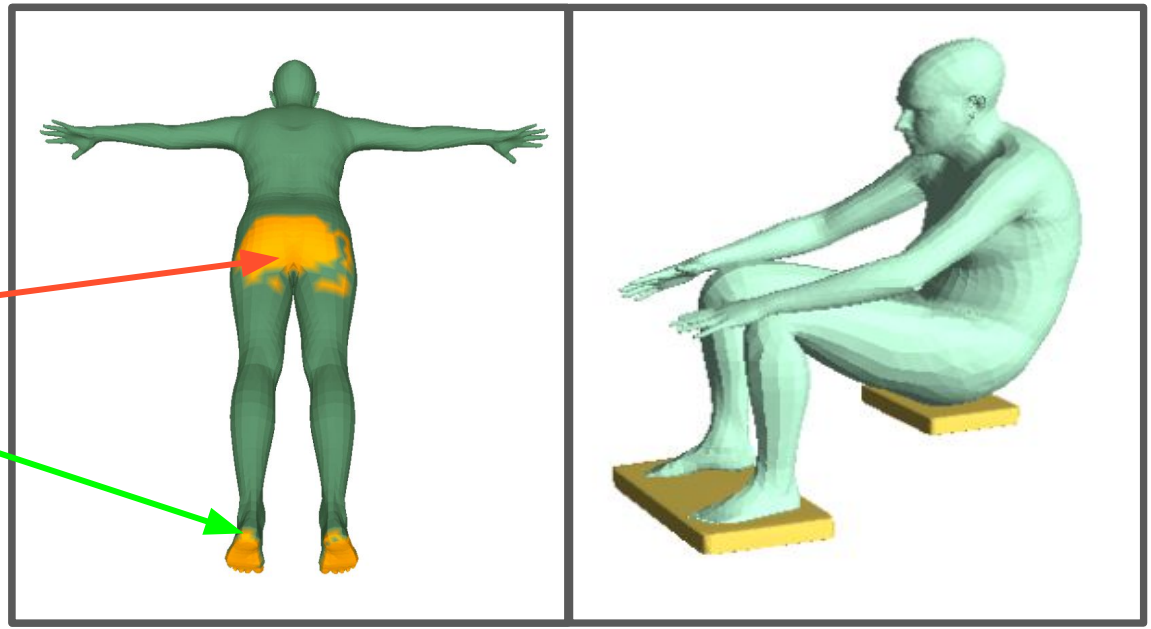
Core challenge 2 – frequent occlusion in monocular video

- Missing occluded scene will cause simulation to fail.

Key Insight 2: contact as clue for scene completion

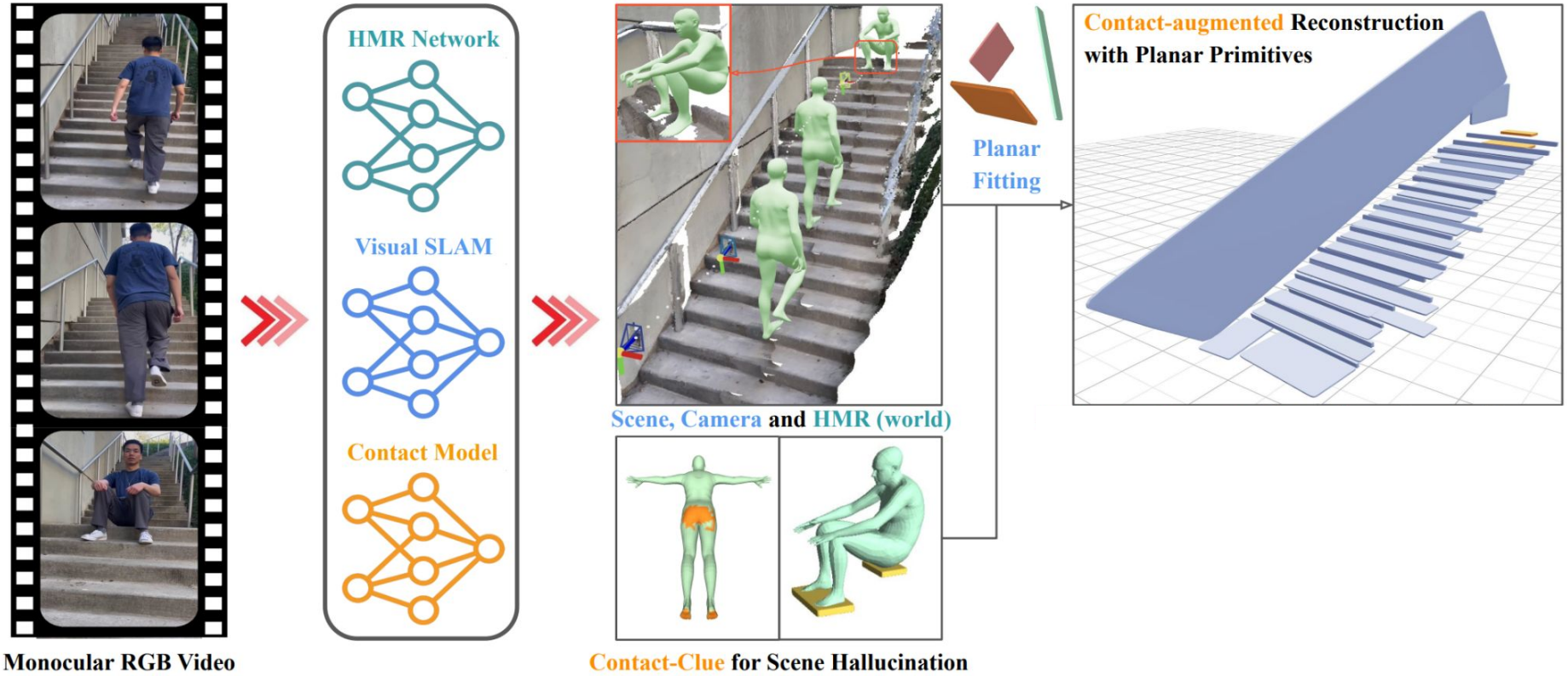
- The platform is occluded, but we know where the butt is.
- How ? Take InteractVLM (input: image → output: binary contact mask on SMPL vertices). For predicted contact points, we fit max 1 planar primitives to each pre-annotated part.



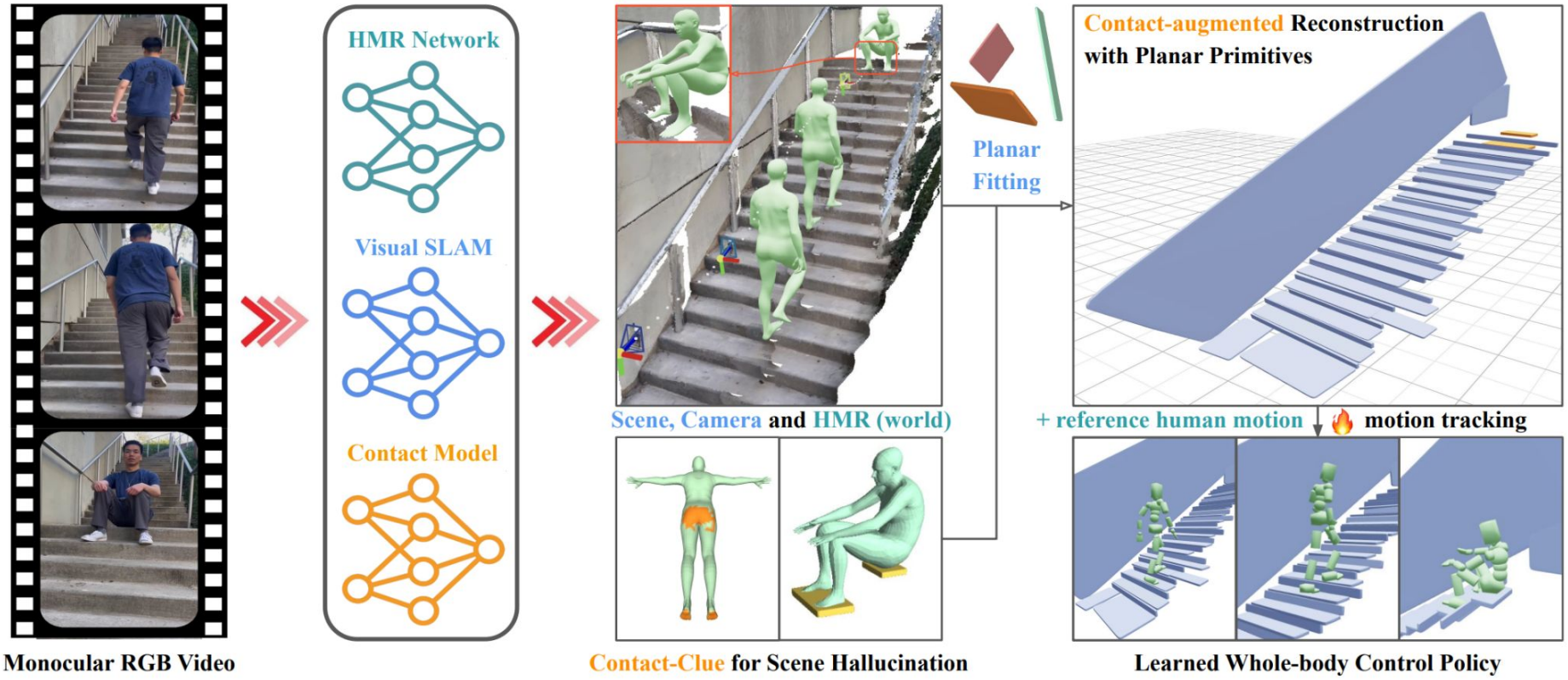


Contact-Clue for Scene Hallucination

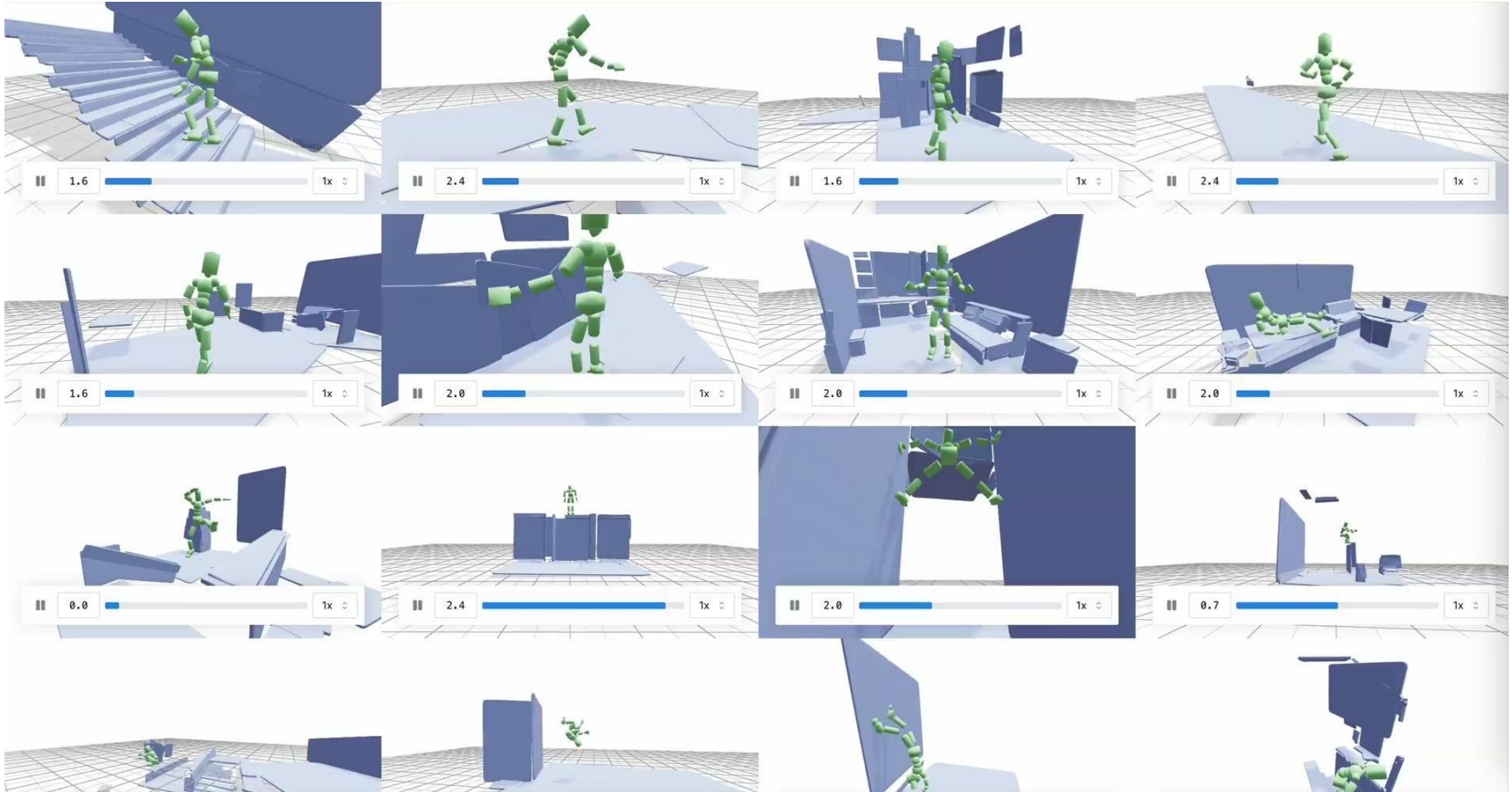
Method Overview



Method Overview

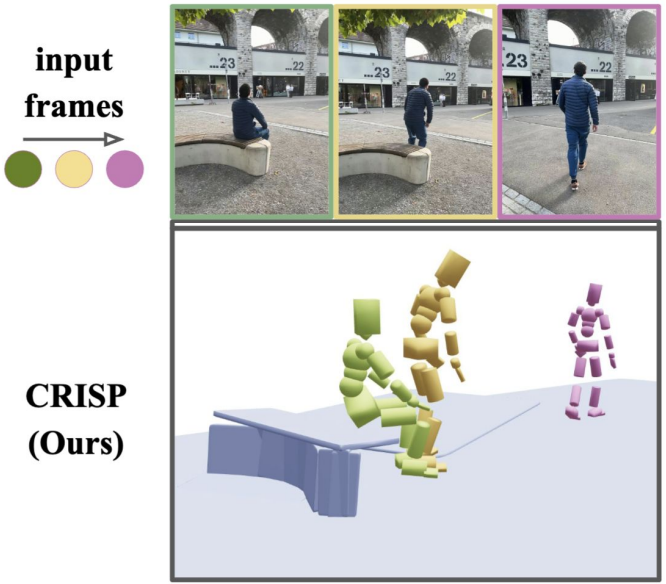


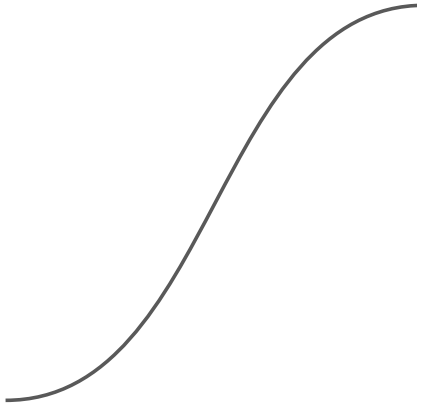
Quick Summary



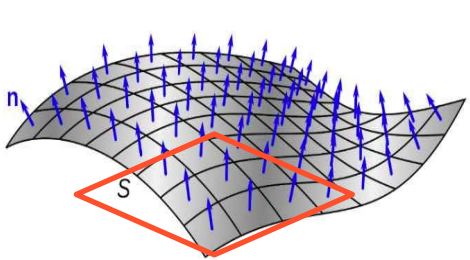
Beyond planar assumption.

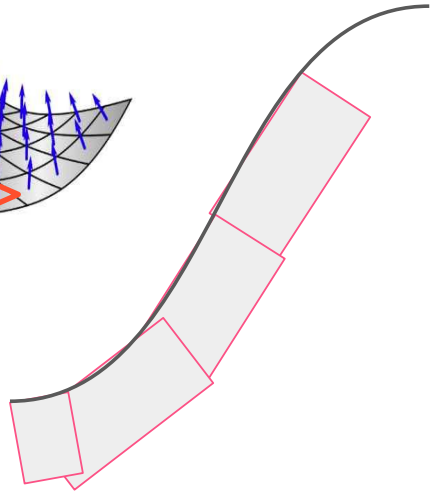
- First, the insight is model scene with convex primitives (planar is a simple design choice)
- Second, fitting planar primitives to normal segments is stronger than you thought:
  - We can fit highly curved structure by segment curves to small pieces
- Finally, we expect more representative primitives for future work





SDF in 2D





planar primitives to fit the surface

Contact modelling ablation

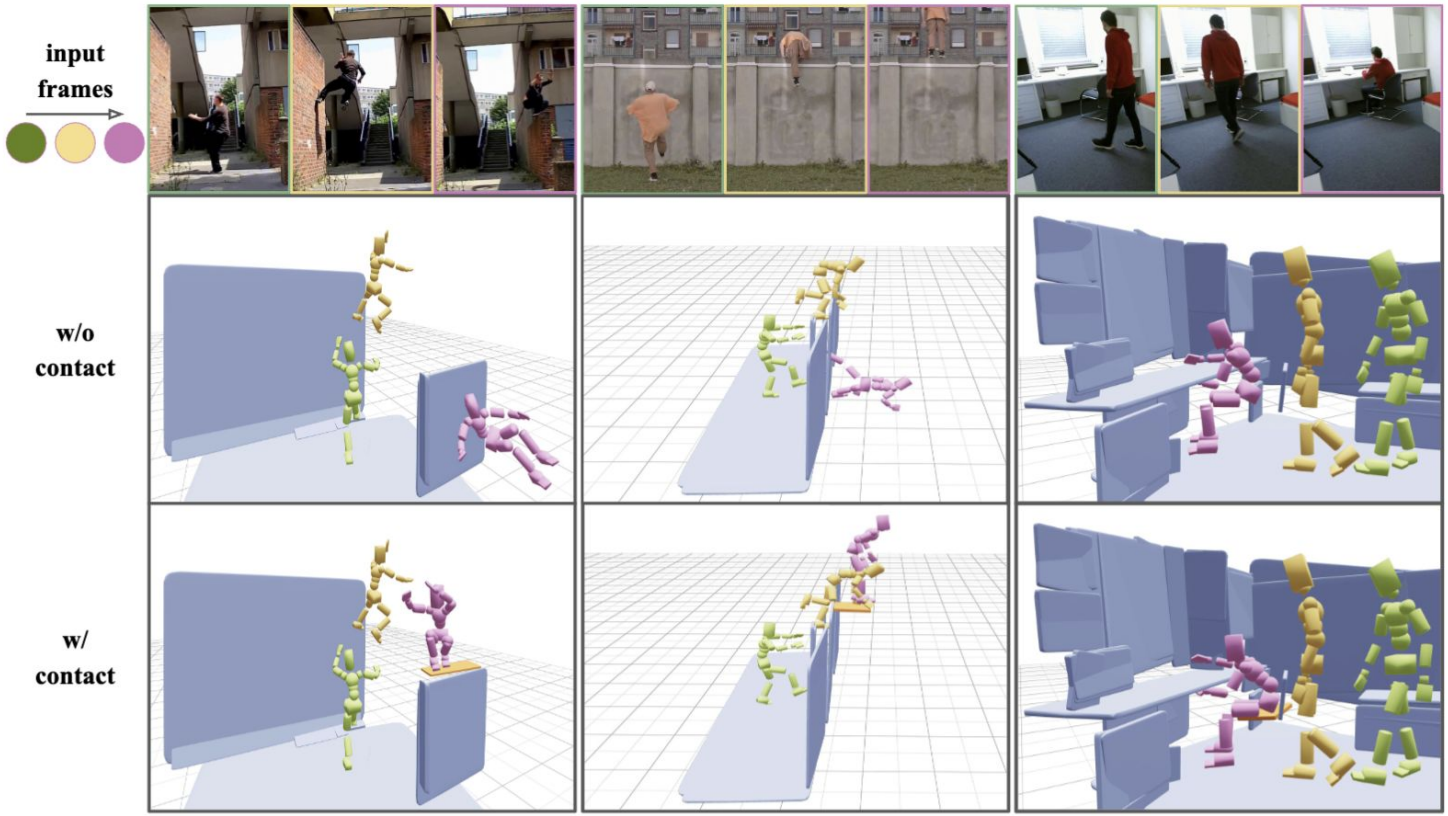


Figure 4: **Ablation on contact-augmented scene completion.** We present 3 sequences in 3 columns. For each sequence, the top strip lists sequential input frames (green → yellow → pink),