Modelling Dynamic 3D Human-Object Interactions From Capture to Synthesis

PhD Defense

Omid Taheri
July 4, 2024

Committee

Prof. Dr. Gerard Pons-Moll
Prof. Dr. Michael J. Black
Prof. Dr. Angela Dai
Prof. Dr. Andreas Geiger

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

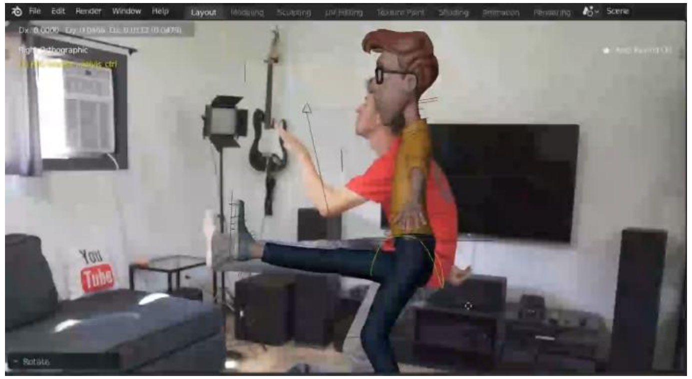
PERCEIVING SYSTEMS
MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

Modelling Humans

Long term goal in Computer Vision/Graphics → **Modelling Virtual Humans**

Movies [1]

Games [2]



Animations [3]

[1] https://youtu.be/LH18nGolUKo?si=OUO_V6wfzRiR_7nV [2] https://youtu.be/gSpGHZ1rqEA?si=Z4LJO3RKZz2rLi1J [3] https://youtu.be/TjJLluFKA20?si=4QnWJ-GvnNjZMN-V

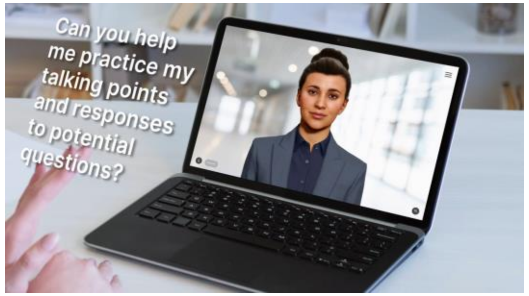
Intro

Virtual World

Things are going virtual:

- AR/VR
- Telepresence
- Metaverse
- Embodied assistants (ChatGPT)

Education [1]



AI Assistants [2]

Entertainment [1]

Learning [1]

Virtual world → Virtual humans

- **Demand for modelling virtual humans**
- **Immersive experience → Realistic humans**

Socializing [1]

[1] https://youtu.be/ycV7iFG_I-U?si=I7VrR9E1IWt8bmDr [2] https://youtu.be/LHbbhakLpBs?si=0KgDVwcaTfCK43QQ

A Fundamental Aspect of Humans

A Fundamental Aspect of Humans

MOTION

Intro

Human Motion

Our motions are:

- At the heart of everything we do
- Complicated and rich
- Fast and diverse

Used for:

- Communication
- Interaction
- Emotion
- Navigation

pexels.com/video/[a-woman-painting-using-a-paintbrush-5670466] - [dinner-meal-breakfast-kitchen-4296854] - [hands-music-musician-piano-4088191] [a-person-working-on-a-computer-and-writing-notes-on-pieces-of-papers-3196061] – [making-a-latte-4929541] – [mixing-sugar-and-egg-6138113] – [mixing-sugar-and-egg-6138113] – [person-working-on-his-laptop-1851768]

The Role of Hands

Primary tool for interaction/manipulation

- Dexterous
- Do delicate or rough motions
- Sense hot or cold
- Feel soft or rough surfaces
- Make gestures

- 70% of daily activities

----
pexels.com/video/[a-woman-painting-using-a-paintbrush-5670466] - [dinner-meal-breakfast-kitchen-4296854] - [hands-music-musician-piano-4088191] [a-person-working-on-a-computer-and-writing-notes-on-pieces-of-papers-3196061] – [making-a-latte-4929541] – [mixing-sugar-and-egg-6138113] – [mixing-sugar-and-egg-6138113] – [person-working-on-his-laptop-1851768]

The Role of Hands

A day without hands – try at home!

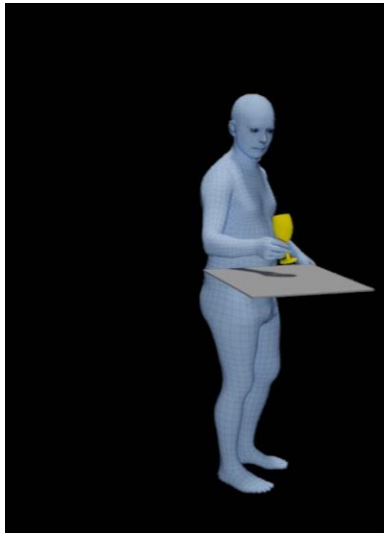
[1] https://youtu.be/ycV7iFG_l-U?si=I7VrR9E1IWt8bmDr [2] https://youtu.be/Sq1QZB5baNw?si=SILfwcgdcPhf_NND

"The hand is the visible part of the brain."

- Immanuel Kant

Intro

The Goal



Train Computers to model realistic virtual-human motions with a focus on hand-object interactions.

Why?

- Applications in various fields:
  - Entertainment, Games, Movies
  - Architecture, Healthcare, Education
- Answer the demand for modelling human motions:
  - Fast
  - Cost-effective
- Computers better understanding human

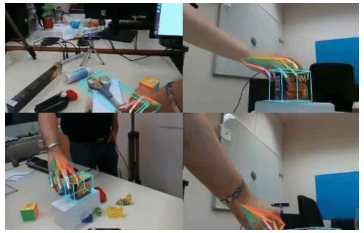
Problems

Separating ***hands*** and ***objects*** from the body motions and vice versa

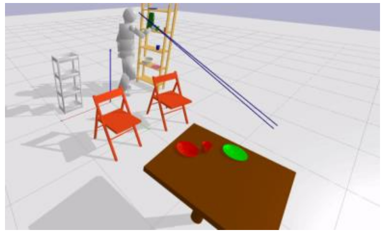
Available datasets:

- Body, no hands or objects
- Body + objects, no hands
- Hands + objects, no body
- No accurate ground truth motions

[Araujo et al. CVPR'2023]



[Hampali et al. CVPR'20]



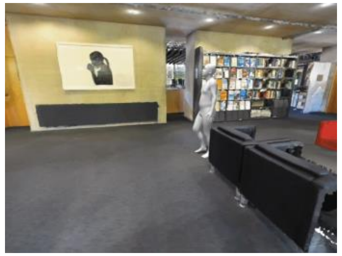
[Kratzer et al. RAL 2020]

[Mandery et al. T-RO'16]

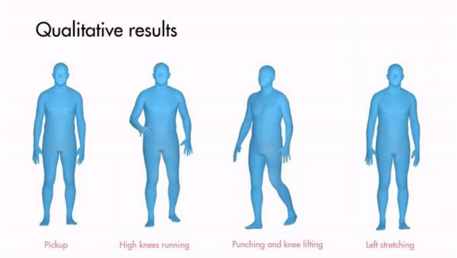
Motion generation methods:

- Body in "isolation", no interactions
- Body + objects, no hands
- Hands + objects, no body

[Li et al. Sigasia'2023]



[Mir et al. 3DV'2024]



[Petrovich et al. ICCV'21]

[Starke et al. TOG'19]

Intro

What are the challenges?

Interactions are essential part of human motions **- high accuracy**

Primary tools for interactions → **Hands**

Hands are:

- Dexterous → Fast or slow motions
- High DOF → Diverse and complex motions
- Small → Are often occluded during interactions

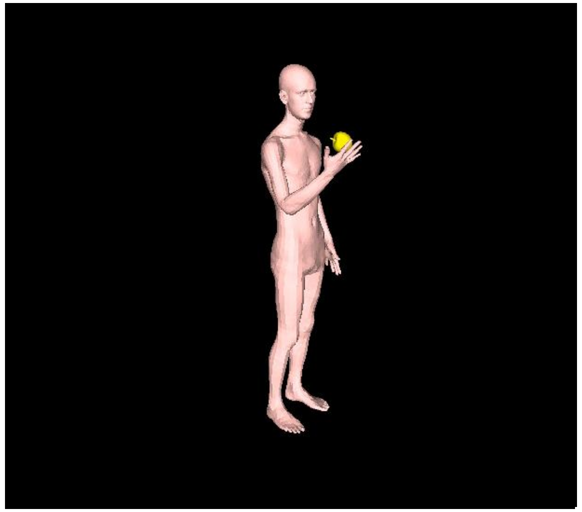
Therefore:

- Challenging to track them → Datasets
- Hard to generate their motion accurately → Modelling

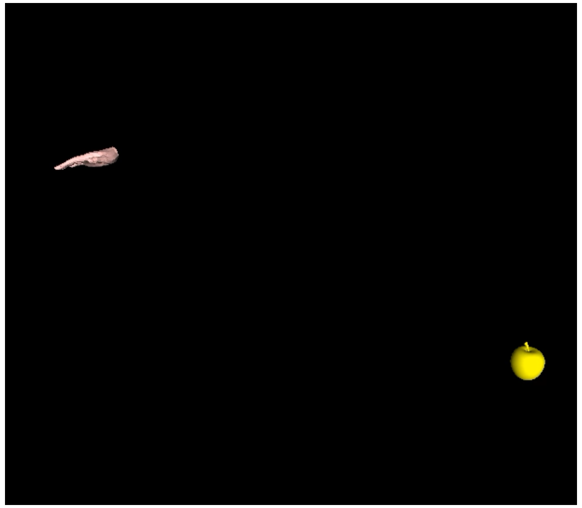
---

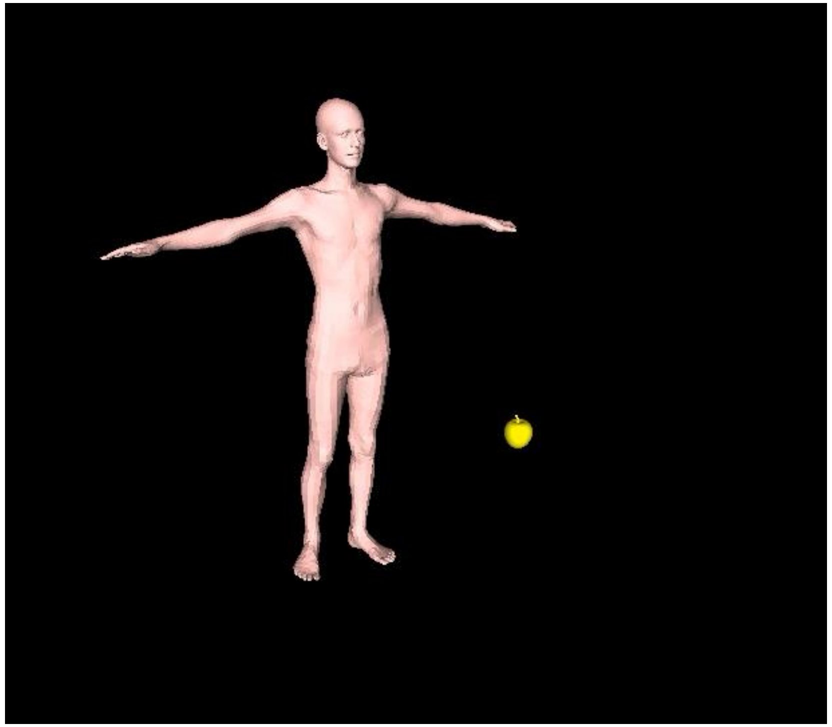
pexels.com/video/[person-using-a-rubik-s-cube-8471394] - [a-person-using-a-chisel-on-wood-5972648] – [

Complete Human Motion



+





Human Motion:

Body and Hands together

Our Contributions

Datasets

Human-Object Interaction

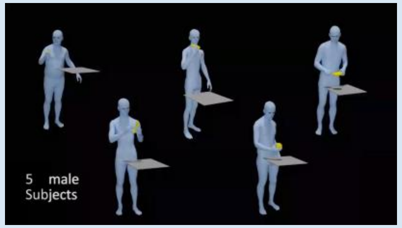
Modelling

Our Contributions

Datasets

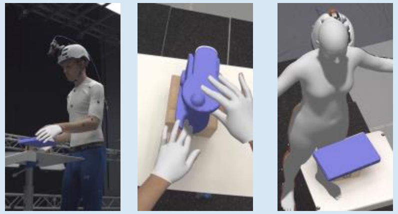
Human-Object Interaction

Modelling



GRAB

Whole-Body Human Grasping of Objects – **3D**



ARCTIC

Dexterous Bimanual Manipulation of **Articulated Objects** – **3D + RGB**

InterCap

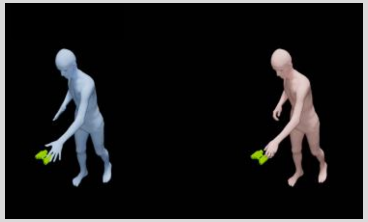
Markerless Tracking of Humans and Objects in Interaction – **3D + RGBD**

Generating **4D Whole-Body Motion** for Hand-Object Grasping

GOAL

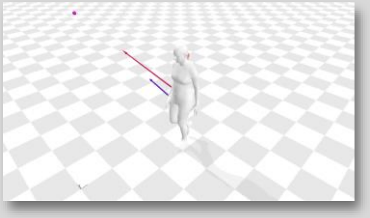
Generating **Interaction Poses** Using Spatial Cues and Latent Consistency

GRIP



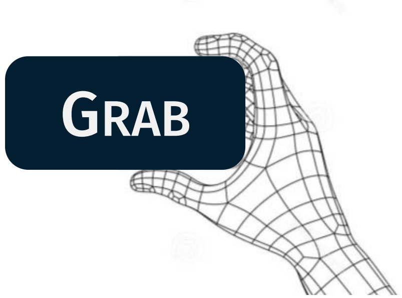
Intention-Guided Human Motion Generation

WANDR



Our Contributions

Intro
GRAB
GOAL
GRIP
Conc



A Dataset of **Whole-Body** Human Grasping of Objects

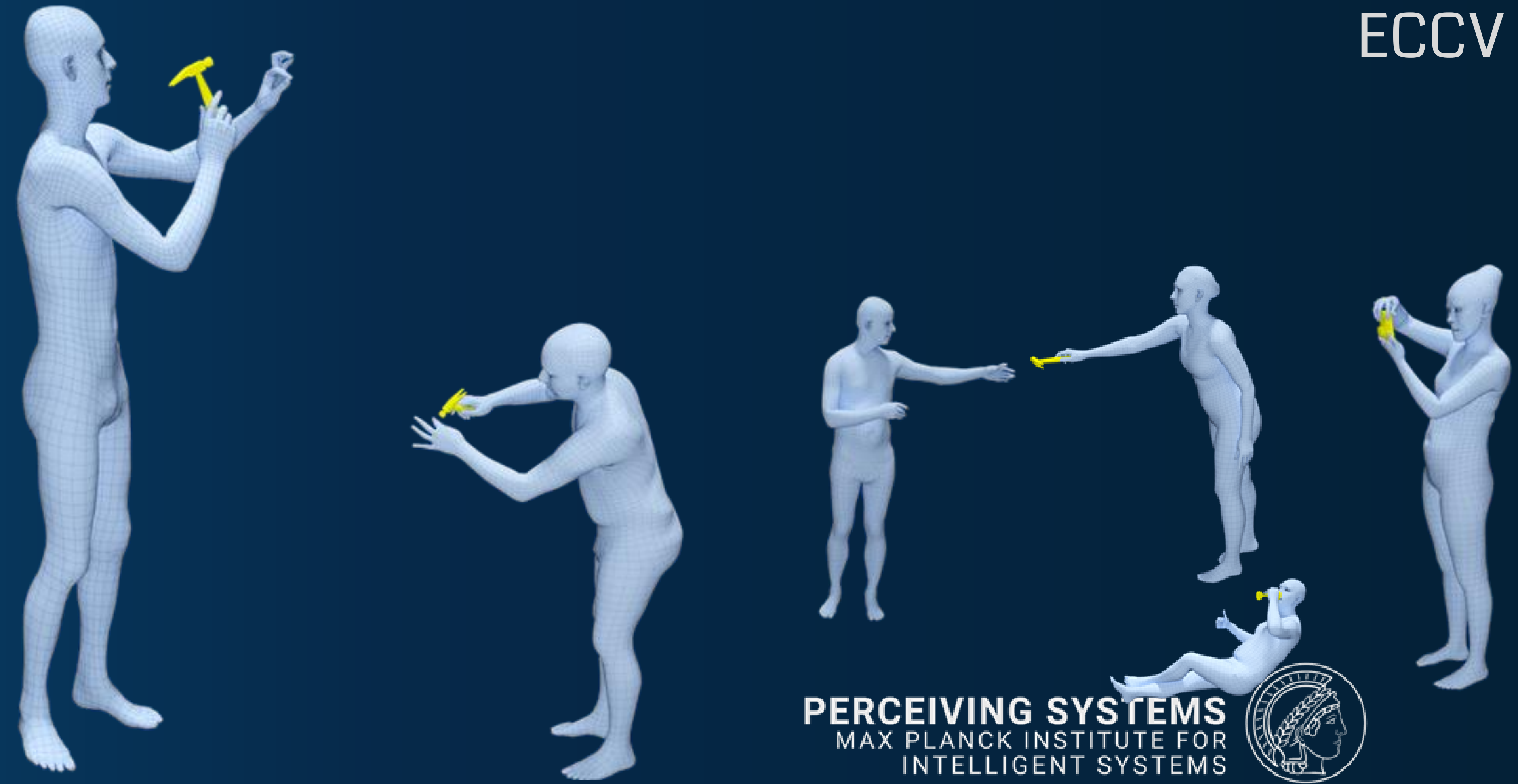
Generating **4D Whole-Body Motion** for Hand-Object Grasping

Generating **Interaction Poses** Using Spatial Cues and Latent Consistency

GRAB: A Dataset of Whole-Body Human Grasping of Objects

Omid Taheri, Nima Ghorbani, Michael J. Black, Dimitrios Tzionas

ECCV 2020

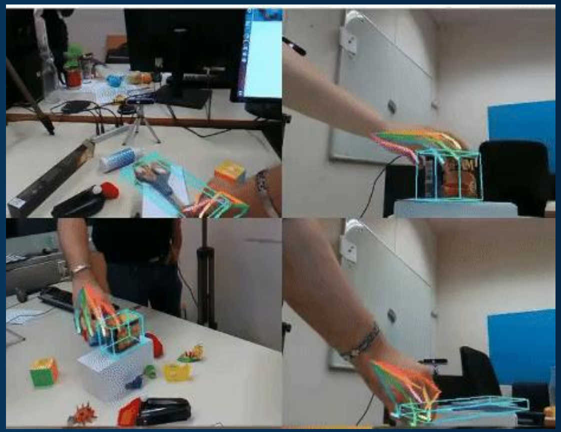


PERCEIVING SYSTEMS
MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

Problems

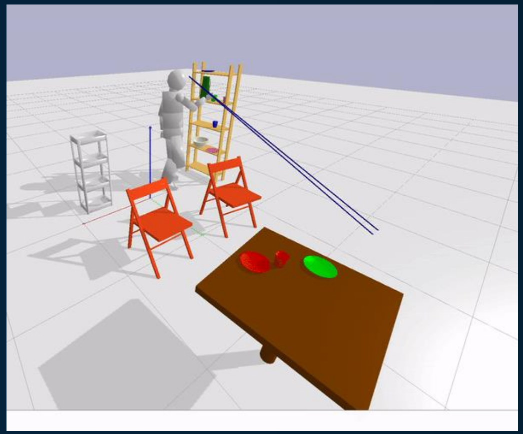
Modelling Human-Object interactions requires

DATA!



[Hampali et al. CVPR'20]

[Mandery et al. T-RO'16]

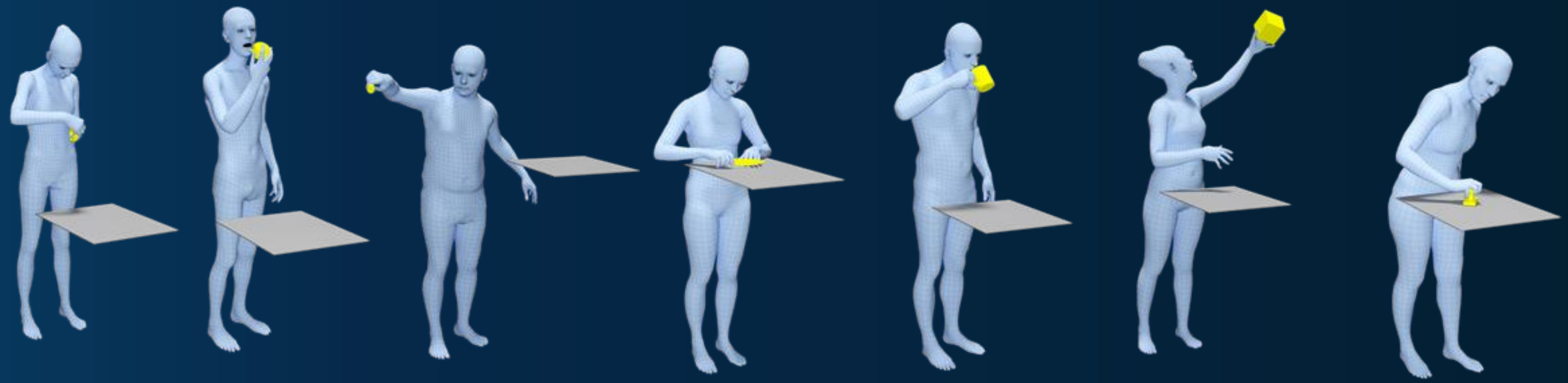


[Kratzer et al. RAL 2020]

Objective

Fill the gap by capturing:

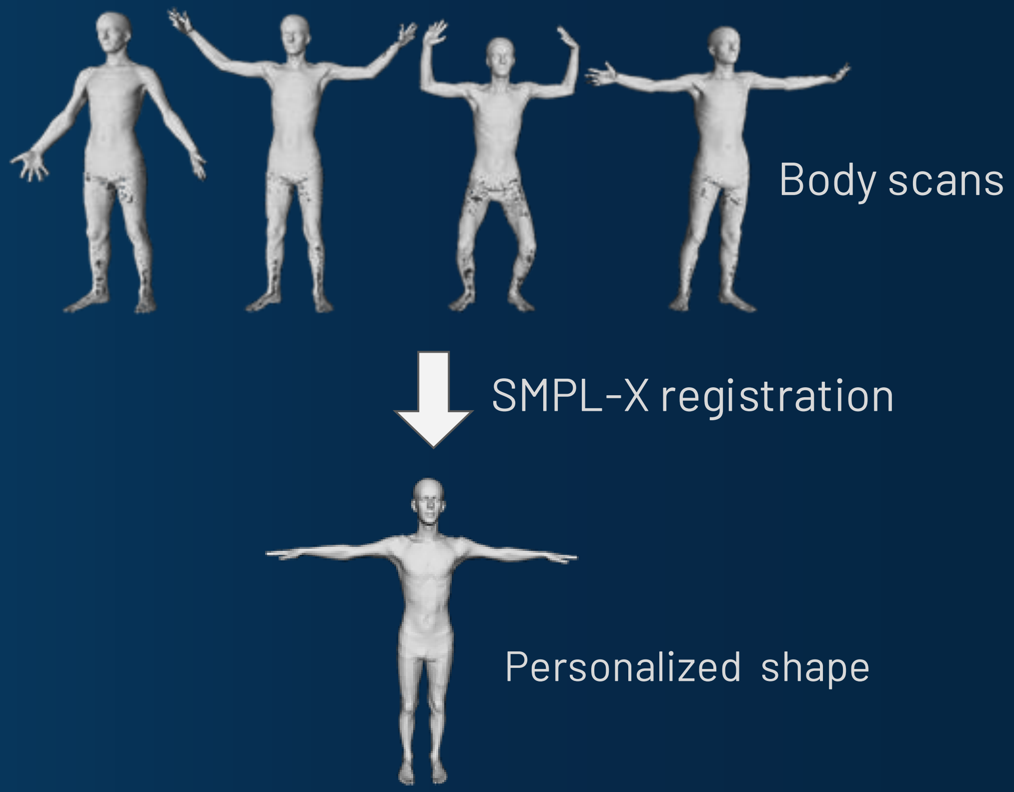
- Accurate **whole-body** motions
- Finger movement and facial expressions
- Accurate objects meshes and poses
- Different motion intents



Capturing Accurate Bodies

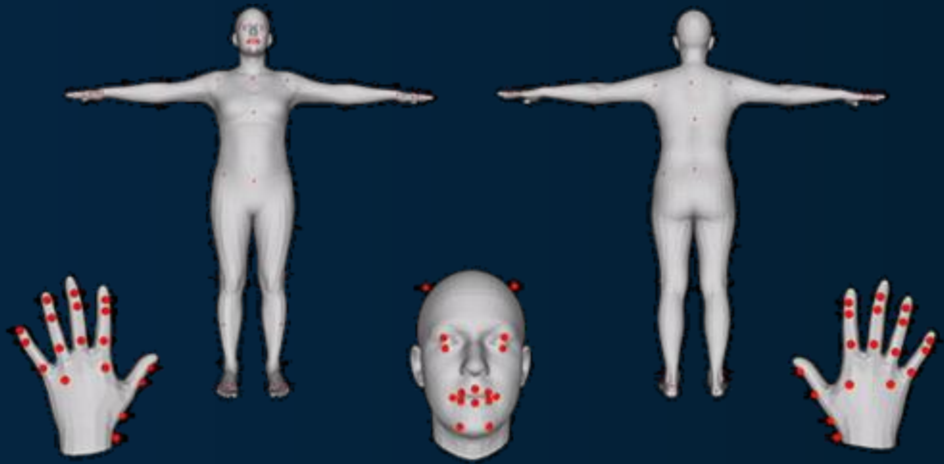
Body Shape

- Accurate body shape (high-res scanner)
- SMPL-X [1] expressive statistical body model



Body Motion

- Accurate MoCap system (Vicon)
- Rich minimally-intrusive marker set:
  - 99 markers on body
  - 14 markers on the face
  - 15 markers on each hand



[1] Pavlakos et al. CVPR 2019

Capturing Objects

Object Shape:

- 3D print objects of ContactDB [2]
- 51 objects

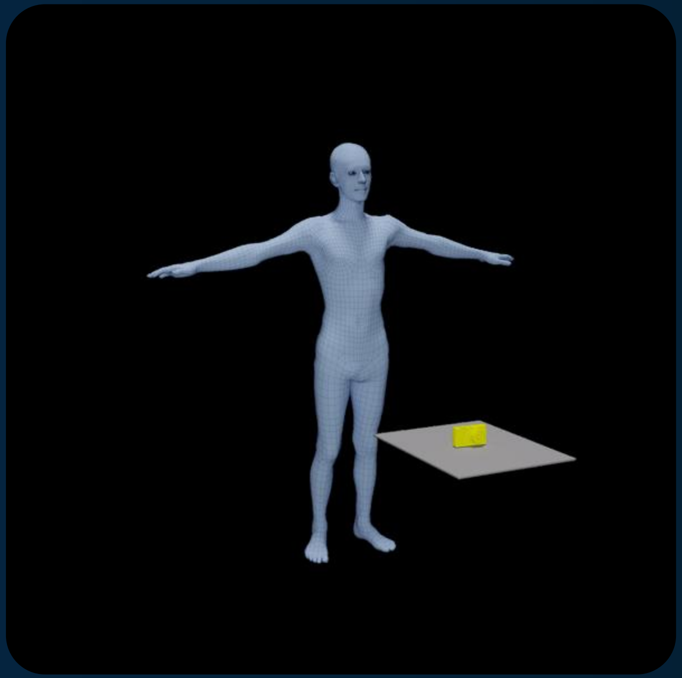
Object Motion:

- Glue 1.5mm-radius markers
- At least 8 markers per object
- Accurate MoCap system (Vicon)

[2] Brahmbhatt et al. CVPR 2019

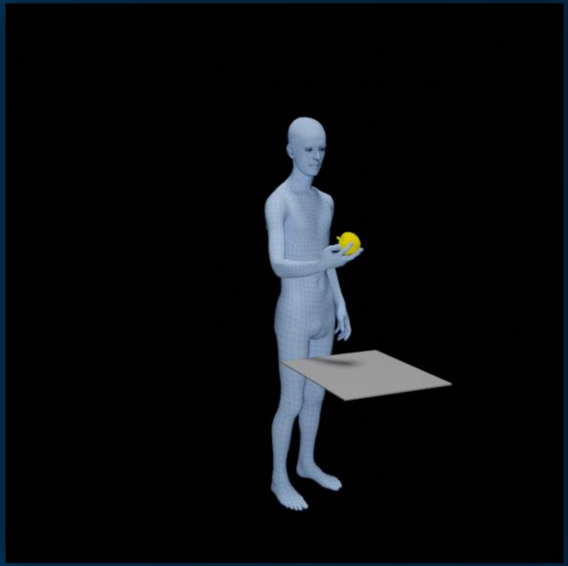
From Markers to 3D Surface

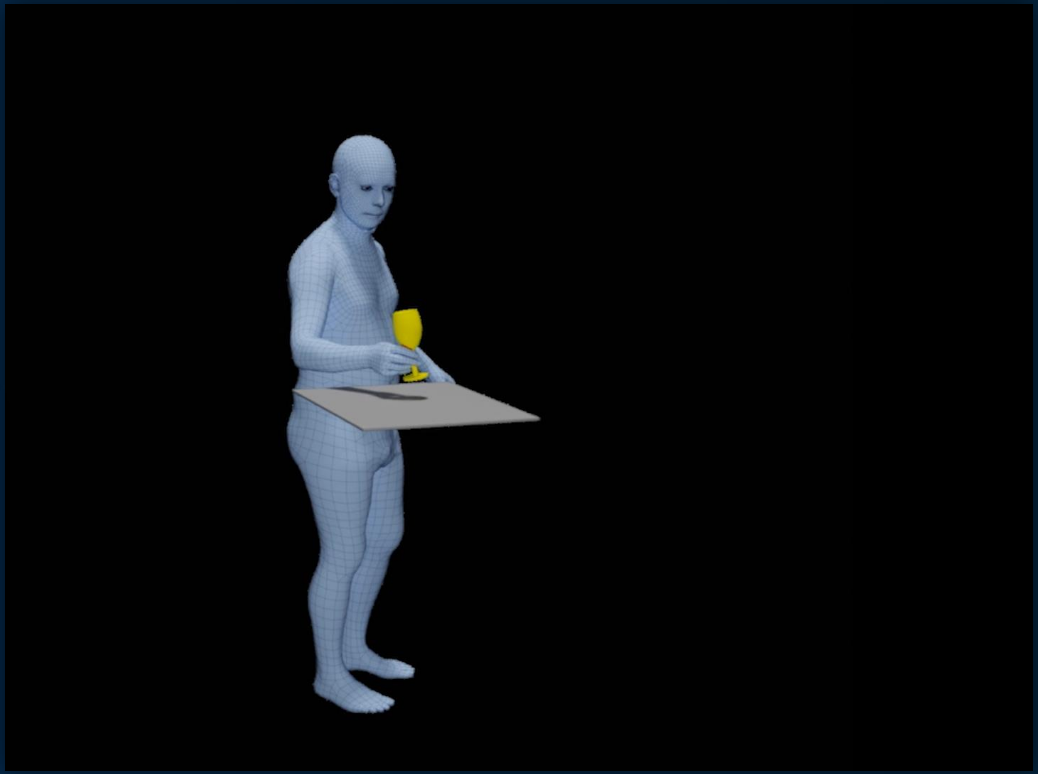
- Track body and object markers
- Label MoCap data (marker IDs)
- Adapt MoSh++ [3] to get body, face and hands surface
- Rigidly fit object meshes to object markers



[3] Mahmood et al. ICCV 2019

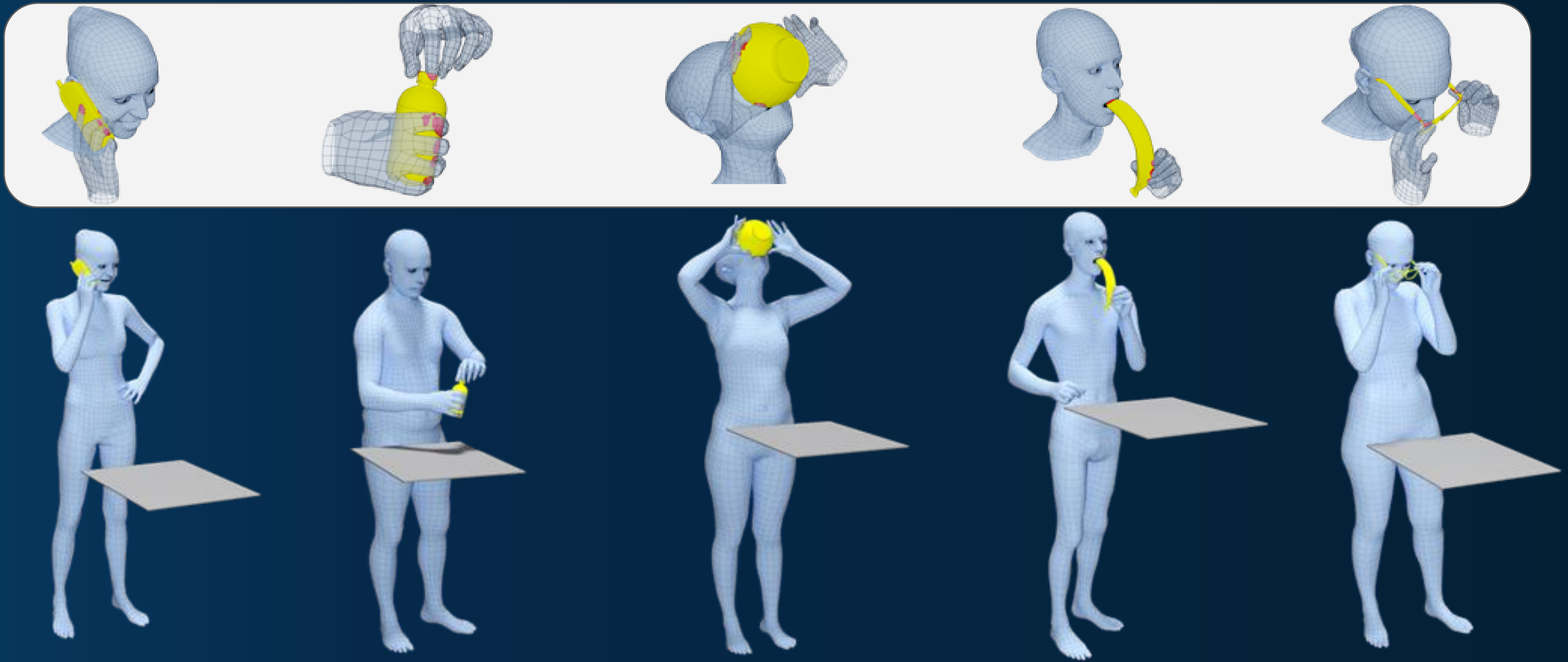
GRAB Motions





GRAB

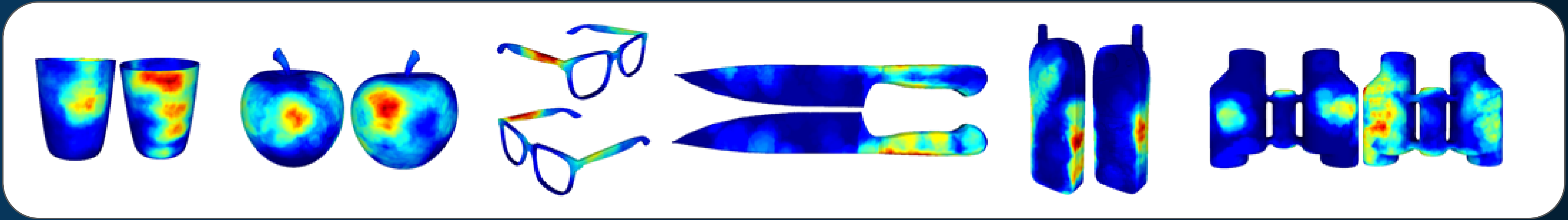
- "Whole-body" grasps
- Detailed body and object 3D meshes
- Accurate hand and face motions
- Accurate contact areas

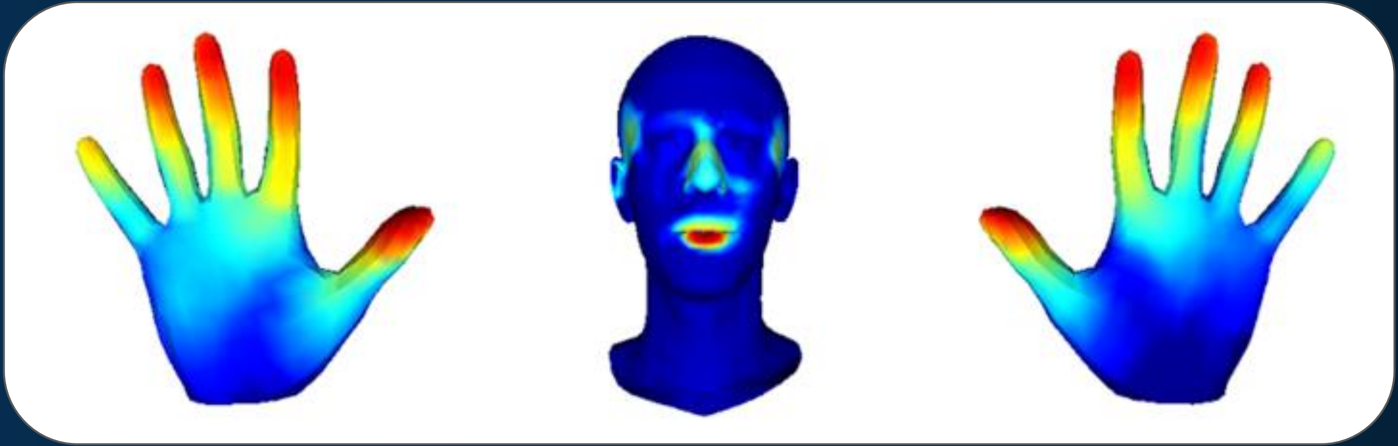


Contact Analysis

Integrate binary contacts over time - “heatmaps”

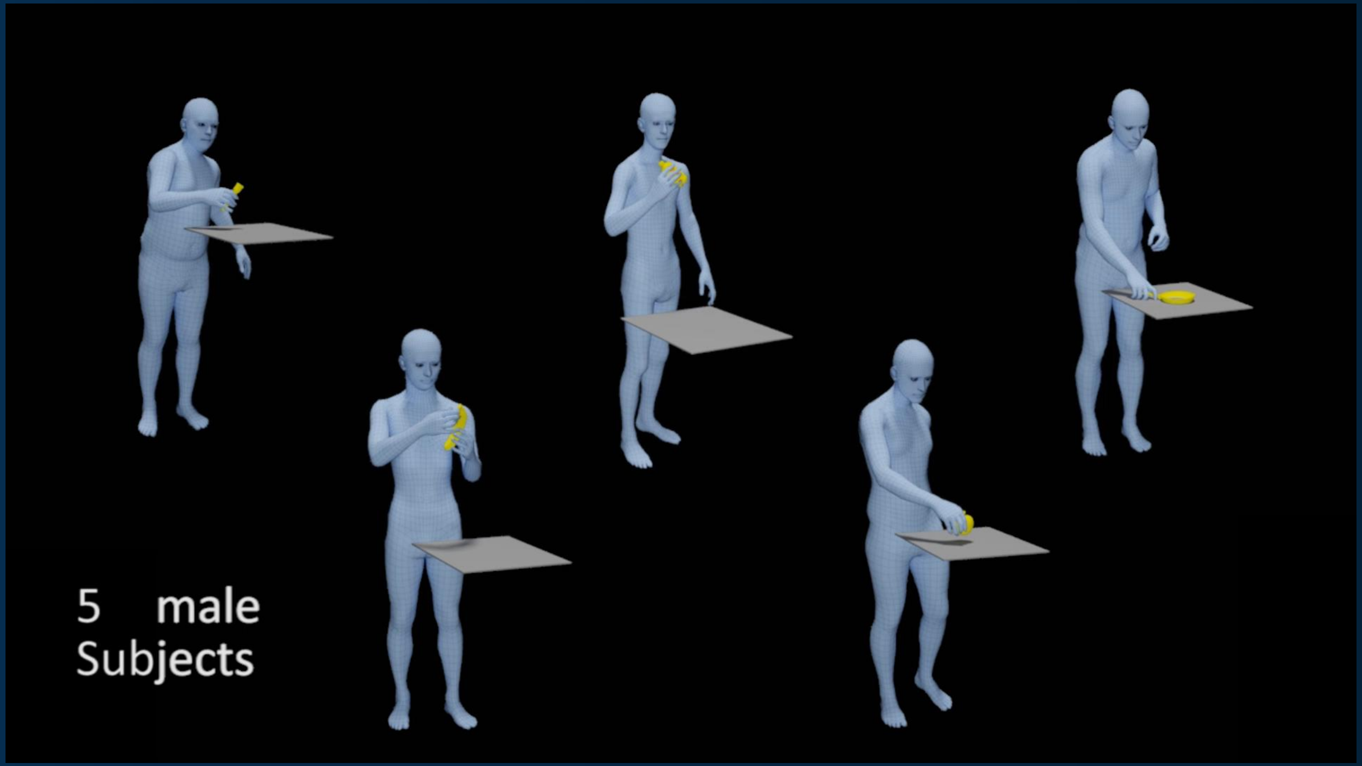
- Important areas for interaction (frequent and rare contact areas)
- Analysis of human grasps





Dataset Stats

- Captures 10 human subjects:
  - 5 males
  - 5 females
- Captures 4 different motion intents:
  - “Use”, “Pass”, “Lift”, “Off-Hand”
- Includes more than 1.6M frames in total
- Includes roughly 1M contact frames

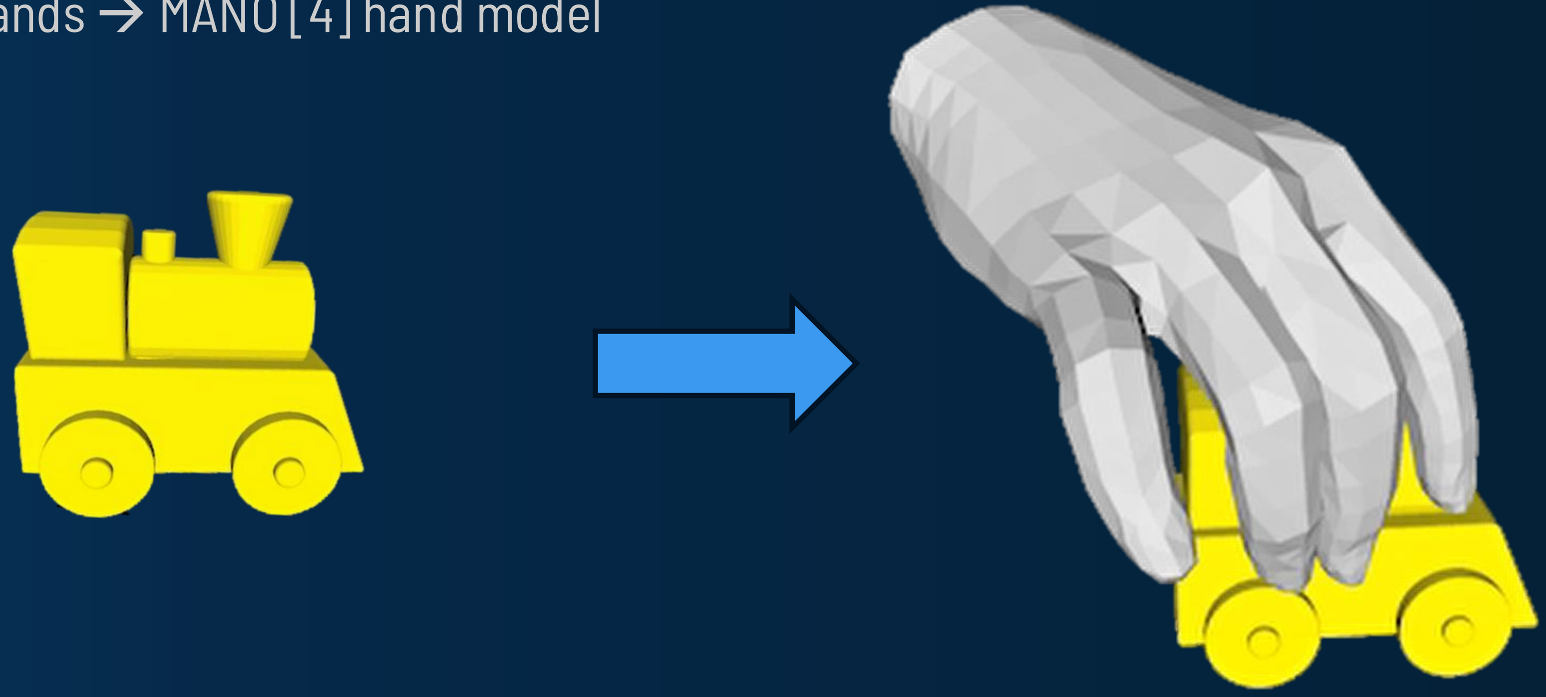


Generating Grasps

Goal:

Given a 3D unseen object as input → Can we generate various 3D hands grasping it?

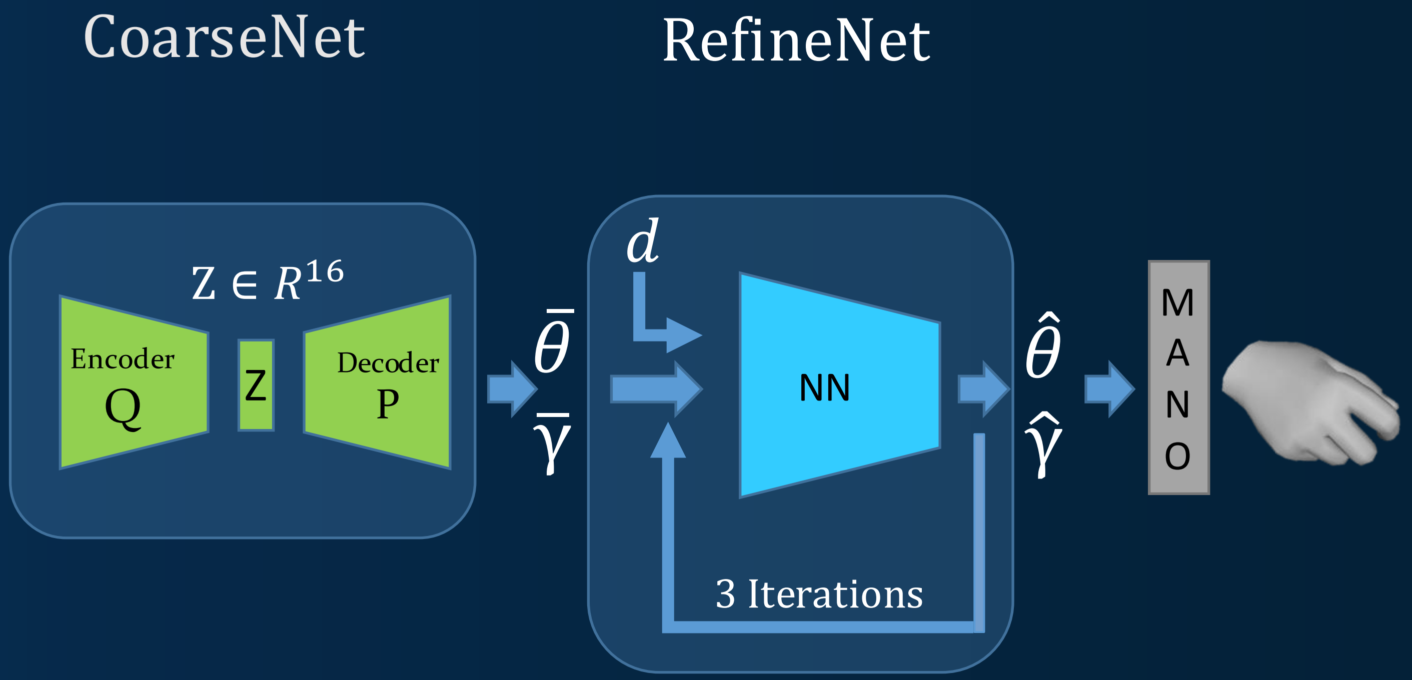
Representing hands → MANO [4] hand model



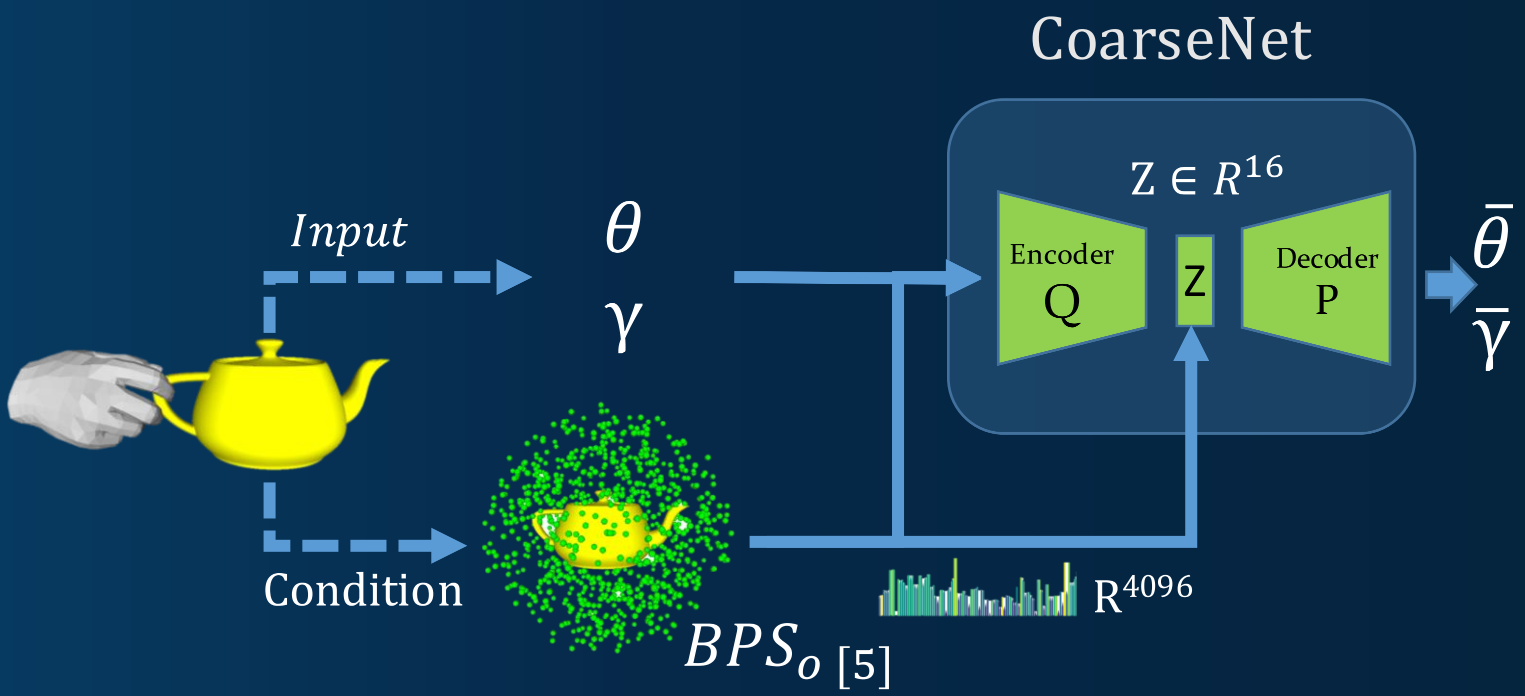
Input **Output**

[4] Romero et al, SIGASIA'2017

GrabNet: A Generative Model for 3D Hand Grasps



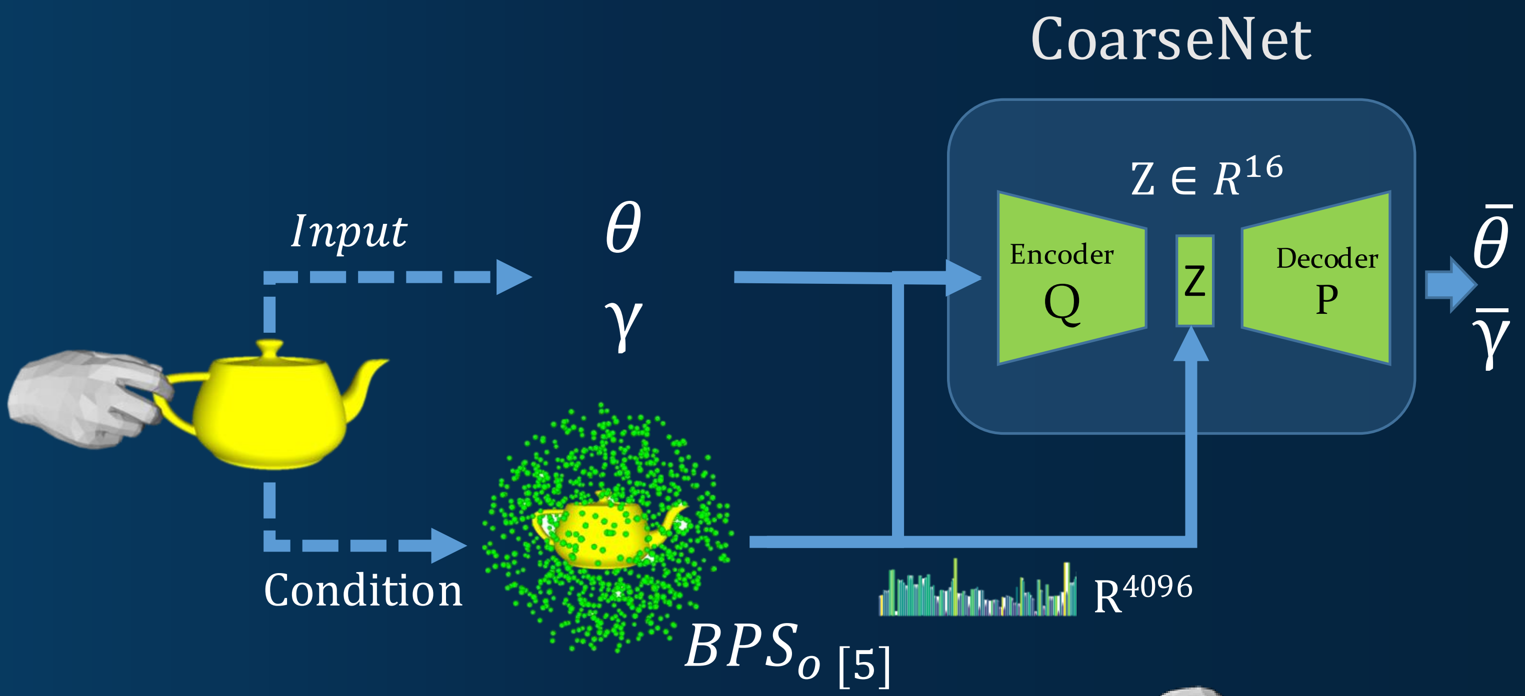
GrabNet: A Generative Model for 3D Hand Grasps



Intro
GRAB
GOAL
GRIP
Conc

[5] Prokudins et al, ICCV 2019

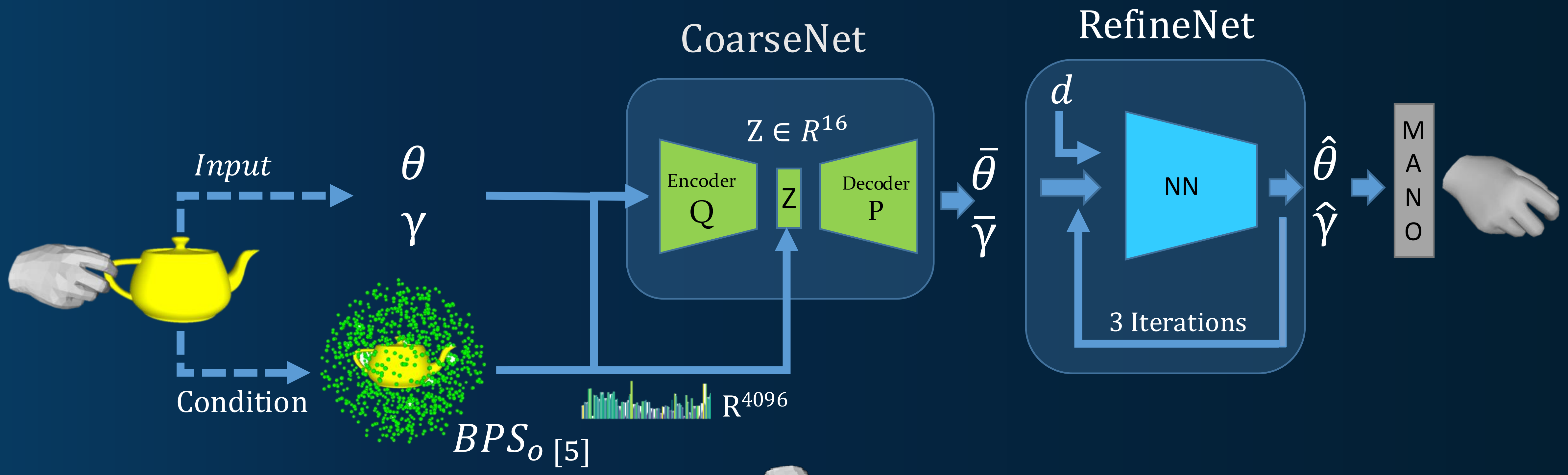
GrabNet: A Generative Model for 3D Hand Grasps



CoarseNet output

[5] Prokudins et al, ICCV 2019

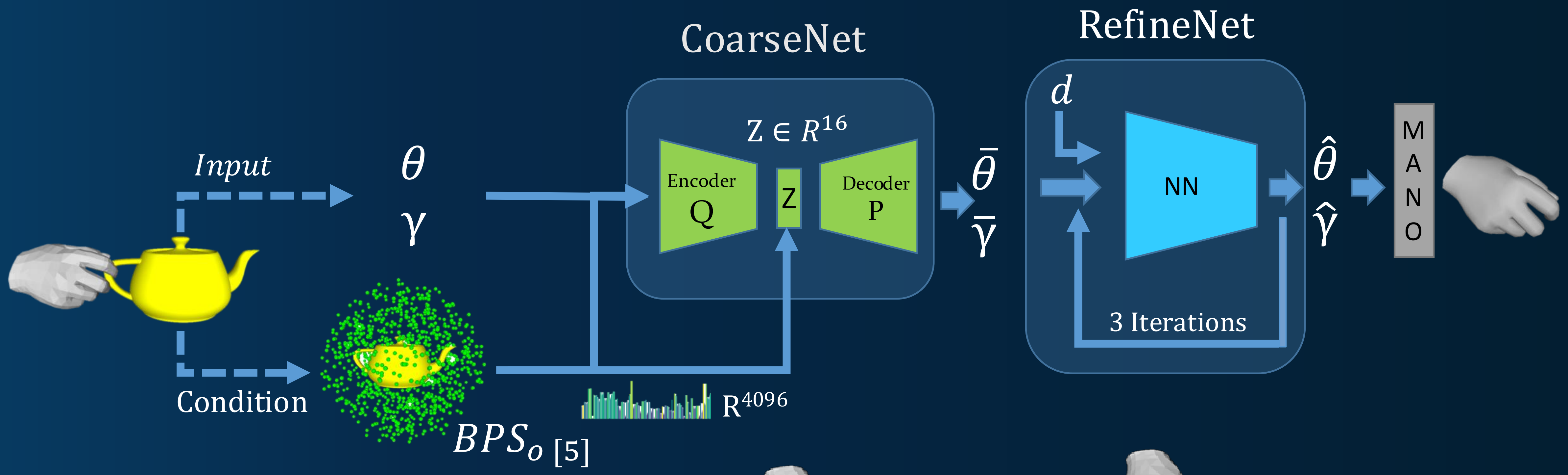
GrabNet: A Generative Model for 3D Hand Grasps



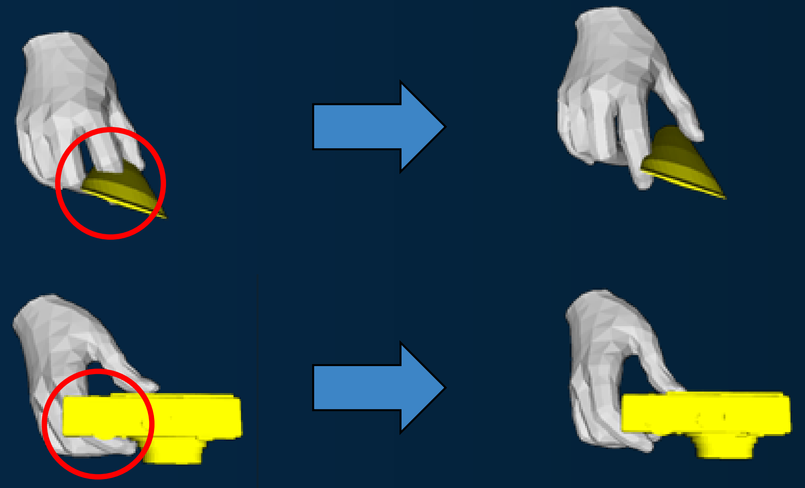
CoarseNet output

[5] Prokudins et al, ICCV 2019

GrabNet: A Generative Model for 3D Hand Grasps



CoarseNet output



RefineNet output

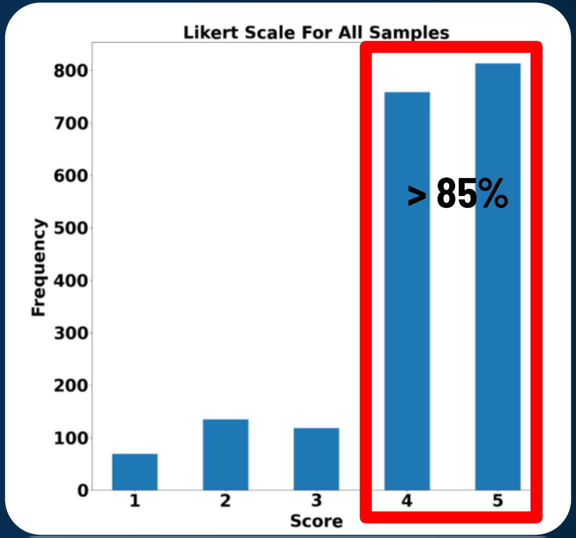
[5] Prokudins et al, ICCV 2019

Results - Unseen Objects

GrabNet Evaluation

Perceptual study:

- 5-level Likert score to:
    - "How realistic are the generated grasps?"
        - 1 → "Very Unrealistic"
        - 5 → "Very Realistic"



	AMT				Vertices	Contact
	Generation		Ground Truth		cm	%
Test Object	mean	std	mean	std	N=100	N=20
binoculars	4.09	0.93	4.27	0.80	2.56	4.00
camera	4.40	0.79	4.34	0.76	2.90	3.75
frying pan	3.19	1.30	4.49	0.67	3.58	4.16
mug	4.13	1.00	4.36	0.78	1.96	3.25
toothpaste	4.56	0.67	4.42	0.77	1.78	5.39
wineglass	4.32	0.88	4.43	0.79	1.92	4.56
Average	4.12	1.04	4.38	0.77	2.45	4.18

Intro
GRAB
GOAL
GRIP
Conc

GOAL: Generating 4D Whole-Body Motion for Hand-Object Grasping

Omid Taheri, Vasileios Choutas, Michael Black, Dimitrios Tzionas

CVPR 2022

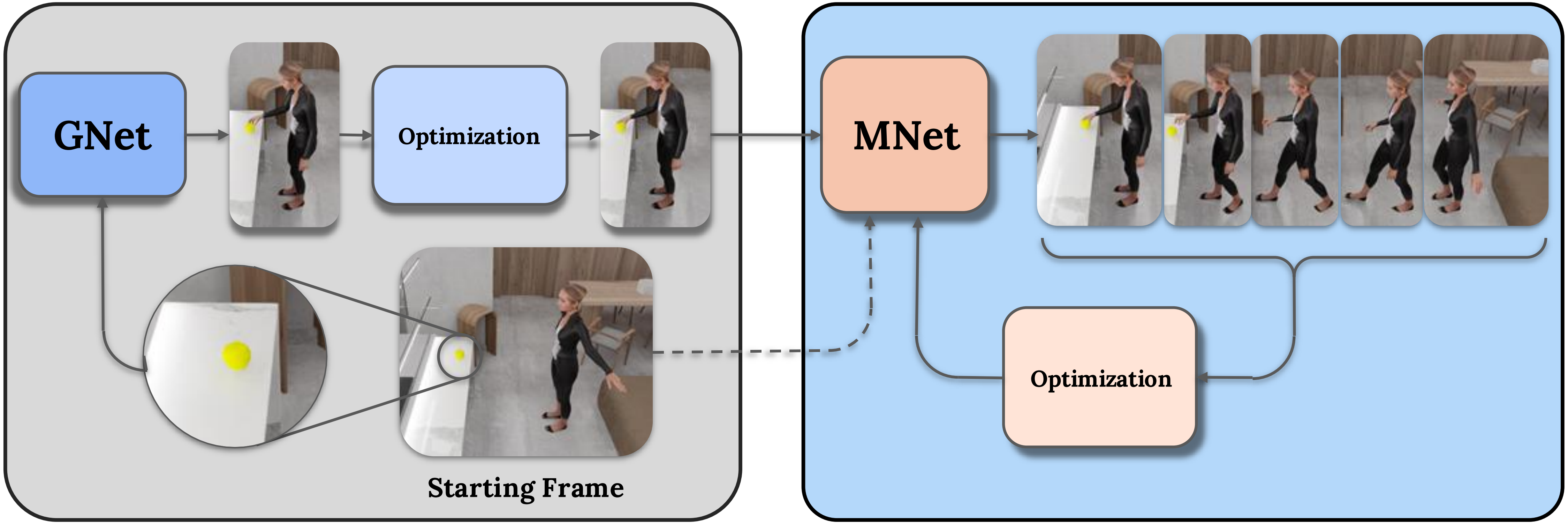
Objective

Common step for most interactions → Walking up to and grasping the object.

Generate full-body motions that:

- Grasp unseen 3D objects
- Have realistic hand grasps
- Realistic foot-ground contact
- Natural head orientation for grasping

GOAL Setup

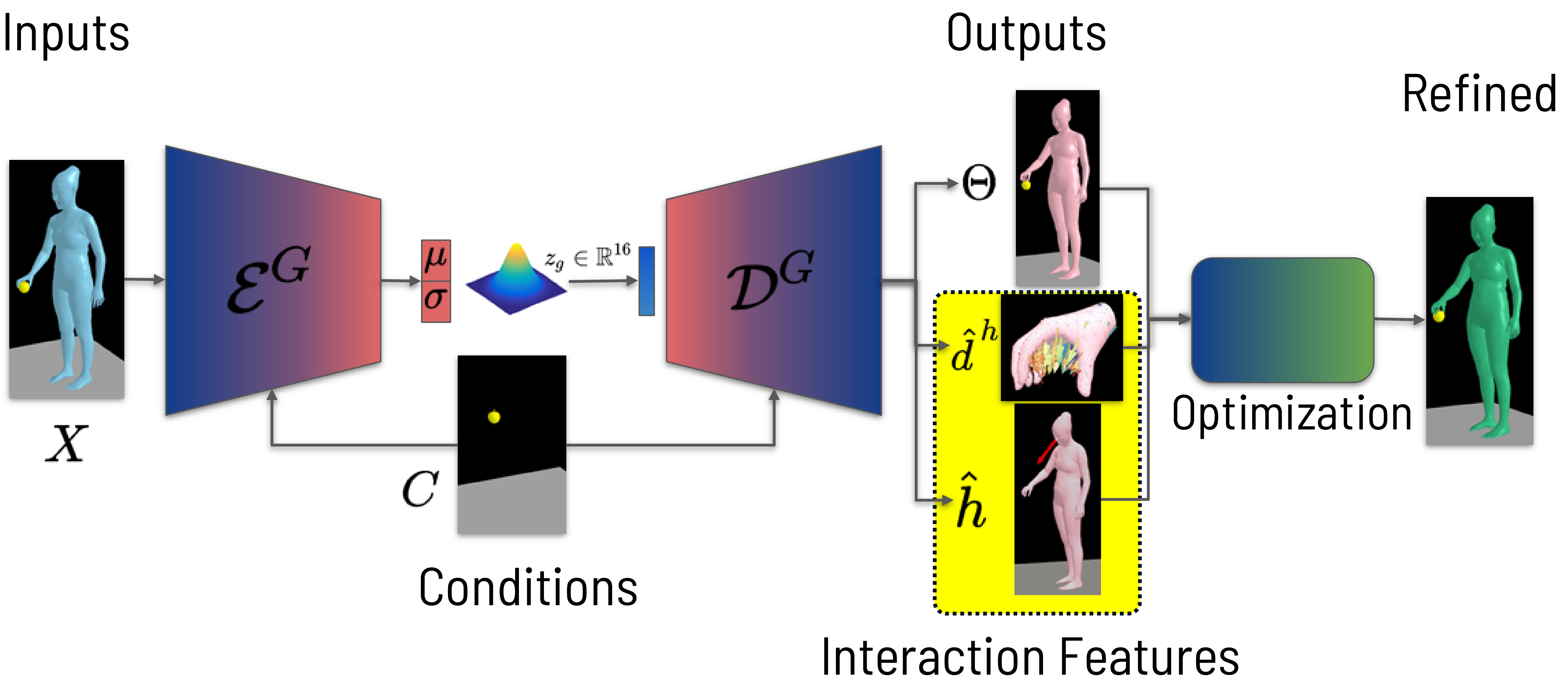


1. Full-body Grasp

2. Body Motion

GNet Architecture

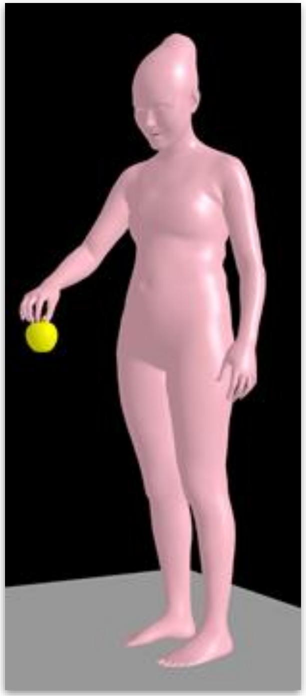
To generate the end frame's full-body grasp.



GNet – Key Idea

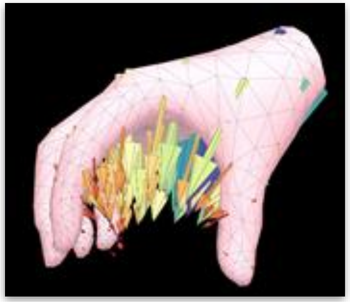
SMPL-X Parameters

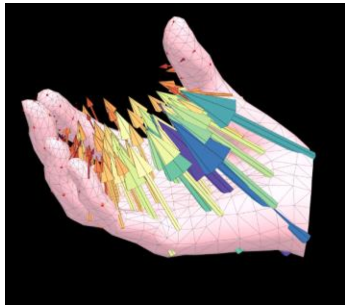
Θ



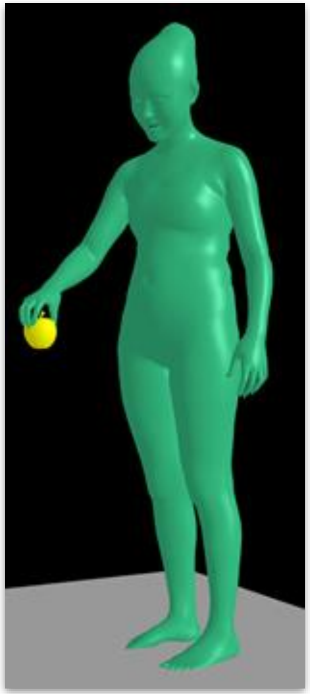
Interaction Features

$\hat{d}^h$ $\hat{h}$



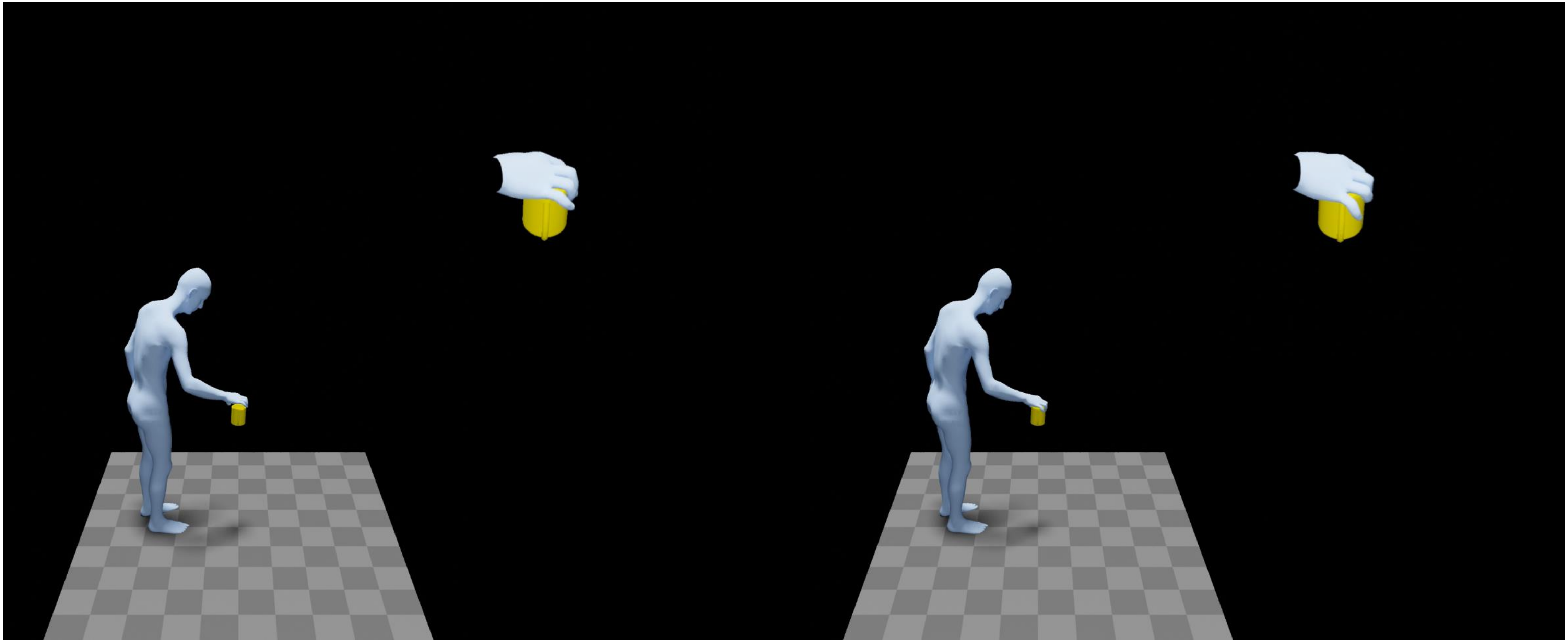


Optimized Grasp



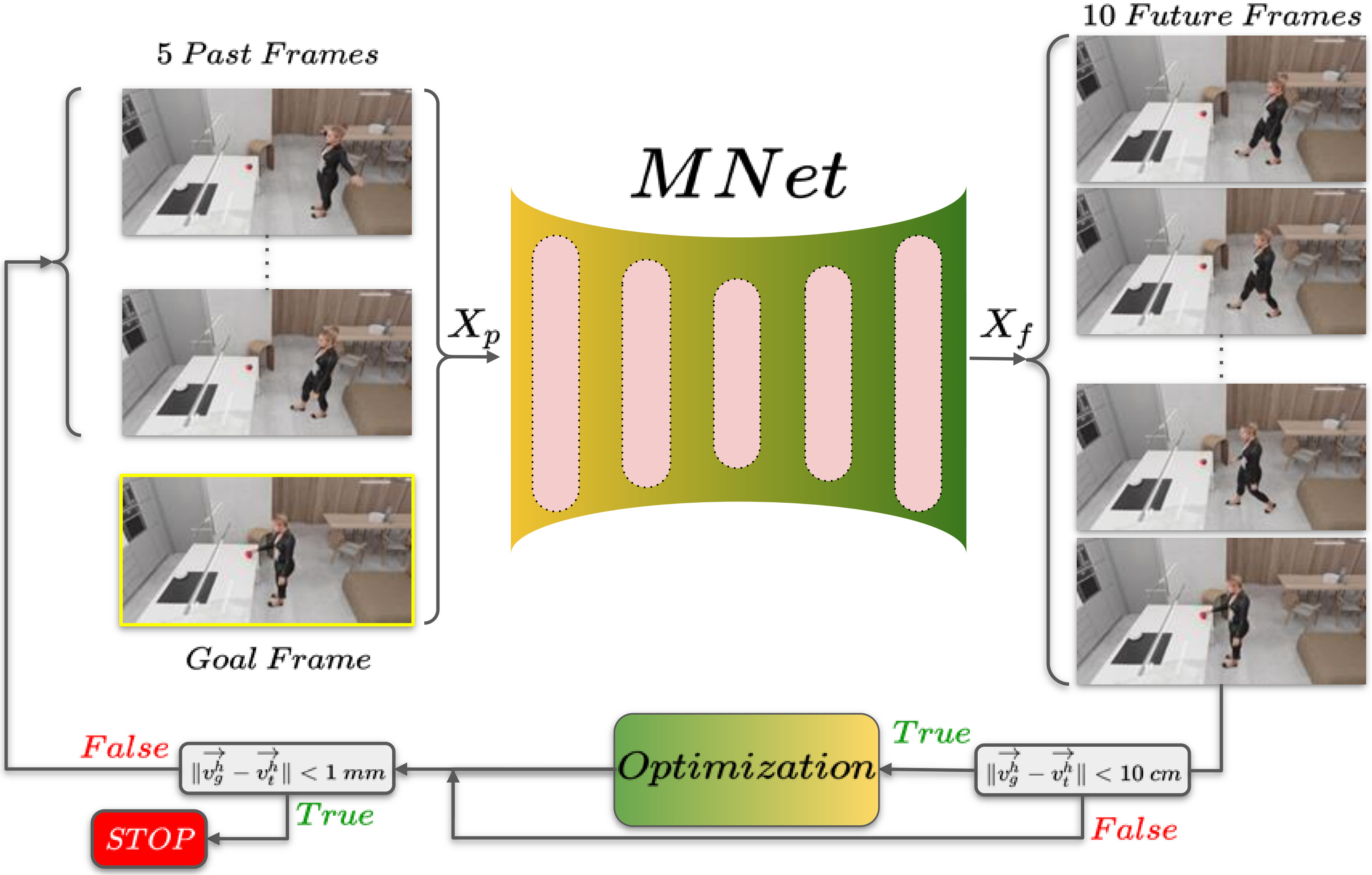
GNet – Optimization Results

Before Optimization **After Optimization**



MNet Architecture

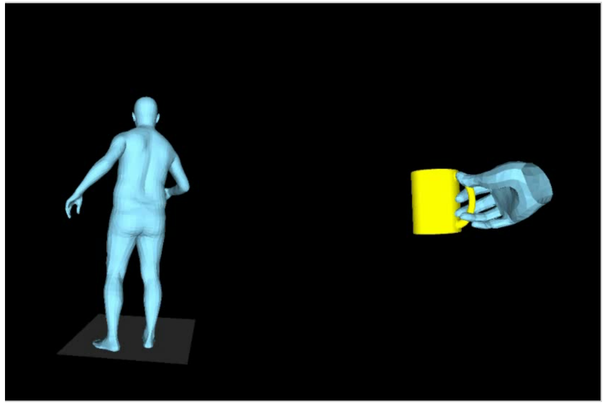
Intro
GRAB
GOAL
GRIP
Conc

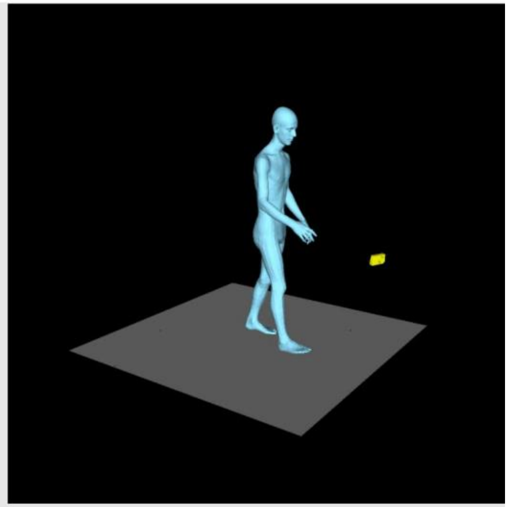


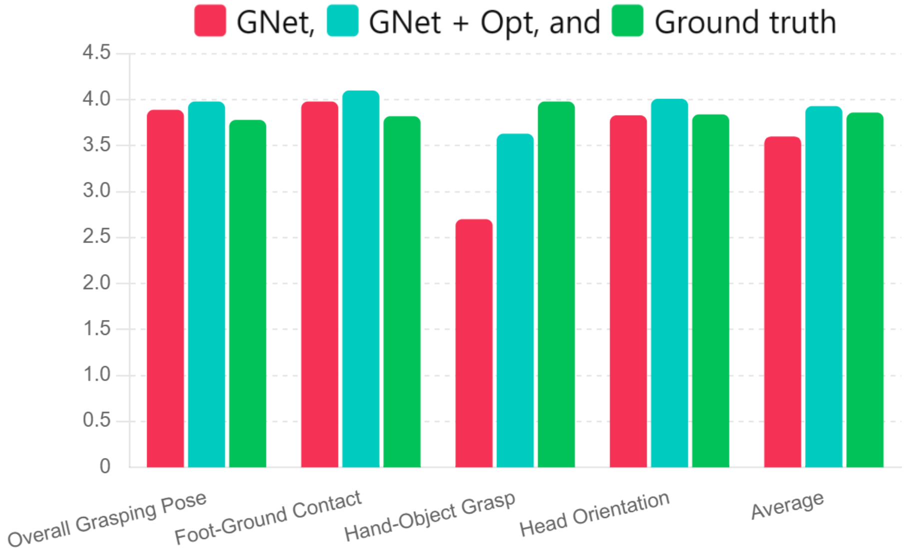
MNet - Results

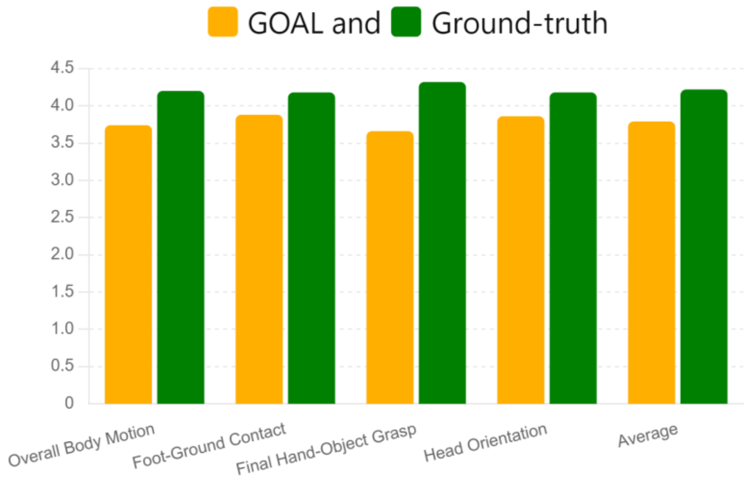
Intro
GRAB
GOAL
GRIP
Conc

Evaluations





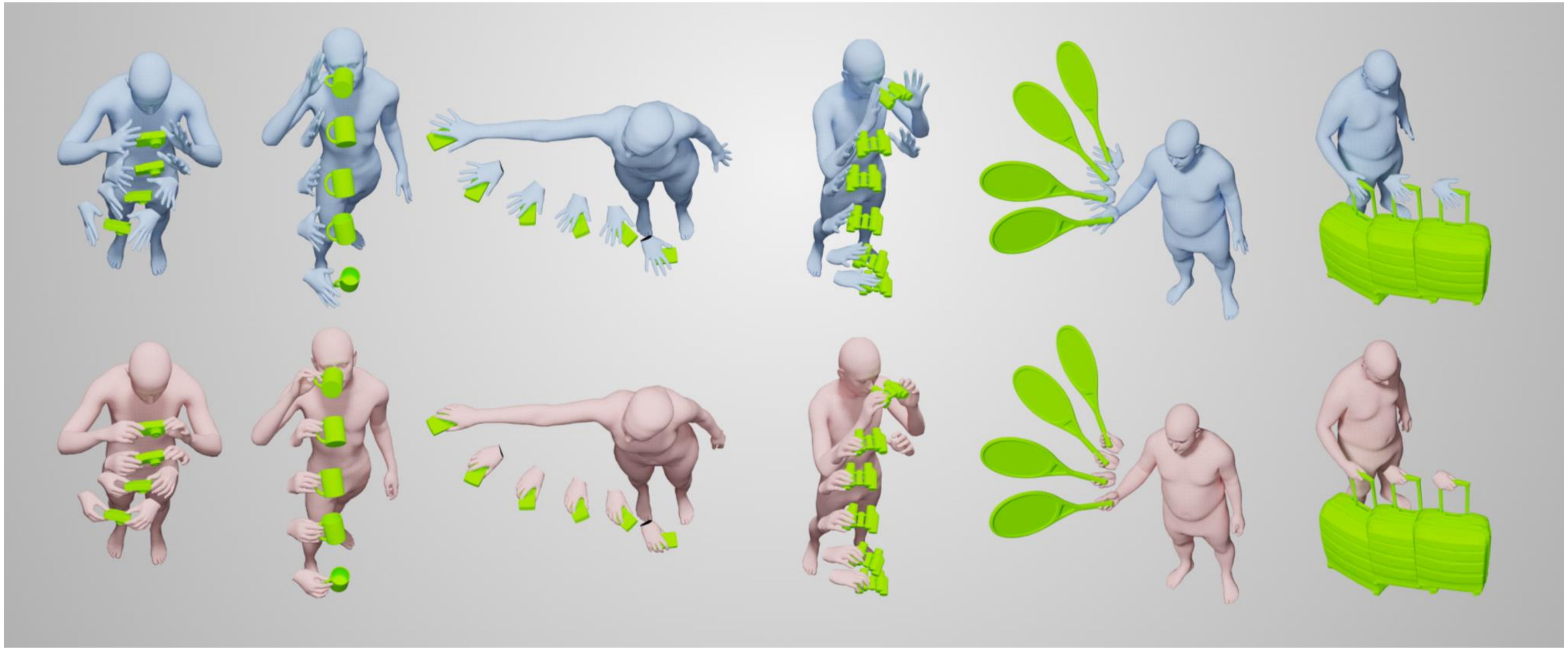




GRIP: Generating Interaction Poses Using Spatial Cues and Latent Consistency

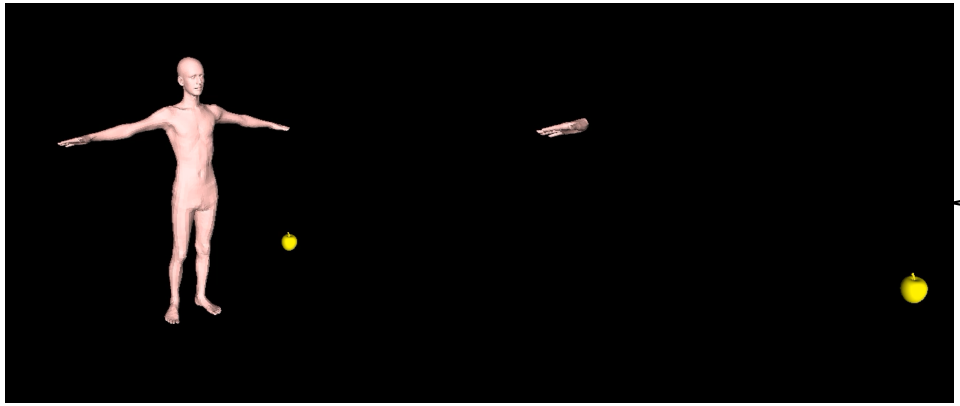
Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Soren Pirk, Michael J. Black

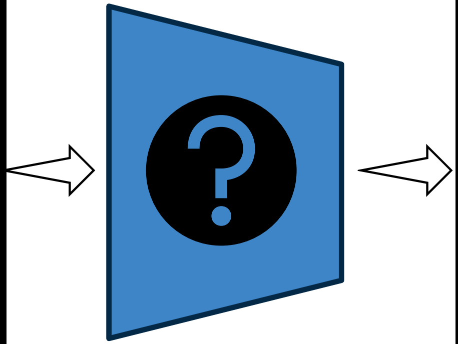
3DV 2024



Goal

Given a sequence of **body** and **object motion** → Accurately generate ***interacting-hand poses***





Input **Output**

Why?

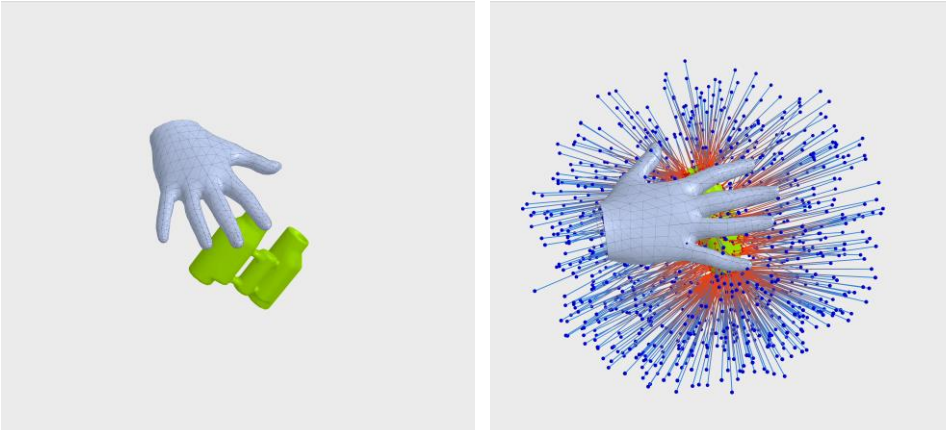
The combination of body and hands has been overlooked in datasets & motion modelling:

- Add hands to new or previous datasets
- Refine the hands generated/reconstructed using other models

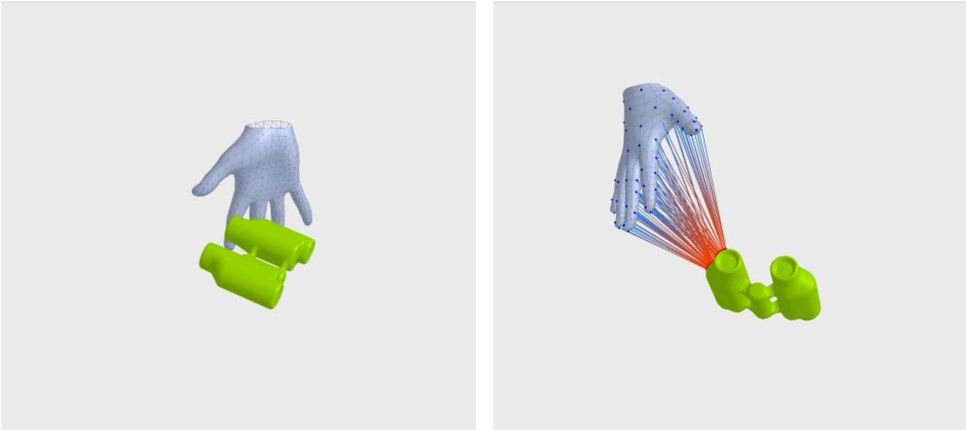
Spatio-Temporal Features-via Hand Sensors

Extract rich features:

- Based on the relative body and object motion
- Bidirectional: body ↔ object
- Generalizable

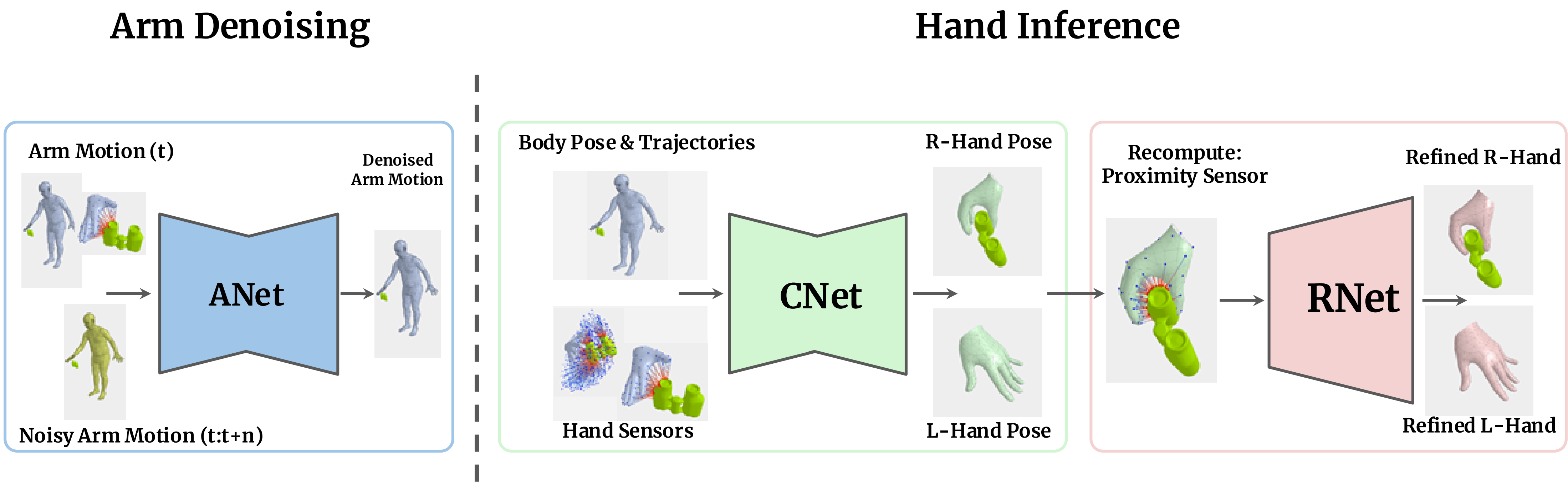


Ambient Sensor

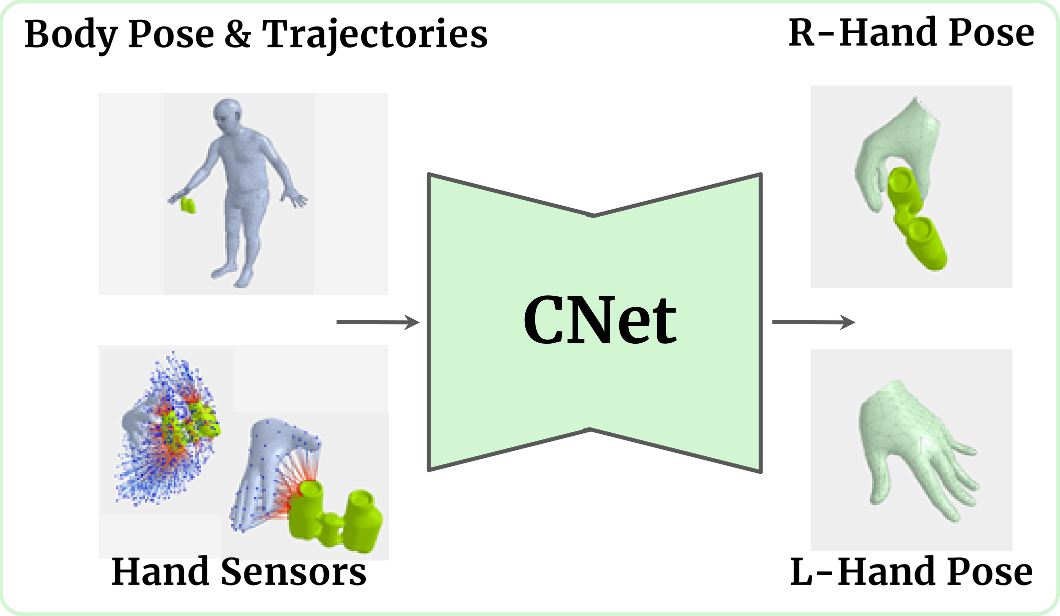


Proximity Sensor

Method - Architecture



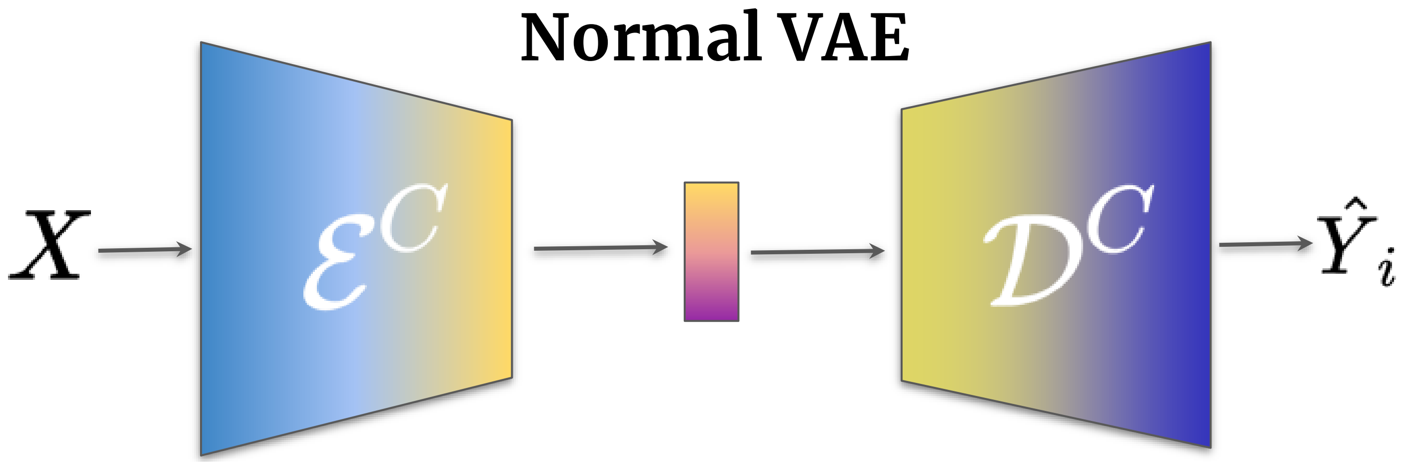
Method - Architecture



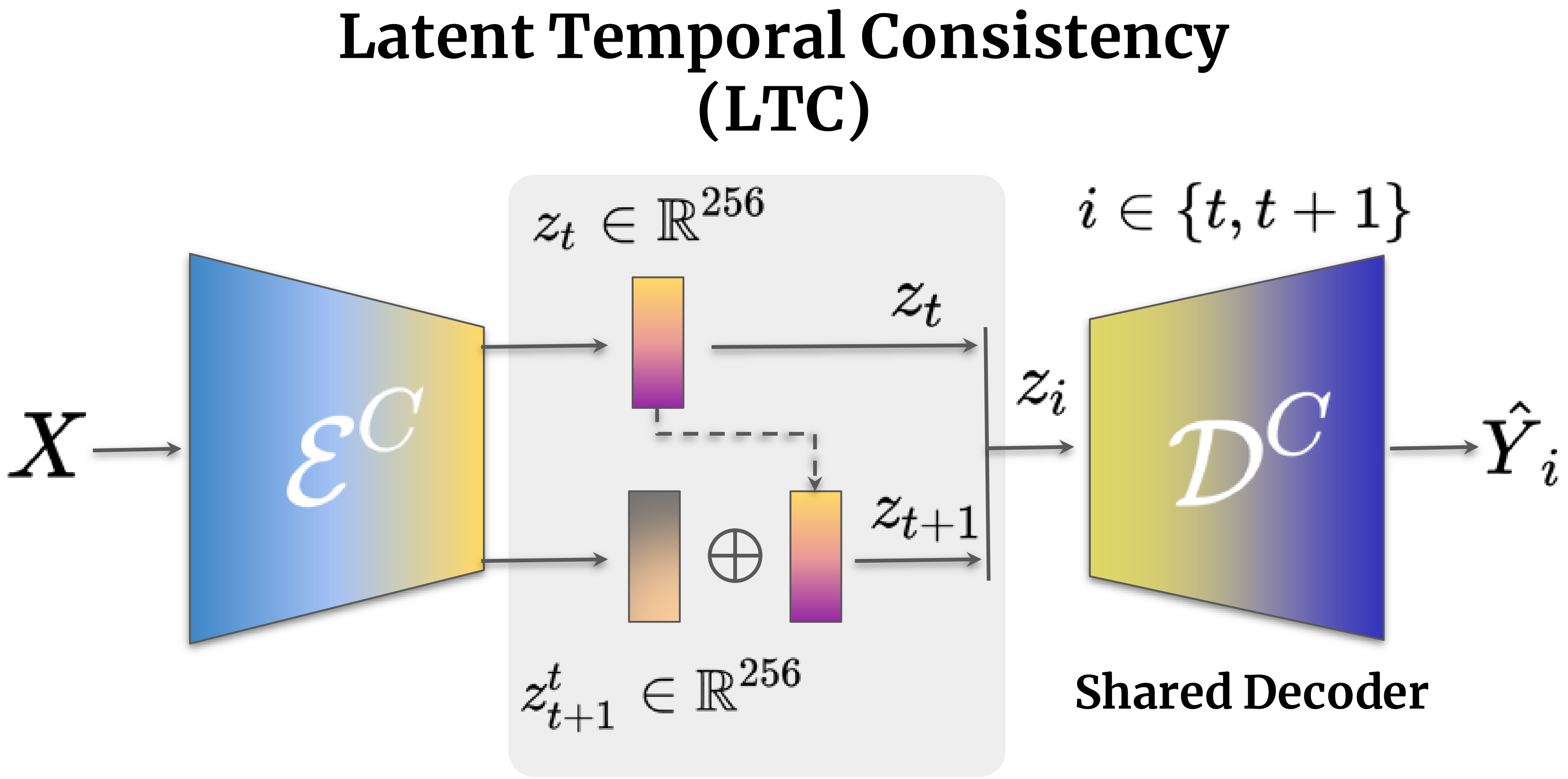
Method – CNet

Goal:

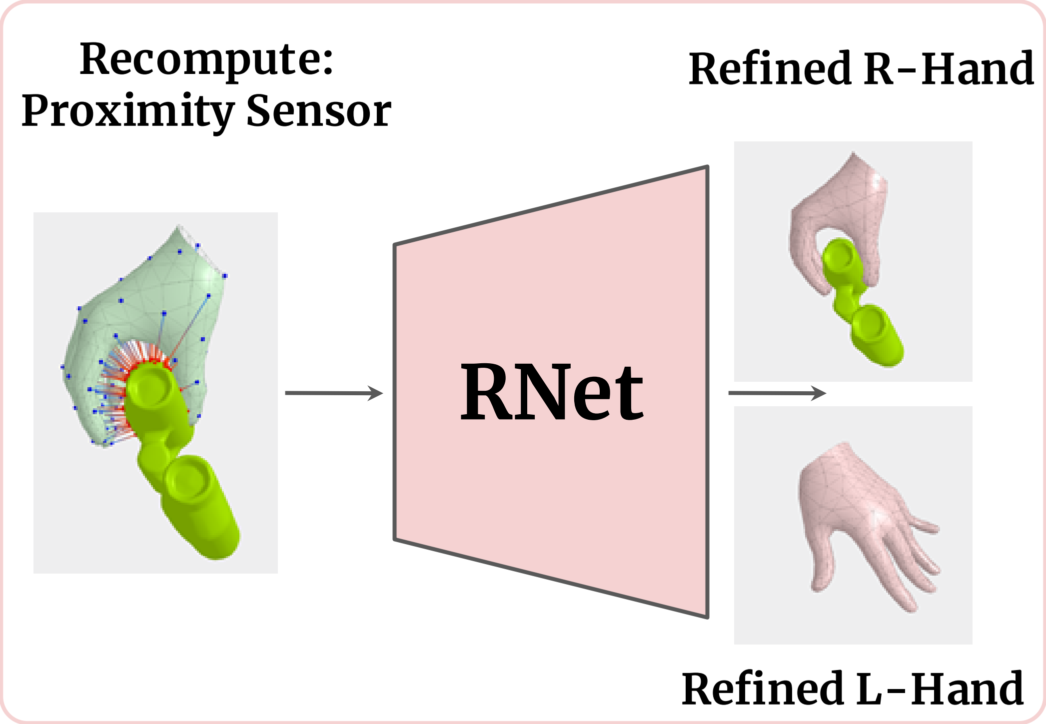
- Generate hand motions in real-time: frame by frame
- Smooth and consistent motion



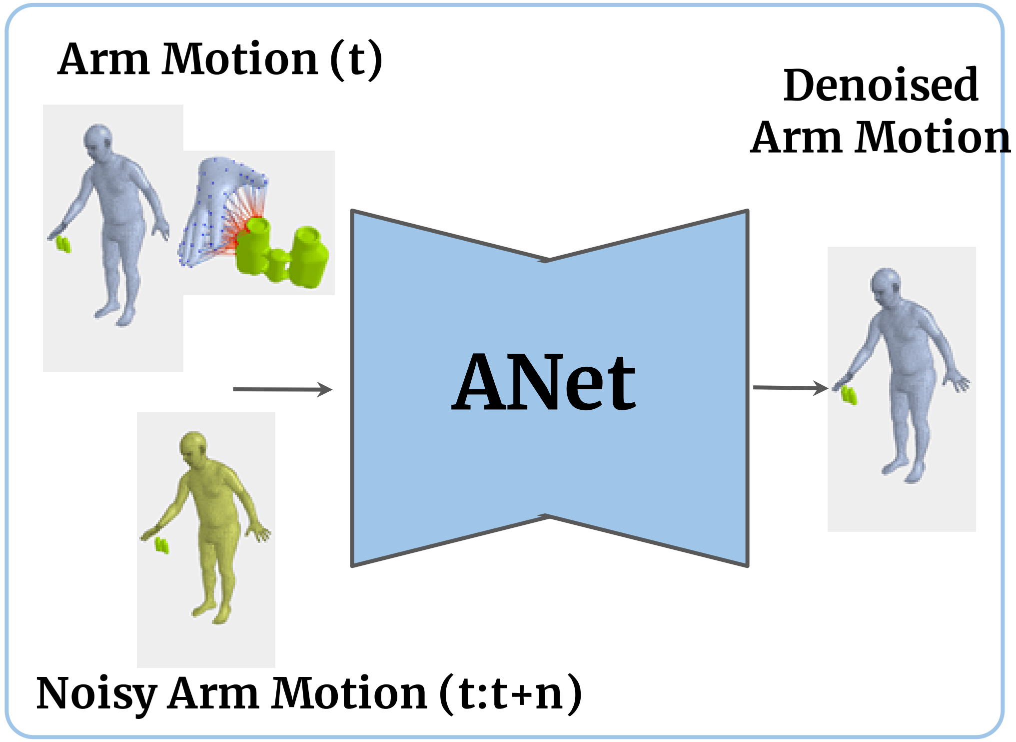
Method – CNet


Latent Temporal Consistency (LTC)
X
$\mathcal{E}^C$
$z_t \in \mathbb{R}^{256}$
z_t
$z^t_{t+1} \in \mathbb{R}^{256}$
$\oplus$
z_{t+1}
z_i
$i \in \{t, t+1\}$
$\mathcal{D}^C$
$\hat{Y}_i$
Shared Decoder

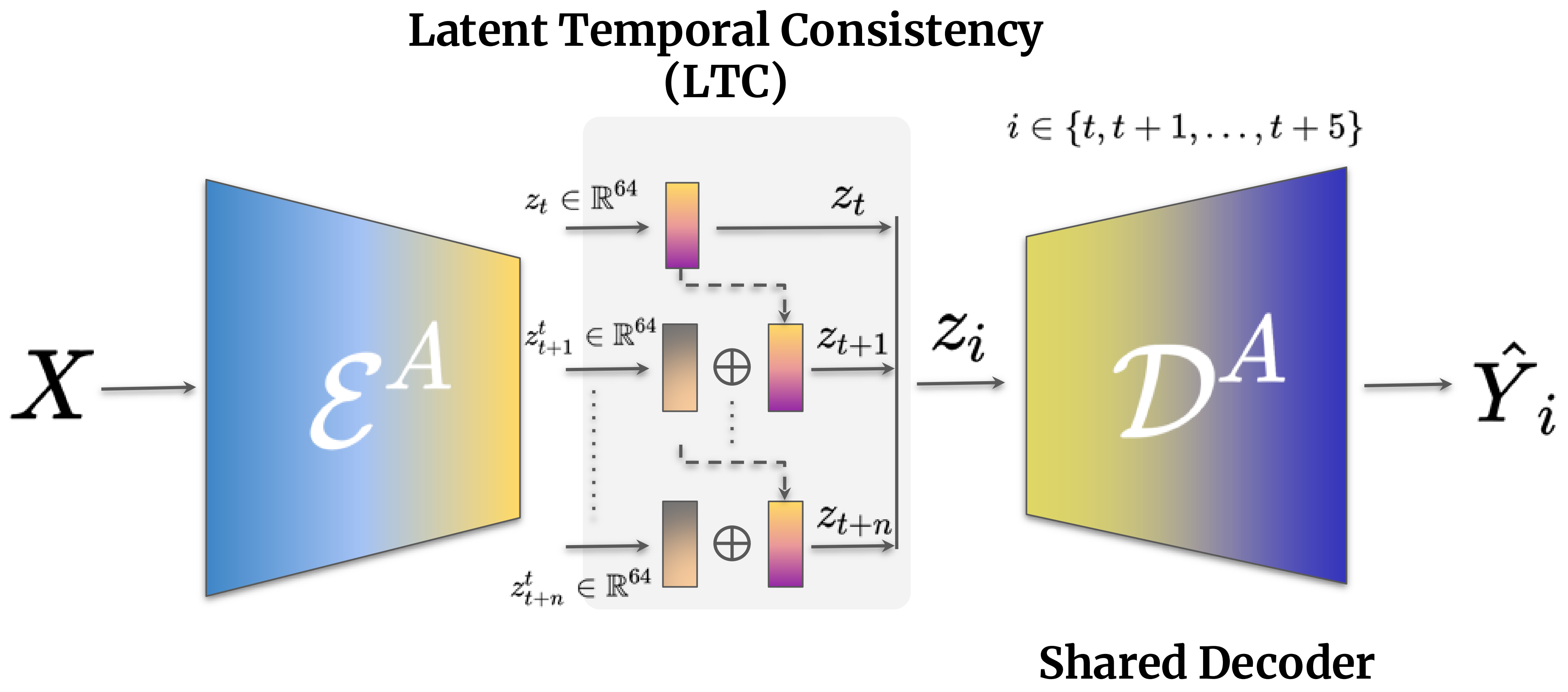
Method – RNet


Recompute:
Proximity Sensor
Refined R-Hand
RNet
Refined L-Hand

Method – ANet



Method – ANet



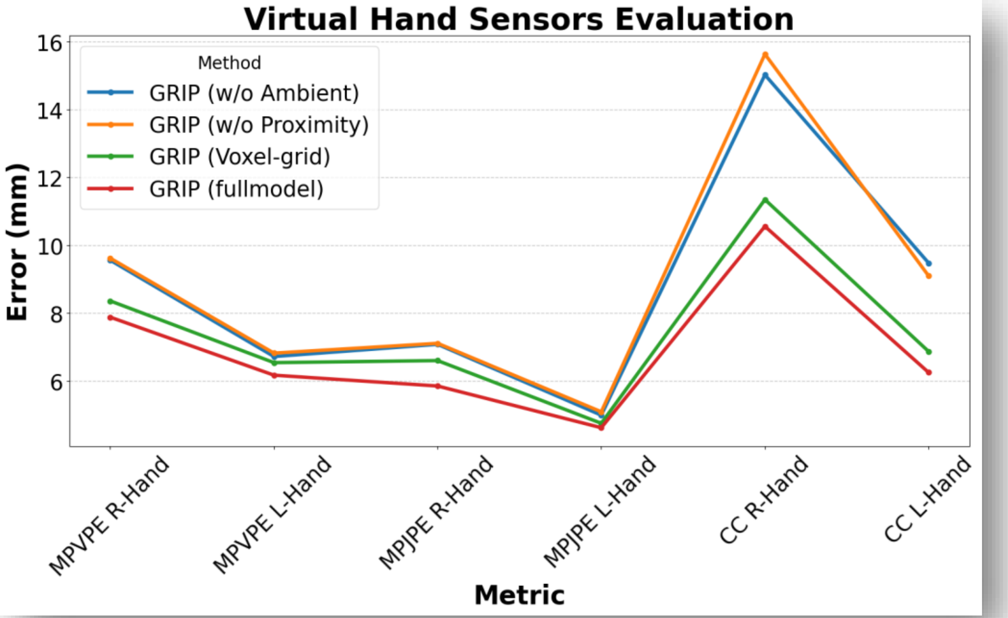
Intro
GRAB
GOAL
GRIP
Conc

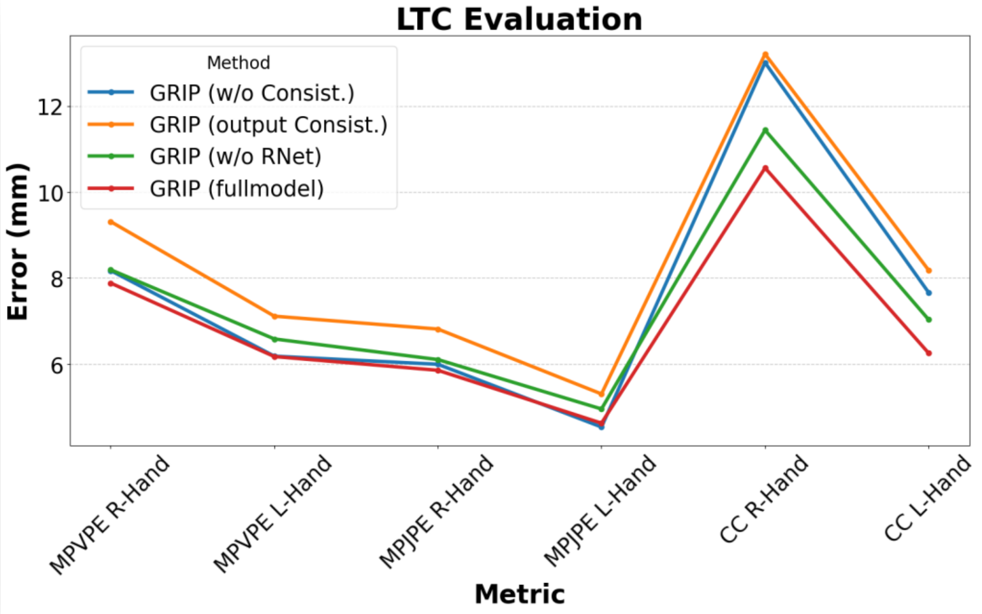
Evaluations

Method ↓	MPVPE (mm) ↓		MPJPE (mm) ↓		CC (mm) ↓	
	R-Hand	L-Hand	R-Hand	L-Hand	R-Hand	L-Hand
Hand Sensors Ablation						
GRIP (w/o Ambient)	9.56	6.72	7.08	4.99	15.03	9.48
GRIP (w/o Proximity)	9.62	6.82	7.11	5.09	15.64	9.10
Latent Temporal Consistency (LTC) Evaluation						
GRIP (w/o Consist.)	8.17	6.18	5.99	**4.53**	13.01	7.66
GRIP (output Consist.)	9.31	7.11	6.81	5.31	13.21	8.18
GRIP (Voxel-grid)	8.36	6.54	6.60	4.75	11.35	6.87
GRIP (w/o RNet)	8.19	6.58	6.10	4.95	11.44	7.03
GRIP (fullmodel)	**7.88**	**6.17**	**5.85**	4.62	**10.56**	**6.25**

Evaluations

Method ↓	MPVPE (mm) ↓		MPJPE (mm) ↓		CC (mm) ↓	
	R-Hand	L-Hand	R-Hand	L-Hand	R-Hand	L-Hand
Hand Sensors Ablation						
GRIP (w/o Ambient)	9.56	6.72	7.08	4.99	15.03	9.48
GRIP (w/o Proximity)	9.62	6.82	7.11	5.09	15.64	9.10
Latent Temporal Consistency (LTC) Evaluation						
GRIP (w/o Consist.)	8.17	6.18	5.99	**4.53**	13.01	7.66
GRIP (output Consist.)	9.31	7.11	6.81	5.31	13.21	8.18
GRIP (Voxel-grid)	8.36	6.54	6.60	4.75	11.35	6.87
GRIP (w/o RNet)	8.19	6.58	6.10	4.95	11.44	7.03
GRIP (fullmodel)	**7.88**	**6.17**	**5.85**	4.62	**10.56**	**6.25**



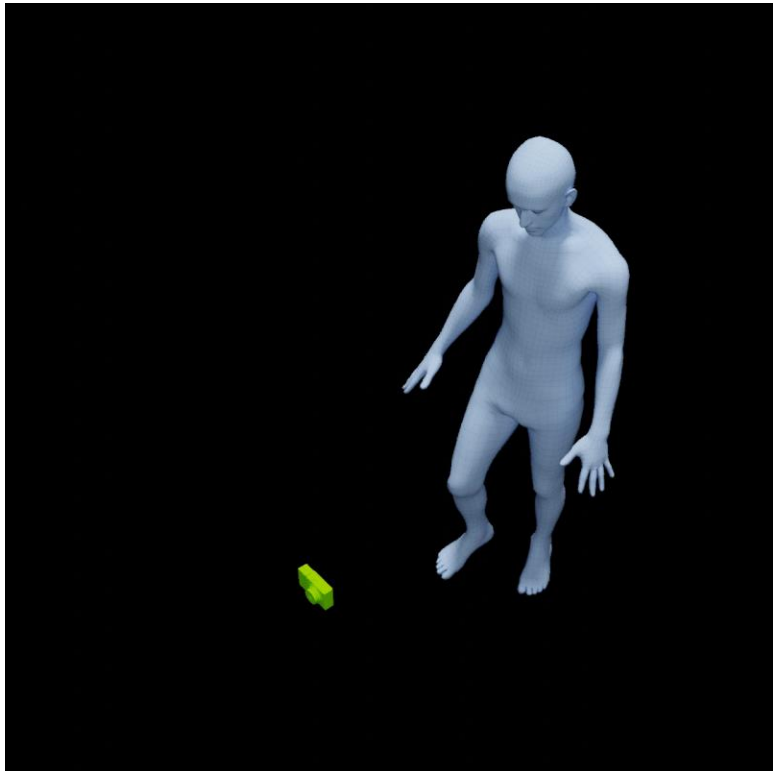


Results

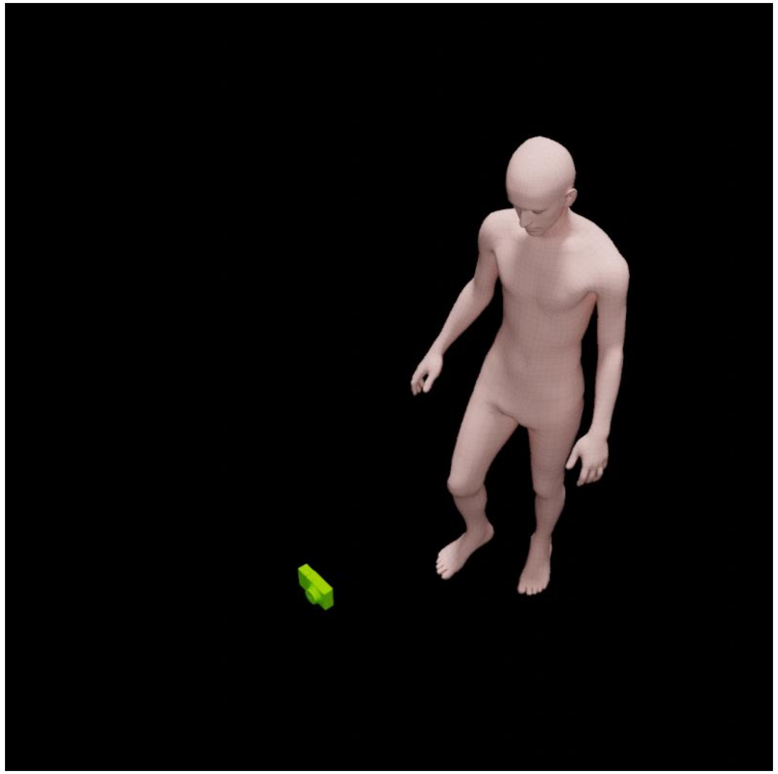
Input

Output

Results



Input



Output

Takeaways

- Data is not just for you, it's for the community:
  - Consider what people would need in 5-10 years
- Big data matters BUT right features matter more:
  - Key to generalization
- Accuracy in interactions is crucial - refinement:
  - Feedback Loop
  - Optimization
  - Diffusion Models
- Interactions need different 3D representation:
  - Different from general 3D object representations
  - Focused on spatial information between body & objects

Limitations & Future Work

GRAB

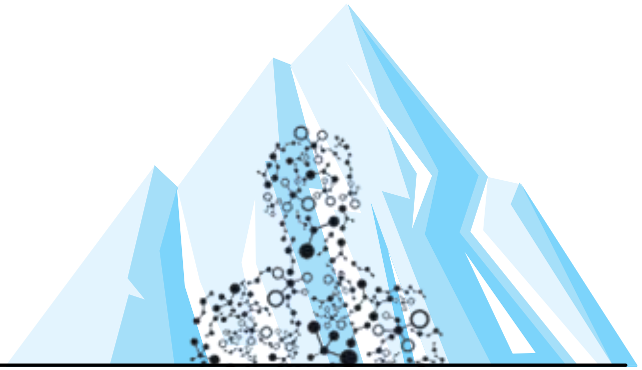
A Dataset of **Whole-Body** Human Grasping of Objects

GOAL

Generating **4D Whole-Body Motion** for Hand-Object Grasping

GRIP

Generating **Interaction Poses** Using Spatial Cues and Latent Consistency



Limitations & Future Work

GRAB

A Dataset of **Whole-Body** Human Grasping of Objects

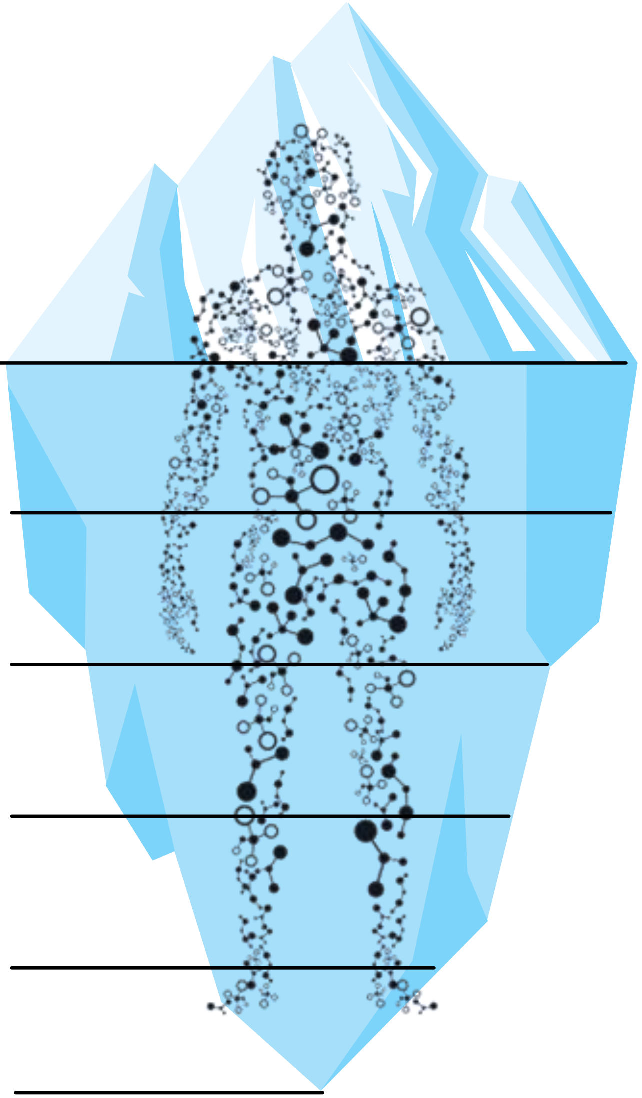
GOAL

Generating **4D Whole-Body Motion** for Hand-Object Grasping

GRIP

Generating **Interaction Poses** Using Spatial Cues and Latent Consistency

- Generate full interaction motions:
  - Action Labels - Text Descriptions
- Scene Interaction & Navigation
- Interaction with large objects
- Human-Object-Interaction Reconstruction from Videos
- Use LLMs for Interaction Motion Synthesis



Thank You!