Language-guided Human Motion Generation in 3D Scenes

梁玮

@ Beijing Institute of Technology (BIT)

Outline

I. Overview

II. Language-guided Human Motion Generation in 3D Scenes

Overview — Motivation

We humans can interact with 3D scenes following instructions.

Overview — Motivation

Education

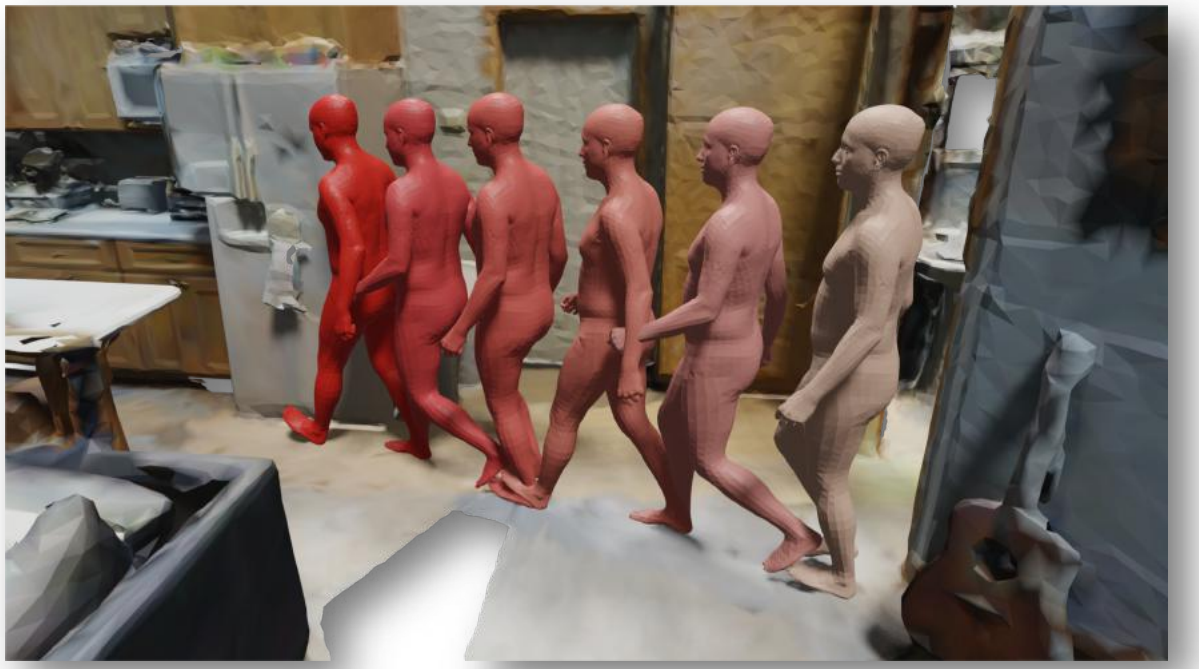
Entertainment

Socialization

Customer Service

Overview — Motivation

Imagine instructing a virtual human to "*walk to the refrigerator*" in a given 3D scene.



A plausible human motion semantically consistent with the given language instruction.

The goal is to generate ***physically plausible*** and ***semantically consistent*** human motions in 3D scenes.

Overview — Motivation

Action *“Jump”*

Text *“kick left leg”*

Object 

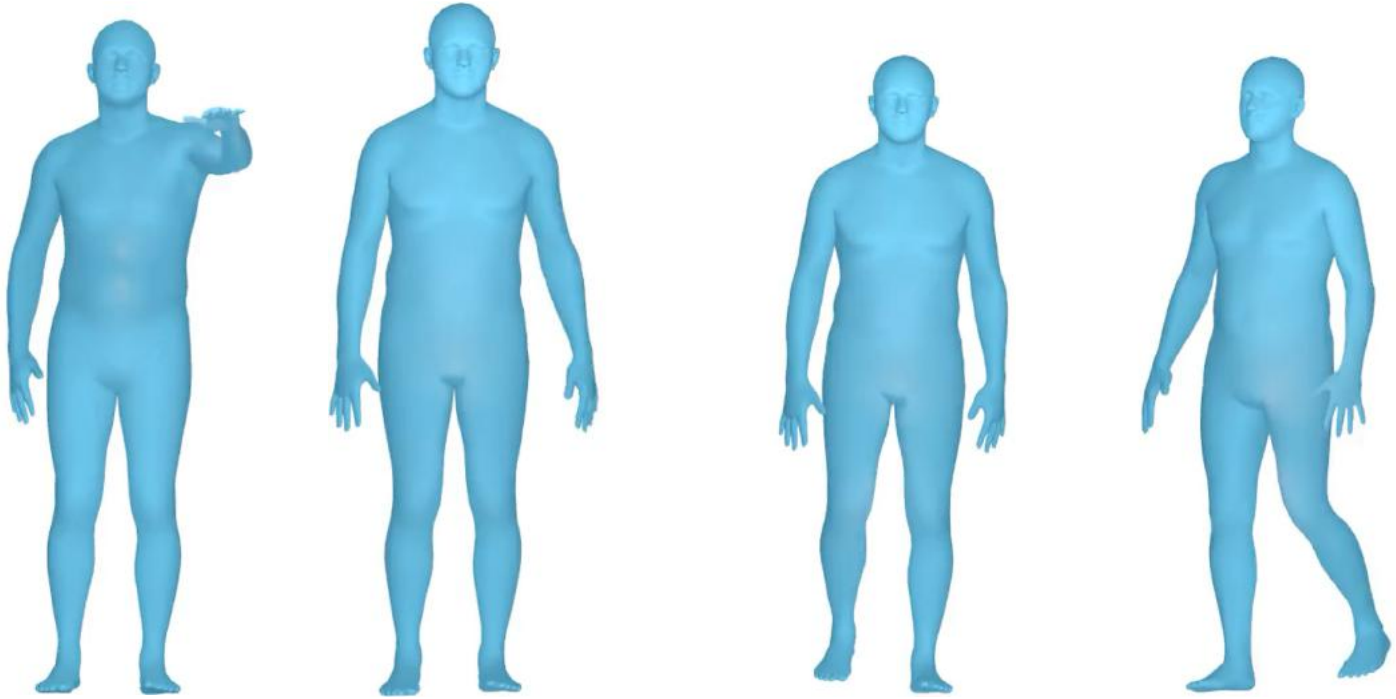
Scene 

Multi-conditional

Overview — Related Work

- **Conditional human motion generation: Action-to-Motion**

Action *“Jump”*

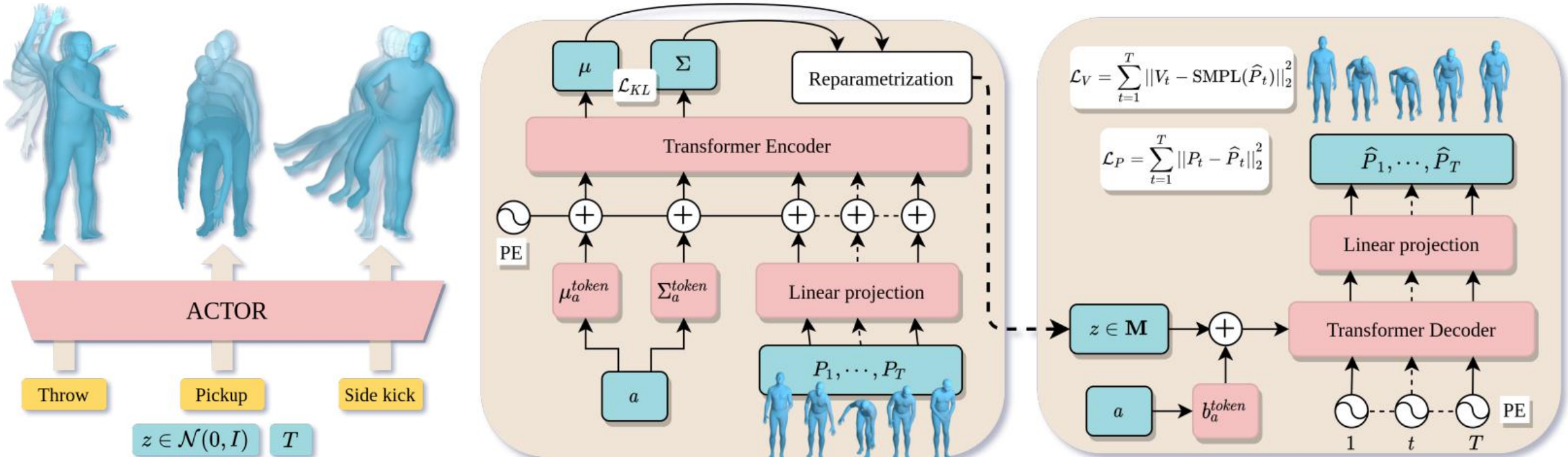


raising arms salute high knee running knee raising

The goal is to generate human motion sequences accurately aligned with the given **action labels**.

Overview — Related Work

- **Action-to-Motion**



Action-Conditioned 3D Human Motion Synthesis with Transformer VAE, ICCV 2021

Overview — Related Work

- **Conditional human motion generation: Text-to-Motion**

Action *"Jump"*

Text *"kick left leg"*

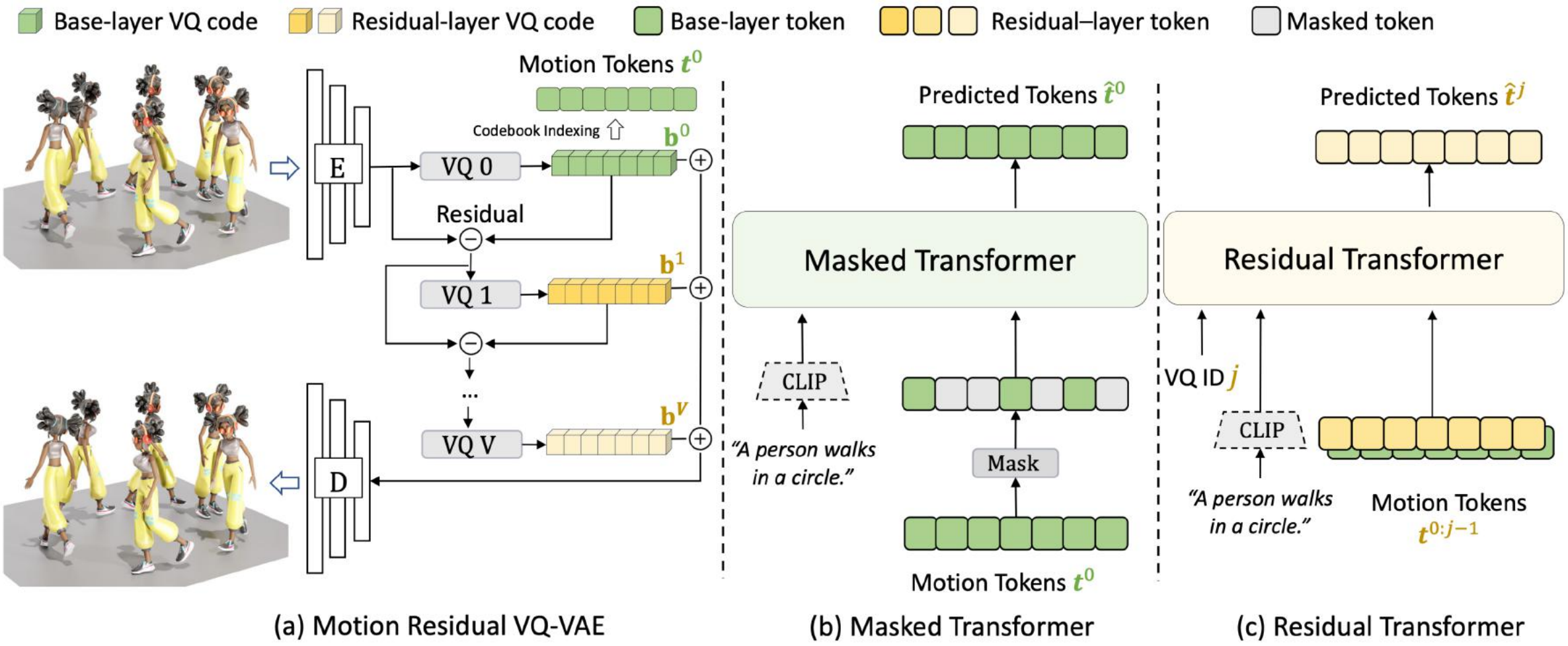
"A person walks forward, bends down to pick something up off the ground."

"A person is skipping rope."

The goal is to generate realistic and contextually appropriate human motions based on **textual descriptions**.

Overview — Related Work

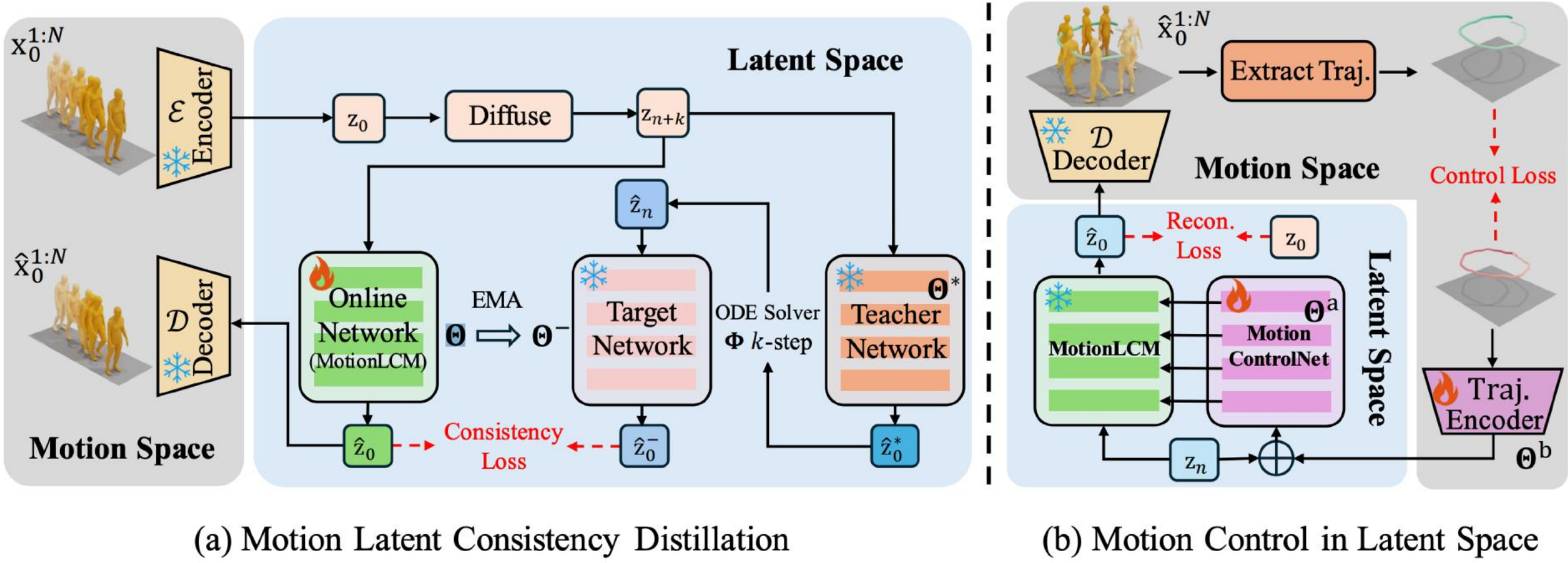
- **Text-to-Motion**



MoMask: Generative Masked Modeling of 3D Human Motions, CVPR 2024

Overview — Related Work

- **Text-to-Motion**



(a) Motion Latent Consistency Distillation

(b) Motion Control in Latent Space

MotionLCM: Real-time Controllable Motion Generation via Latent Consistency Model, ECCV 2024

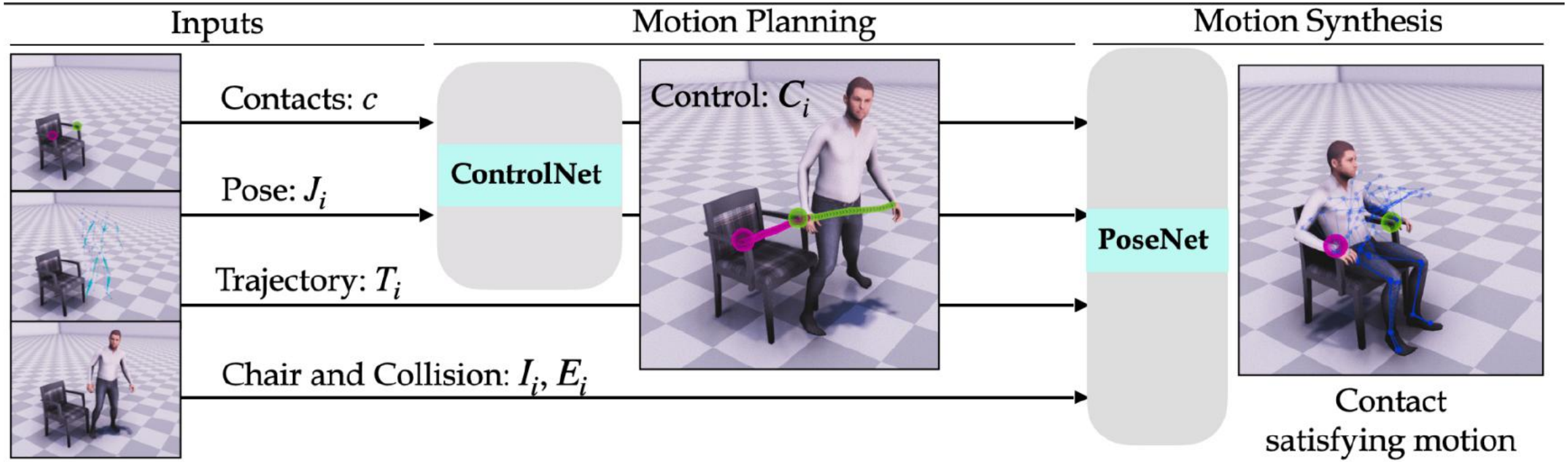
Overview — Related Work

- **Conditional human motion generation: Object-to-Motion**

The goal is to generate realistic human motions tailored to interactions with **specified objects** in 3D scenes.

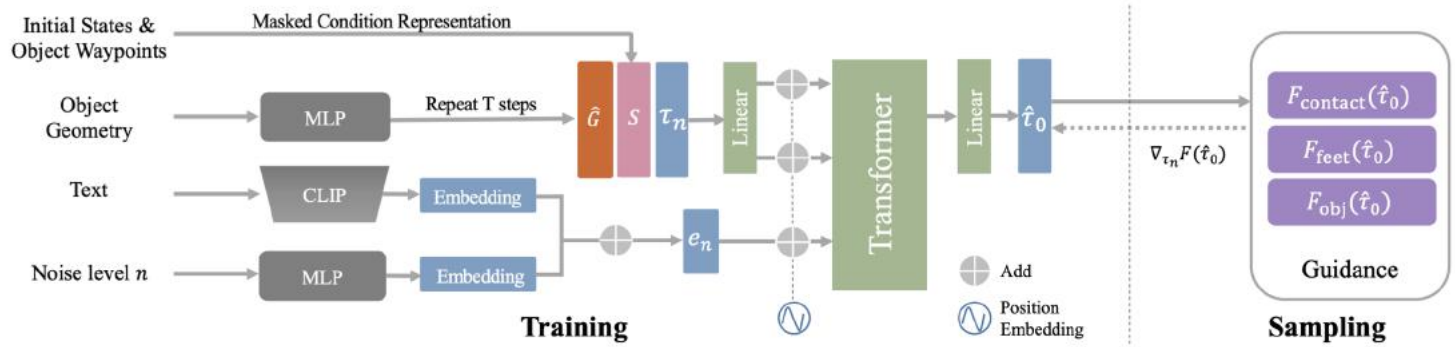
Overview — Related Work

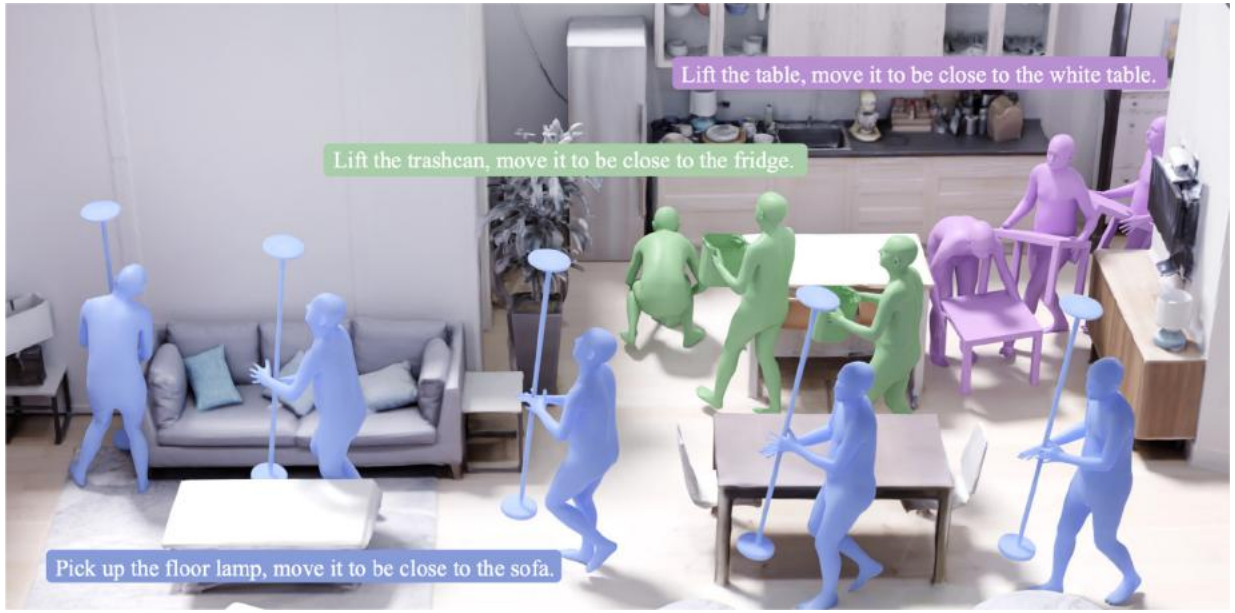
- **Object-to-Motion**



Overview — Related Work

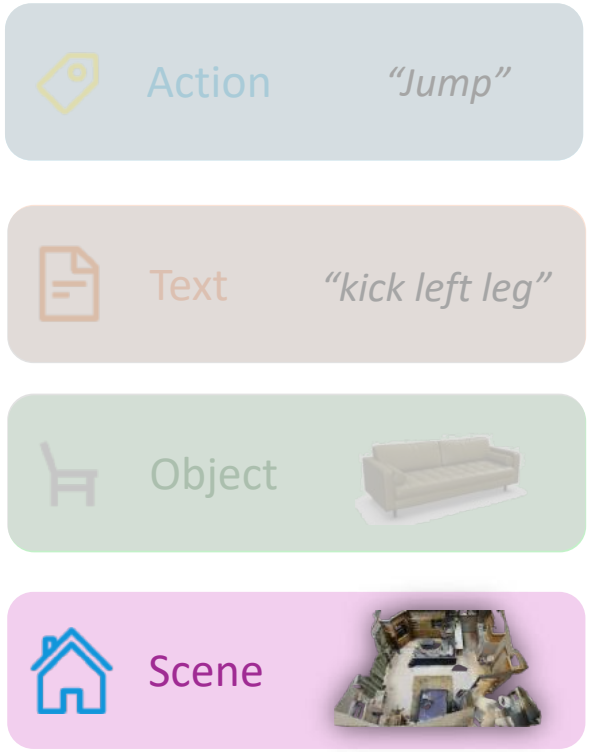
- **Object-to-Motion**





Overview — Related Work

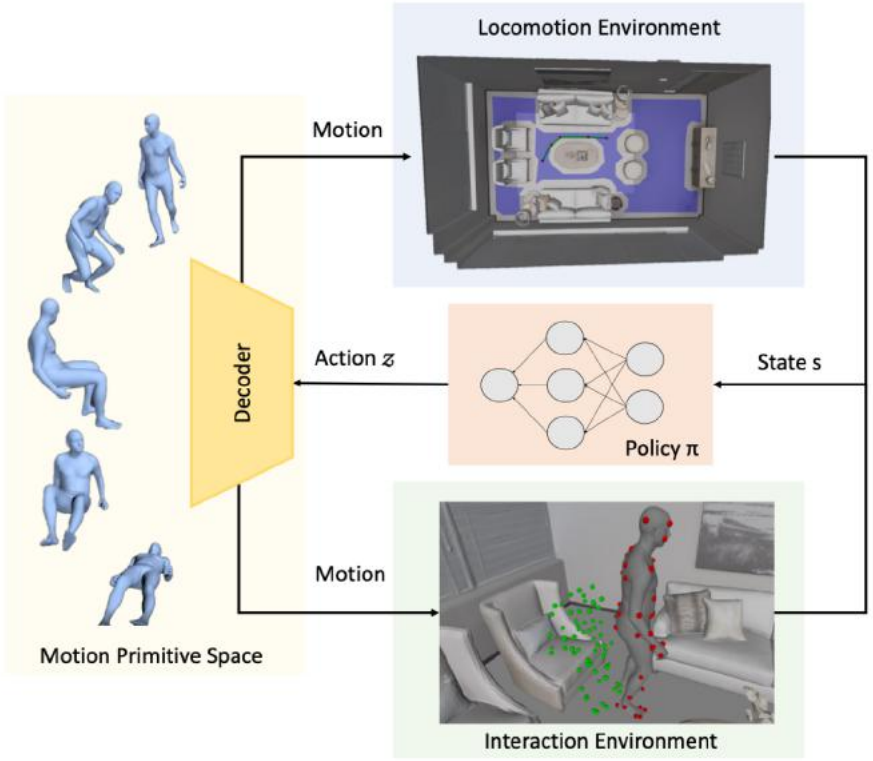
- **Conditional human motion generation: Scene-to-Motion**



The goal is to generate physically plausible human motions that interact with **3D scenes**.

Overview — Related Work

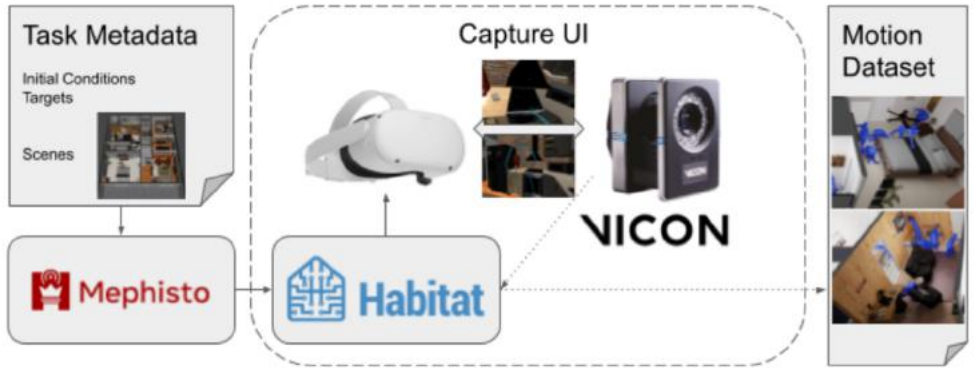
- **Scene-to-Motion**

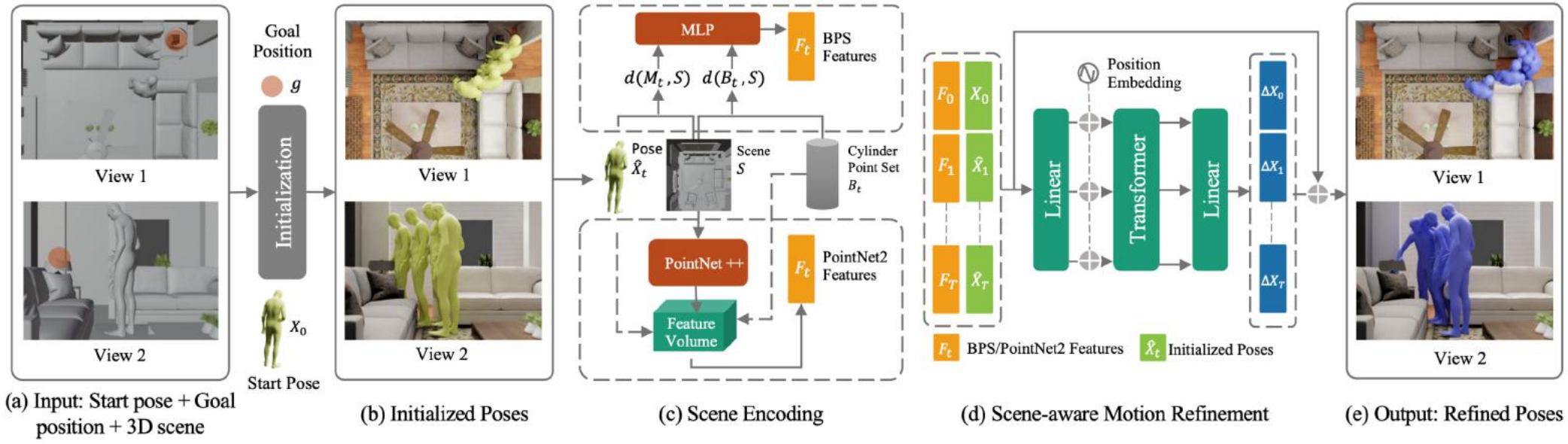


Synthesizing Diverse Human Motions in 3D Indoor Scenes, ICCV 2023

Overview — Related Work

- **Scene-to-Motion**





Overview — Related Work

- **Conditional human motion generation: Multi-conditional**

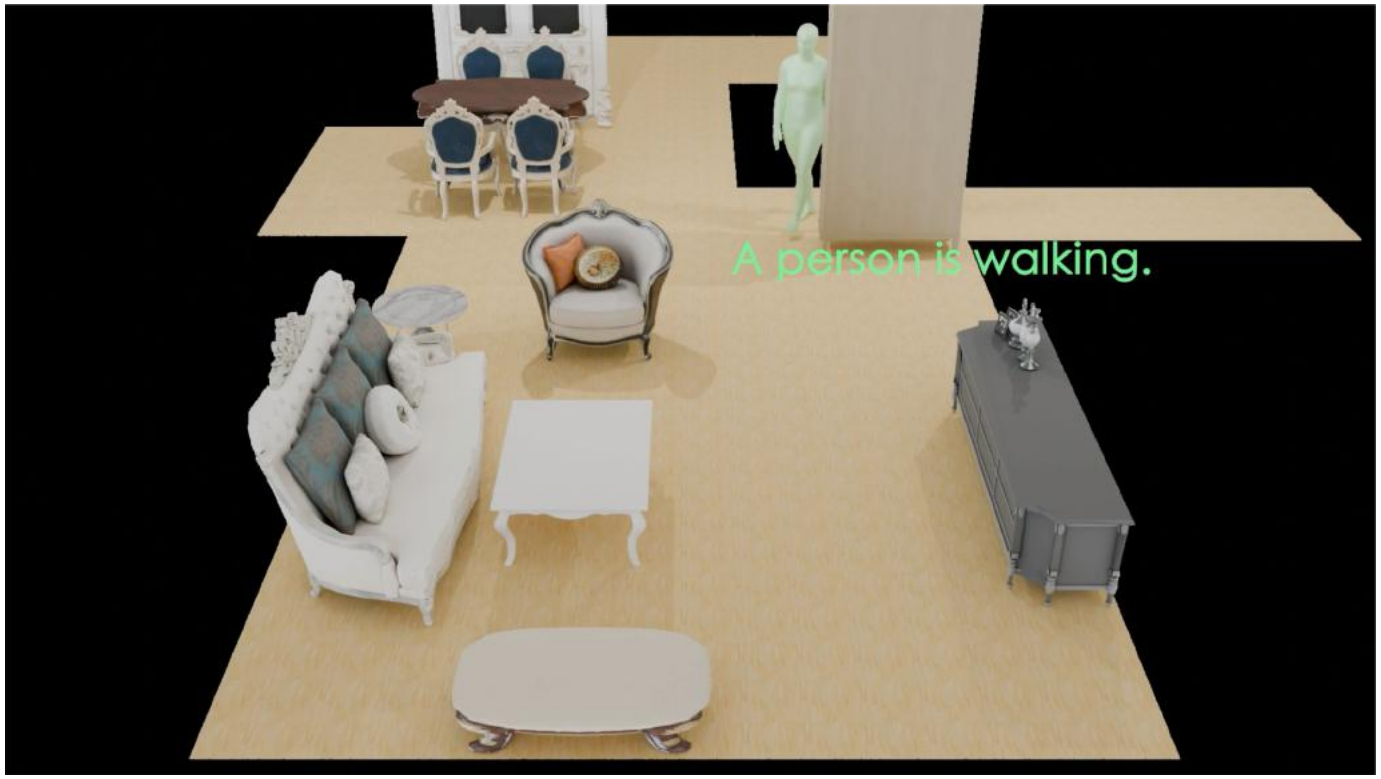
Action "Jump"

Text "kick left leg"

Object 

Scene 

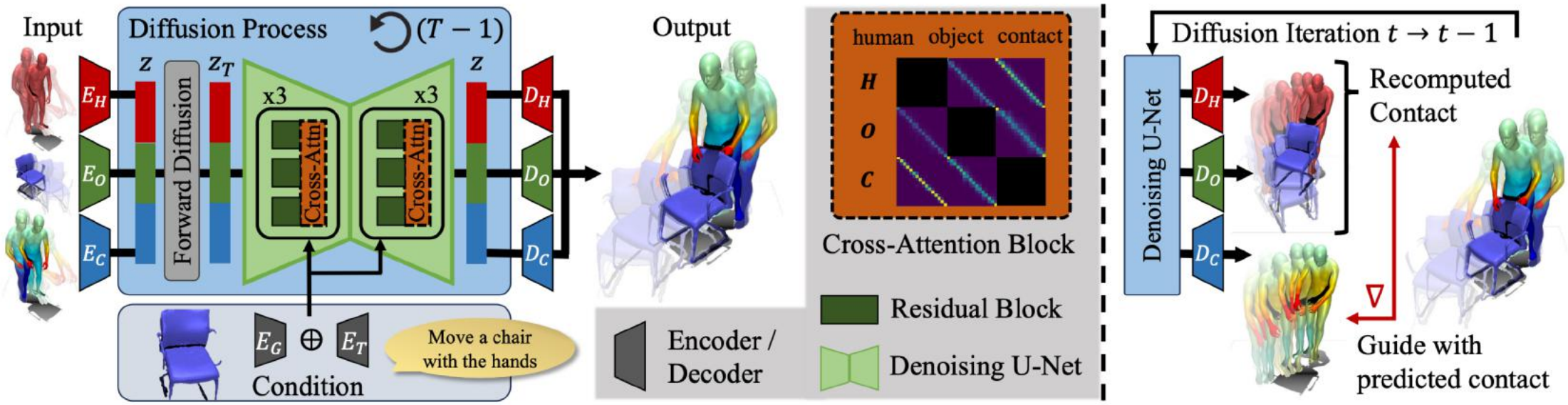
Multi-conditional



The goal is to generate human motions by integrating **multiple conditions**, such as text, object, and goal position, to produce coherent and context-aware movements.

Overview — Related Work

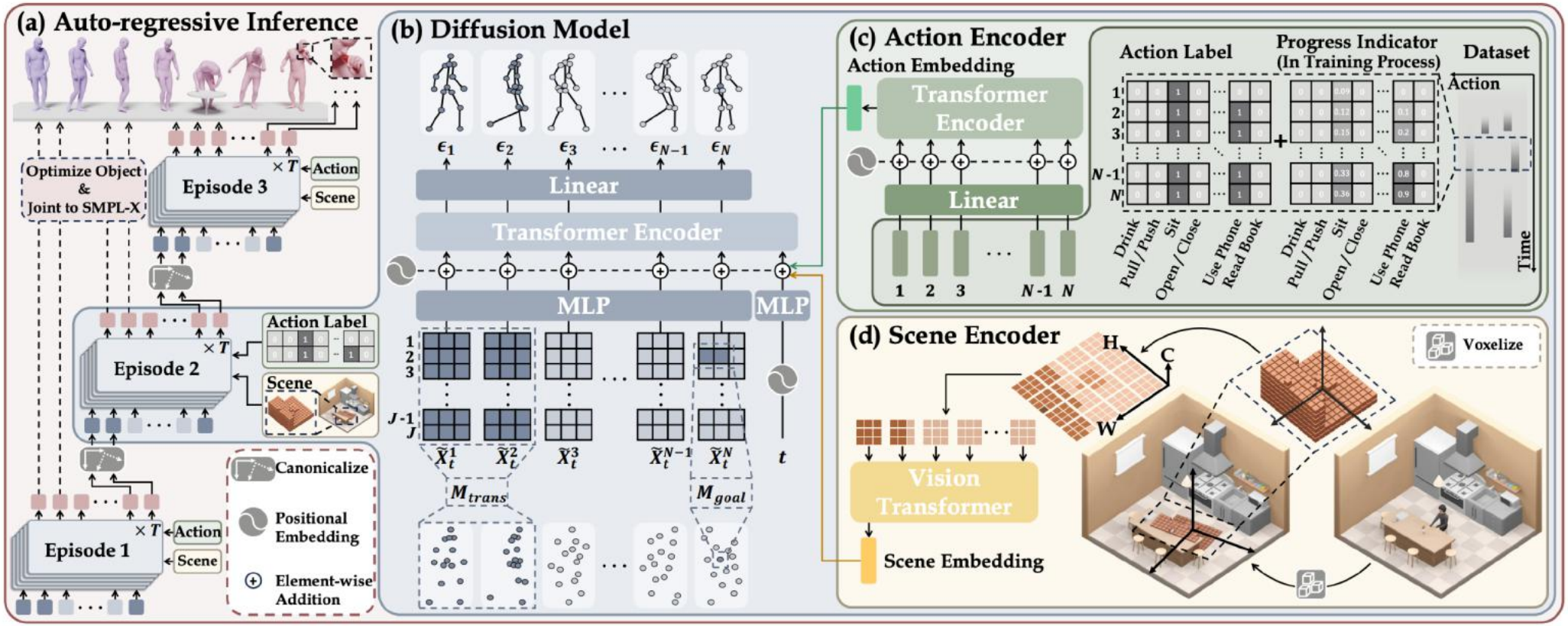
- **Multi-conditional Motion Generation**



Text + Object

Overview — Related Work

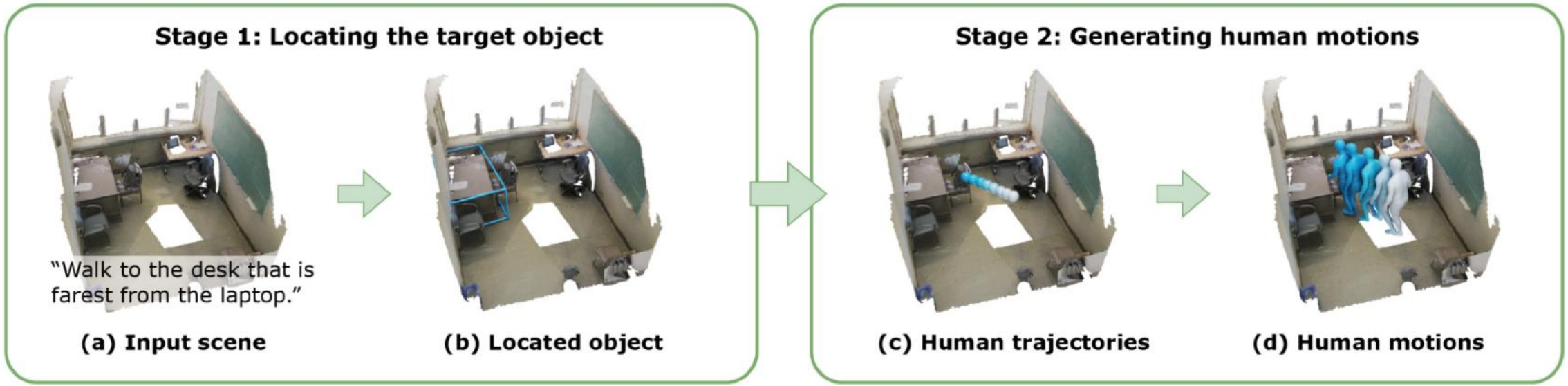
- **Multi-conditional Motion Generation**



Scene + Action

Overview — Related Work

- **Multi-conditional Motion Generation**



Scene + Text

Overview — Related Work

- **Conditional human motion generation**

Text *"kick left leg"*

Action *"Jump"*

Object 

Scene 

Multi-conditional

Regression Model

GAN
VAE
Diffusion
Normalizing Flow

Physics-based RL Model

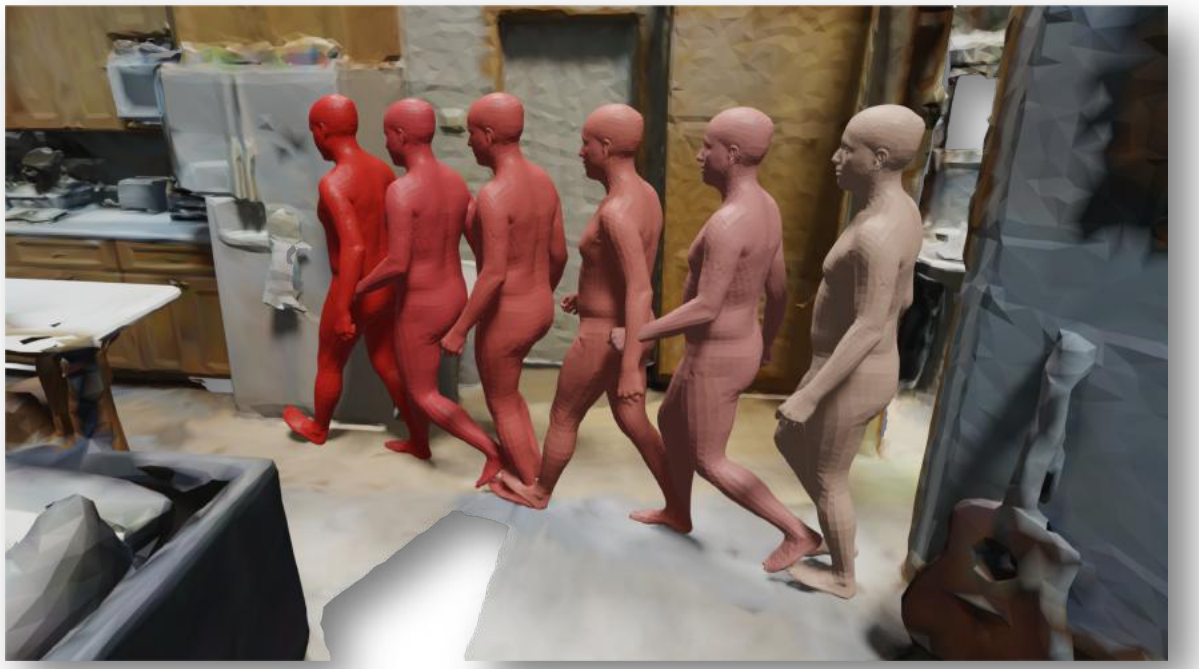
Outline

I. Overview

II. Language-guided Human Motion Generation in 3D Scenes

Recall

Imagine instructing a virtual human to "*walk to the refrigerator*" in a given 3D scene.



A plausible human motion semantically consistent with the given language instruction.

The goal is to generate ***physically plausible*** and ***semantically consistent*** human motions in 3D scenes.

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

NeurIPS 2022

HUMANISE — Motivation

Two fundamental limitations:

- Limitated scale and quality
- Absence of scene and language semantics

Mohamed Hassan, et al., CVPR 2019.

Zhe Cao, et al., ECCV 2020.

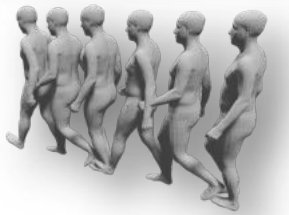
HUMANISE — Dataset

19.6k motion clips
643 3D scenes
1.2M frames
4 action types
Semantics annotations

Motion alignment pipeline

Specify interacting objects

Sample $\mathcal{T}, \mathcal{R}$



Transform motion with $\mathcal{T}, \mathcal{R}$

Collision
Contact
Constraints

Satisfy?

Language description synthesis

< action >< target−class > [< spatial−relation >< anchor−class(es) >].

e.g., sit on the armchair near the desk.

HUMANISE — Dataset

HUMANISE — Task Definition

"walk to the refrigerator"

Input

Generative Model

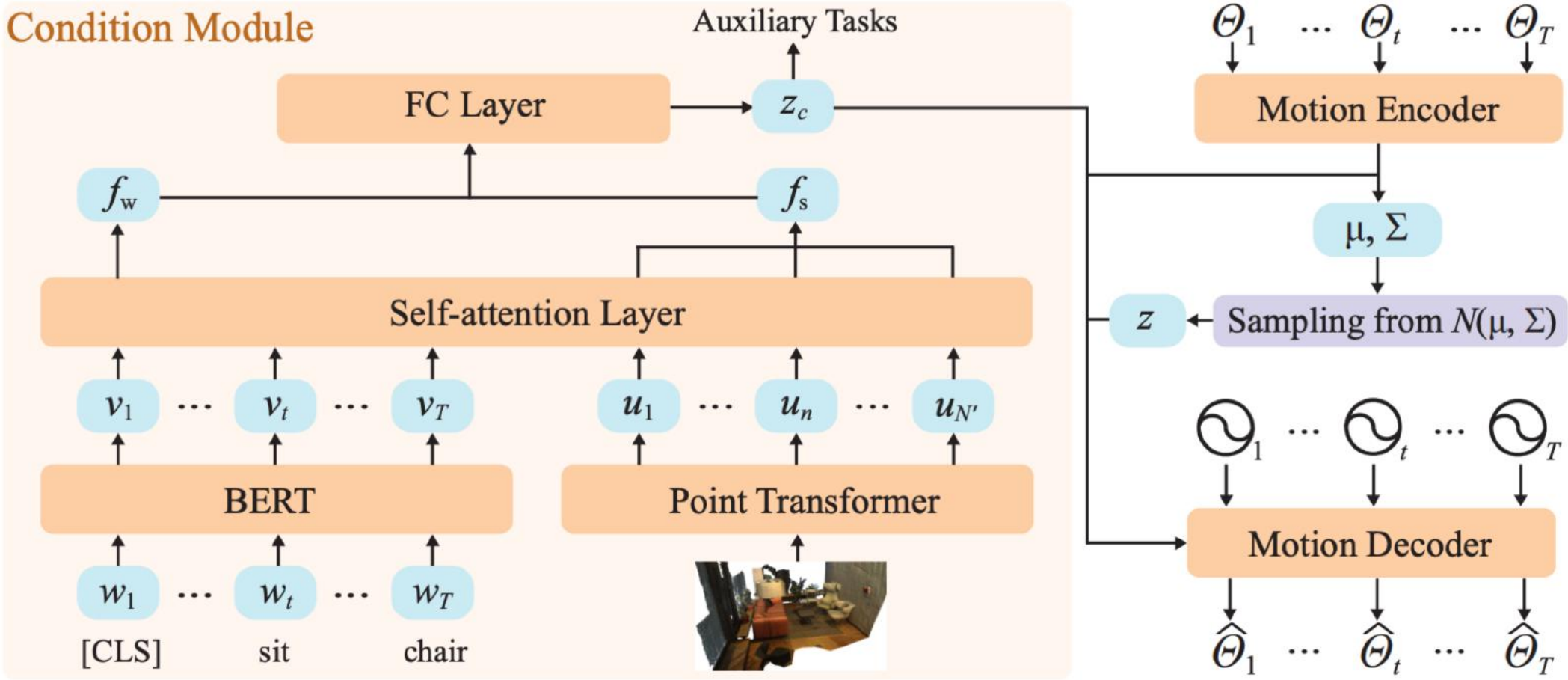
$p_\theta(X \mid S, L)$

Output

Multi-modal conditions: ***3D scene*** **and** ***language description***

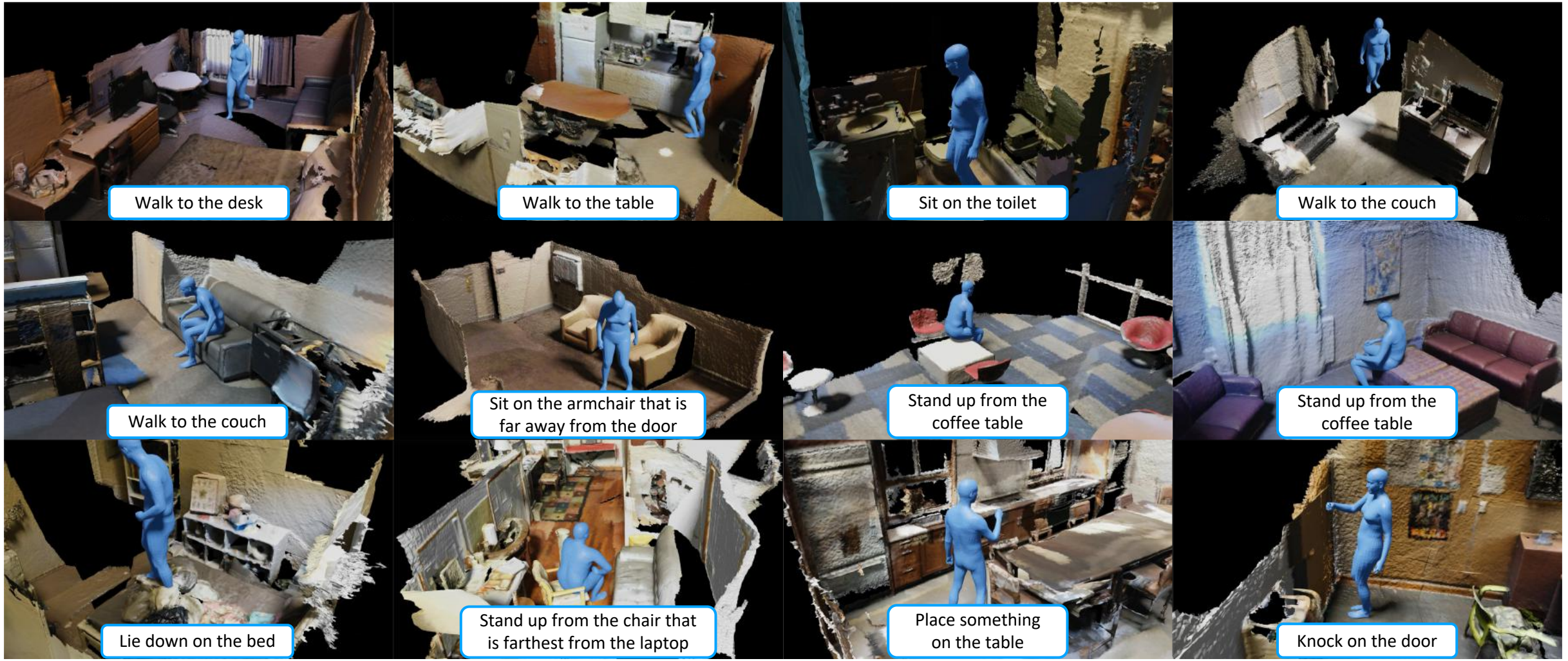
- The generated human motions should perform the correct action and be precisely grounded near the target location according to the language descriptions.
- The generated human motions should be realistic and physically plausible within the 3D scenes.

HUMANISE — Model Framework



HUMANISE — Results

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



HUMANISE — Conclusion

- We propose a **large-scale** and **semantic-rich** synthetic HSI dataset, ***HUMANISE***, that contains human motions aligned with 3D scenes and corresponding language descriptions.

- We introduce a new task of ***language-conditioned human motion generaiton in 3D scenes*** that requires a holistic and joint understanding of 3D scenes, human motions, and language.

- We develop a generative model that can produces **diverse** and **semantically consistent** human motions conditioned on the 3D scene and language description.

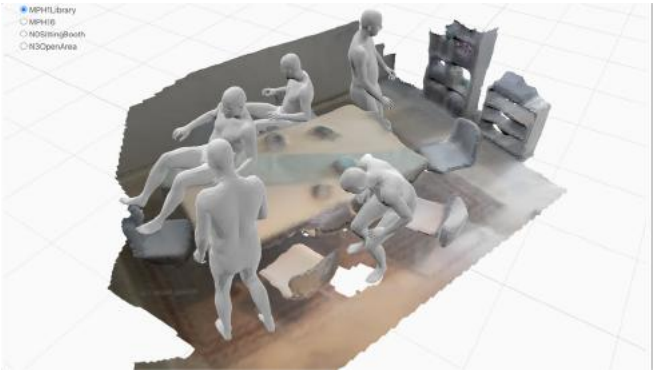
Diffusion-based Generation, Optimization, and Planning in 3D Scenes (CVPR2023)

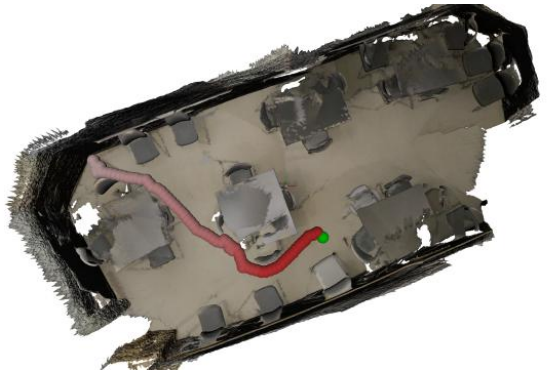
SceneDiffuser *is a conditional generative model for 3D scene understanding.*

It is applicable to various scene-conditioned 3D tasks.

SceneDiffuser — Motivation

- The long-standing goal for 3D scene understanding
  - Generation – Scene-aware
  - Optimization – Physics-based
  - Planning – Goal oritened
- Two fundamental limitations
  - Lack of ***powerful*** generative model
  - Lack of ***unified*** framework

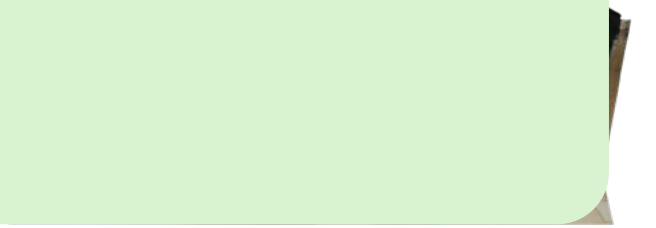




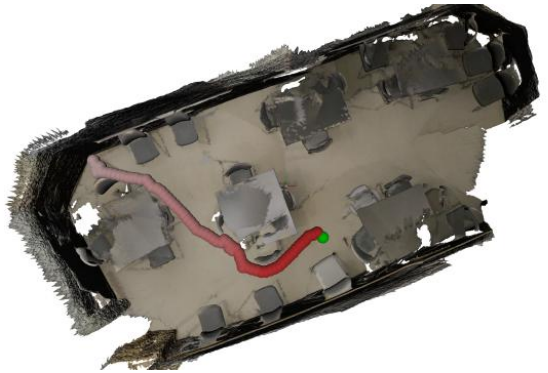
SceneDiffuser — Motivation

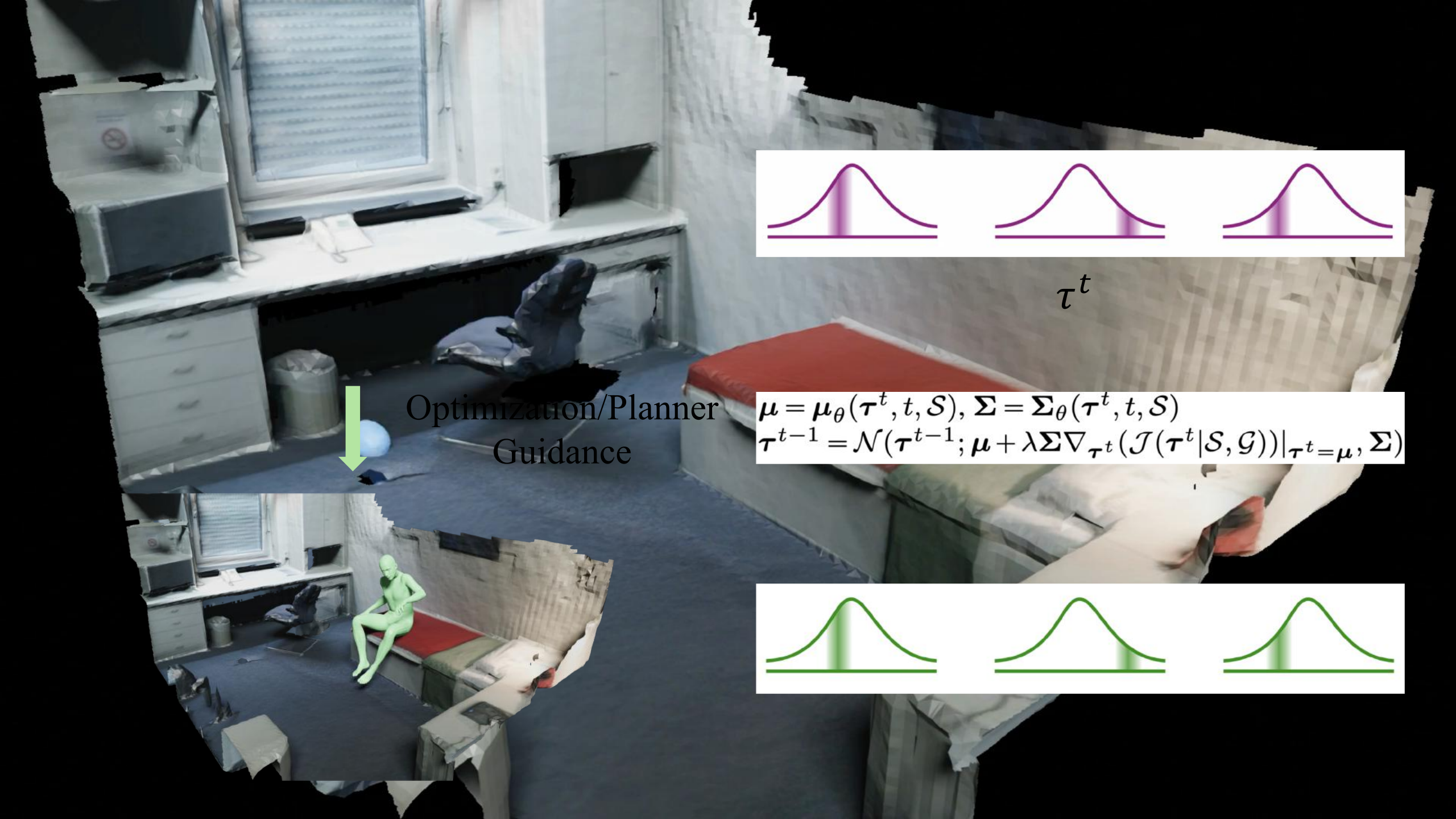
- The long-standing goal for 3D scene understanding
  - Generation of S

SceneDiffuser is the **first** framework that models the *3D scene-conditioned* generation with a diffusion model and integrates the *generation*, *optimization*, and *planning* into a **unified** framework.

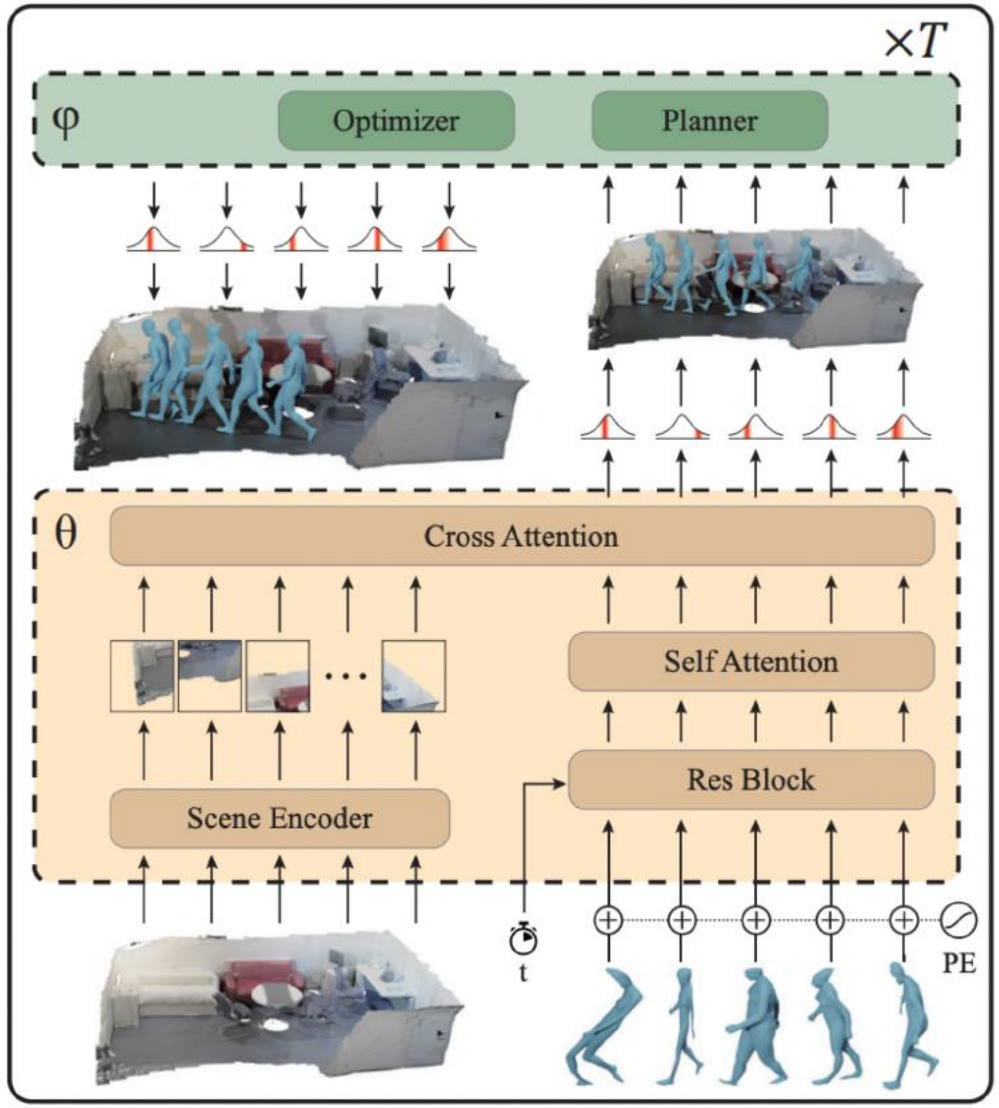


  - Lack of ***unified*** framework

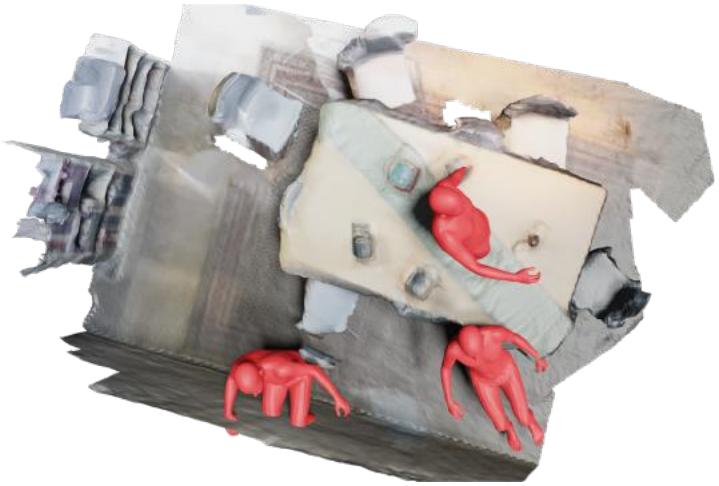



Optimization/Planner
Guidance
τ^t
$\mu = \mu_\theta(\tau^t, t, \mathcal{S}), \Sigma = \Sigma_\theta(\tau^t, t, \mathcal{S})$
$\tau^{t-1} = \mathcal{N}(\tau^{t-1}; \mu + \lambda \Sigma \nabla_{\tau^t}(\mathcal{J}(\tau^t | \mathcal{S}, \mathcal{G}))|_{\tau^t = \mu}, \Sigma)$

SceneDiffuser — Architecture



SceneDiffuser — Human Pose Generation

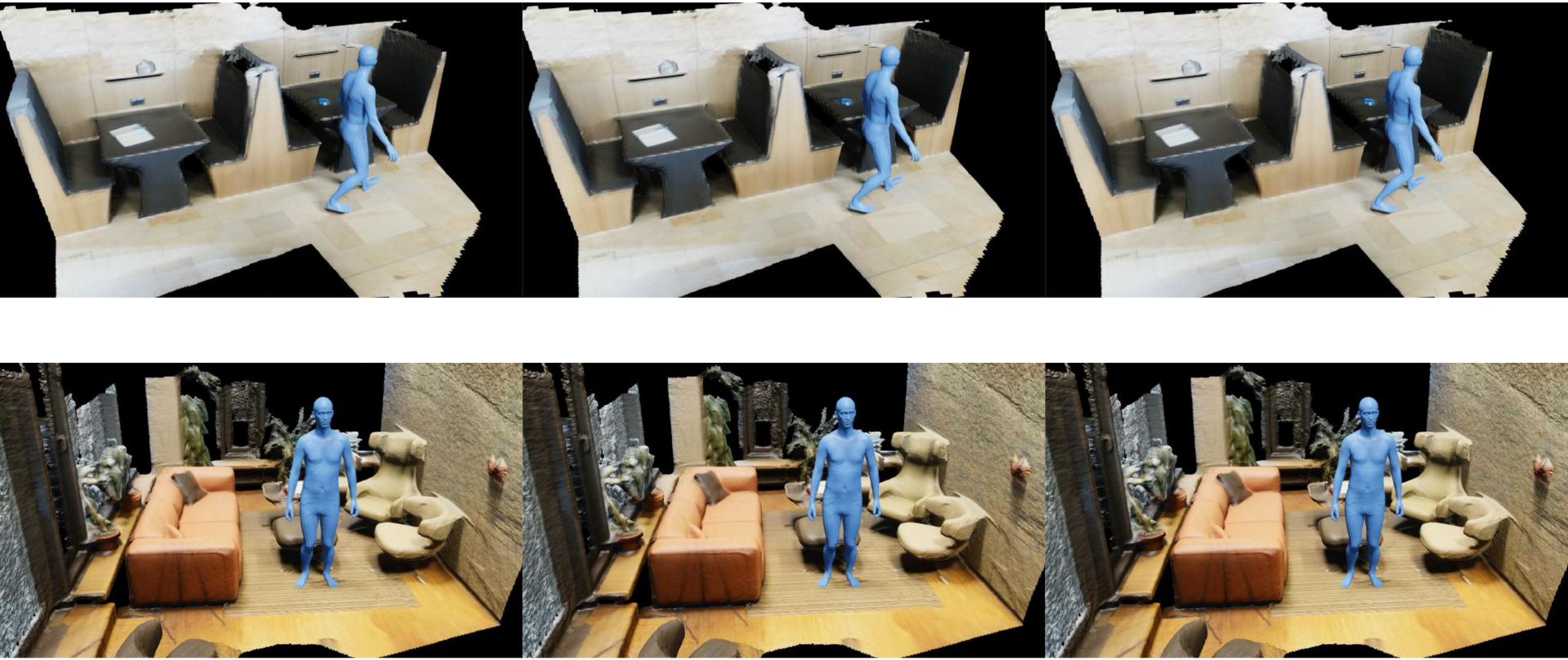


cVAE (baseline)

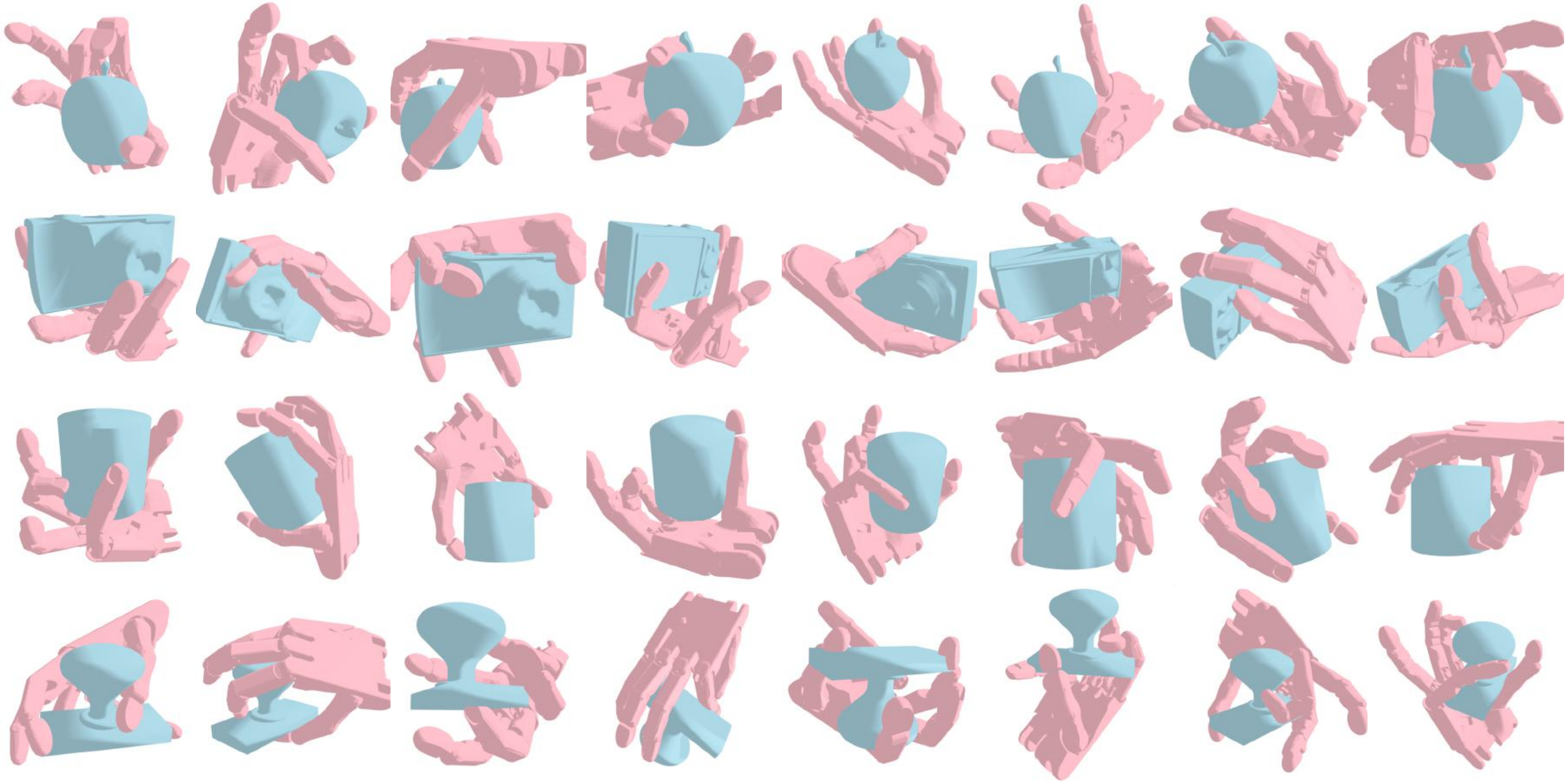
SceneDiffuser w/o opt.

SceneDiffuser w/ opt.

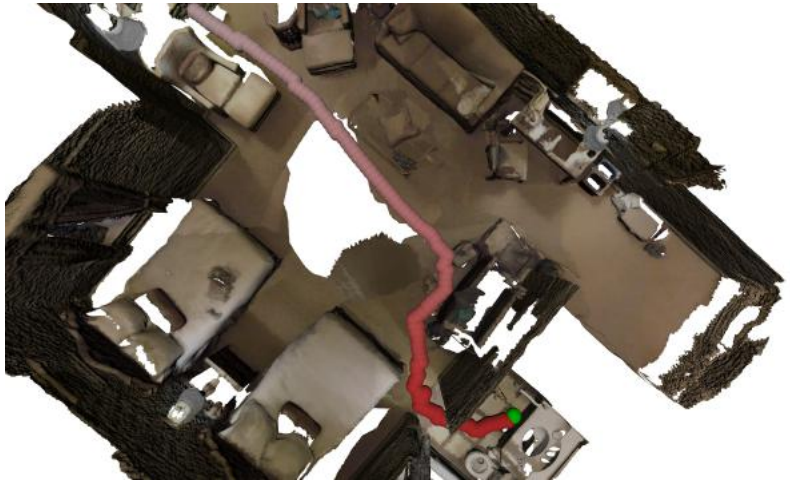
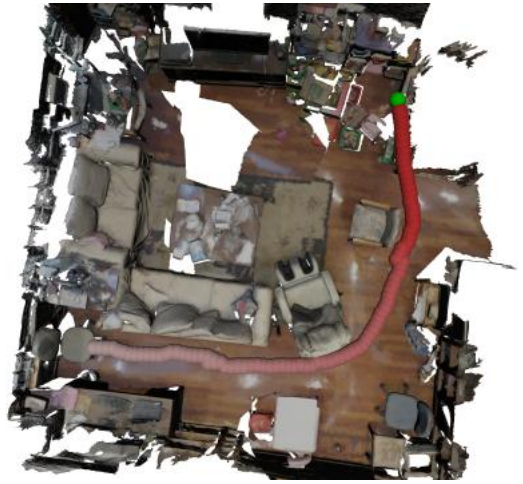
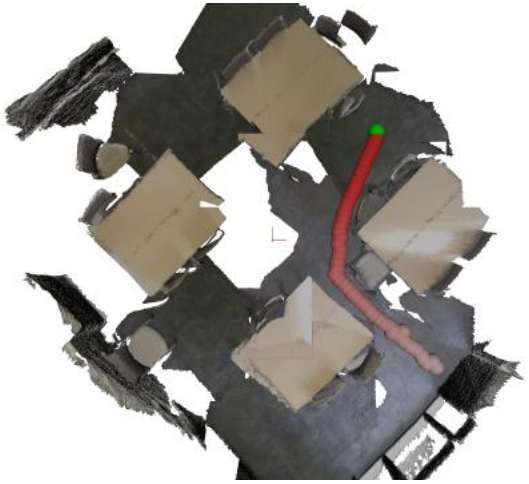
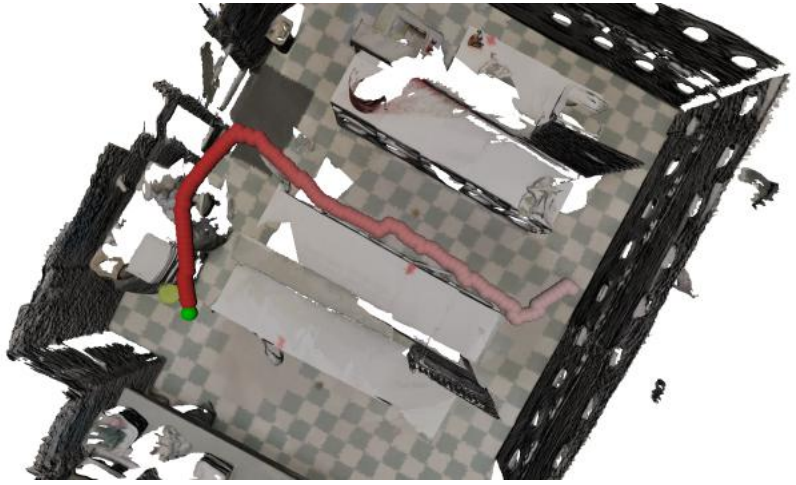
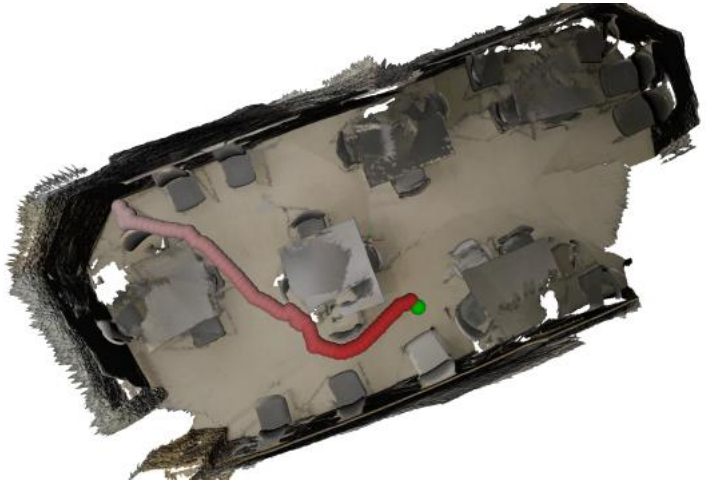
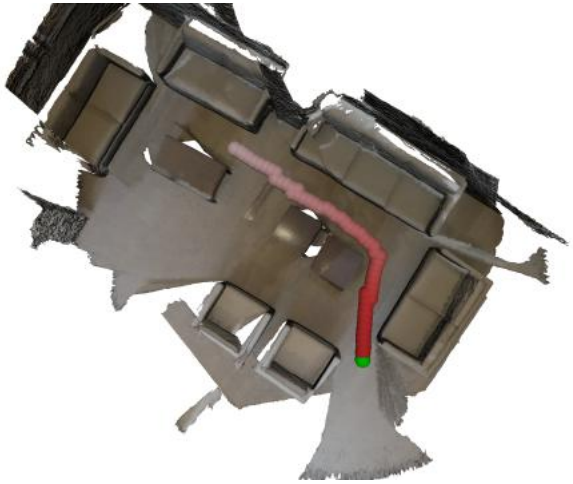
SceneDiffuser — Human Motion Generation



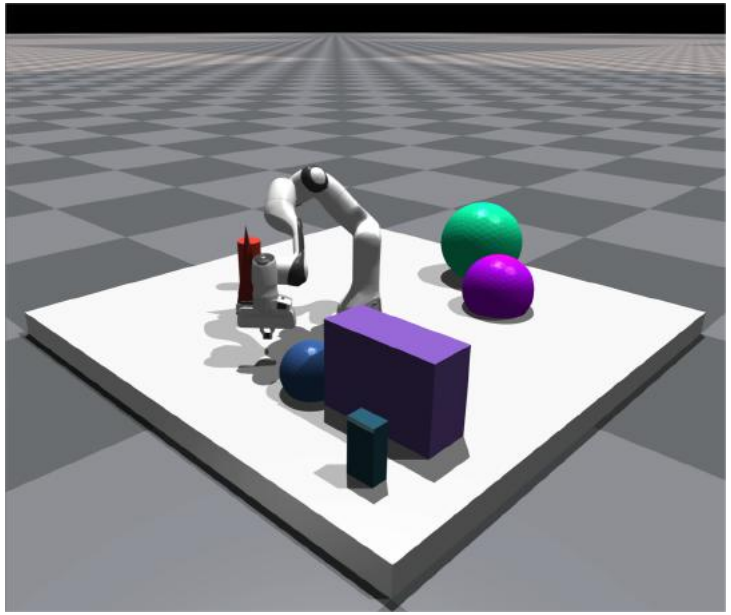
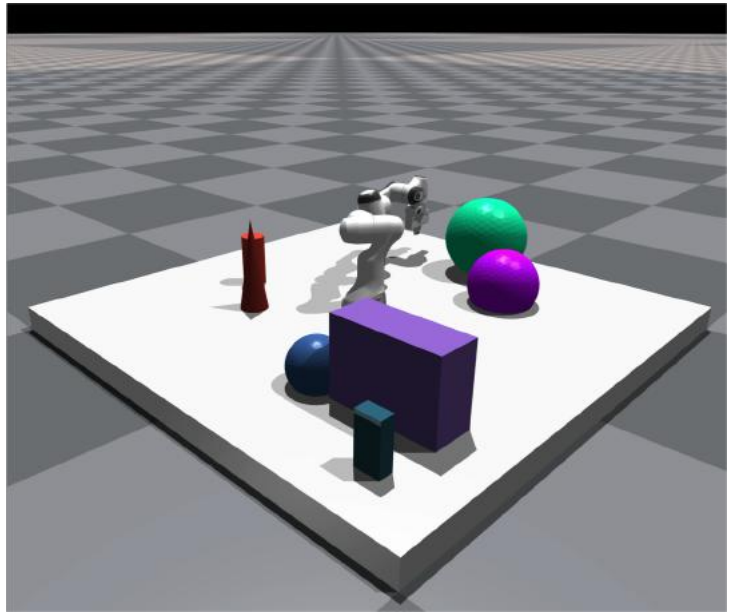
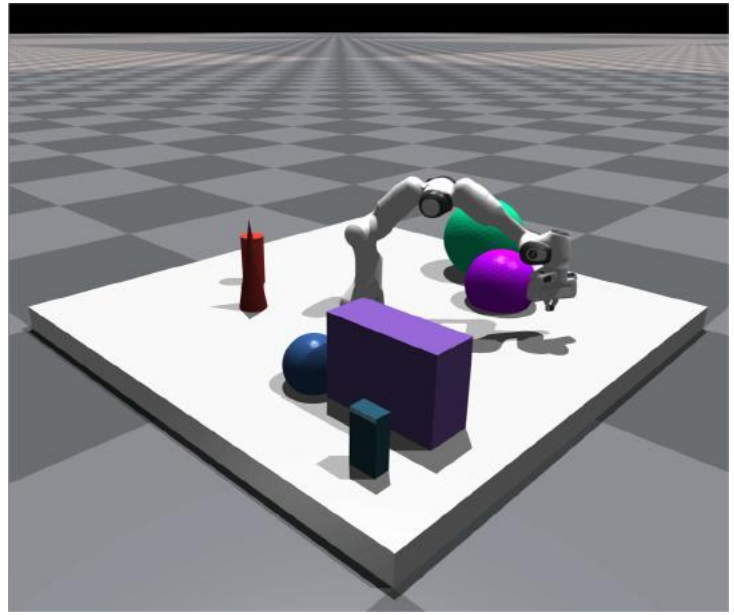
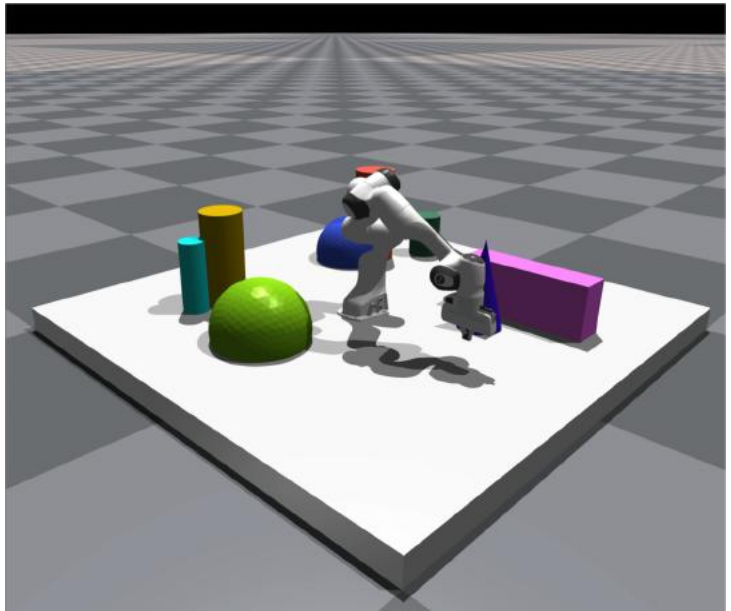
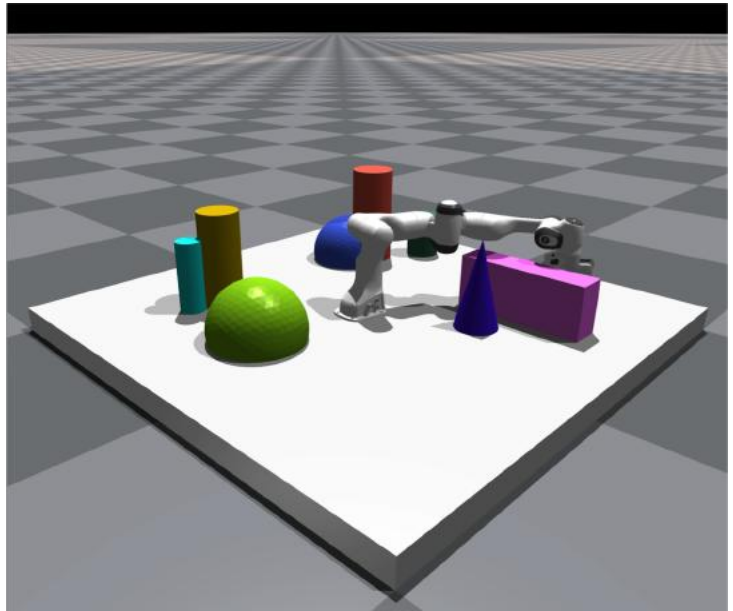
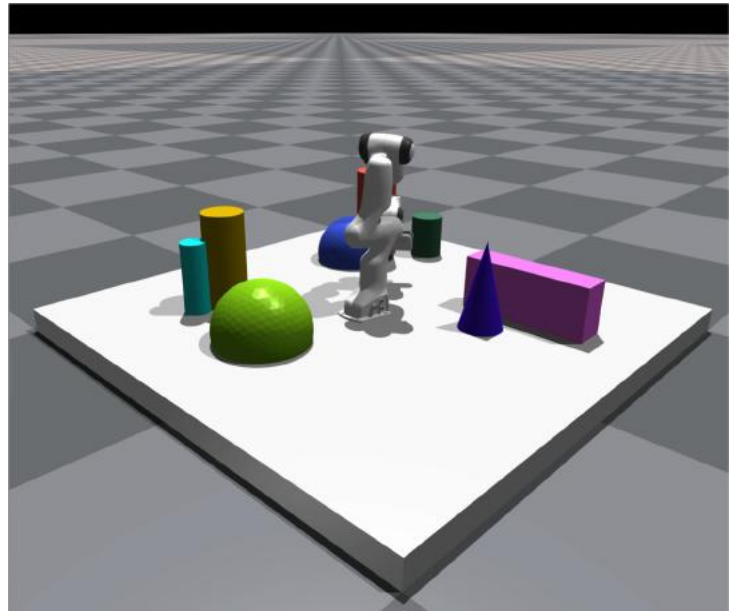
SceneDiffuser — Dexterous Grasp Generation



SceneDiffuser — Path Planning for 3D Navigation

SceneDiffuser — Motion Planning for Robot Arms

SceneDiffuser — Conclusion

- We propose the SceneDiffuser as a general conditional generative model for *generation*, *optimization*, and *planning* in 3D scenes.
- SceneDiffuser is intrinsically *scene-aware*, *physics-based*, and *goal-oriented*, applicable to various scene-conditioned 3D tasks.
- We demonstrate that the SceneDiffuser outperforms previous models by a *large margin* on **5** scene understanding tasks, establishing its **efficacy** and **flexibility**.

Move as You Say, Interact as You Can: Language-guided Human Motion Generation with Scene Affordance

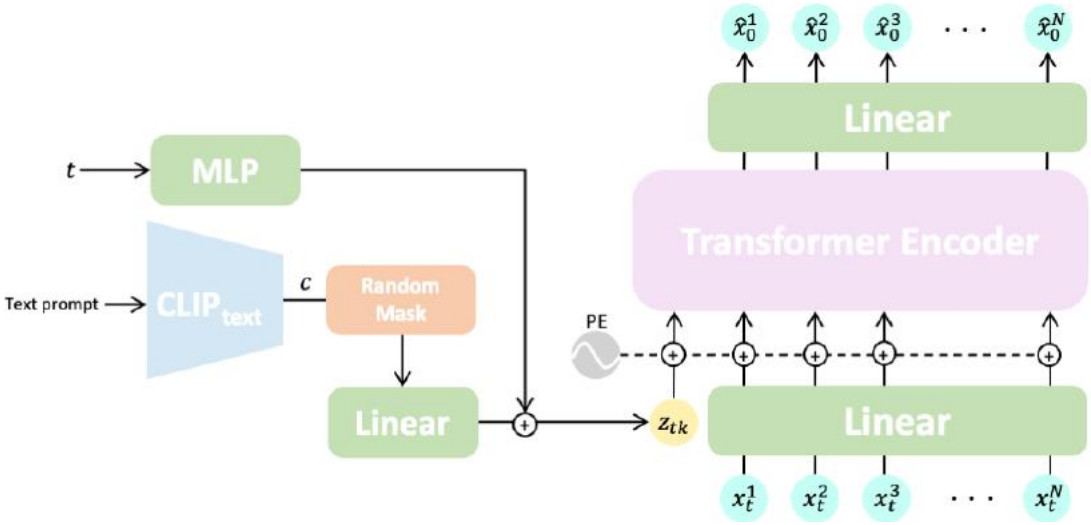
CVPR 2024, Highlight 2.8%

*We propose to **leverage scene affordance as an intermediate representation** to facilitate language-guided human motion generation in 3D scenes.*

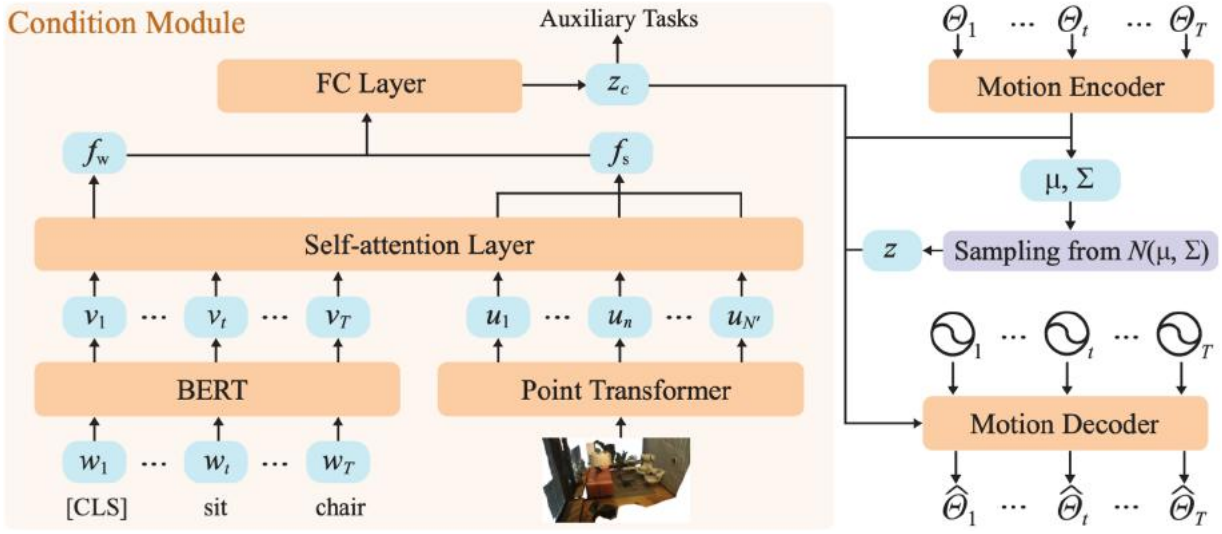
AffordMotion — Motivation

Challenge 1

- The inherent complexity of marrying 3D scene grounding and conditional motion generation
  - ➢ This complexity impedes the model's ability *to generalize* to novel scenes and descriptions.



Tevet *et al.*, ICLR 2023



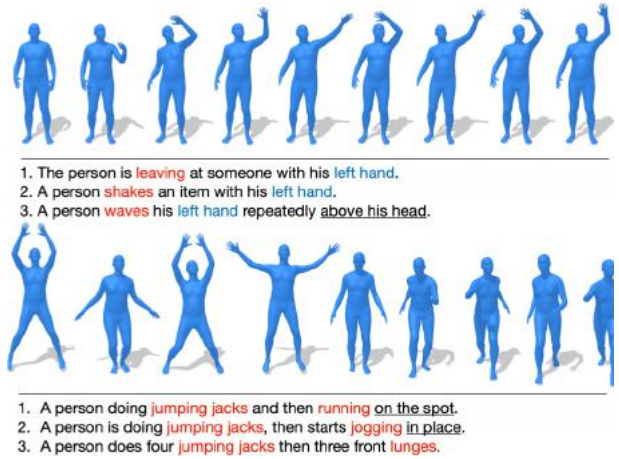
Wang *et al.*, NeurIPS 2022

AffordMotion — Motivation

Challenge 2

- The generative models' dependency on large volumes of high-quality paired data
  - ➢ Existing datasets lack large-scale, motion-diverse, and semantic-rich human-scene interactions.

Hassan *et al.*, ICCV 2019

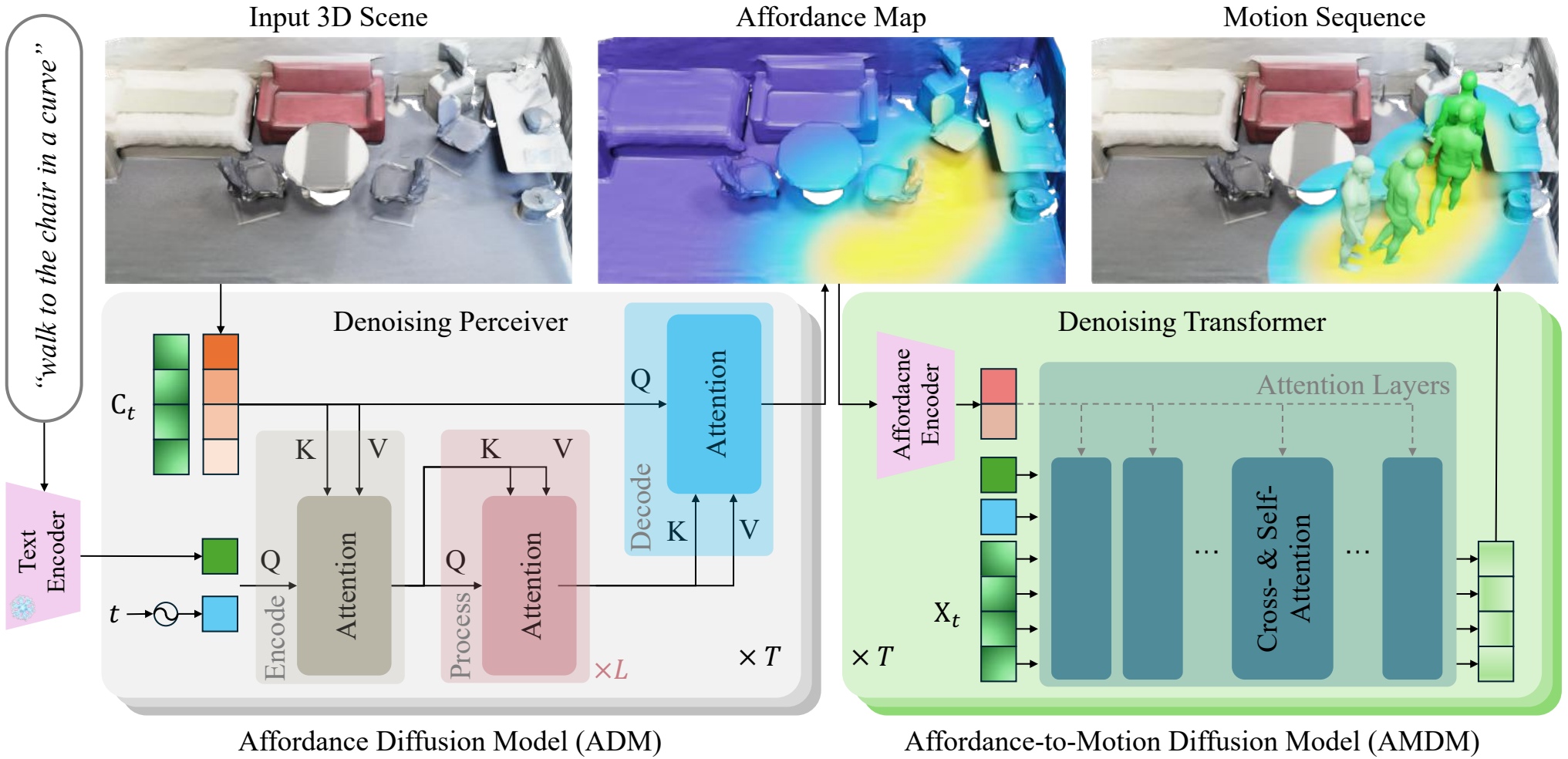


Guo *et al.*, CVPR 2022

Wang *et al.*, NeurIPS 2022

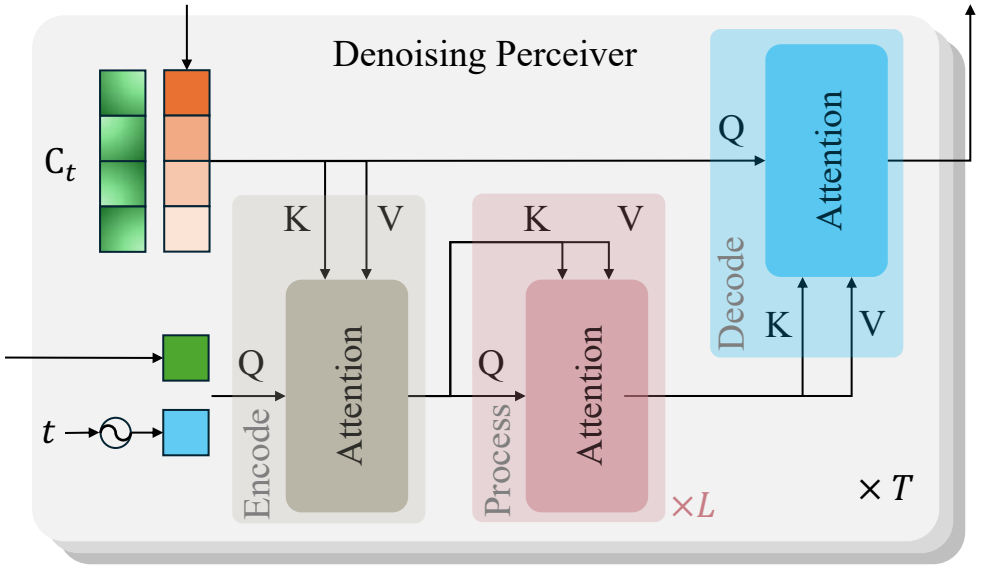
Araújo *et al.*, CVPR 2023

AffordMotion — Method



AffordMotion — Method

Affordance Diffusion Model (ADM)



- **Input:** scene point, language, noisy affordance map
- **Output:** denoised affordance map
- **Model Design:** A Perceiver architecture

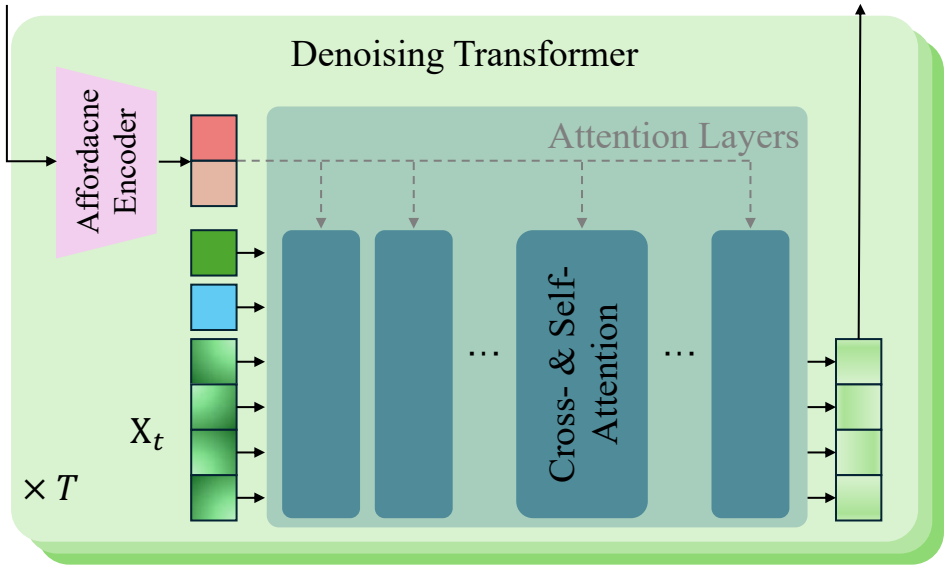
$$p_\theta(\mathbf{C}_{0:T} \mid \mathcal{S}, \mathcal{L}) = p(\mathbf{C}_T) \prod_{t=1}^{T} p_\theta(\mathbf{C}_{t-1} \mid \mathbf{C}_t, \mathcal{S}, \mathcal{L})$$

AffordMotion — Method

Affordance-to-Motion Diffusion Model (AMDM)

- **Input:** affordance, language, noisy motion
- **Output:** denoised motion
- **Model Design:** stack of attention layers

$$p_\phi(\mathbf{X}_{0:T} \mid \mathbf{C}, \mathcal{S}, \mathcal{L}) = p(\mathbf{X}_T) \prod_{t=1}^{T} p_\phi(\mathbf{X}_{t-1} \mid \mathbf{X}_t, \mathbf{C}, \mathcal{S}, \mathcal{L})$$



A person waves with his left hand.


Walk to the chair

AffordMotion — Results

Novel Evaluation Set – curated for assessing generalization capability

Quantitative Results

- Our method **enhances *FID*** while maintaining comparable *R-Precision* and *Multimodal-Dist* scores.
- The results highlight our model's effectiveness in producing ***physically plausible*** and ***semantically consistent*** results.

Table 4. **Qualitative results on our novel evaluation set.** "Real" indicates that we compute these metrics as a reference using the language-motion pairs within the test set of HumanML3D. Of note, our novel evaluation set does not contain ground truth motions.

Model	R-Precision (Top 3)↑	FID↓	MultiModal Dist.↓	Diversity→	MultiModality↑	contact↑	non-collision↑	quality score↑	action score↑
Real	$0.875^{\pm.002}$	$0.000^{\pm.000}$	$3.342^{\pm.004}$	$9.442^{\pm.301}$	-	-	-	-	-
one-stage @ Enc	$\mathbf{0.500}^{\pm.044}$	$11.848^{\pm1.634}$	$\mathbf{5.954}^{\pm.235}$	$\mathbf{8.395}^{\pm.850}$	$4.966^{\pm.321}$	$46.64^{\pm4.024}$	$99.88^{\pm.018}$	1.94 ± 1.15	2.61 ± 1.45
one-stage @ Dec	$0.403^{\pm.044}$	$12.268^{\pm.900}$	$6.611^{\pm.227}$	$8.049^{\pm.708}$	$5.031^{\pm.423}$	$26.75^{\pm4.264}$	$\mathbf{99.93}^{\pm.023}$	1.44 ± 0.83	1.96 ± 1.27
Ours @ Enc	$0.478^{\pm.069}$	$\mathbf{7.887}^{\pm1.189}$	$6.226^{\pm.261}$	$7.935^{\pm.857}$	$\mathbf{5.159}^{\pm.356}$	$71.98^{\pm2.542}$	$99.83^{\pm.006}$	$\mathbf{2.06 \pm 1.23}$	$\mathbf{2.63 \pm 1.47}$
Ours @ Dec	$0.428^{\pm.023}$	$12.027^{\pm3.164}$	$6.412^{\pm.204}$	$7.603^{\pm.715}$	$4.966^{\pm.353}$	$\mathbf{88.63}^{\pm2.975}$	$99.82^{\pm.015}$	1.99 ± 1.24	2.49 ± 1.40


A person lies down on the floor.

AffordMotion — Conclusion

- We introduce a novel two-stage model that *incorporates scene affordance as an intermediate representation*, facilitating language-guided human motion synthesis in 3D environments.
- We demonstrate our method's superiority over existing motion generation models across the HumanML3D and HUMANISE benchmarks.
- Our model showcases remarkable **generalization capabilities**, performing impressively in generating human motions within **unseen** scenarios.

Thanks!

HUMANISE

SceneDiffuser

AffordMotion