清華大學 i-VisionGroup

Personal Research and Research Proposal

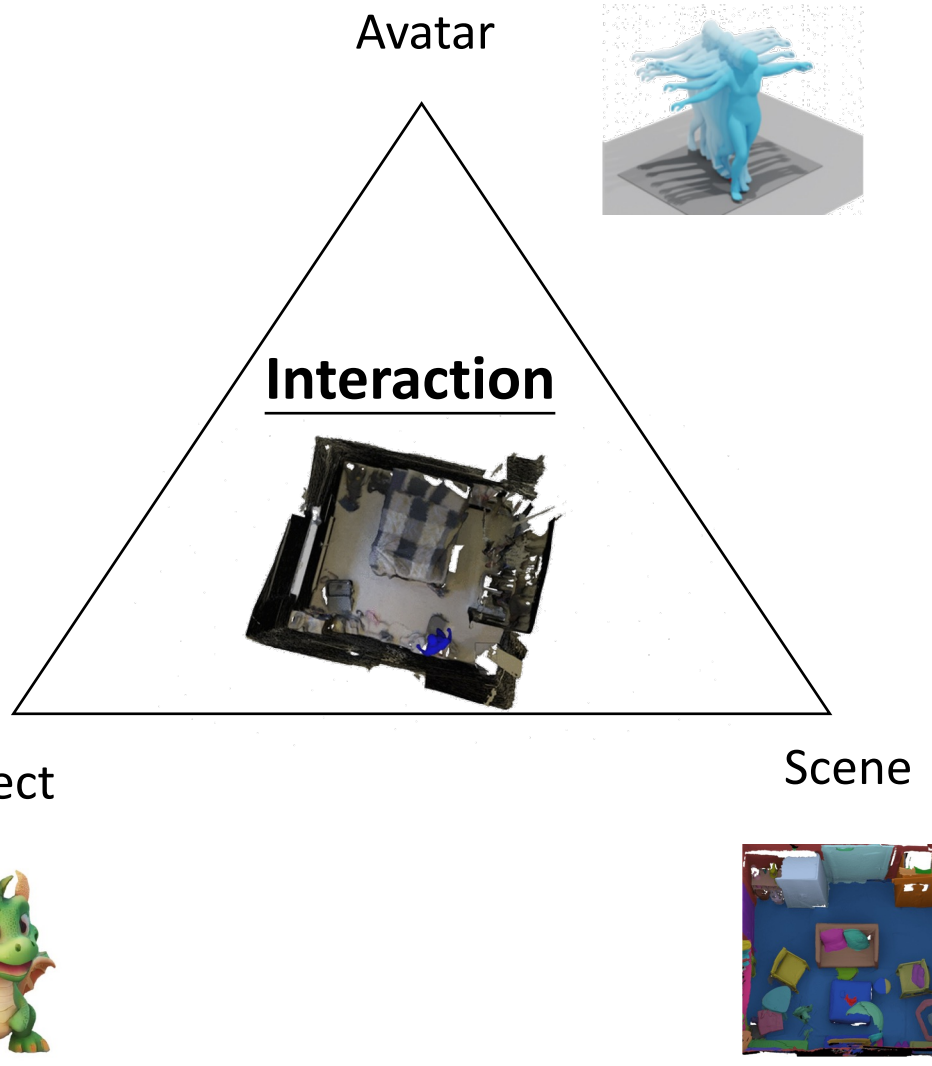
Jinpeng Liu

liujp22@mails.tsinghua.edu.cn

2024/07/05

Outline

- **Research Projects**
  - Avatar: Controllable & Generable
  - Object: Efficient & Diverse
- **Future Research Proposal**
  - Think deeper about “avatar & object”
    - Avatar-object-scene interaction



Broad Application

- Demand for creative digital products is increasing
- Research results are expected to promote digital life system

Microsoft Minecraft

Meta Quest 2

Apple Vision Pro

i-VisionGroup

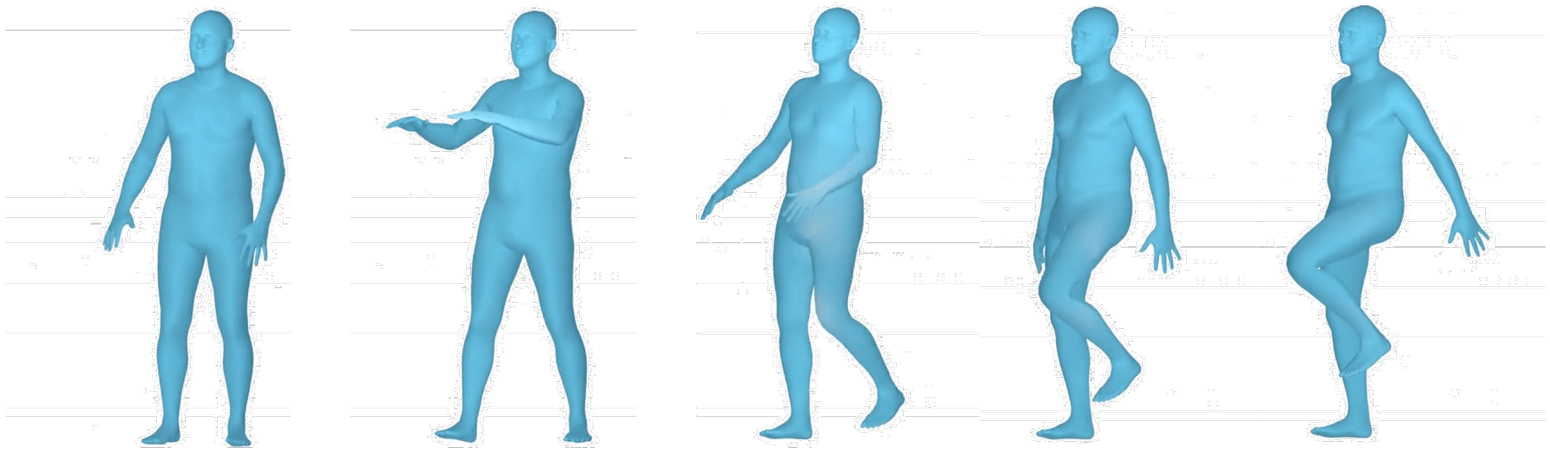
FLAG3D: A 3D Fitness Activity Dataset with Language Instruction

Jinpeng Liu*, Yansong Tang*, Aoyang Liu*,

Bin Yang, Wenxun Dai, Yongming Rao, Jiwen Lu, Jie Zhou, Xiu Li

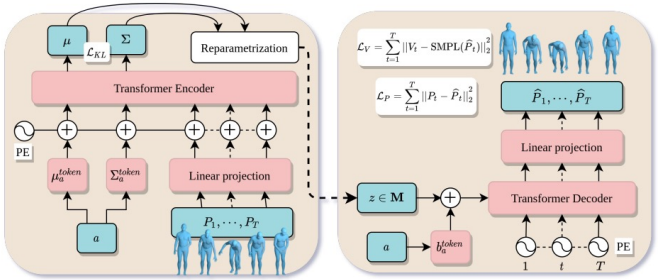
Tsinghua University

Task Description

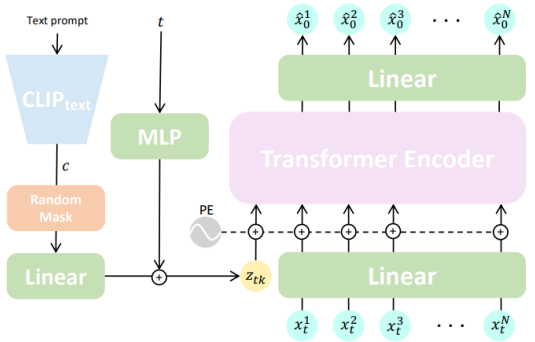
“Knee raising” → 

Language-guided Motion Generation

Low Quantity and Poor Quality of Data

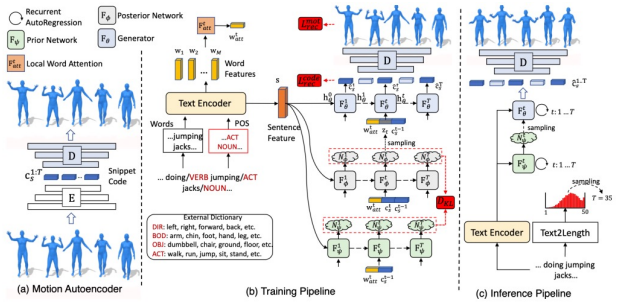


ACTOR
[Petrovich et al. ICCV2021]

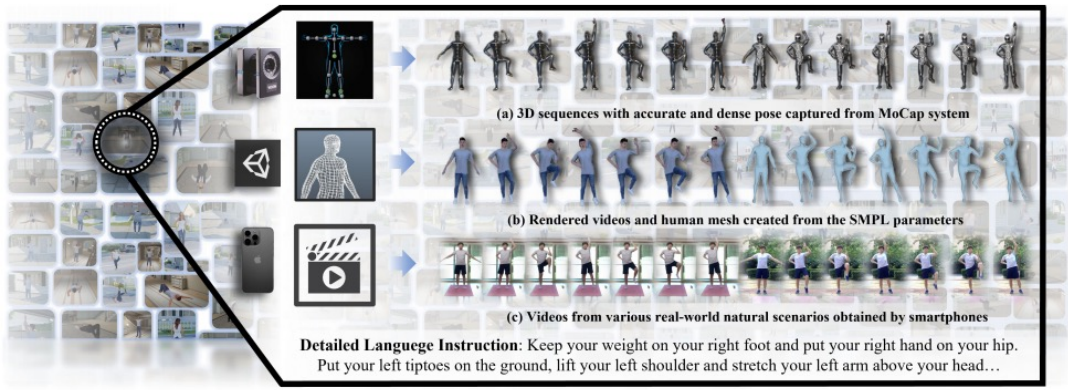


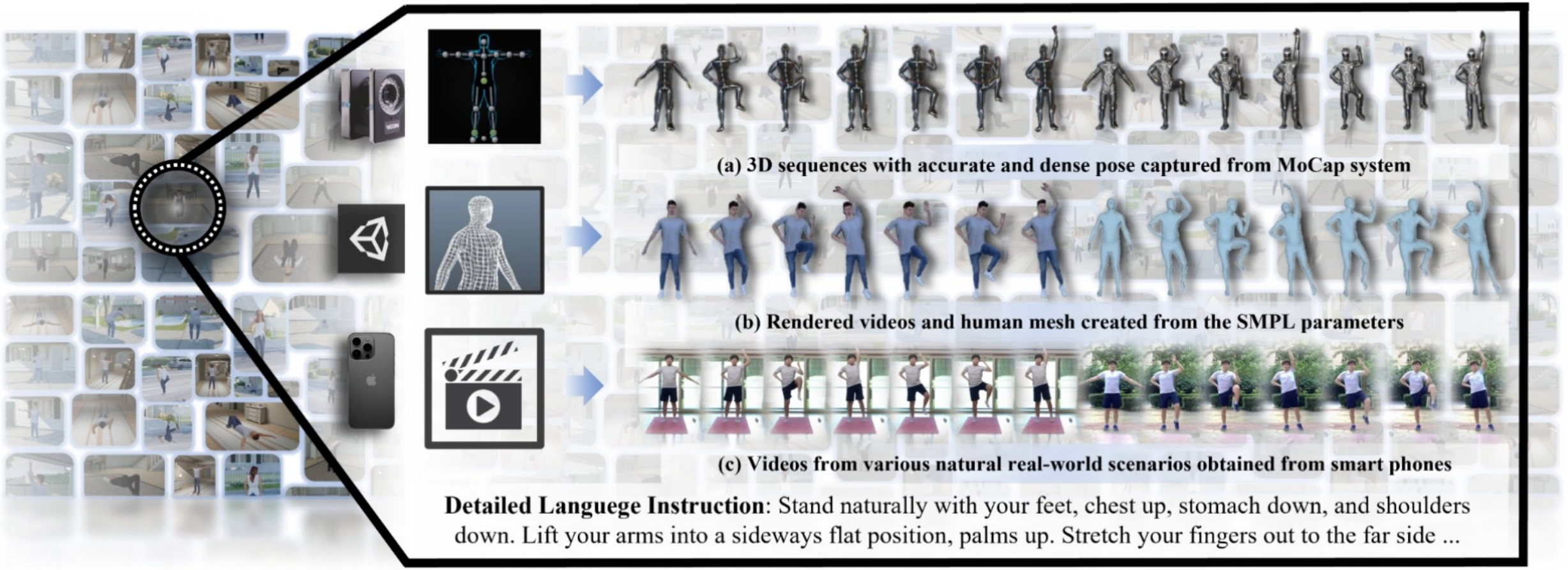
MDM
[Tevet et al. ICLR2023]

Text-to-motion
[Guo et al. CVPR2022]



Ours. CVPR2023





FLAG3D is a large-scale 3D fitness activity dataset with language instruction.

Data

Dataset	Subjs	Cats	Seqs	Frames	LA	K3D	SMPL	Resource	Task
PoseTrack [7]	-	-	550	66K	×	×	×	Nat.	HPE
Human3.6M [33]	11	17	839	3.6M	×	✓	-	Lab	HAR,HPE,HMR
CMU Panoptic [37]	8	5	65	594K	×	✓	-	Lab	HPE
MPI-INF-3DHP [57]	8	8	-	>1.3M	×	✓	-	Lab+Nat.	HPE,HMR
3DPW [96]	7	-	60	51k	×	×	✓	Nat.	HMR
ZJU-MoCap [68]	6	6	9	>1k	×	✓	✓	Lab	HAR,HMR
NTU RGB+D 120 [51]	106	120	114k	-	×	✓	-	Lab	HAR,HAG
HuMMan [11]	1000	500	400K	60M	×	✓	✓	Lab	HAR,HMR
HumanML3D [26]	-	-	14K	-	✓	✓	✓	Lab	HAG
KIT Motion Language [71]	111	-	3911	-	✓	✓	-	Lab	HAG
HumanAct12 [28]	12	12	1191	90K	×	×	✓	Lab	HAG
UESTC [35]	118	40	25K	> 5M	×	✓	-	Lab	HAR,HAG
Fit3D [22]	13	37	-	> 3M	×	✓	✓	Lab	HPE,RAC
EC3D [115]	4	3	362	-	×	✓	-	Lab	HAR
Yoga-82 [95]	-	82	-	29K	×	×	×	Nat.	HAR,HPE
FLAG3D (Ours)	10+10+4	60	180K	20M	✓	✓	✓	Lab+Syn.+Nat.	HAR,HMR,HAG

FLAG3D **20M** << LAION-5B **2.3B**

In-distribution

Out-of-distribution

[1] Tevet G, Raab S, Gordon B, et al. Human Motion Diffusion Model[C]//The Eleventh International Conference on Learning Representations. 2022.

i-VisionGroup

Plan, Posture and Go:
Towards Open-vocabulary Text-to-Motion Generation

Jinpeng Liu[1], Wenxun Dai[1], Chunyu Wang[2], Yiji Cheng[1], Yansong Tang[1], Xin Tong[2]

[1]Tsinghua University [2]Microsoft

Formulation

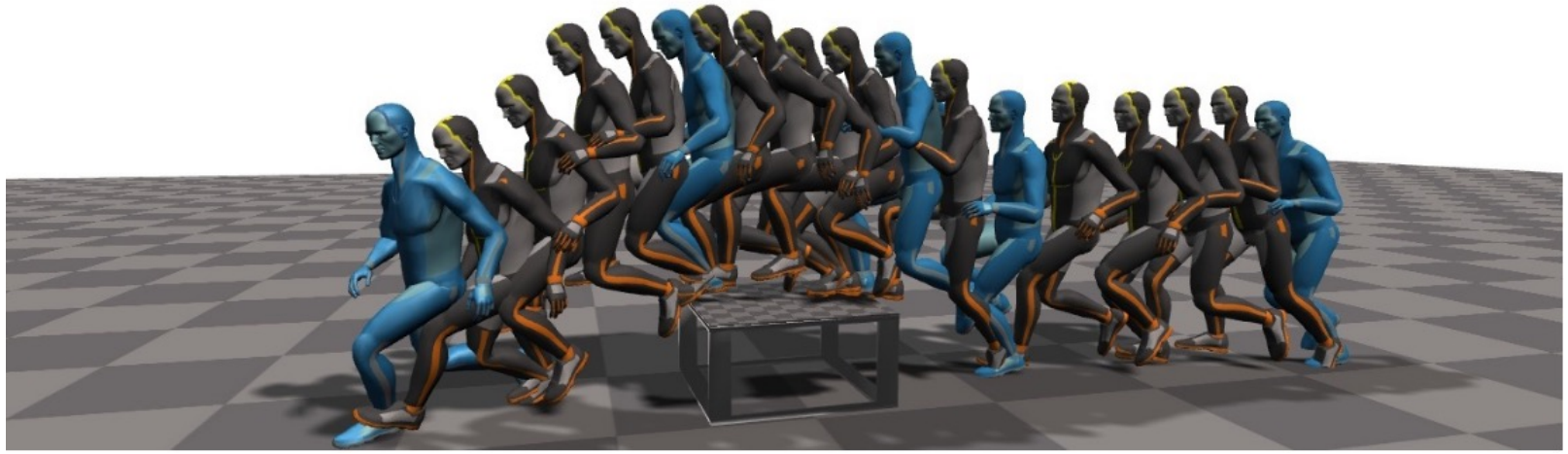
"The language of movement cannot be translated into words."

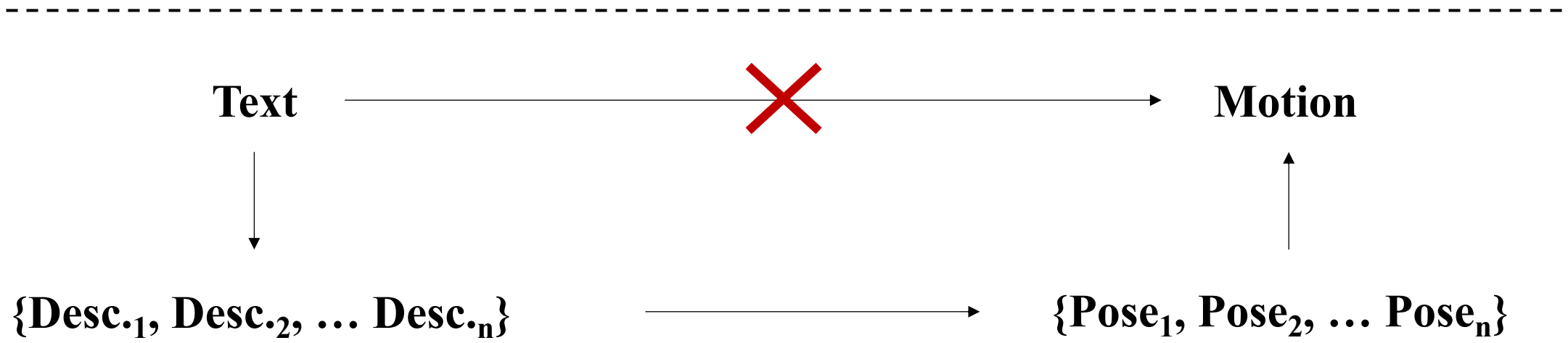
——Barbara Mettler(Dancer)

Is there a novel formulation of the motion generation task that can address general text-to-motion problem without relying on paired text-motion data?

Formulation

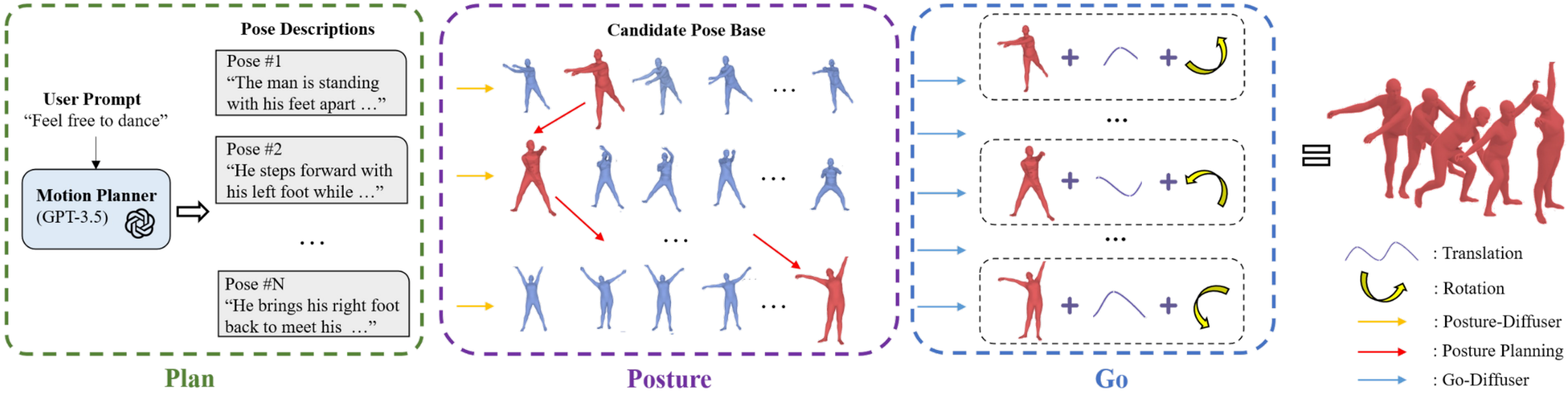
Motion ? ⇨ **Pose sequence** + **Global Information**





Pipeline

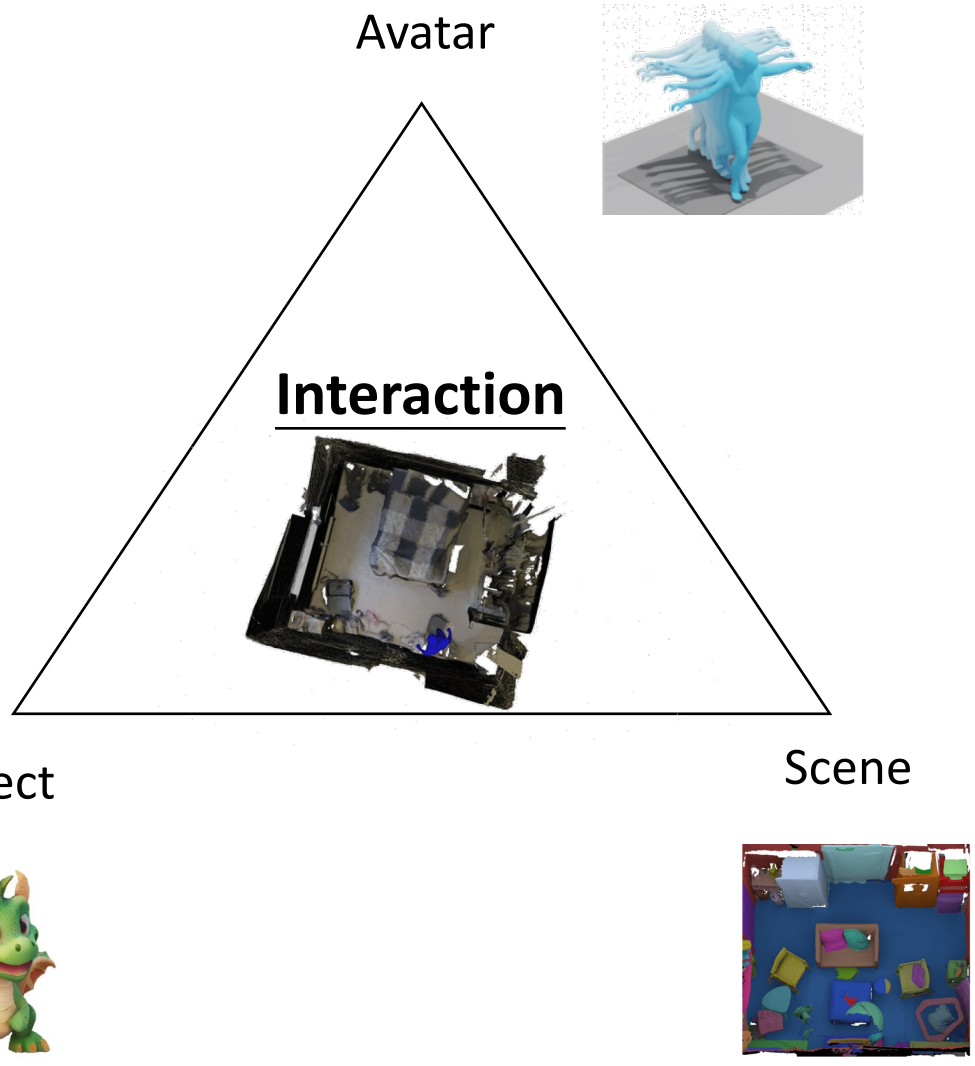
divide-and-conquer



Motion Generated by Our Model

Outline

- **Research Projects**
  - Avatar: Controllable & Generable
  - **Object: Efficient & Diverse**
- **Future Research Proposal**
  - Think deeper about “avatar & object”
    - Avatar-object-world interaction



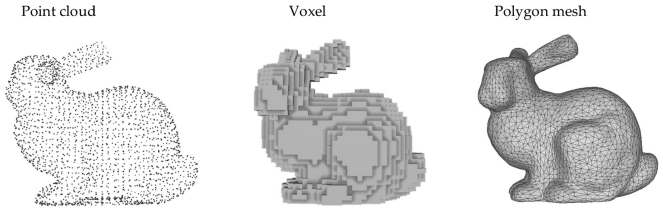
Task Description

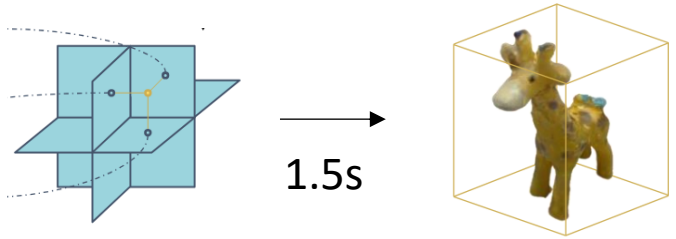
Image to 3D

Motivation

3D Representation

- ❌ Point Cloud: Poor visual result
- ❌ Voxel, Mesh : GPU-unfriendly
- ❌ Triplane: Time consuming
- ✅ Gaussian: Real time and easy to scale up

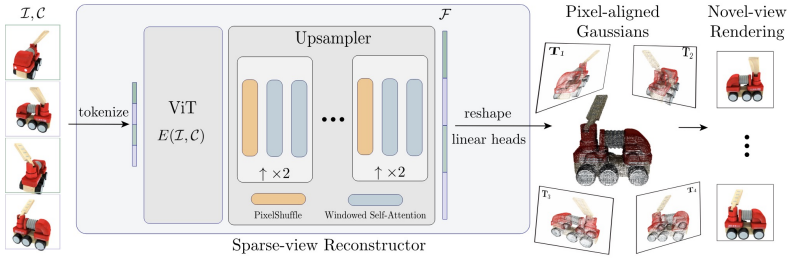




Rendering 2s (60 frames) video cost **1.5min!**

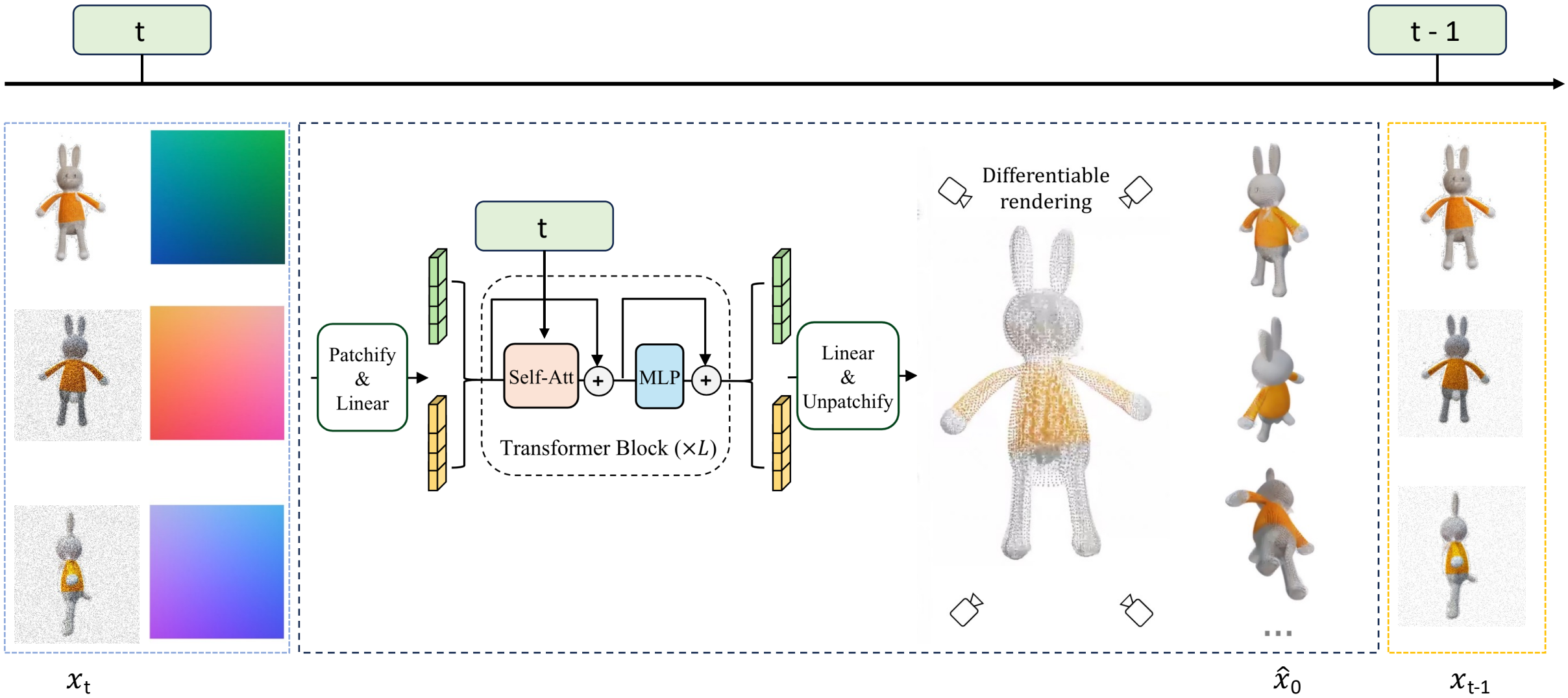
Formulation

- Reconstruction
- Generation
  - Utilizing diffusion to model the probability distribution



GRM. Yinghao Xu, et al. Stanford University.

Pipeline

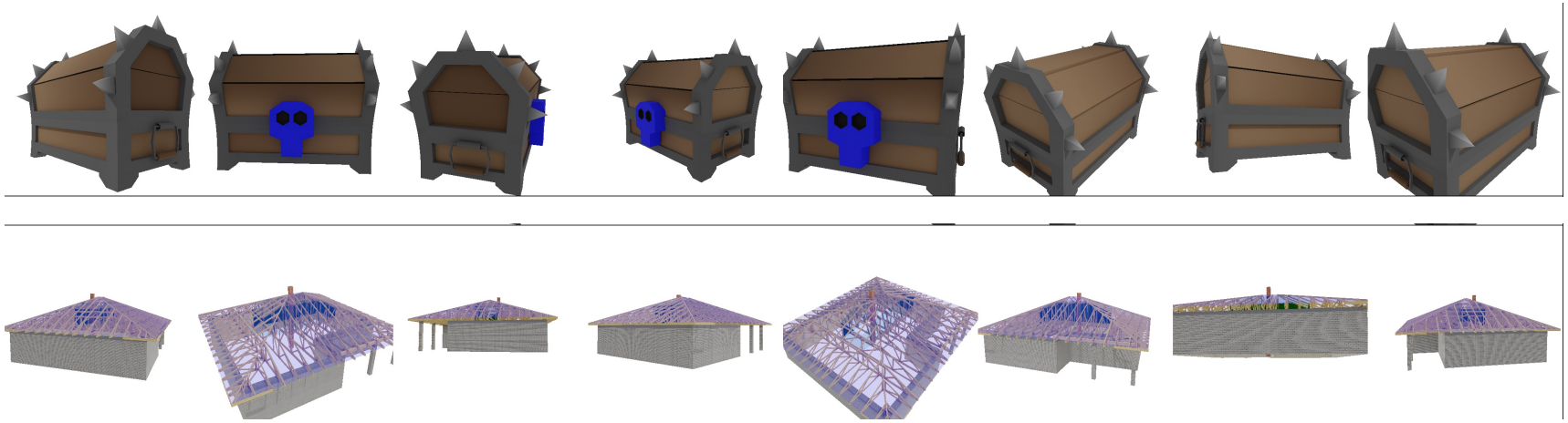


Images + Plücker rays

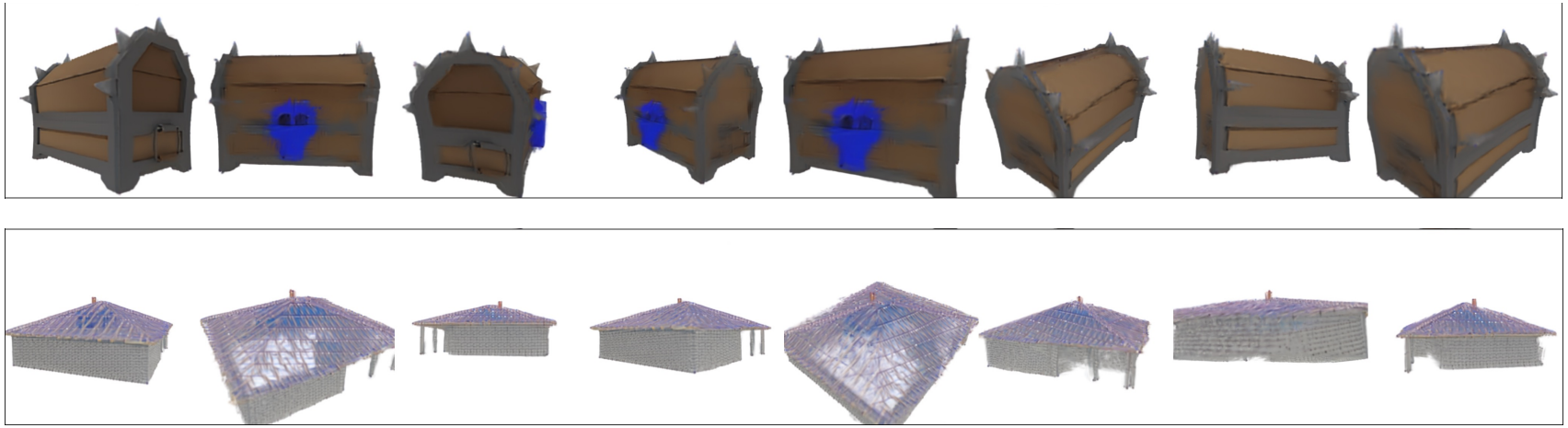
Results

Visualization

GT

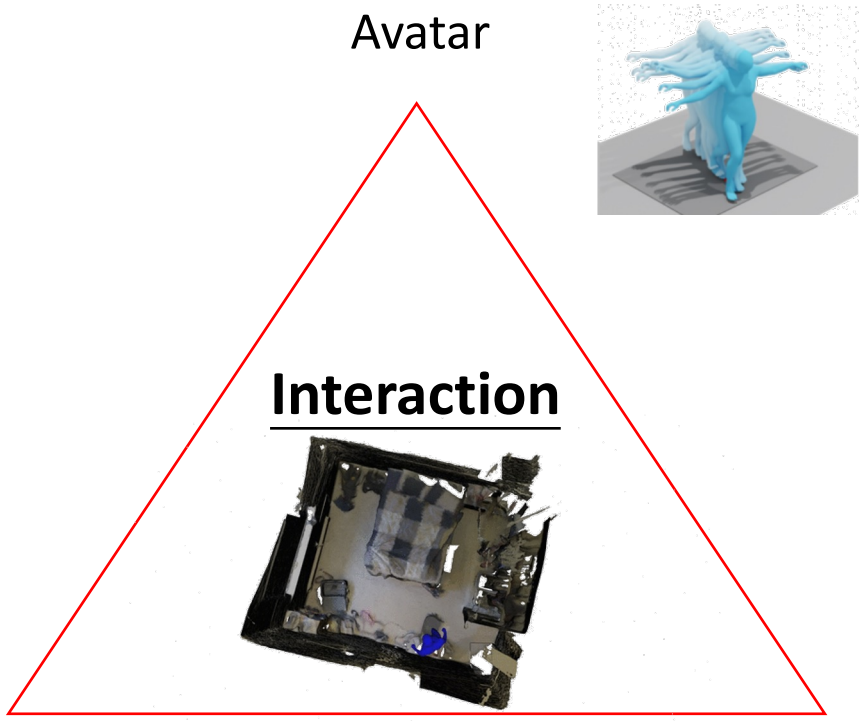


Pred



Outline

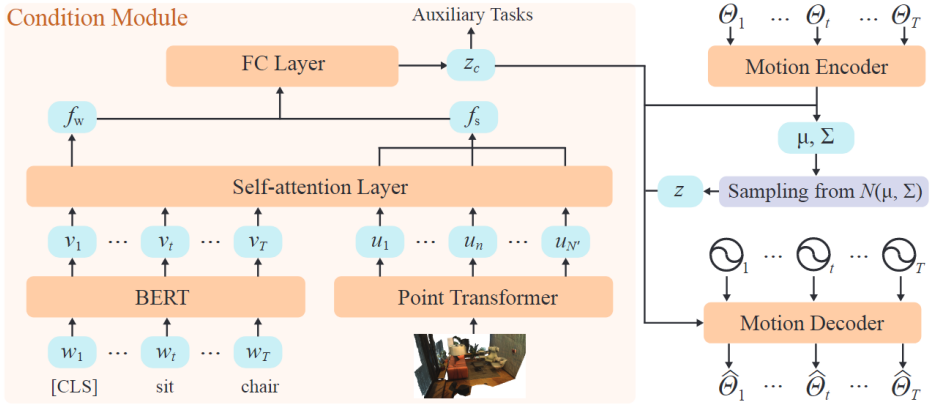
- **Research Projects**
  - Avatar: Controllable & Generable
  - **Object: Efficient & Diverse**
- **Future Research Proposal**
  - Think deeper about “avatar & object”
    - Avatar-object-scene interaction

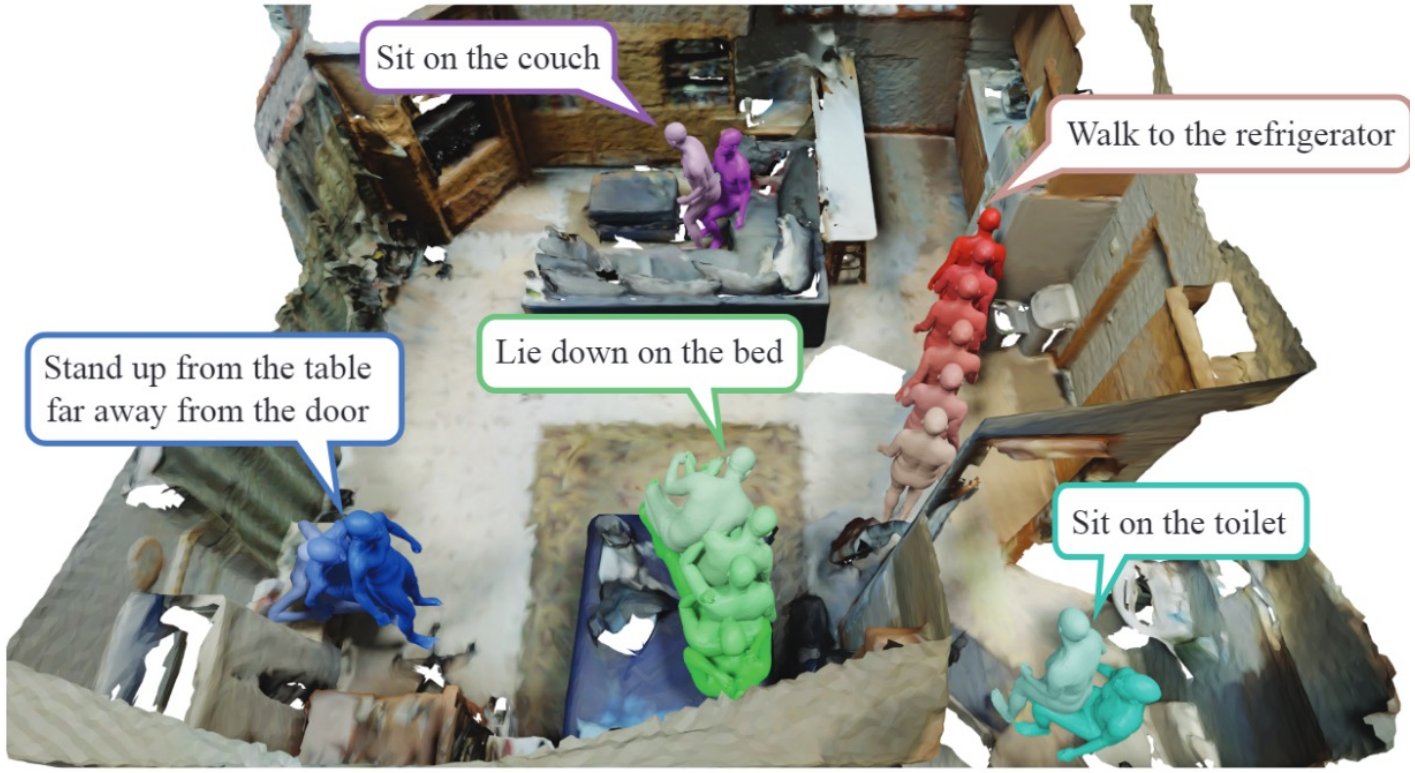


Task Description

Input: Language & Scene

Output: Interaction

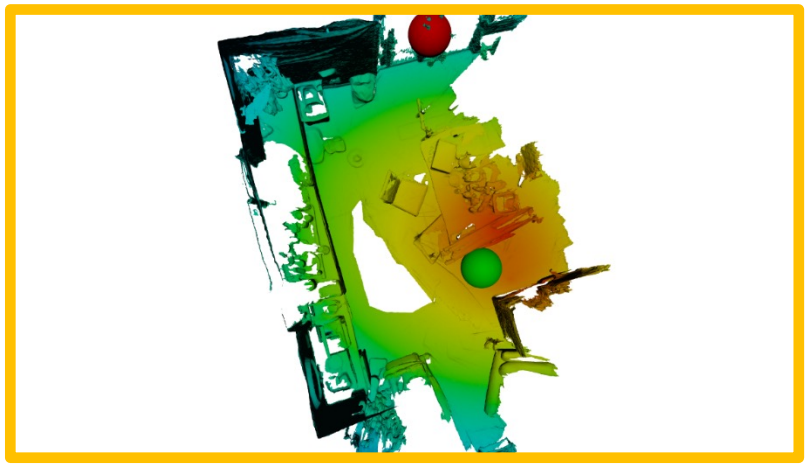
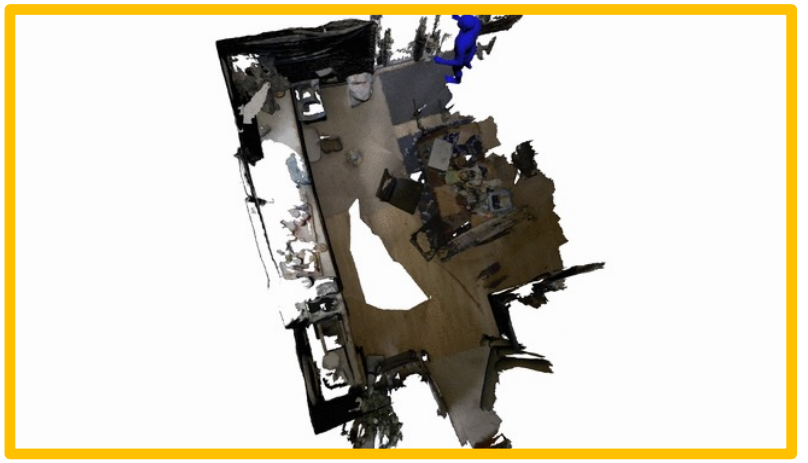
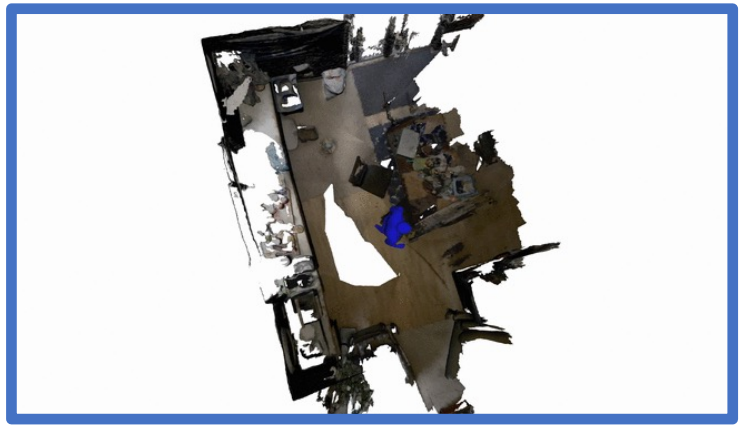
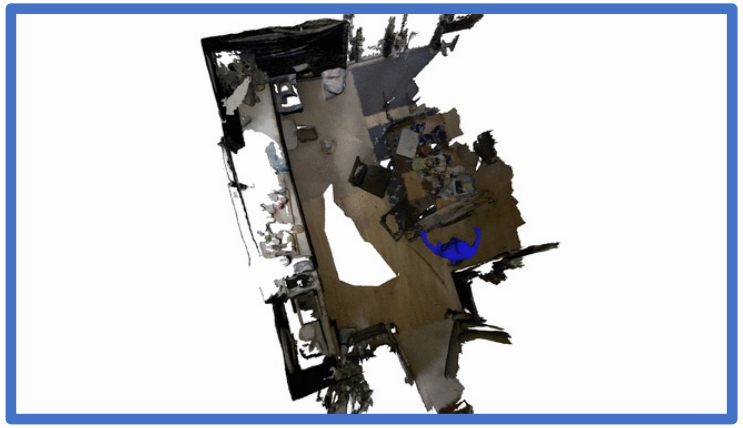
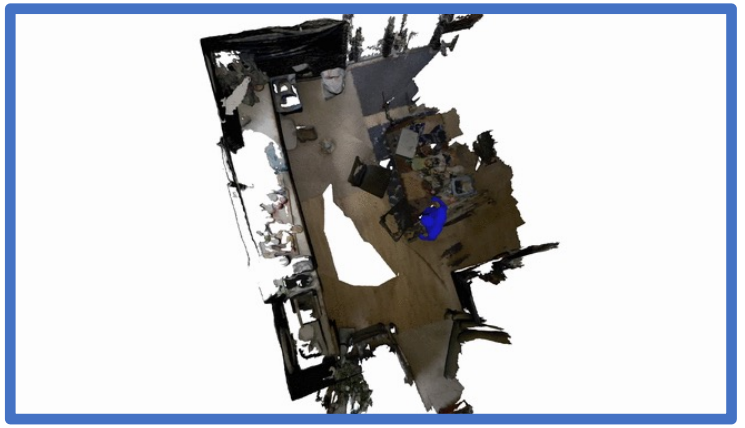




HUMANISE: Language-conditioned Human Motion Generation in 3D Scenes. NeurIPS2023

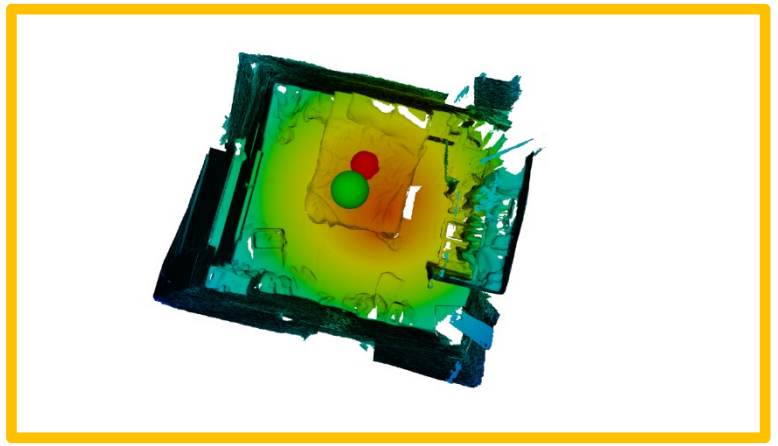
Localization Error

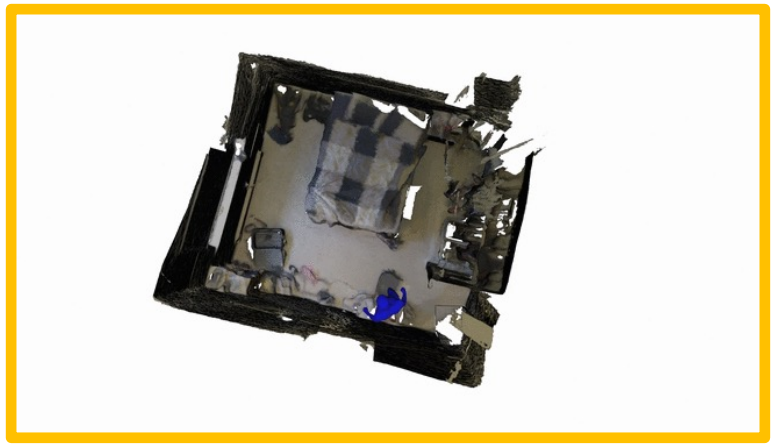
Walk to the glass doors

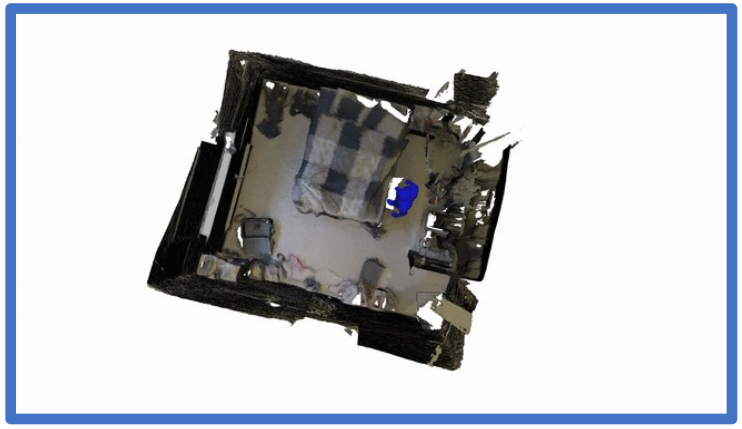






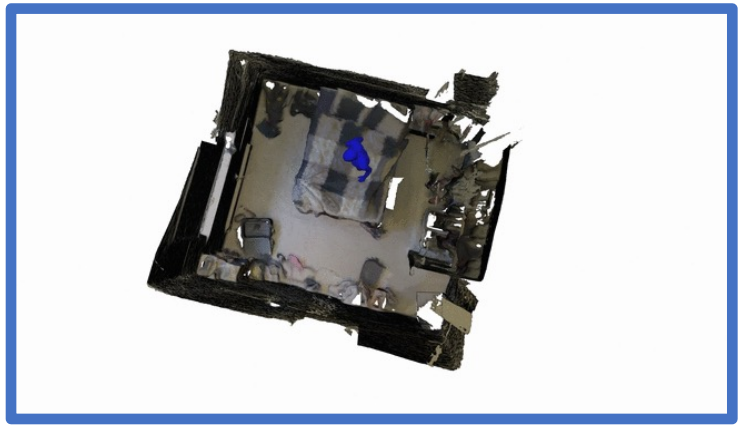
Physical Error

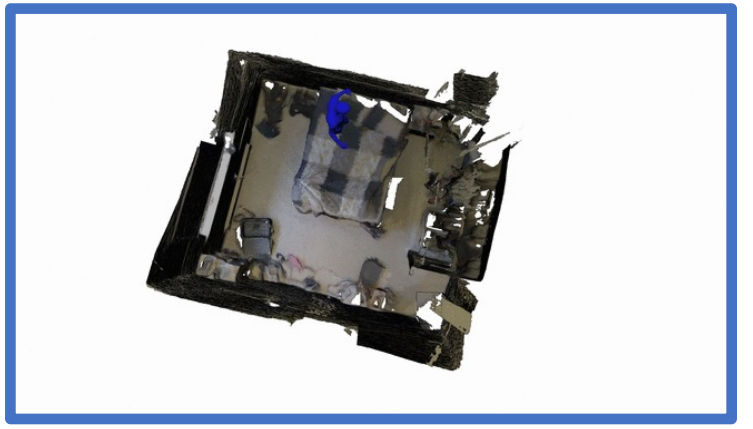
Walk to the bed











Potential Solutions

- Physical Error & Localization Error

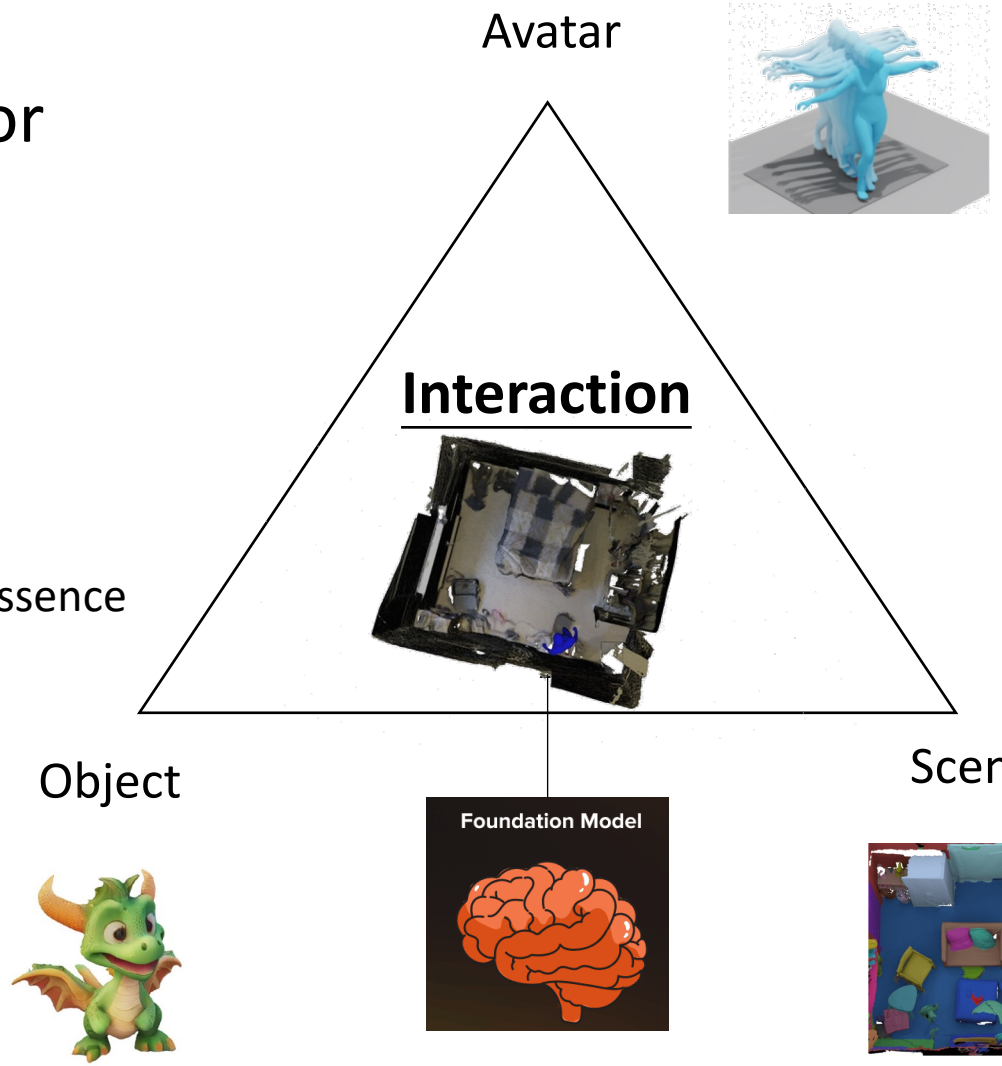
Physics-based Optimization[1] — Difficult to design and inflexible

More Data — Hard to require

Early Fusion[2] — It's not the essence

We need the foundation model !

Physical Knowledge & Object Properties



[1] Diffusion-based Generation, Optimization, and Planning in 3D Scenes. CVPR23
[2] LAVT: Language-Aware Vision Transformer for Referring Image Segmentation. CVPR22

Some Thoughts

- **Thoughts**
  - In the last ten years
    - Recommendation beat Search
  - In the future ten years
    - Generation beat Recommendation

- **Basis**
  - Vision Pro is the iPhone 1. When iPhone 4 will arrive?

Education Background

- 2022.08-(exp. 2025) Tsinghua University, M.E. in Data Science

  Advisor: Prof. Yansong Tang

- 2018.08-2022.07 Sun Yat-sen University, B. E. in Intelligent Science and Technology

Industrial Experience

Tencent ARC Lab, Shenzhen, China. December-Now, 2023.

- Project: Text to 3D Object Generation.
- Work with Dr. Xintao Wang, Dr. Ying Shan.

Microsoft Research - Asia (MSRA), Beijing, China. April-December, 2023.

- Project: Human Motion Understanding and Generation.
- Work with Dr. Chunyu Wang, Dr. Xin Tong.

Webset: https://moonsliu.github.io/ Github: https://github.com/moonsliu