Digital Humans – Winter 24/25

Lecture 13_4 – Interaction Reconstruction with Diffusion

Prof. Dr. Gerard Pons-Moll

University of Tübingen / MPI-Informatics

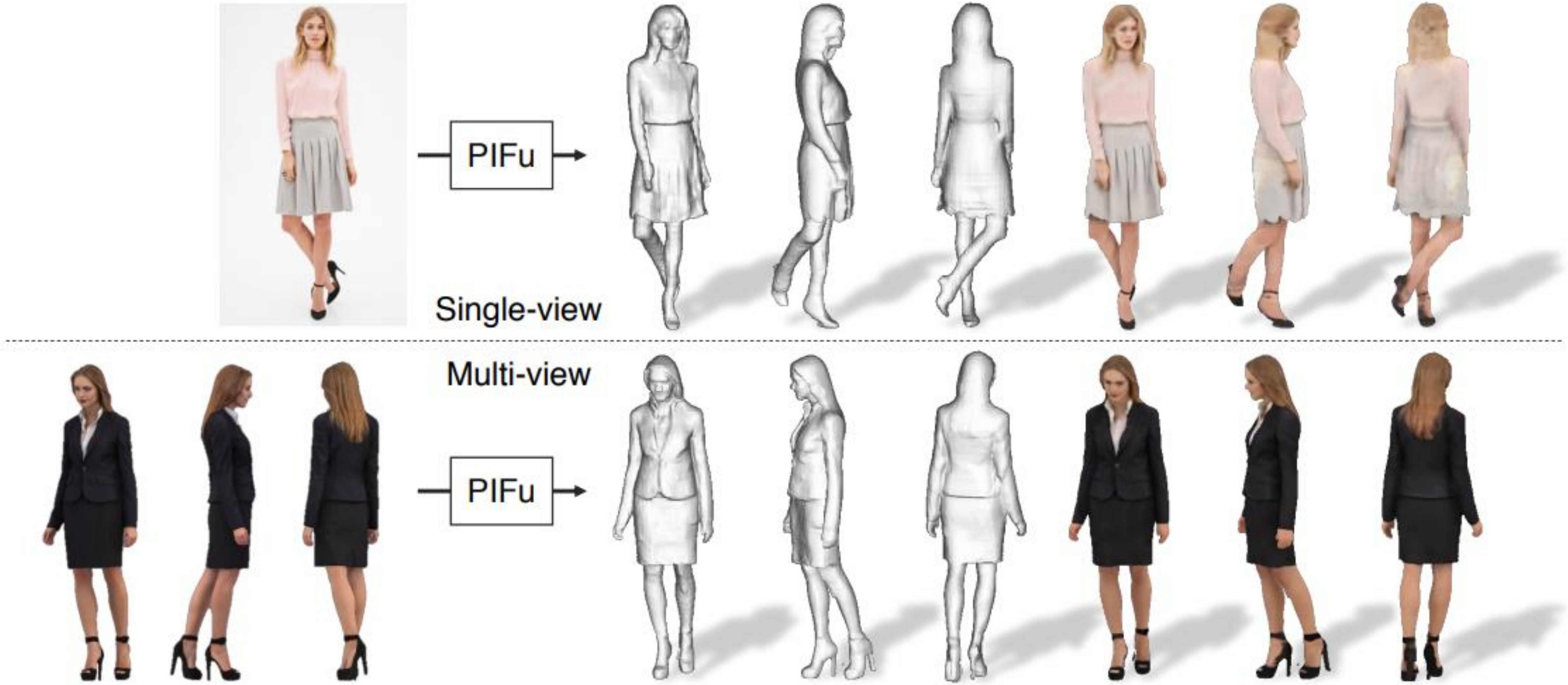
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Main contents

- **Pixel aligned reconstruction and diffusion.**
  - **PiFU revisited.**
  - **Projection conditioned diffusion.**
- Hierarchical diffusion model for interaction.
  - Hierarchical model.
  - Training data preparation.
  - Interaction tracking.

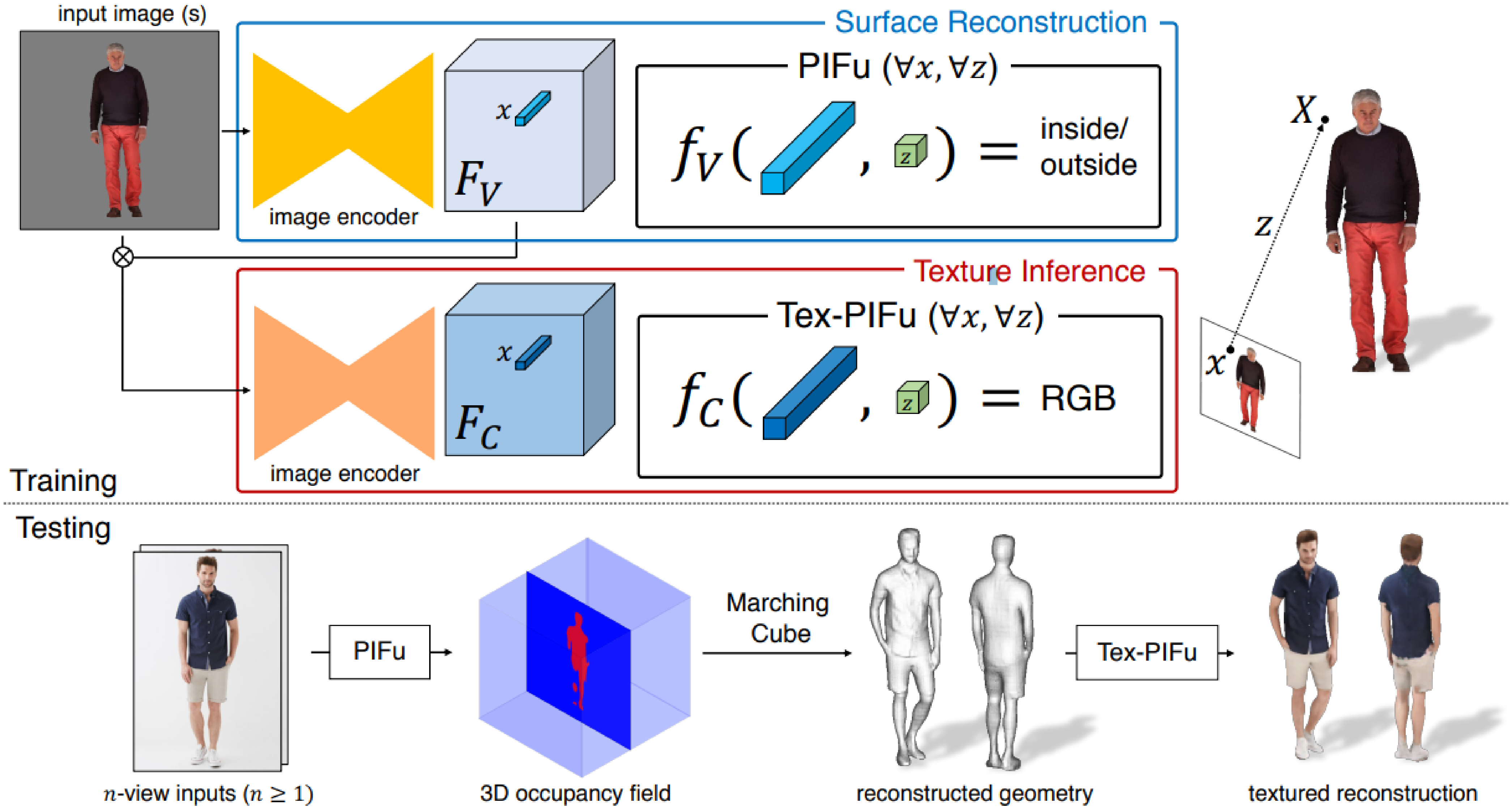
PiFU: pixel-aligned implicit function

- Goal: given single/multiple images, reconstruct the 3D human.



PiFU method overview

- Predicting continuous occupancy and color fields.

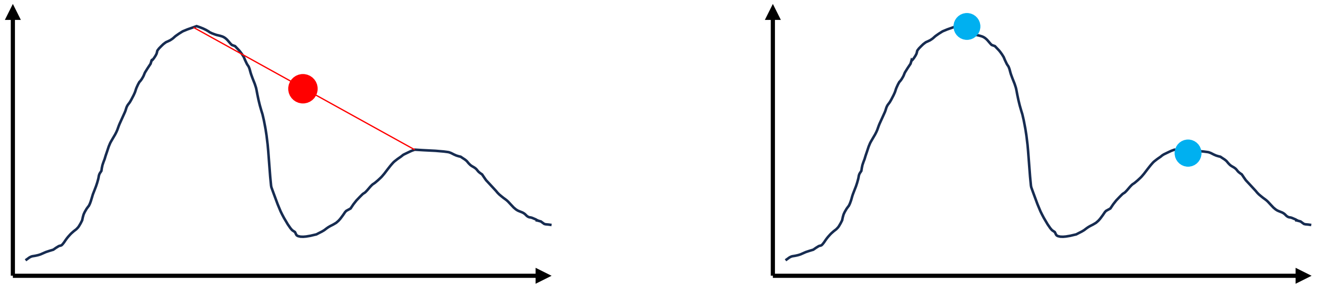


PiFU results: single RGB image to 3D human

- Limitation: deterministic prediction.

Motivation for generative model

- Goal: single view reconstruction is ill-posed.
- Deterministic model might collapse to average value.

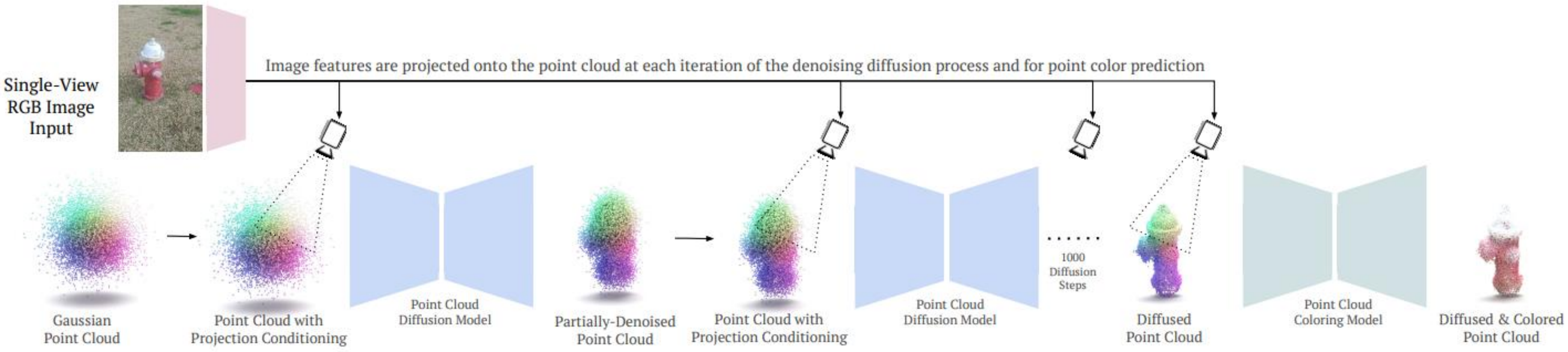


Deterministic: learn an average　　Generative model: learn a distribution

- We should learn a distribution of all possible configurations instead of one prediction.
- Diffusion model for conditional generation!

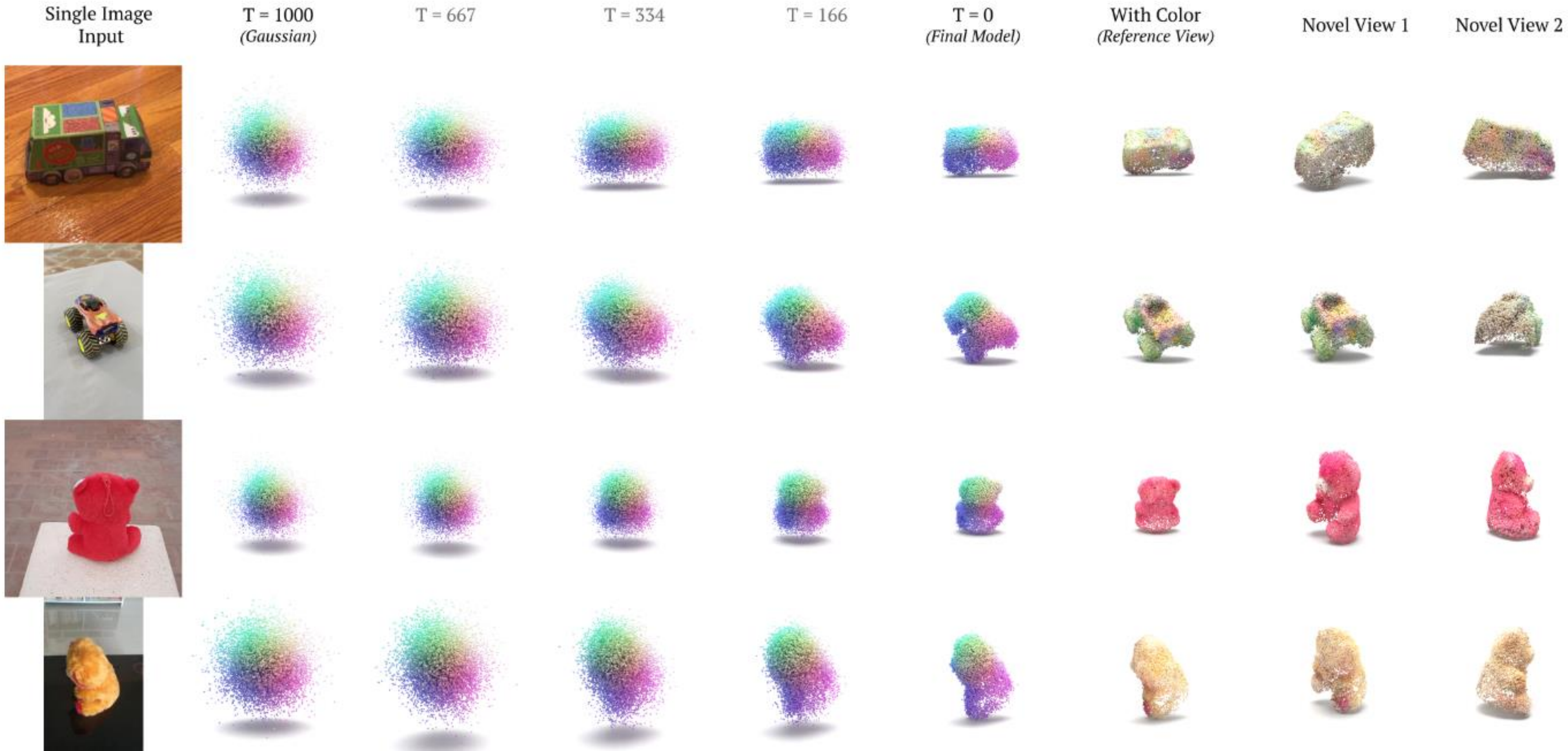
PC2: projection conditioned diffusion model

- Random Gaussian point $p \in \mathbb{R}^3$, perspective projection: $\pi: \mathbb{R}^3 \to \mathbb{R}^2$
- Image encoder $f_\phi: I \in \mathbb{R}^{3\times H\times W} \to \mathbb{R}^{D\times H'\times W'}$
- Pixel aligned feature: $F_\mathrm{p} = f_\theta(I)[\pi(p)]$, $[\cdot]$: bilinear interpolation.
- Diffusion model $\epsilon_\theta: \left(F_p, t\right) \to \mathbb{R}^3$. Predicts update for next step.



PC^2: Melas-Kyriazi et al. CVPR'23.

PC2 results



Main contents

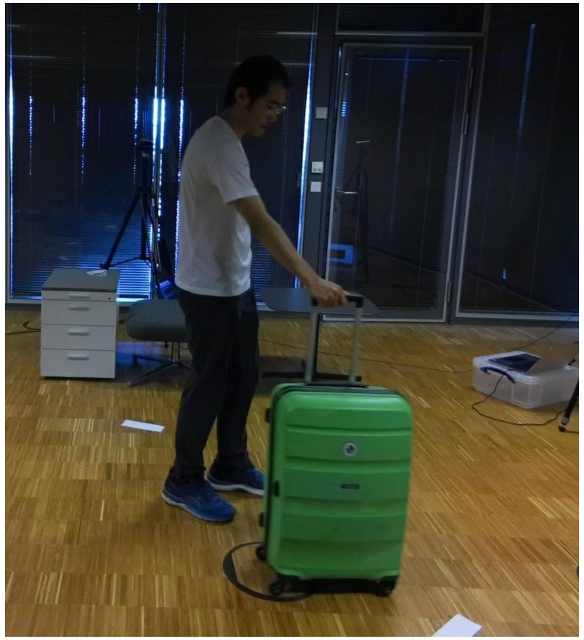
- Pixel aligned reconstruction and diffusion.
  - PiFU revisited.
  - Projection conditioned diffusion.
- **Hierarchical diffusion model for interaction.**
  - **Hierarchical model.**
  - **Training data preparation.**
  - **Interaction tracking.**

PC^2: Melas-Kyriazi et al. CVPR'23.

Interaction reconstruction with diffusion

- PC2 is a general method for 3D reconstruction.
- Can we train it directly to reconstruct human + object?
  - No, interaction is a complex combinatorial space!
  - Interaction = human pose & shape space × object pose & shape space.

Input image

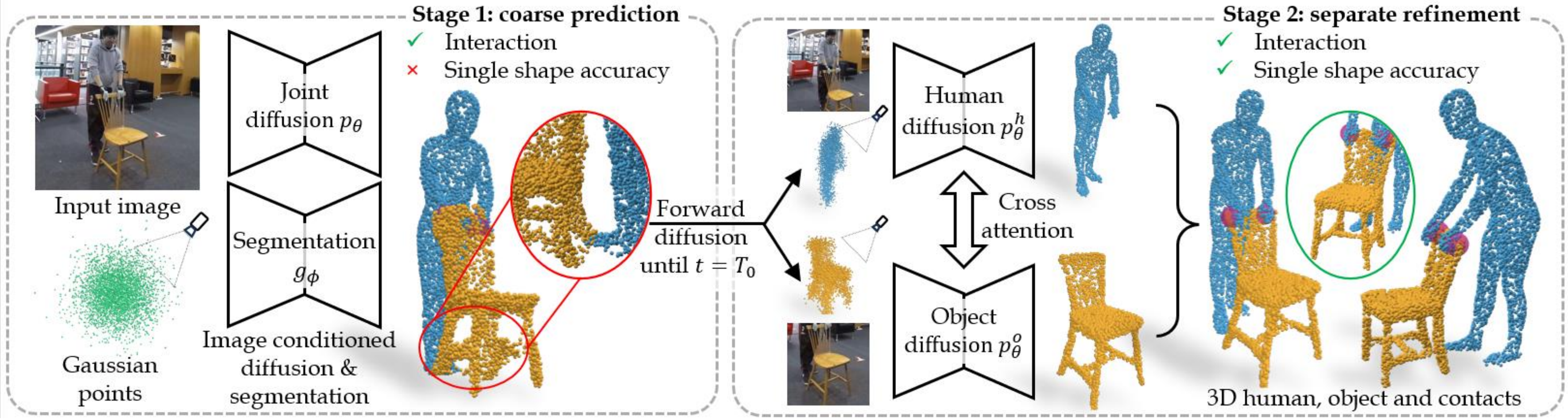


PC2 results

Hierarchical diffusion model

- Key idea: learn subspaces for interaction and individual shapes separately.
  - Three models to learn human, object and interactions.
  - One segmentation model to separate human and object in stage 1.
  - Communicate between human and object using cross attention.

$$\mathbf{F}_l^{h \mapsto o} = \mathrm{Attn}(\mathrm{enc}(\mathbf{P}_l^o), \mathrm{enc}(\mathbf{P}_l^h), \mathbf{F}_{\mathbf{P}_l^h})$$



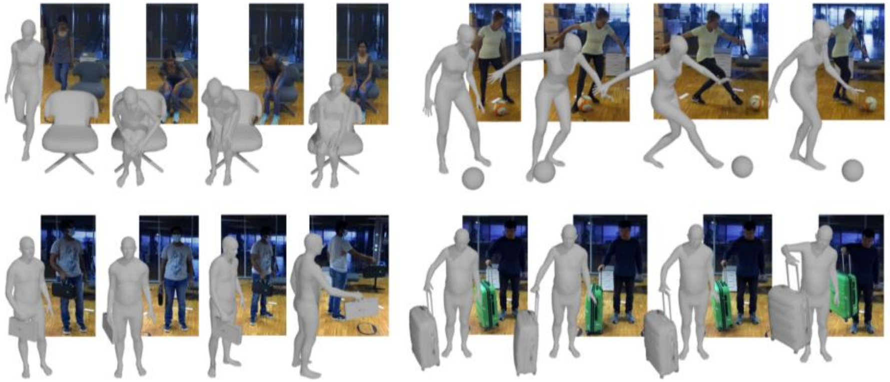
BEHAVE: Bhatnagar et al. CVPR’23.
InterCap: Huang et al. GCPR’22.
BEDLEM: Black et al. CVPR’23.
Objaverse-XL: Deitke et al. AriXiv’23.

Challenge: lack of data

- Diffusion models are data hungry: stable diffusion trained on 5B images.
- Existing real interaction data is limited: only 10-20 object shapes.
- Human or object shapes are much more diverse.

BEHAVE: 7 humans, 20 objects, 5 scenes

BEDLEM: 271 bodies, 1691 clothing

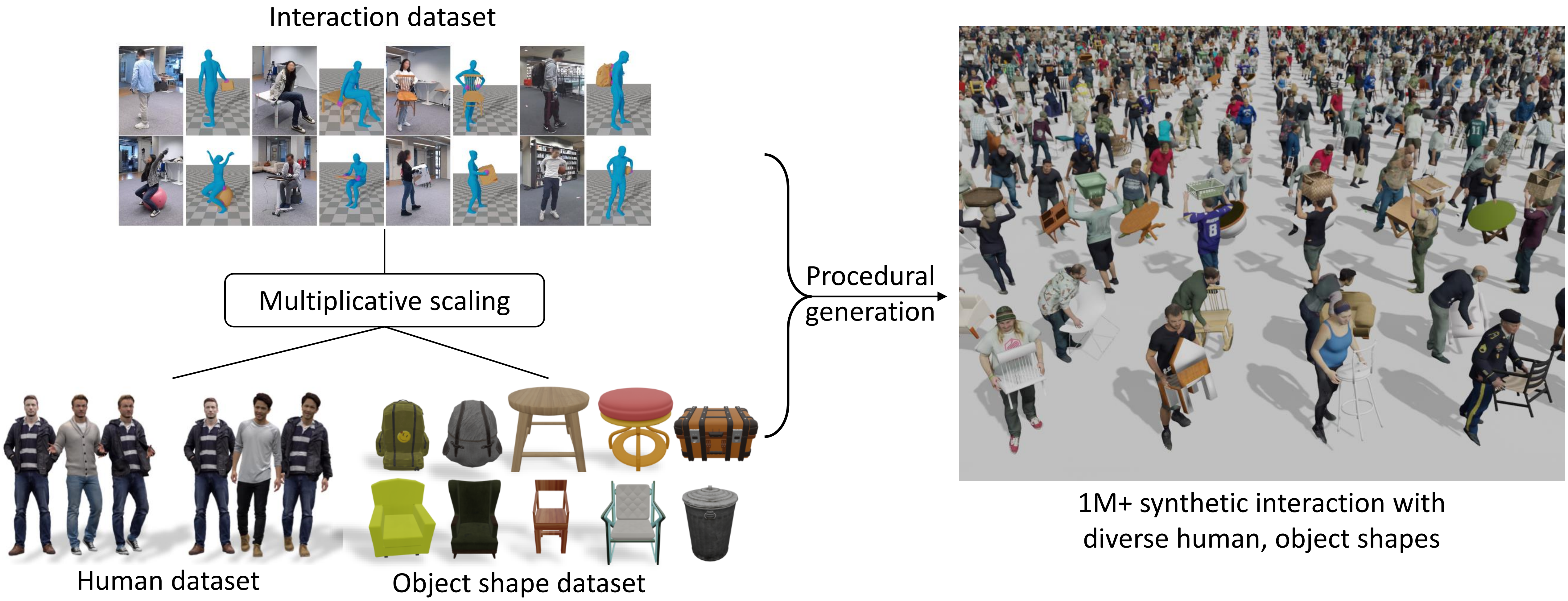


InterCap: 10 humans, 10 objects, 1 scene

Objaverse-XL: 10M+ 3D objects

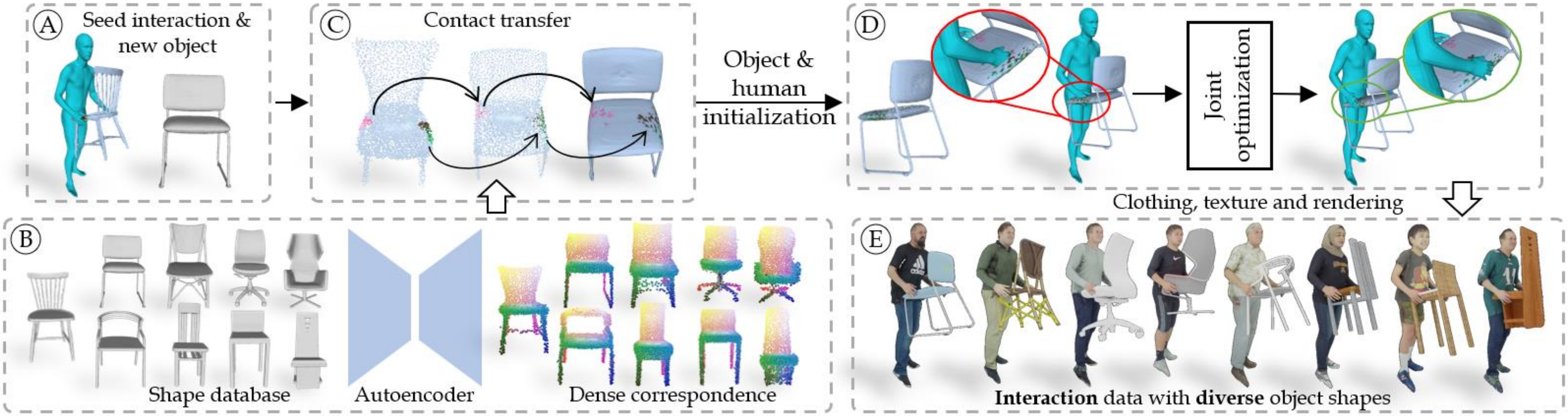
Key idea: generate synthetic data

- Procedurally generate interaction data with diverse human object shapes.



ProciGen: **Proc**edural **i**nteraction **Gen**eration

- Key idea: humans interact similarly with objects of the same category.
- Autoencoder training: self-supervised with Chamfer distance.
  - Requires object to be aligned in the canonical space.



Contact based optimization

- Optimize: human pose shape θ, β, object pose: $T \in SE(3)$
- Loss: $L(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{T}) = \lambda_c L_c + \lambda_n L_n + \lambda_{\text{colli}} + \lambda_{\text{init}} L_{\text{init}},$
- Define contacts: $\mathcal{C} = \{(i, j) \mid ||\mathbf{H}_i - \mathbf{T}^{-1} f(\mathbf{TP})_j)||_2^2 < \sigma\}$
  - P: object mesh surface samples. T: object pose, from interaction to canonical space. $f: \mathbb{R}^{N\times3} \rightarrow \mathbb{R}^{M\times3}$, unordered points to ordered points.
  - $f(TP)$ has semantic correspondence with new shape P'.

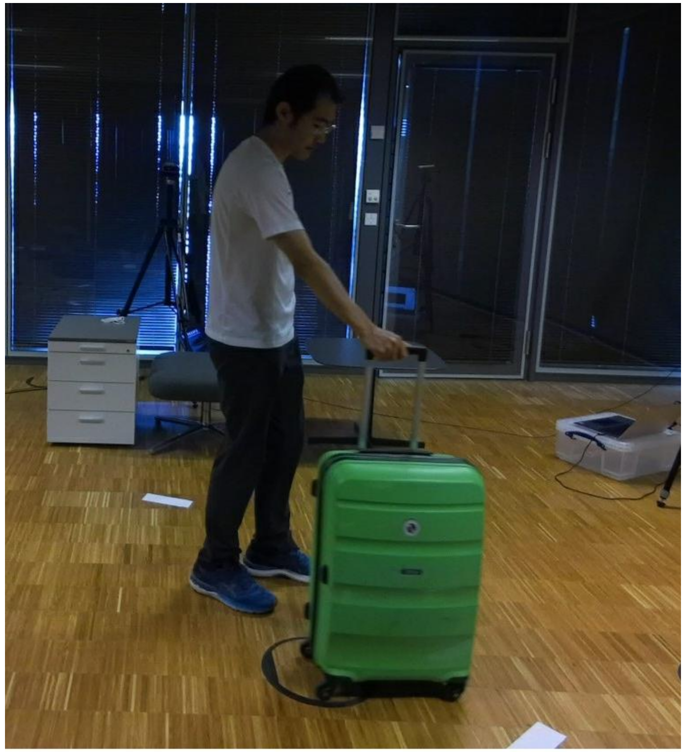
- **Contact:** $L_c = \sum_{(i,j)\in\mathcal{C}} ||\mathbf{H}_i - \mathbf{P}'_j||_2^2$, minimizing the distance between contact points.
- **Normal:** $L_n = \sum_{(i,j)\in\mathcal{C}} ||1 + \mathbf{n}_i^T \mathbf{n}_j||_2^2$, ensuring that normals $\mathbf{n}_i, \mathbf{n}_j$ of contacting faces point in opposite directions.
- **Interpenetration:** L_{colli} penalizing interpenetration based on the bounding volume hierarchy [88].
- **Initialization:** L_{init} is the L2 distance between new and original human pose, regularizing the deformation.

ProciGen dataset: **1M+ interaction** images with **21k+ objects**

Our method reconstruct highly accurate shapes

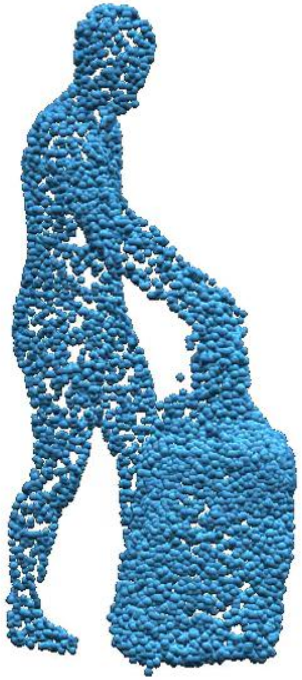
- Our method obtains high quality interaction reconstruction.

Input image



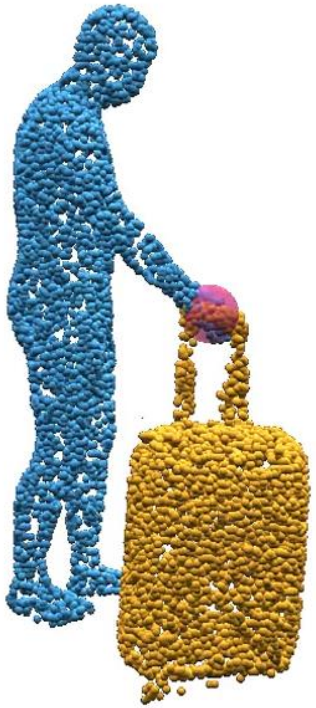
PC^2

√ Template-free
× Shape accuracy
× Interaction semantics
× Generalization



Ours

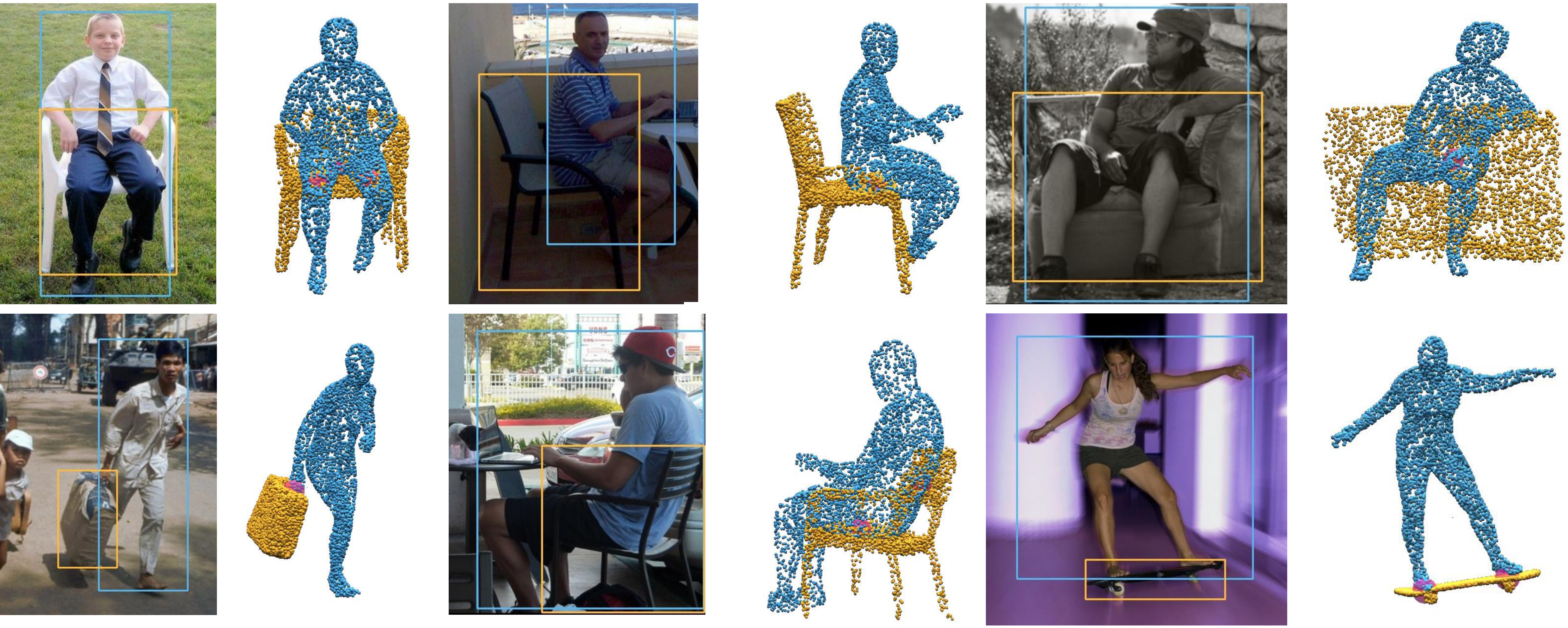
√ Template-free
√ Shape accuracy
√ Interaction semantics
√ Generalization



COCO dataset: Lin et al. ECCV’14.
Xie et al. ProciGen, CVPR’24.

Our method generalizes to COCO dataset

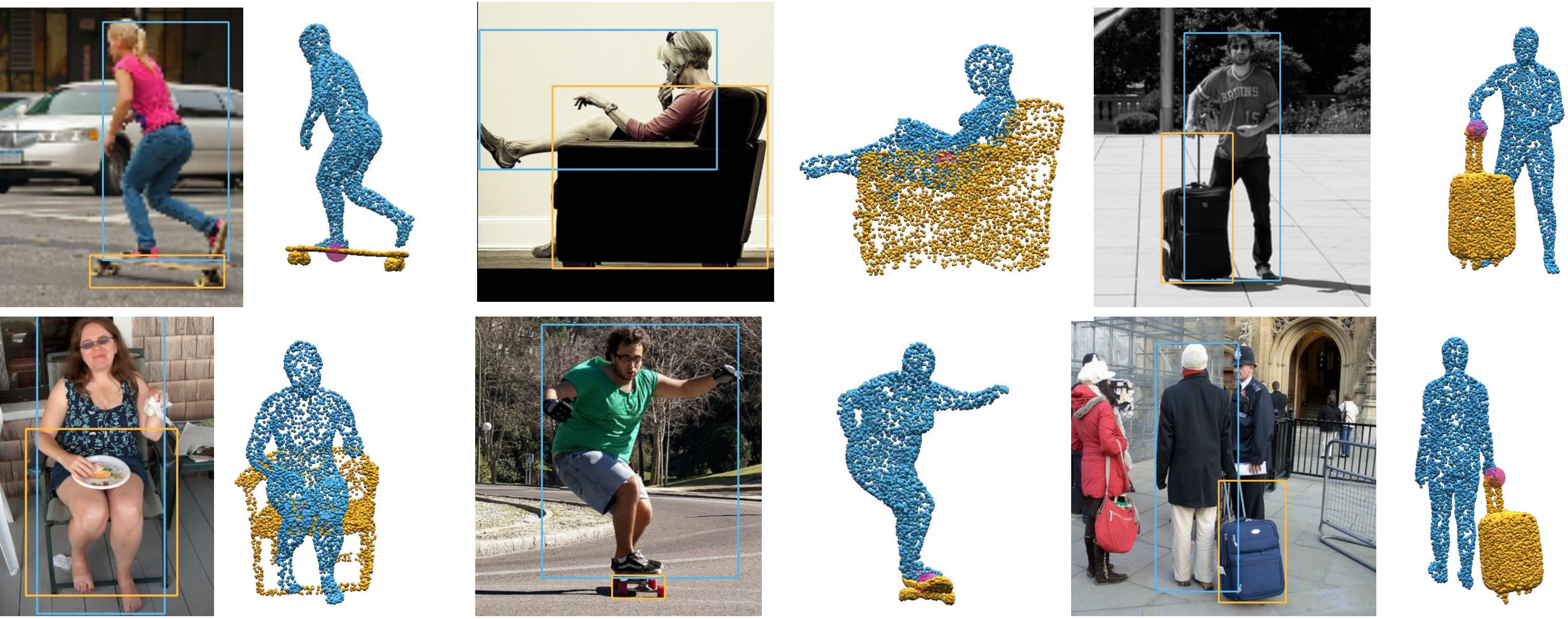
Input image | Our result | Input image | Our result | Input image | Our result



COCO dataset: Lin et al. ECCV'14.
Xie et al. ProciGen, CVPR'24.

Our method generalizes to COCO dataset

Input image Our result Input image Our result Input image Our result



InterTrack: Tracking Human Object Interaction without Object Templates

Input RGB video

Our tracking results

Challenges.

- Heavy occlusion and dynamic motion.
- No template: need to reason both shape and pose at the same time.

…

…

HDM: Xie et al. CVPR'24.
BEHAVE: Bhatnagar et al. CVPR'22.

Challenge: no correspondence across frames.

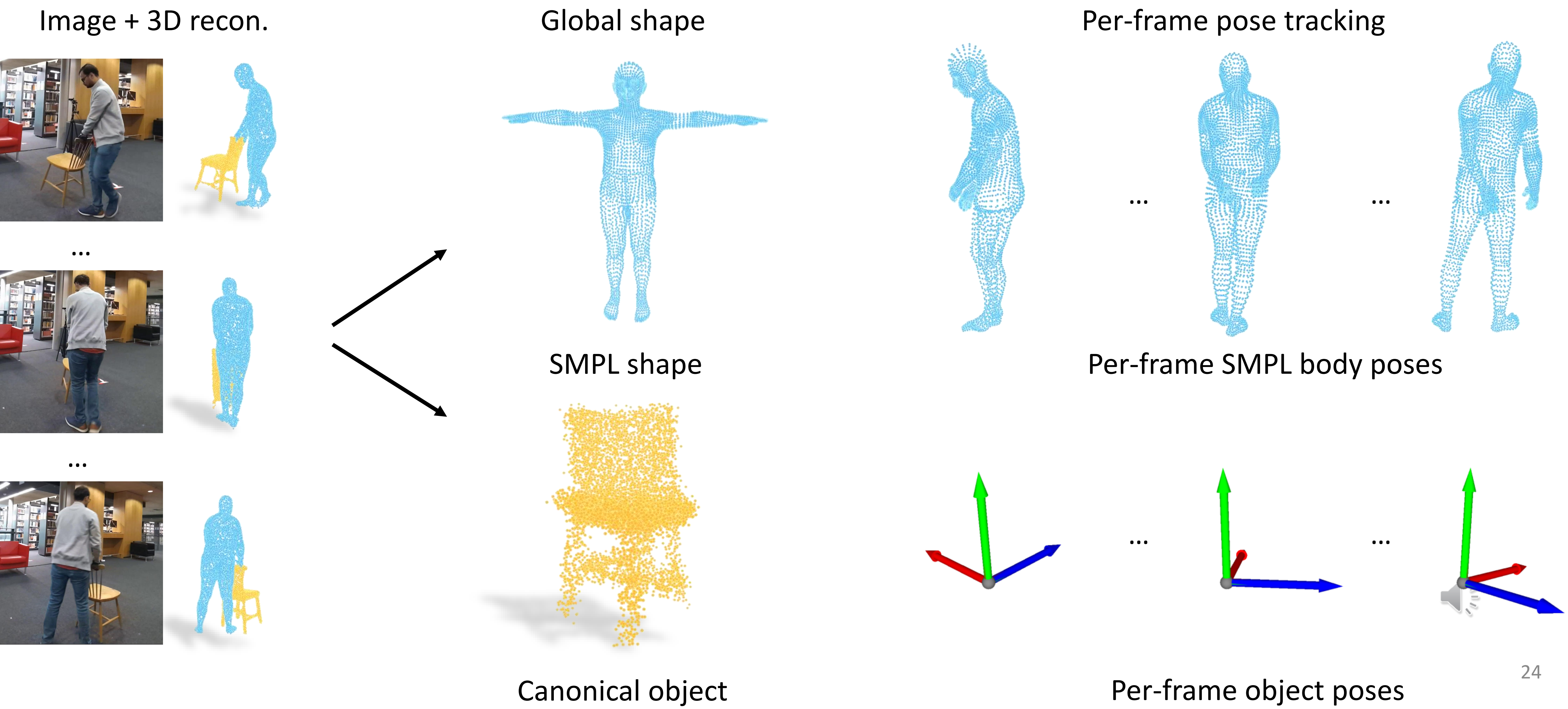
- HDM: image-based interaction reconstruction.
  - ✓ Template free reconstruction.
  - × No temporal information: inconsistent shapes.
  - × No correspondence across frames.

Input sequence

HDM result

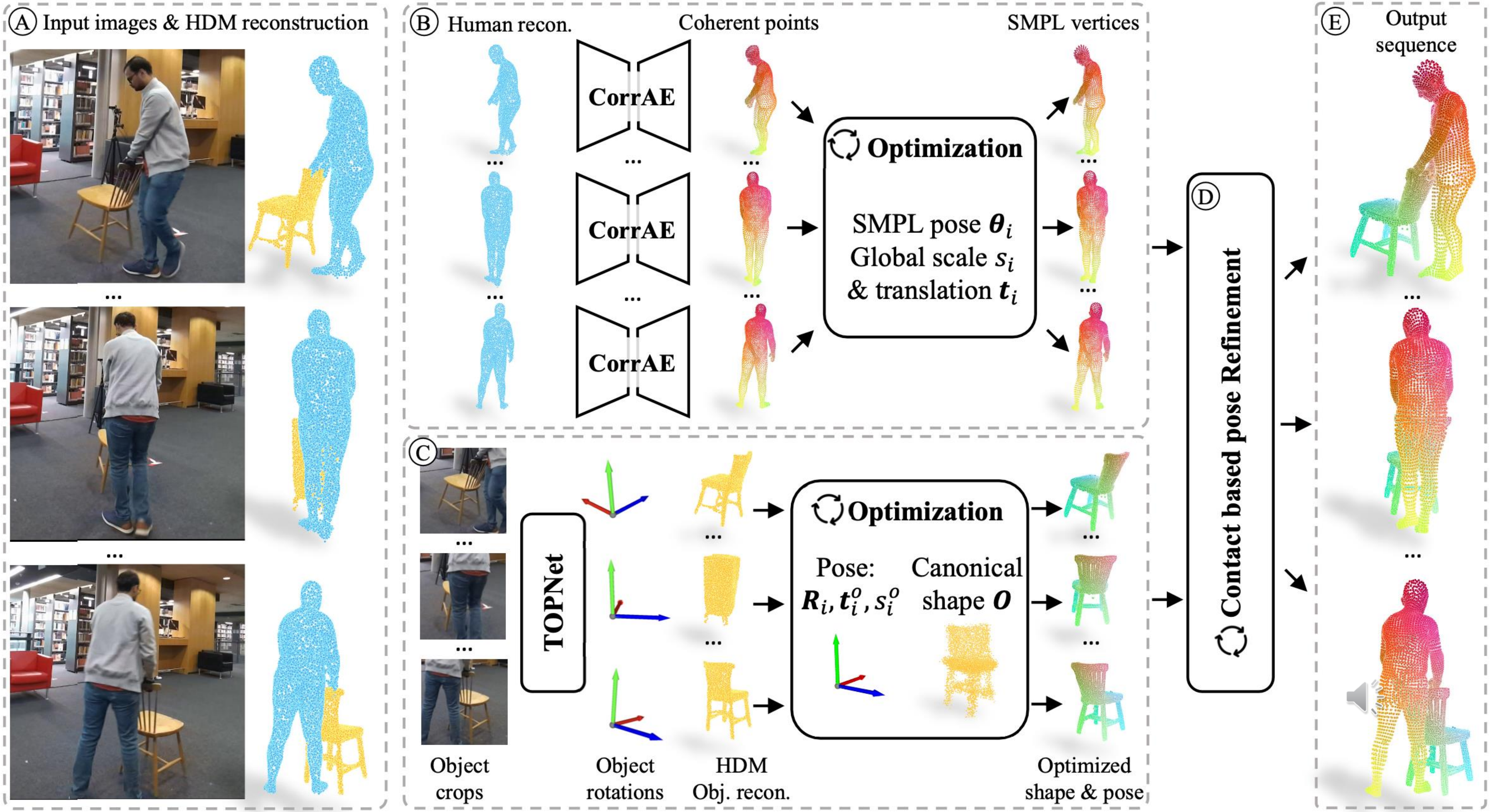
Key idea: constrain solution space to shape & pose.

- 4D tracking = one global shape + per-frame poses.



HDM: Xie et al. CVPR'24.
InterTrack: Xie et al. 3DV'25.

InterTrack: method overview.

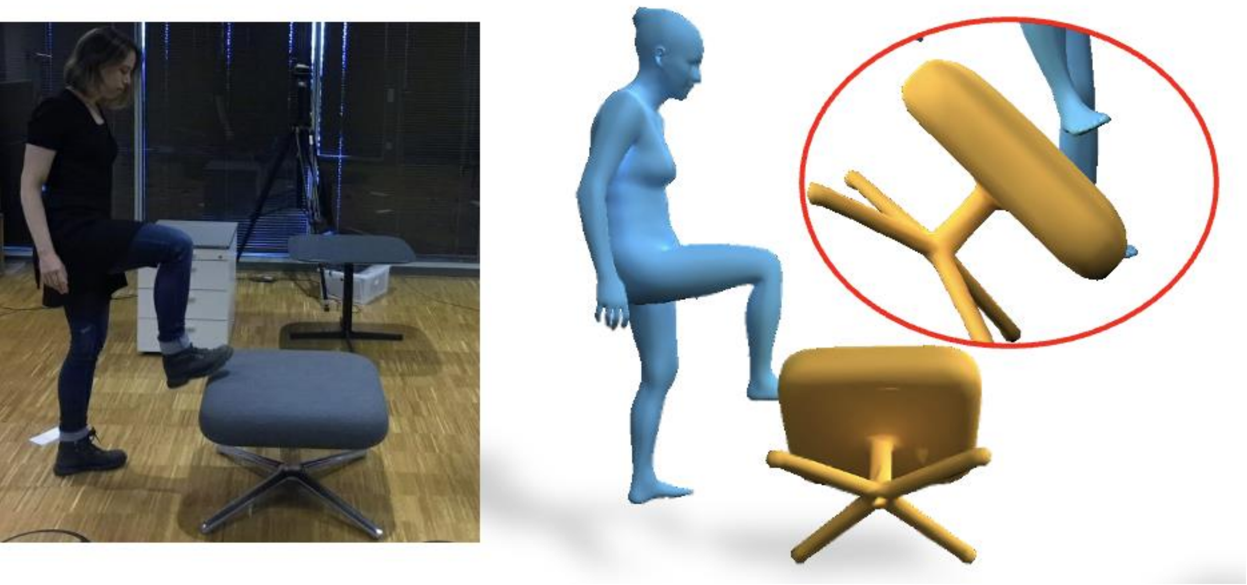


CHORE: Xie et al. ECCV'22.

Training data problem.

- Our pose estimator TOPNet requires video data to train.
- Prior method trained on real data has limited generalization ability.
  - E.g.: CHORE trained on BEHAVE cannot work on InterCap dataset.
- Solution: generate synthetic data.

Input image CHORE results



Input image CHORE results

ProciGen-Video: 8.5k videos with 4.1k different objects.

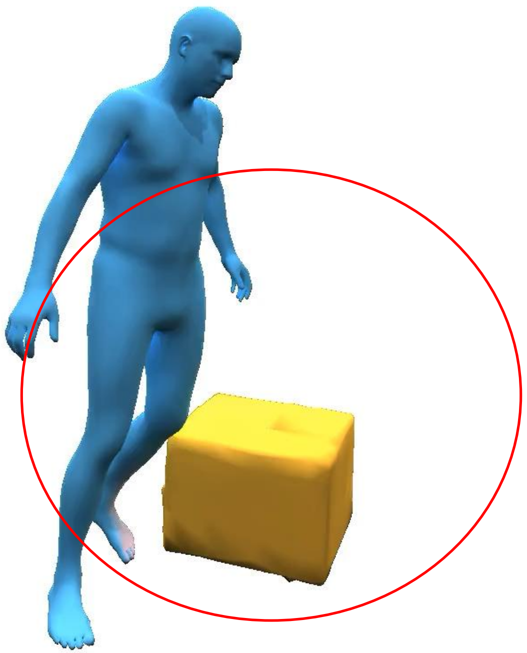
VisTracker: Xie et al. CVPR'23.
BEHAVE: Bhatnagar et al. CVPR'22.
InterTrack: Xie et al. 3DV'25.

Comparison with VisTracker on BEHAVE.

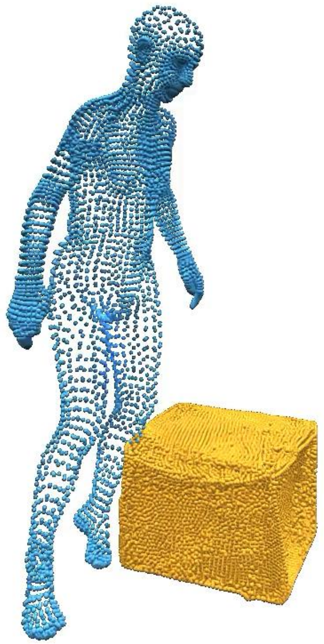
- Our method produces more stable tracking.

Input sequence

VisTracker result



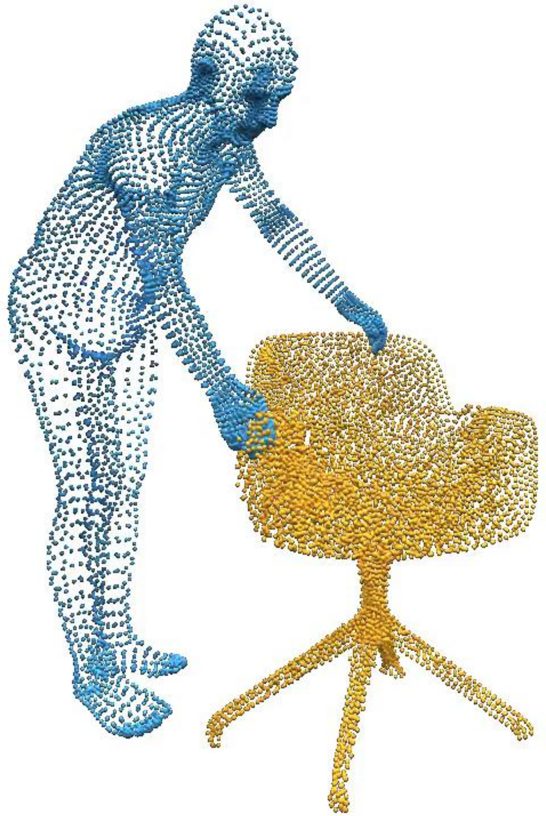
Our result



Our method generalizes to mobile phone videos.

Input video

Tracking result



Take away messages

- Pixel-aligned features are important for detailed reconstruction (PiFU).
- Generative models are better suited for ill-posed problems (monocular reconstruction PC2).
- For interaction, we can decompose the combinatorial space into human, object subspaces and learn them separately (HDM).
- Procedural synthetic generation is the way to scale up interaction/combinatorial data (ProciGen).
- Complex 4D tracking can be decomposed into global shape reconstruction + per-frame pose estimation (InterTrack).