Personal Research and Research Proposal

Jinpeng Liu

liujp22@mails.tsinghua.edu.cn

Education Background

- 2022.08-(exp. 2025) Tsinghua University, M.E. in Data Science
  - Supervisor: Yansong Tang
  - **Leader of Multimodal Generation and Interaction Team at IVG@SZ**
- 2018.08-2022.07 Sun Yat-sen University, B. E. in Intelligent Science and Technology

Industrial Experience

TEXAS
The University of Texas at Austin

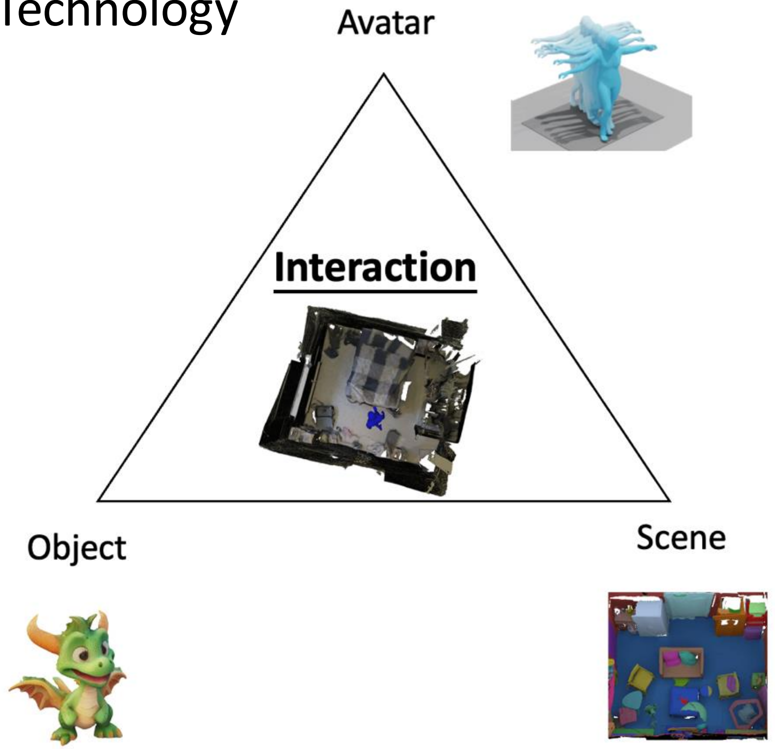
- Project: **3D Avatar Mesh Recovery**
- Works with Dr. Georgios Pavlakos, Dr. Qixing Huang

Tencent腾讯

- Project: **3D Object & Scene Generation**
- Works with Dr. Xintao Wang, Dr. Ying Shan

Microsoft

- Project: **3D Avatar Animation**
- Works with Dr. Chunyu Wang, Dr. Xin Tong



Outline

□ Timeline of previous research

- 2022.06 - 2023.02 (3D-Avatar). **Jinpeng Liu***, Yansong Tang*, Aoyang Liu*, Bin Yang, Wenxun Dai, Yongming Rao, Jiwen Lu, Jie Zhou, Xiu Li. *FLAG3D: A 3D Fitness Activity Dataset with Language Instruction.* **[Accepted by CVPR 2023]**
- 2023.02 - 2023.11 (3D-Avatar). **Jinpeng Liu**, Wenxun Dai, Chunyu Wang, Yiji Cheng, Yansong Tang, Xin Tong. *Posture, Plan and Go: Towards Open-world Motion Generation.* **[Accepted by ECCV 2024]**
- 2023.11 - 2024.04 (3D-Avatar). Wenxun Dai, Ling-Hao Chen, Jingbo Wang, **Jinpeng Liu**, Bo Dai, Yansong Tang. *MotionLCM: Realtime Controllable Motion Generation via Latent Consistency Model.* **[Accepted by ECCV 2024]**
- 2024.04 - 2024.10 (3D-Object). **Jinpeng Liu**, Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Ying Shan, Yansong Tang. *NovelGS: Consistent Novel-view Denoising via Large Gaussian Reconstruction Model.* **[Submitted to CVPR]**
- 2023.02 - 2024.03 (3D&Video). Yansong Tang, Aoyang Liu, **Jinpeng Liu**, Shiyi Zhang, Wenxun Dai, Jie Zhou, Xiu Li, Jiwen Lu. *FLAG3D++: A Benchmark for 3D Fitness Activity Comprehension with Language Instruction.* **[Submitted to TPAMI]**

- **Others**

- Kun Xiang, Xing Zhang, Jinwen She, **Jinpeng Liu**, Haohan Wang, Shiqi Deng, Shancheng Jiang. Toward robust diagnosis: a contour attention preserving adversarial defense for COVID-19 detection. **[Accepted by AAAI 2023]**
- Jinwen She, **Jinpeng Liu**, et. al. A universal computer-aided diagnosis system for cross-regional skin lesion recognition using deep graph-based network. **[Submitted to NAT COMMUN.]**

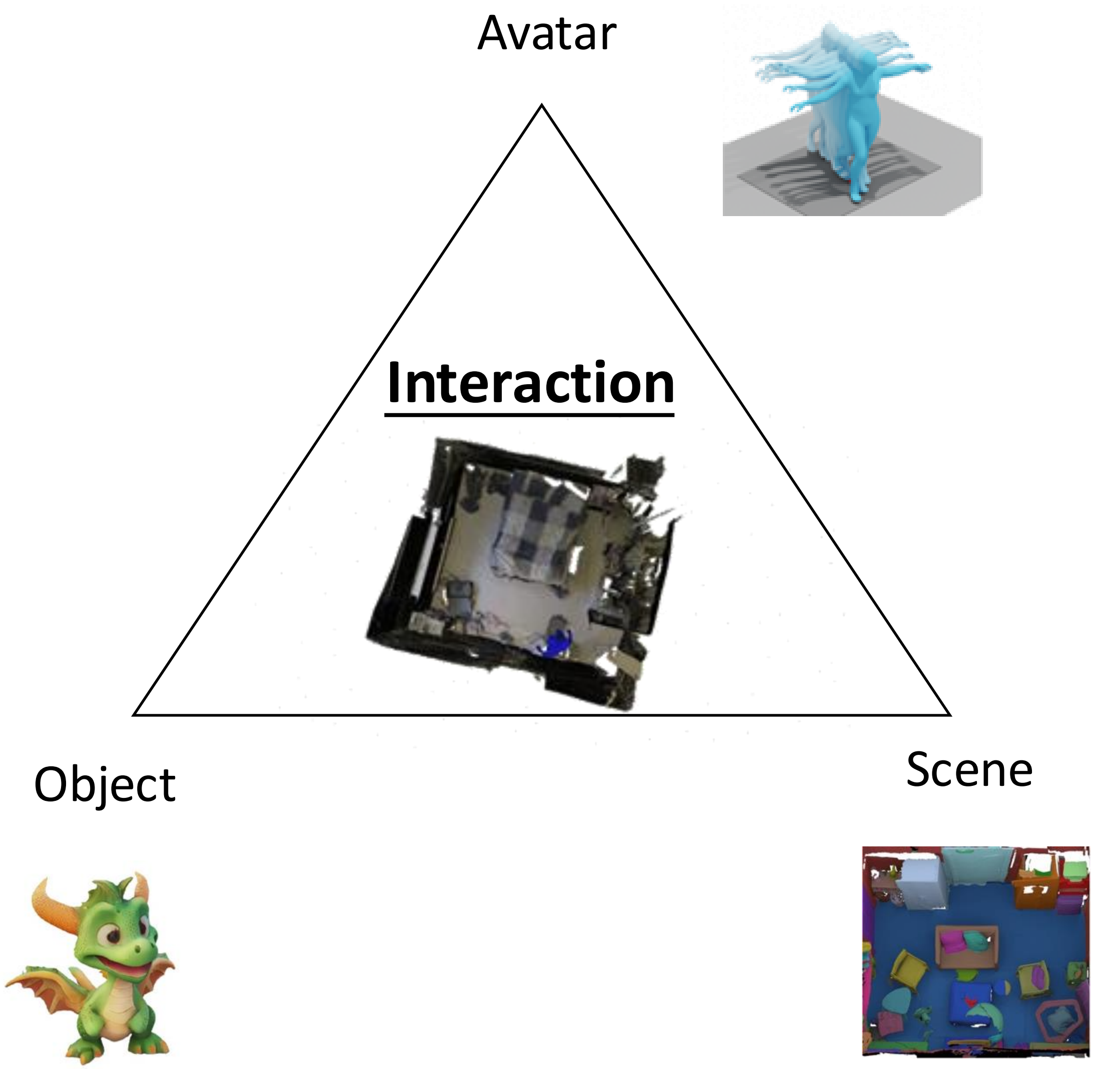
Outline

□ 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

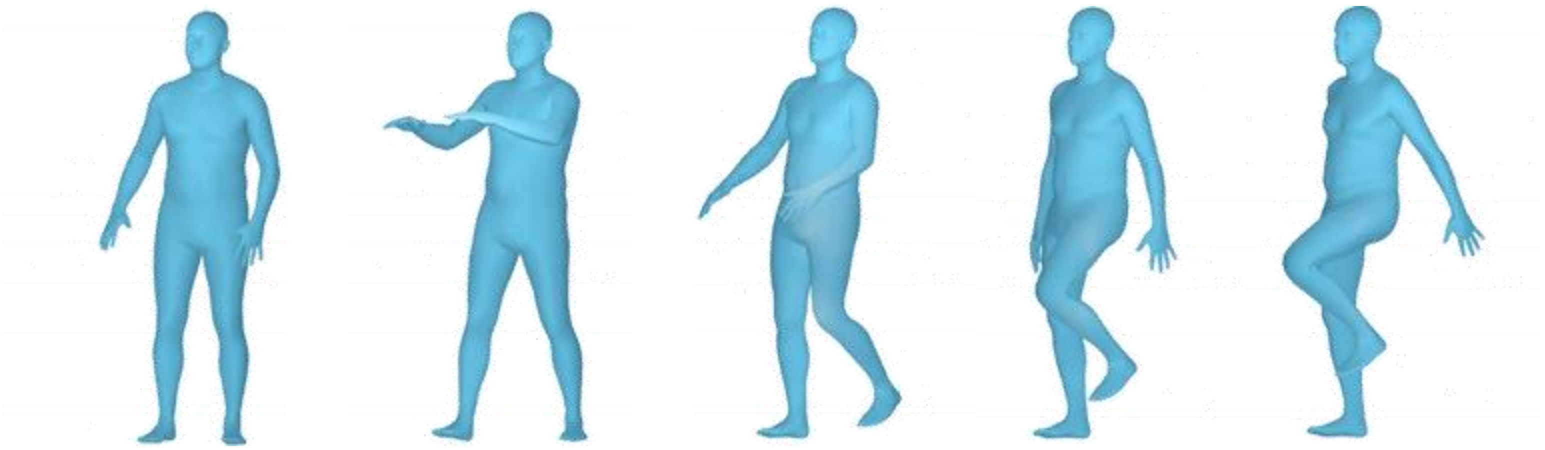
Microsoft Minecraft

Meta Quest 2

Apple Vision Pro

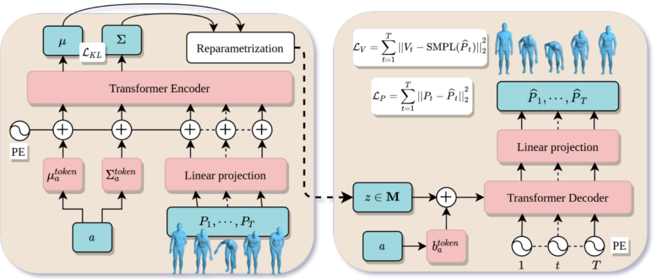
Task Description

“Knee raising” →

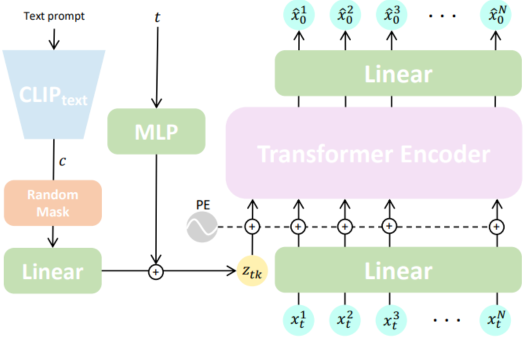


Language-guided Motion Generation

Low Quantity and Poor Quality of Data

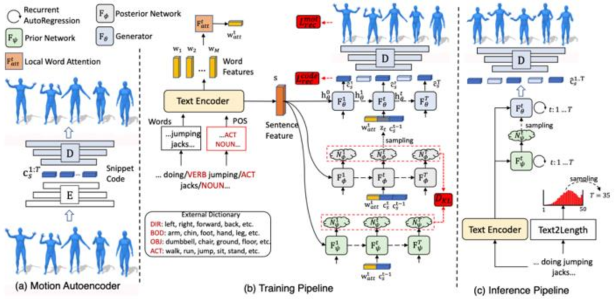


ACTOR
[Petrovich et al. ICCV2021]

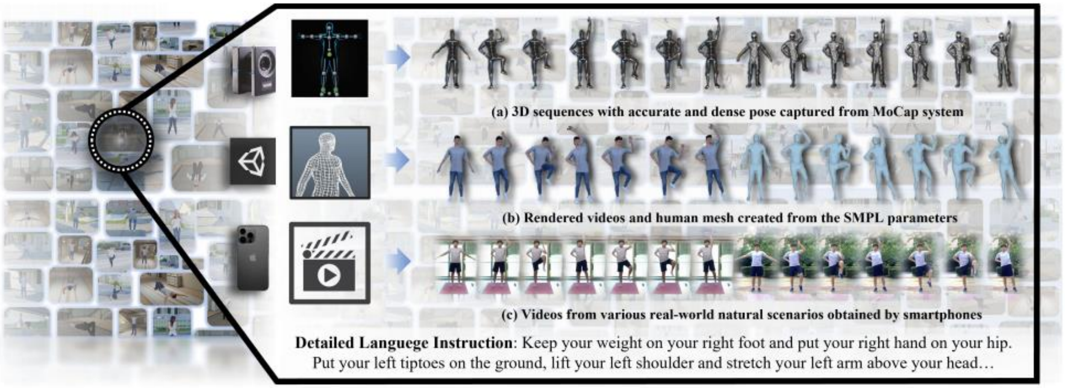


MDM
[Tevet et al. ICLR2023]

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



Ours. CVPR2023



(a) 3D sequences with accurate and dense pose captured from MoCap system

(b) Rendered videos and human mesh created from the SMPL parameters

(c) Videos from various real-world natural scenarios obtained by smartphones

Detailed Languege Instruction: Keep your weight on your right foot and put your right hand on your hip. Put your left tiptoes on the ground, lift your left shoulder and stretch your left arm above your head…

FLAG3D: A 3D Fitness Activity Dataset with Language Instruction

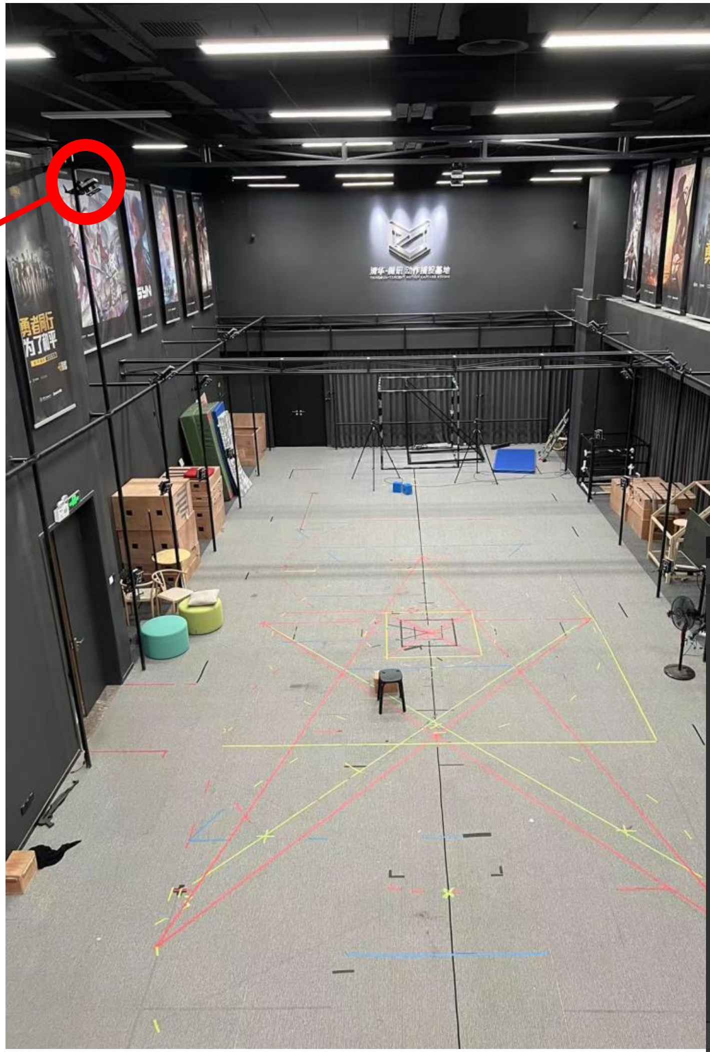
Jinpeng Liu*, Yansong Tang*, Aoyang Liu*,

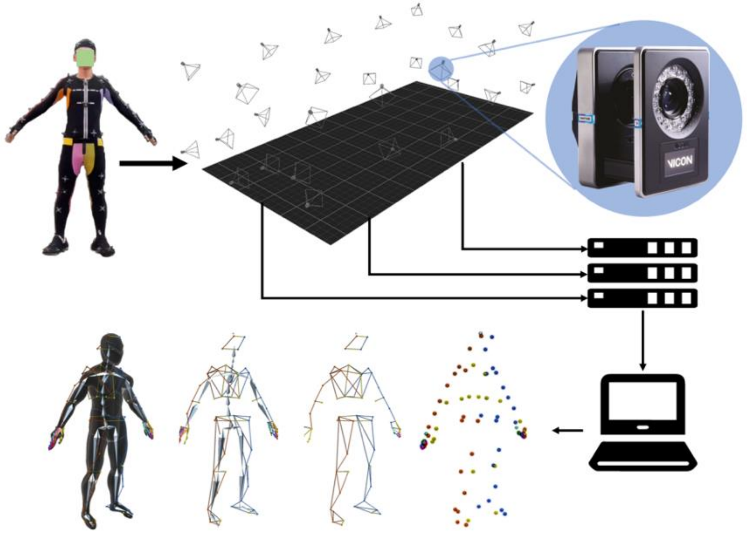
Bin Yang, Wenxun Dai, Yongming Rao, Jiwen Lu, Jie Zhou, Xiu Li

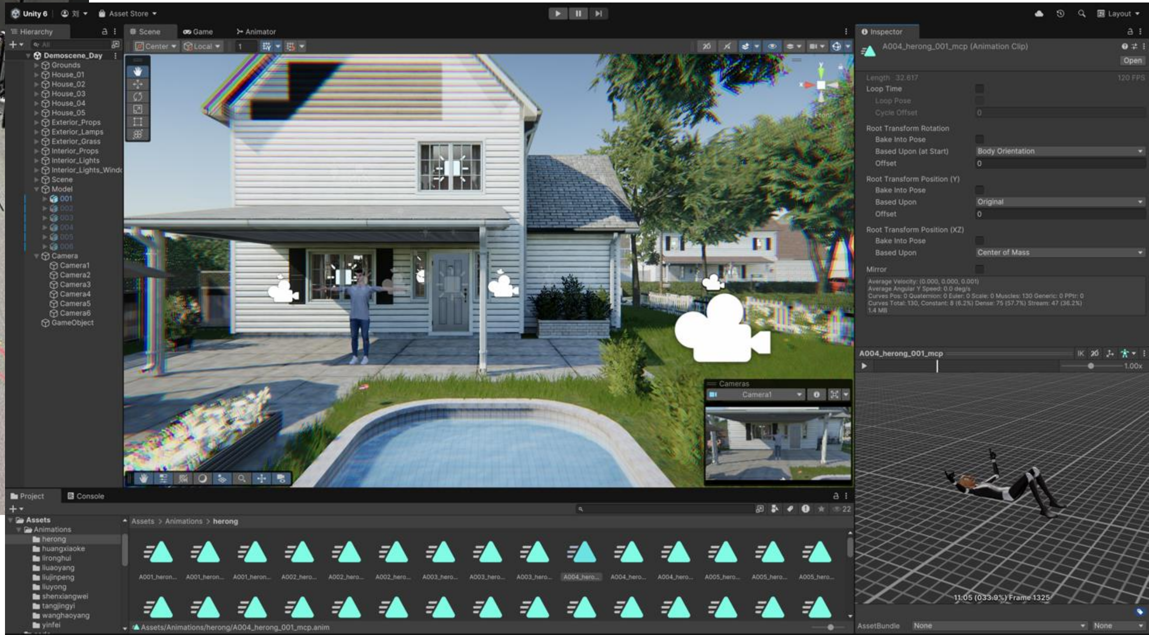
Tsinghua University

清華大學

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA







Hierarchy

- Demoscene_Day
  - Grounds
  - House_01
  - House_02
  - House_03
  - House_04
  - House_05
  - Exterior_Props
  - Exterior_Lamps
  - Exterior_Grass
  - Interior_Props
  - Interior_Lights
  - Interior_Lights_Windc
  - Scene
  - Model
    - 001
    - 002
    - 003
    - 004
    - 005
    - 006
  - Camera
    - Camera1
    - Camera2
    - Camera3
    - Camera4
    - Camera5
    - Camera6
  - GameObject

Inspector

A004_herong_001_mcp (Animation Clip)

Open

Length 32.617 120 FPS

Loop Time □
Loop Pose □
Cycle Offset 0

Root Transform Rotation
Bake Into Pose □
Based Upon (at Start) Body Orientation
Offset 0

Root Transform Position (Y)
Bake Into Pose □
Based Upon Original
Offset 0

Root Transform Position (XZ)
Bake Into Pose □
Based Upon Center of Mass

Mirror □

```
Average Velocity: (0.000, 0.000, 0.001)
Average Angular Y Speed: 0.0 deg/s
Curves Pos: 0 Quaternion: 0 Euler: 0 Scale: 0 Muscles: 130 Generic: 0 PPtr: 0
Curves Total: 130, Constant: 8 (6.2%) Dense: 75 (57.7%) Stream: 47 (36.2%)
1.4 MB
```

A004_herong_001_mcp
1.00x

11:05 (033.0%) Frame 1325

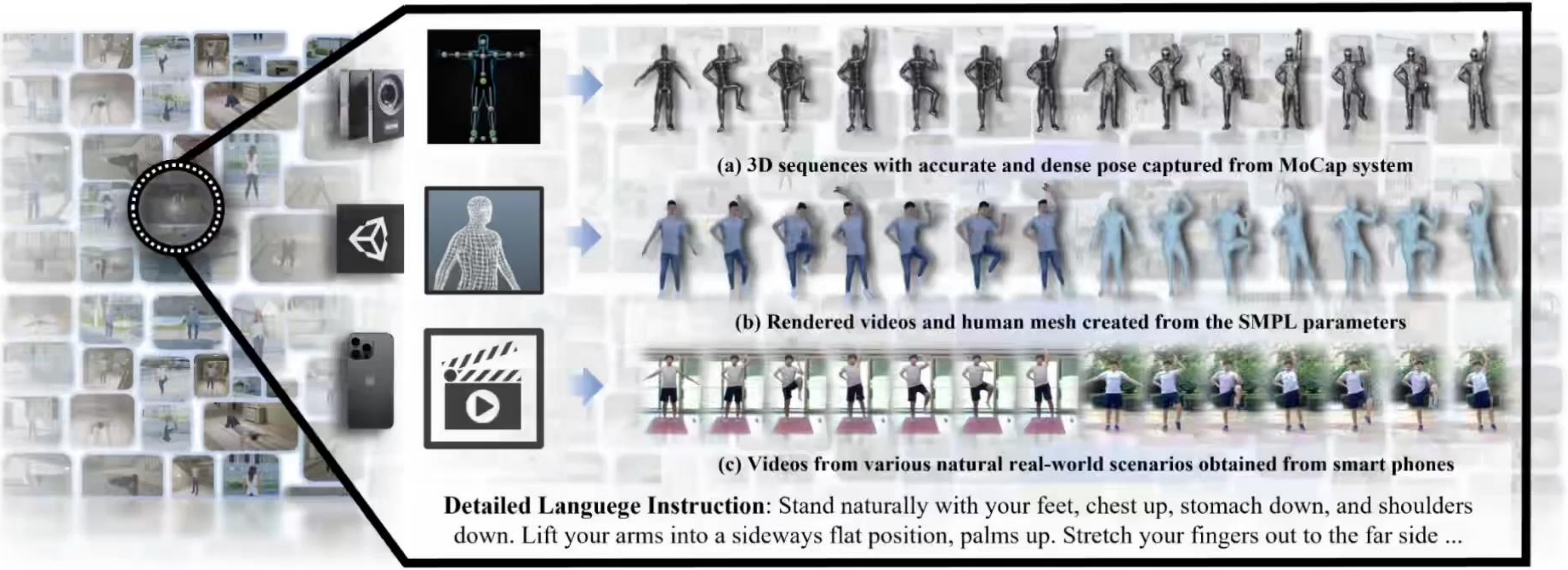
Project

- Assets
  - Animations
    - herong
    - huangxiaoke
    - lironghui
    - liuaoyang
    - liujinpeng
    - liuyong
    - shenxiangwei
    - tangjingyi
    - wanghaoyang
    - yinfei

Assets > Animations > herong

A001_heron... A001_heron... A001_heron... A002_hero... A002_hero... A002_hero... A003_hero... A003_hero... A003_hero... A004_hero... A004_hero... A004_hero... A005_hero... A005_hero... A005_hero...

Assets/Animations/herong/A004_herong_001_mcp.anim



(a) 3D sequences with accurate and dense pose captured from MoCap system

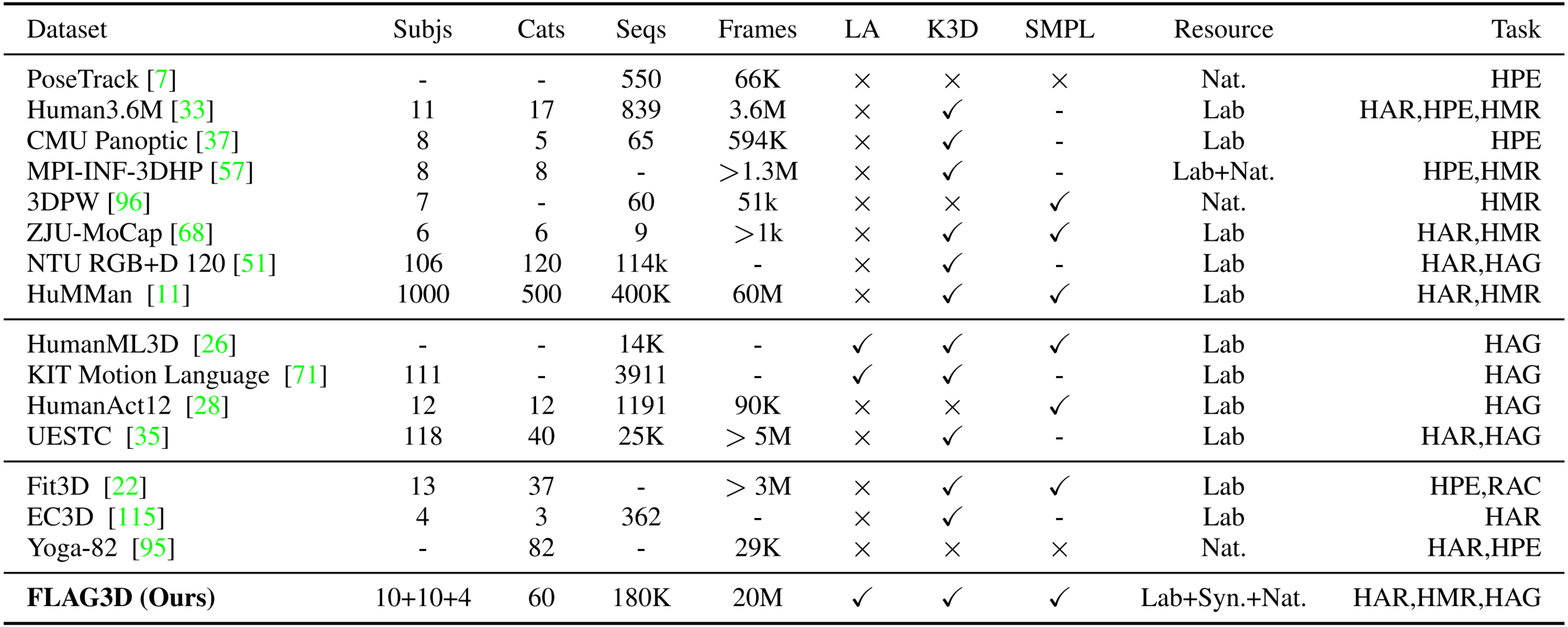
(b) Rendered videos and human mesh created from the SMPL parameters

(c) Videos from various natural real-world scenarios obtained from smart phones

Detailed Languege Instruction: Stand naturally with your feet, chest up, stomach down, and shoulders down. Lift your arms into a sideways flat position, palms up. Stretch your fingers out to the far side ...

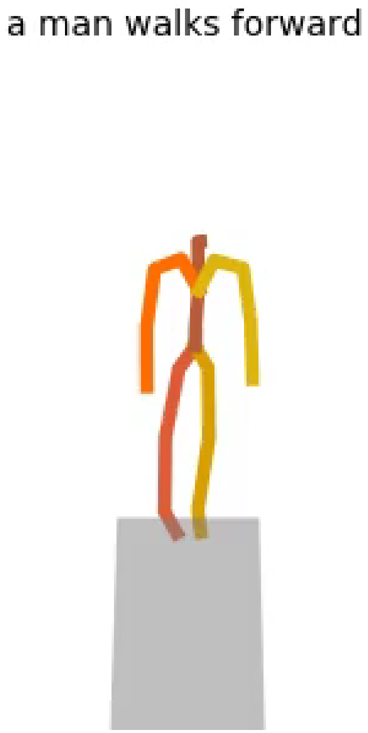
FLAG3D features the following three aspects:

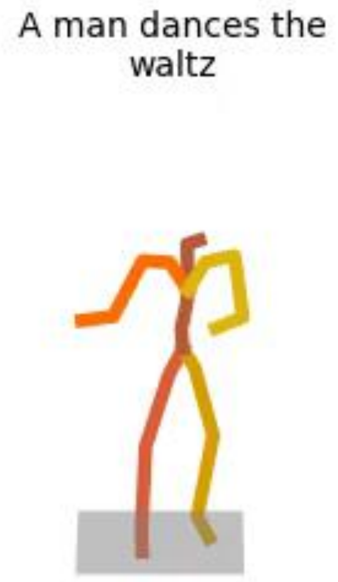
Data



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		LAION-5B
20M	<<	**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.

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

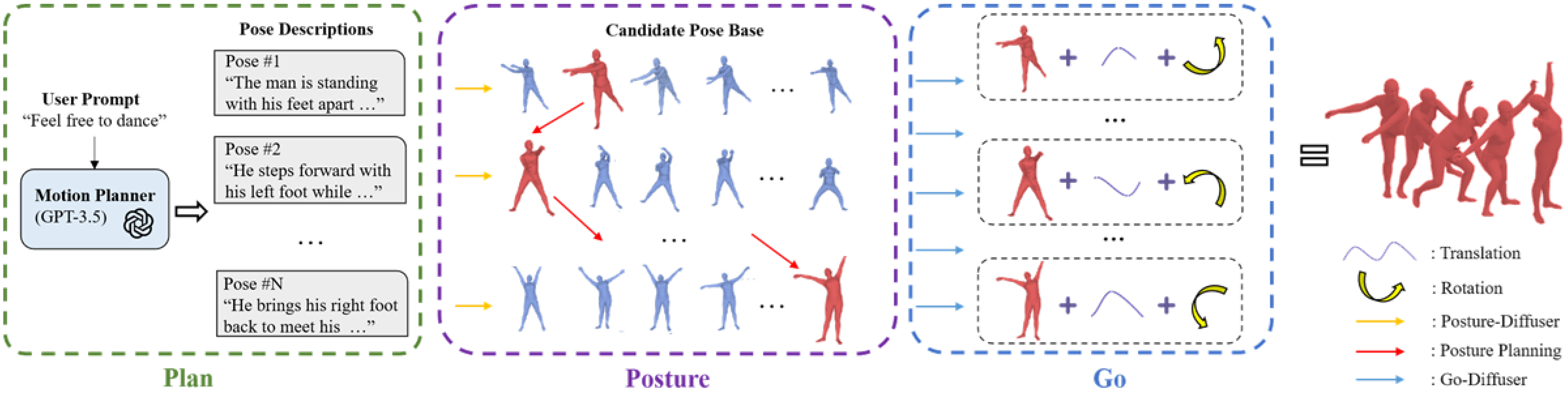
Motion ? ⇨ Pose sequence + Global Information

$$\begin{tikzcd}
\text{Text} \arrow[r, "\times" description] \arrow[d] & \text{Motion} \\
\{\text{Desc.}_1, \text{Desc.}_2, \ldots \text{Desc.}_n\} \arrow[r] & \{\text{Pose}_1, \text{Pose}_2, \ldots \text{Pose}_n\} \arrow[u]
\end{tikzcd}$$

Robust Motion In-betweening. SIGGRAPH 20.

Pipeline

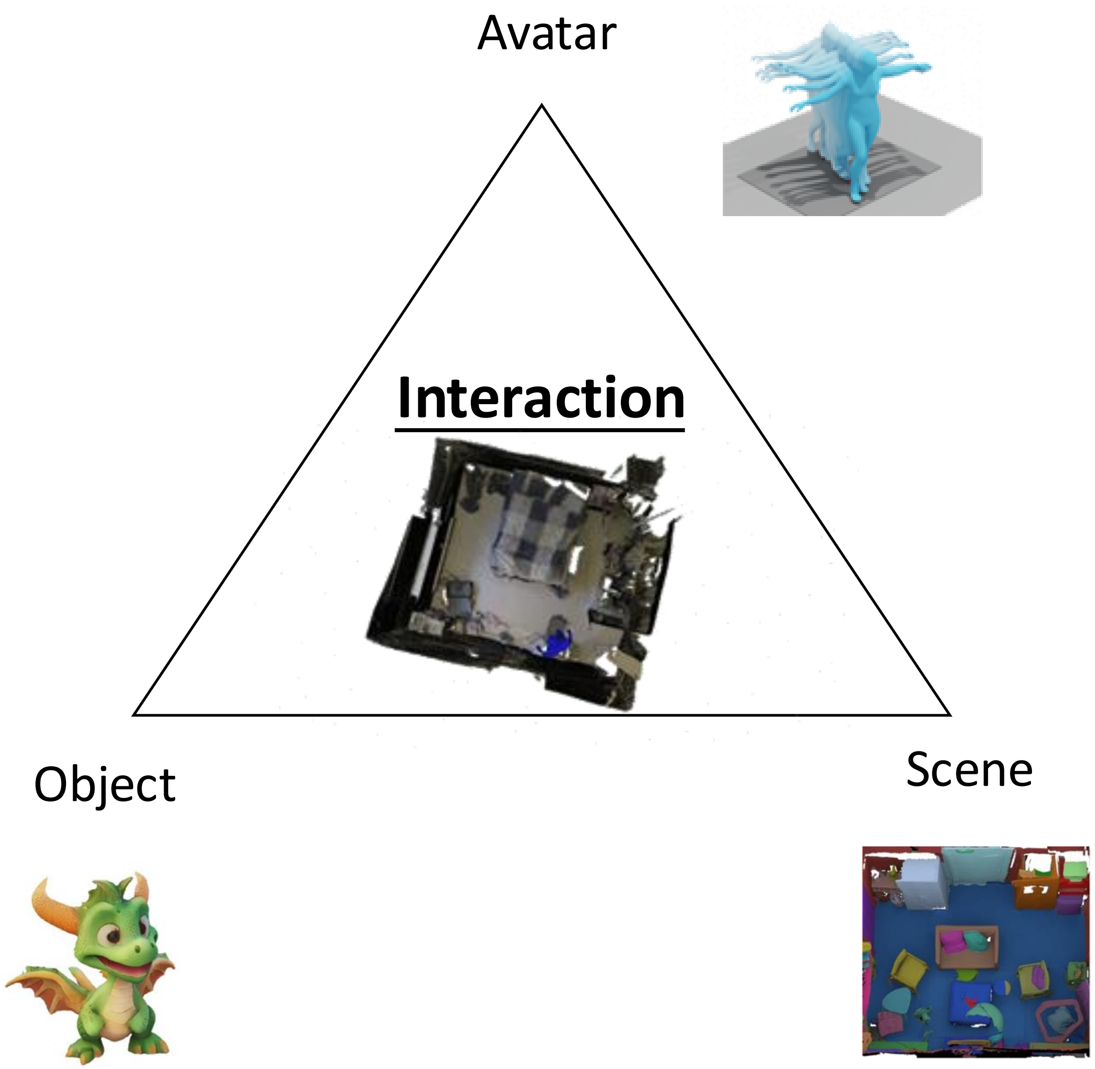
divide-and-conquer



Motion Generated by Our Model

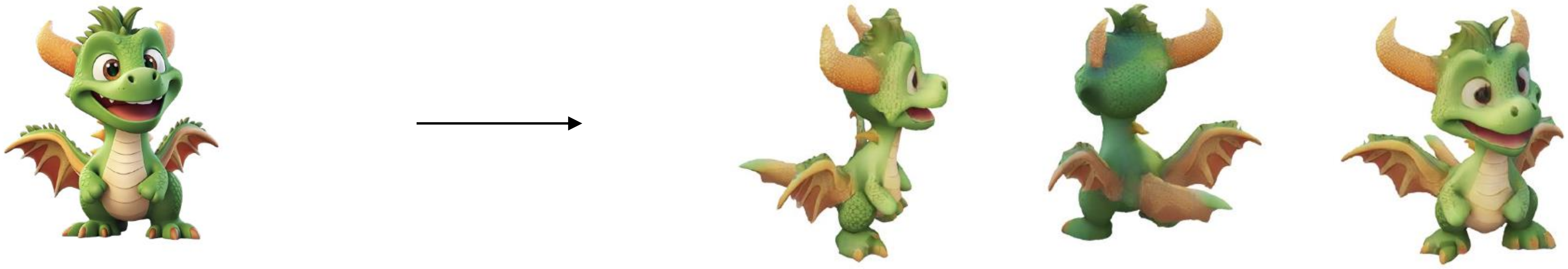
Outline

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



Task Description

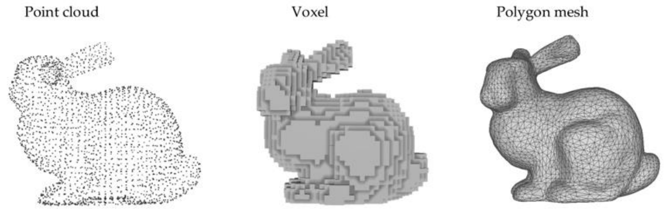
Image to 3D

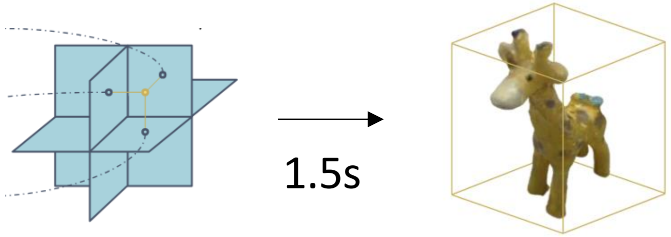


Motivation

□ 3D Representation

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

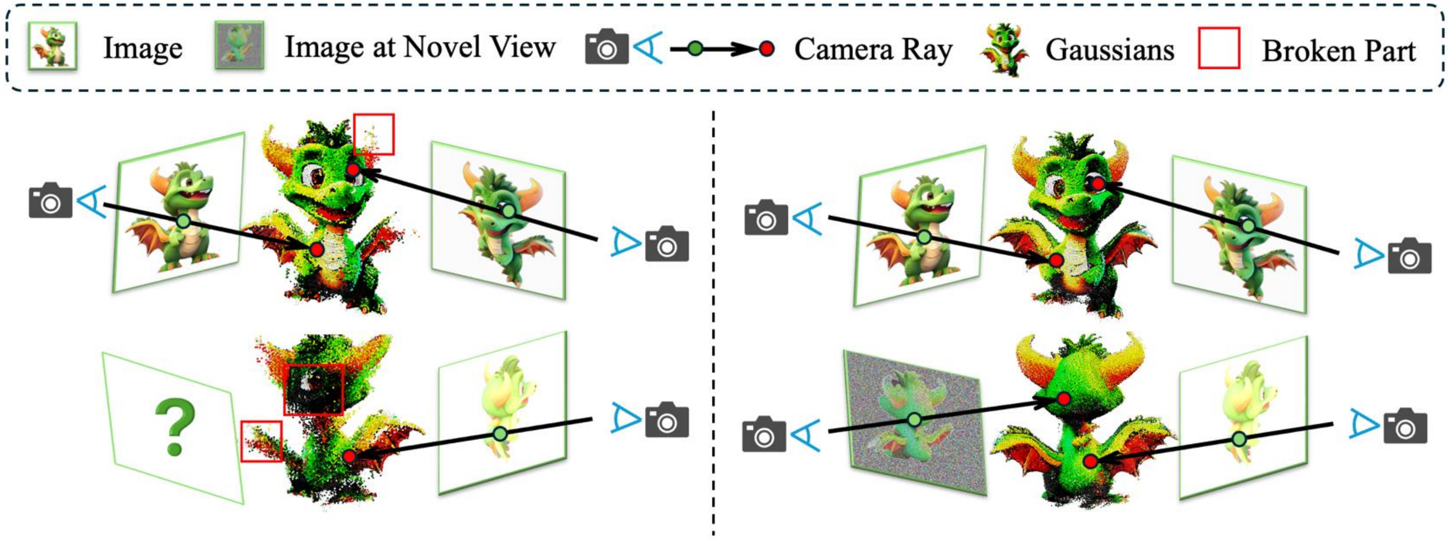




Rendering 2s (60 frames) video cost **1.5min!**

□ Formulation

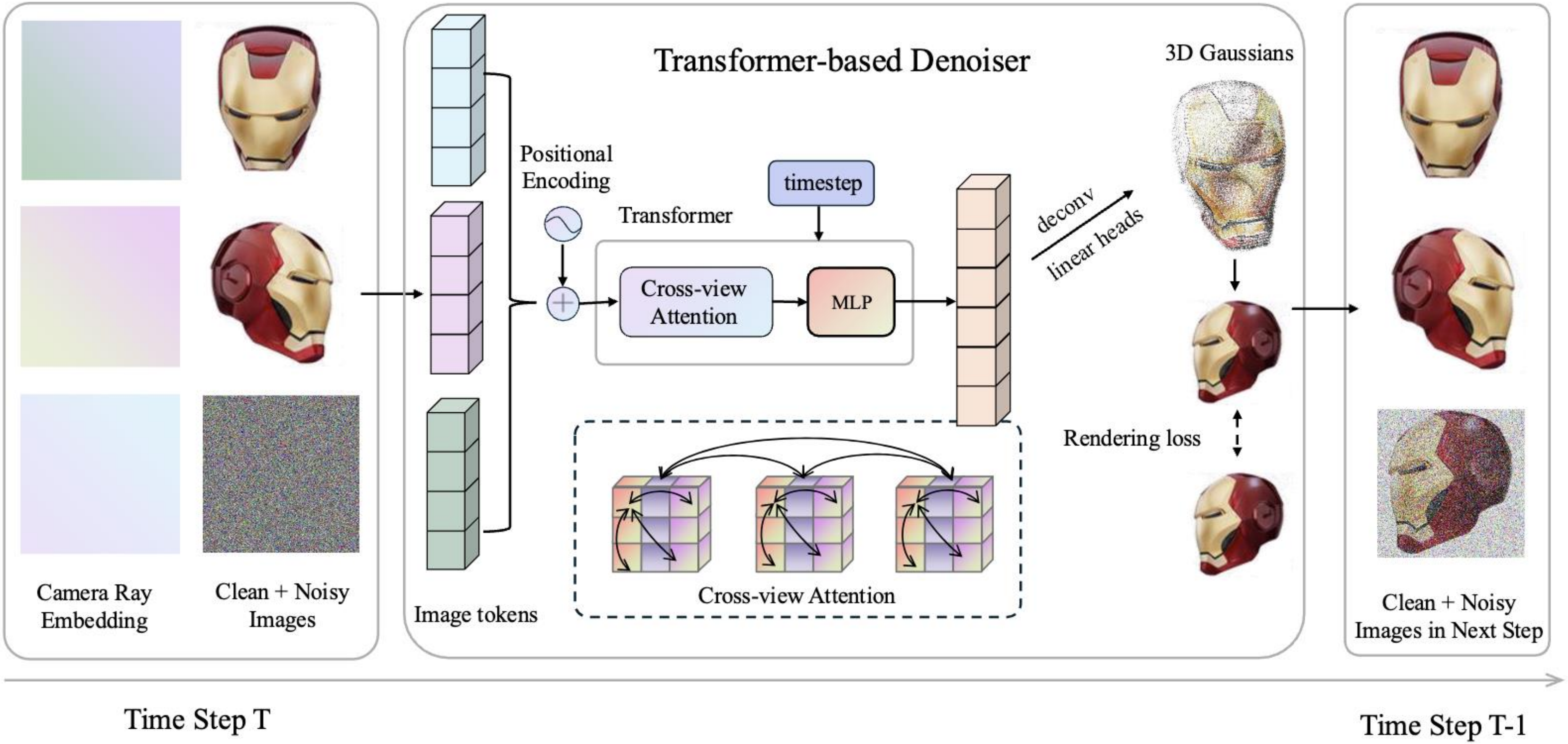
- Pixel-aligned Gaussian



(a) Pixel-aligned Gaussians Reconstruction

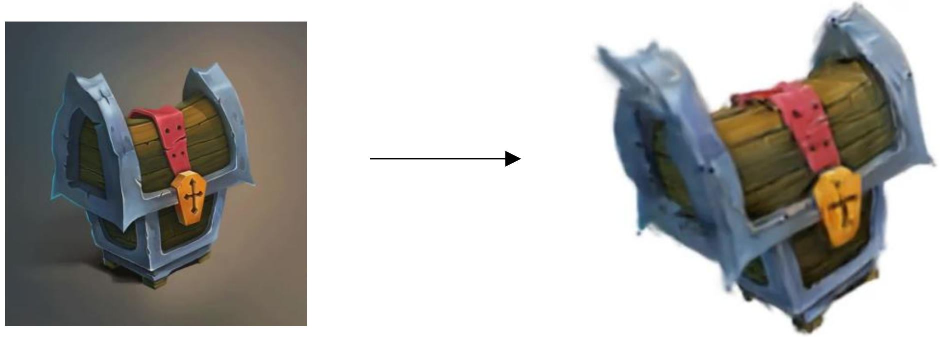
(b) NovelGS (Ours)

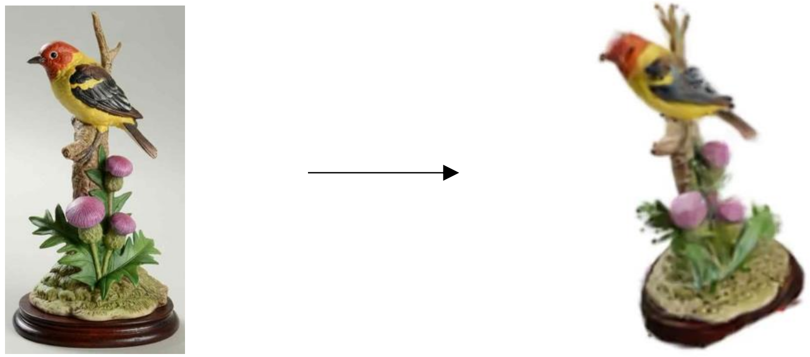
Pipeline

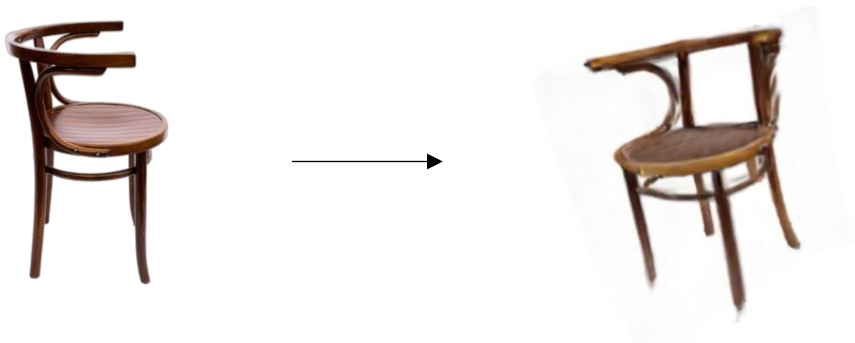

Camera Ray Embedding
Clean + Noisy Images
Image tokens
Positional Encoding
Transformer-based Denoiser
Transformer
timestep
Cross-view Attention
MLP
Cross-view Attention
deconv
linear heads
3D Gaussians
Rendering loss
Clean + Noisy Images in Next Step
Time Step T
Time Step T-1

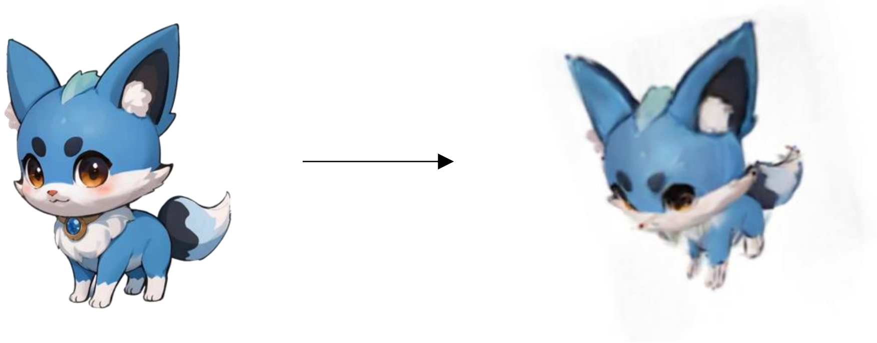
Results

Visualization









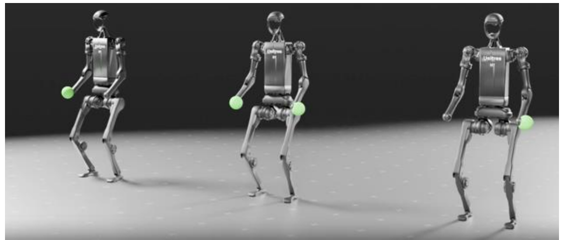
Outline

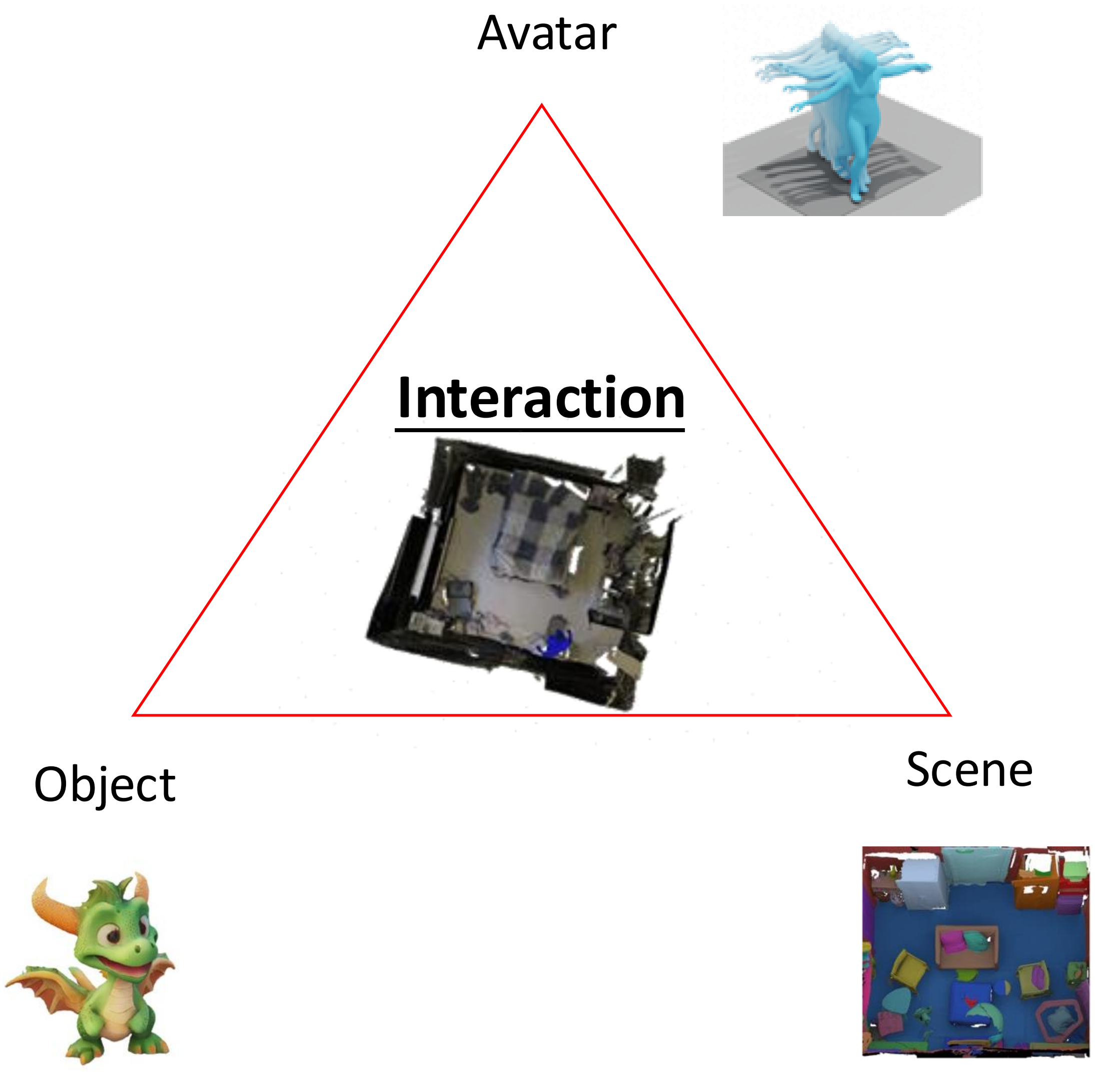
□ Research Projects

- ■ Avatar: Controllable & Generable
- ■ **Object: Efficient & Diverse**

□ Future Research Proposal

- ■ Think deeper about "avatar & object"
  - ■ Avatar-object-scene interaction





Avatar

Interaction

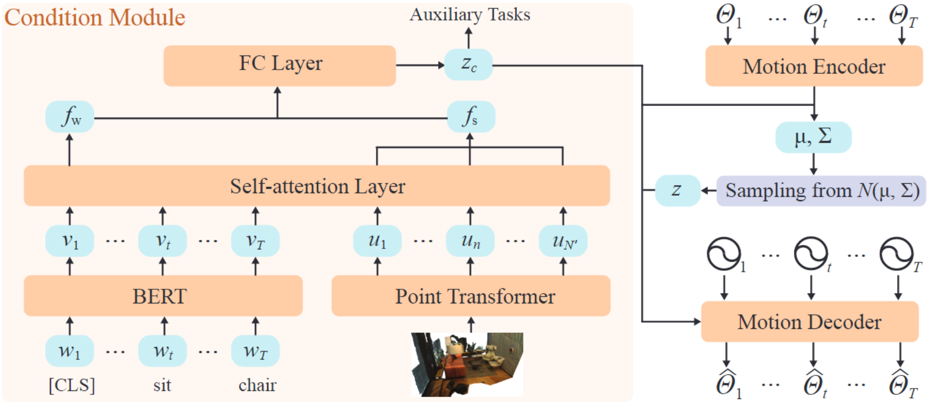
Object

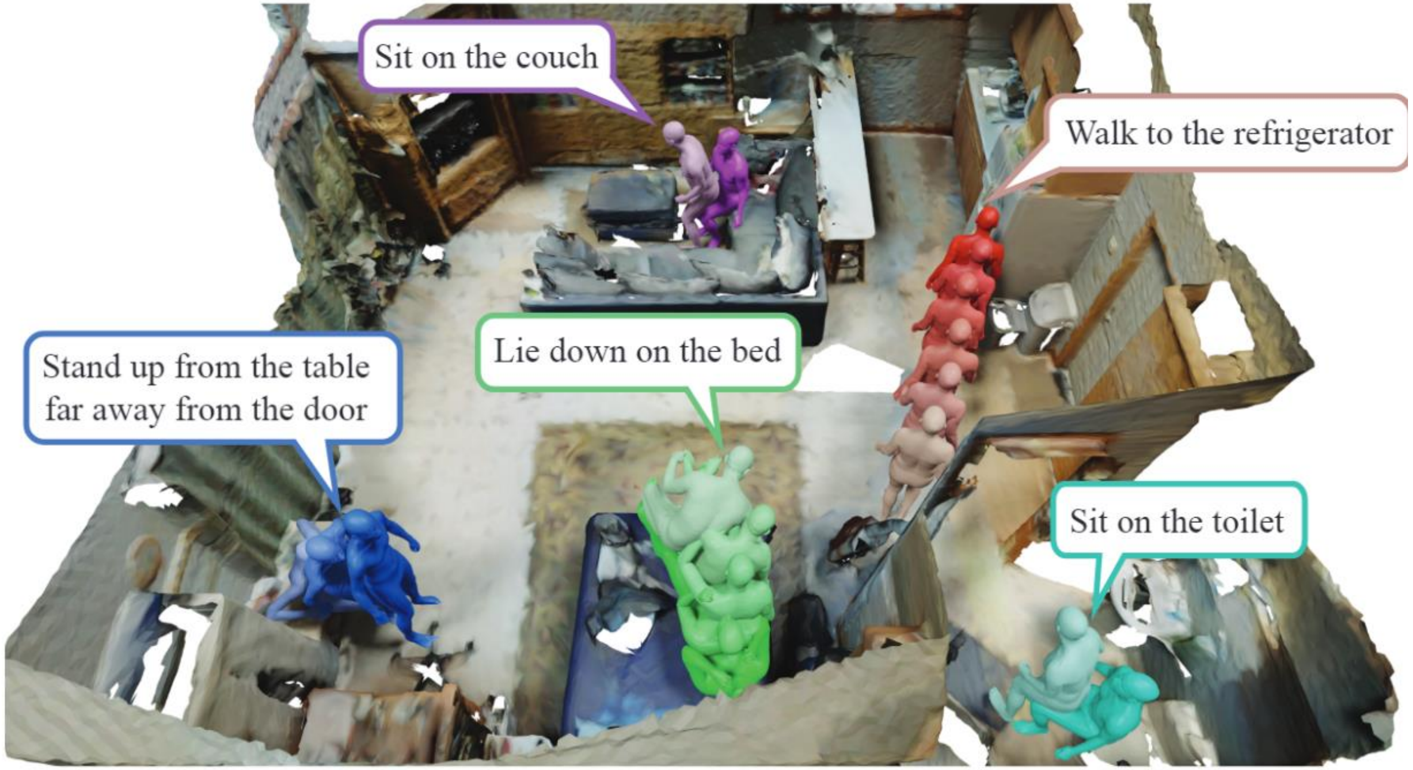
Scene

Task Description

Input: Language & Scene

Output: Interaction

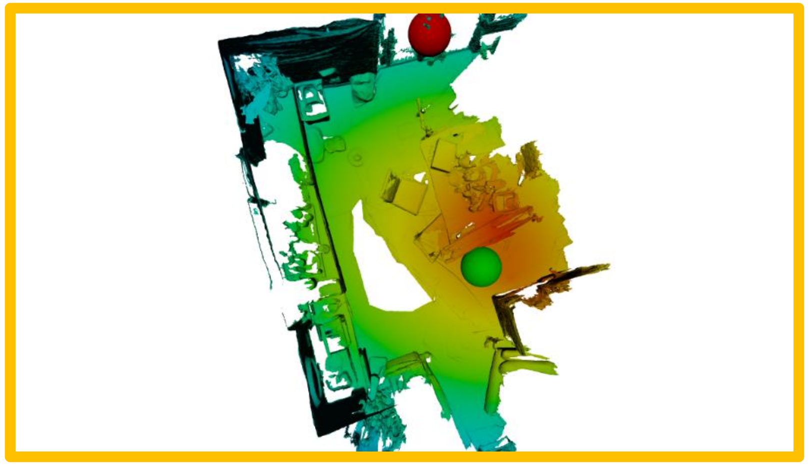


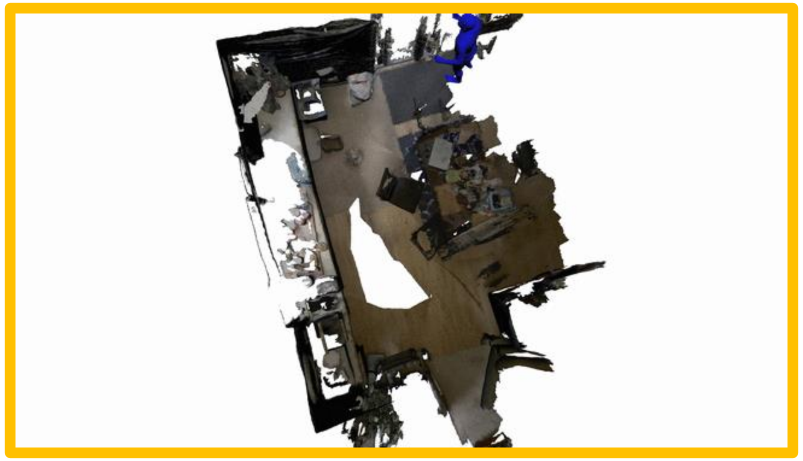


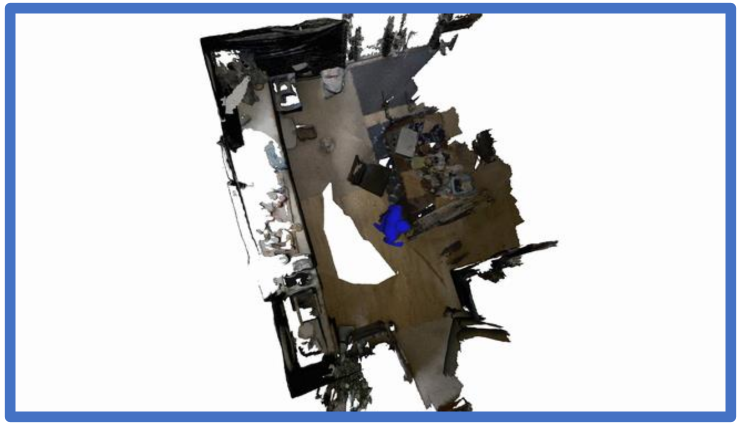
HUMANISE: Language-conditioned Human Motion Generation in 3D Scenes. NeurIPS2023

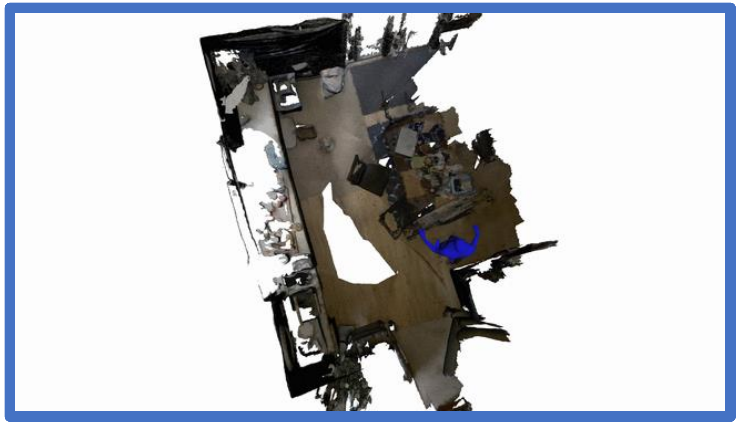
Localization Error

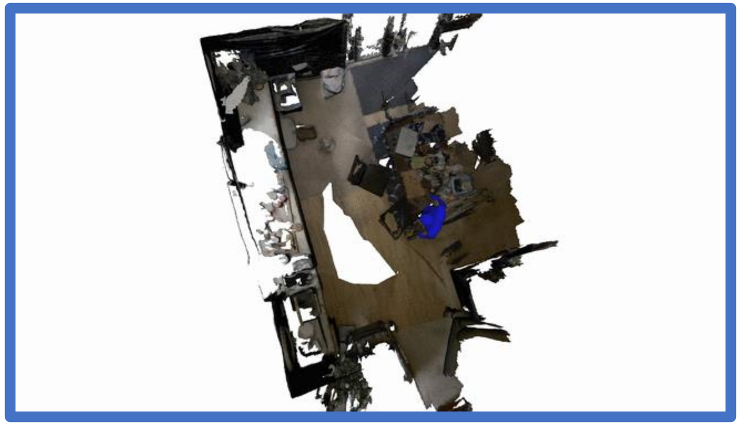
Walk to the glass doors





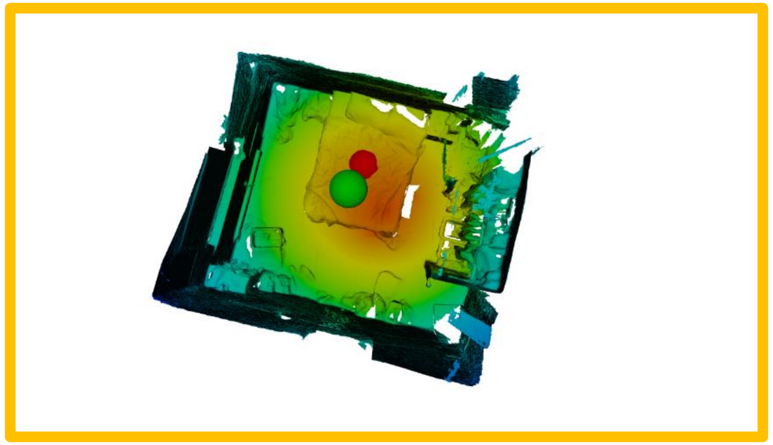


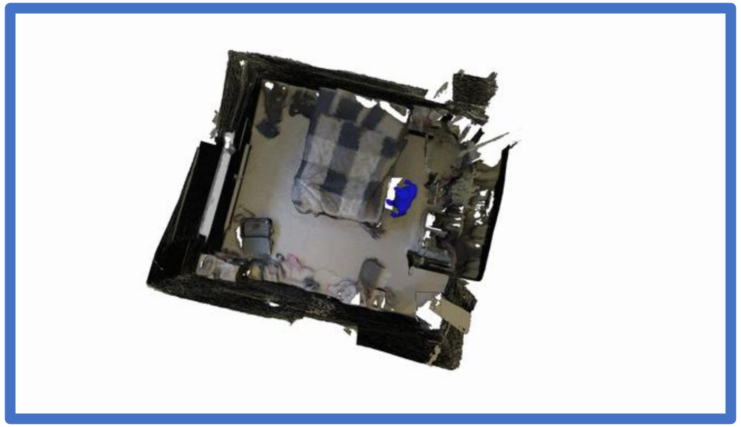


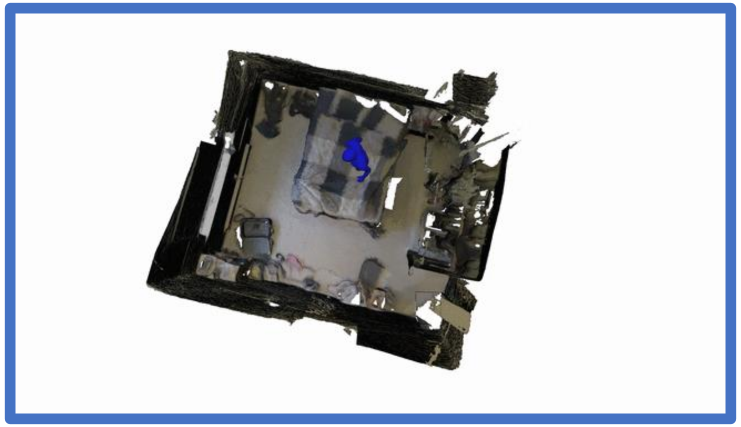


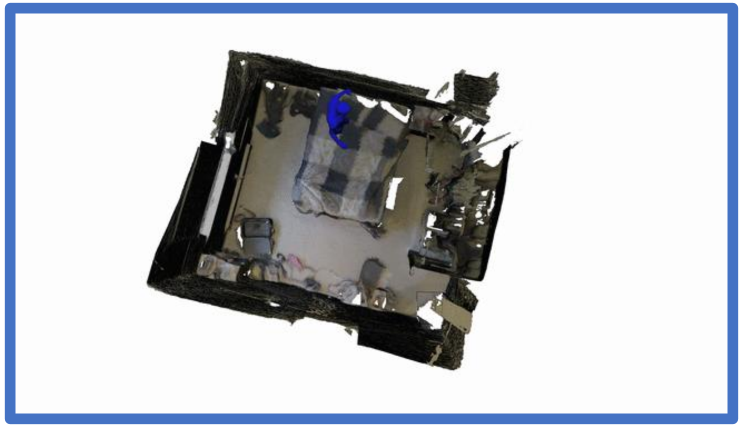
Physical Error

Walk to the bed









Potential Solutions

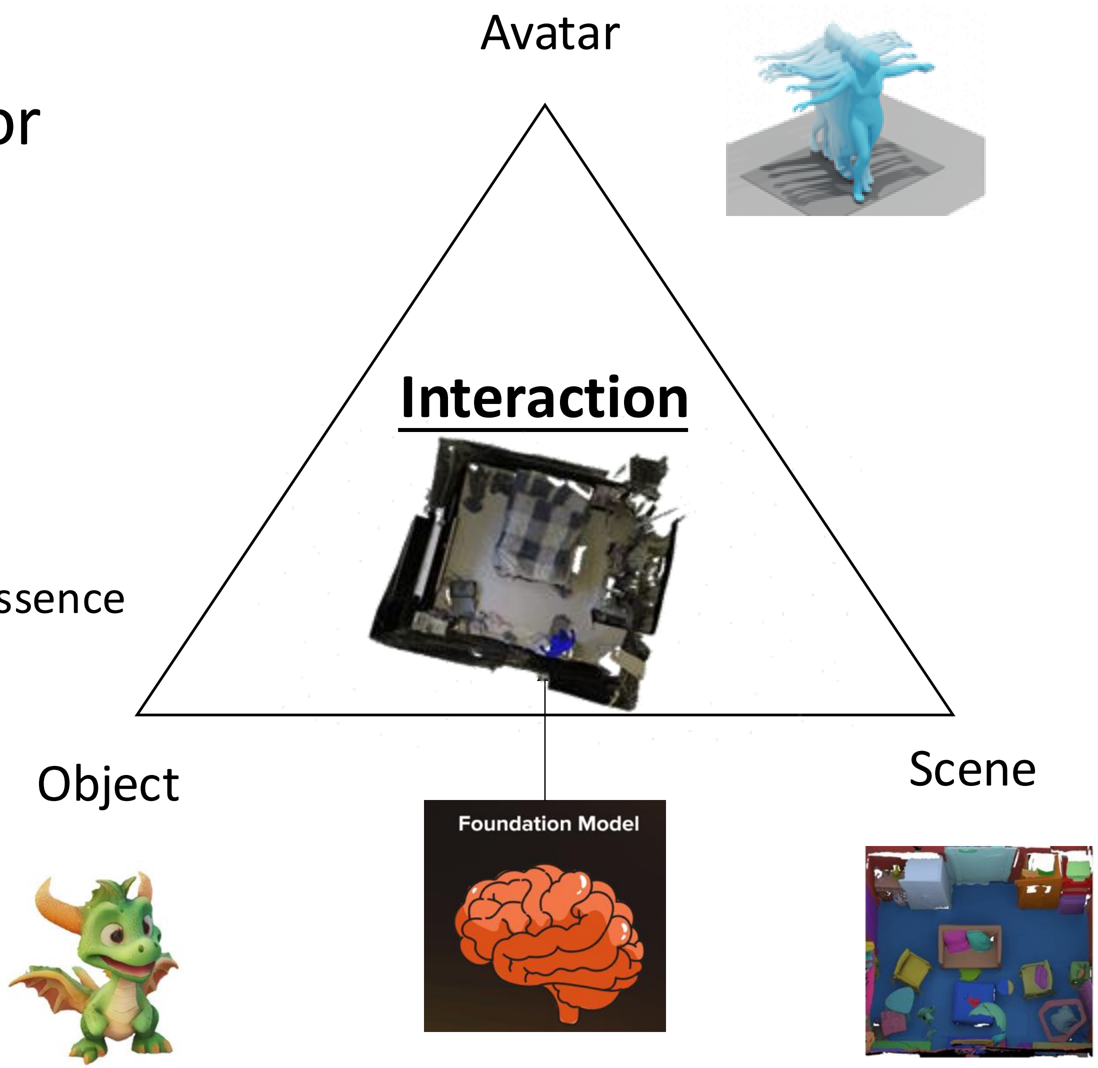
- Physical Error & Localization Error

Physics-based Optimization[1] *More Data* Early Fusion[2]

Difficult to design and inflexible Hard to require It's not the essence

We need the foundation model !

Physical Knowledge & Object Properties

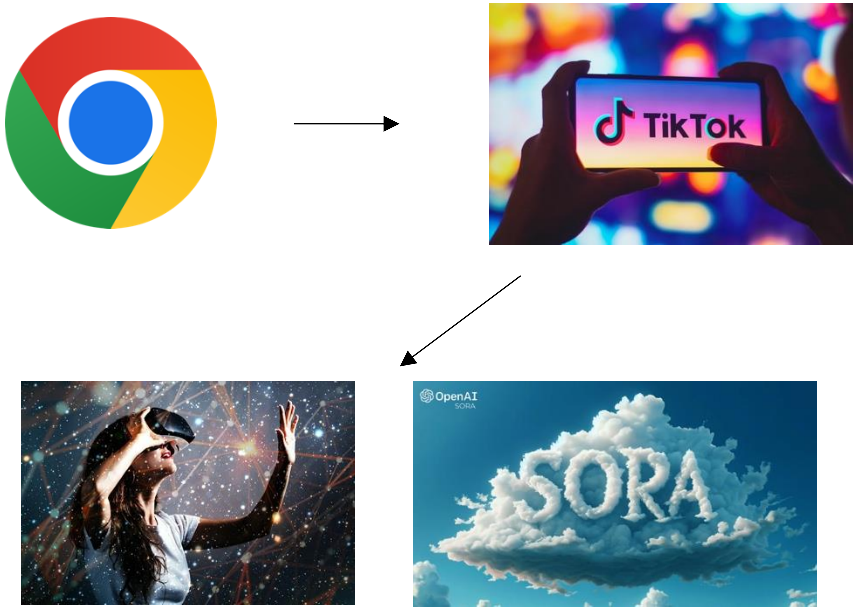


[1] Diffusion-based Generation, Optimization, and Planning in 3D Scenes. CVPR23
[2] LAVT: Language-Aware Vision Transformer for Referring Image Segmentation. CVPR22

Some Thoughts

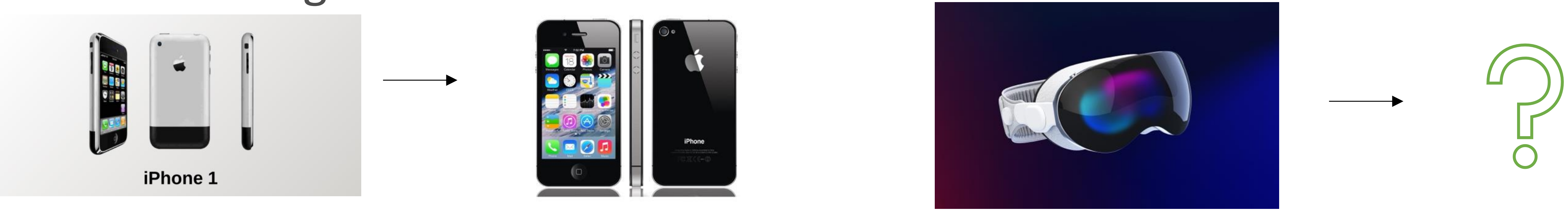
□ Thoughts

- 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?
- Aim high. Settled down.


iPhone 1
iPhone
?