InteracTalker: Unified Prompt-Based Human-Object Interaction with Co-Speech Gesture Generation

Sreehari Rajan*, Kunal Bhosikar*, Charu Sharma
Machine Learning Lab, IIIT Hyderabad

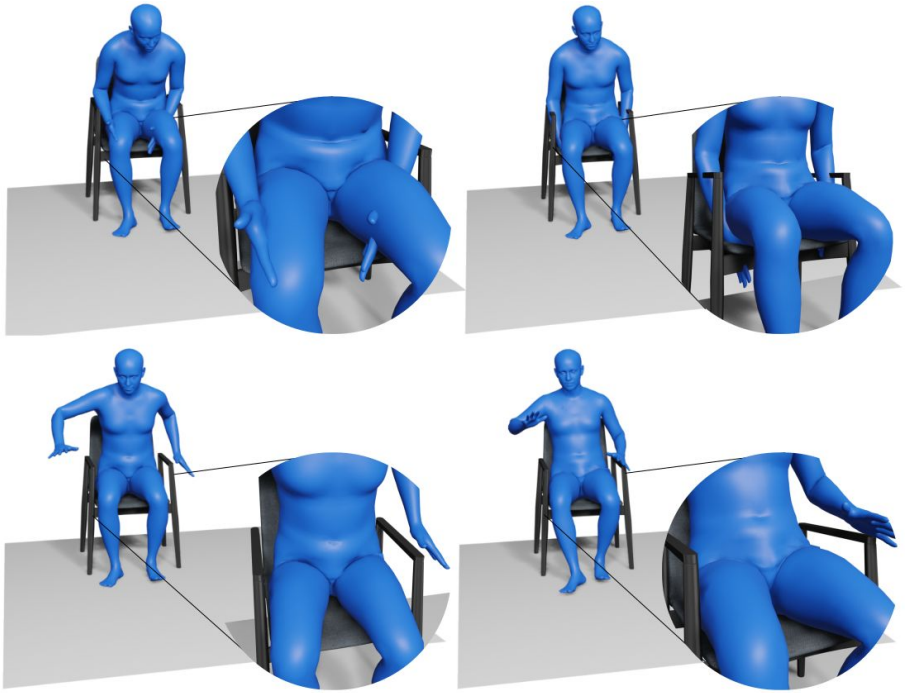
First Unified Framework for Co-Speech + Object-Aware Motion Generation

The Missing Piece in Motion Generation

- Speech-driven gestures (upper body)
- Object interactions (full body)
- Never unified.

What happens when we try to combine them?

Naive Concatenation Breaks Physical Realism

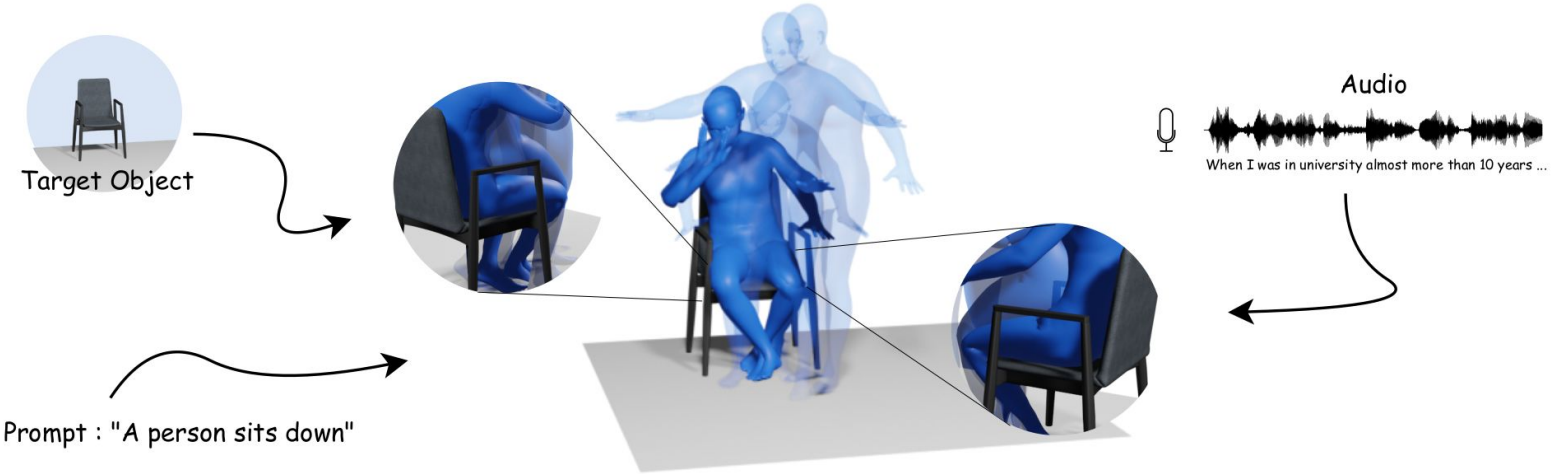


Unifying Speech and Object Interaction in Diffusion

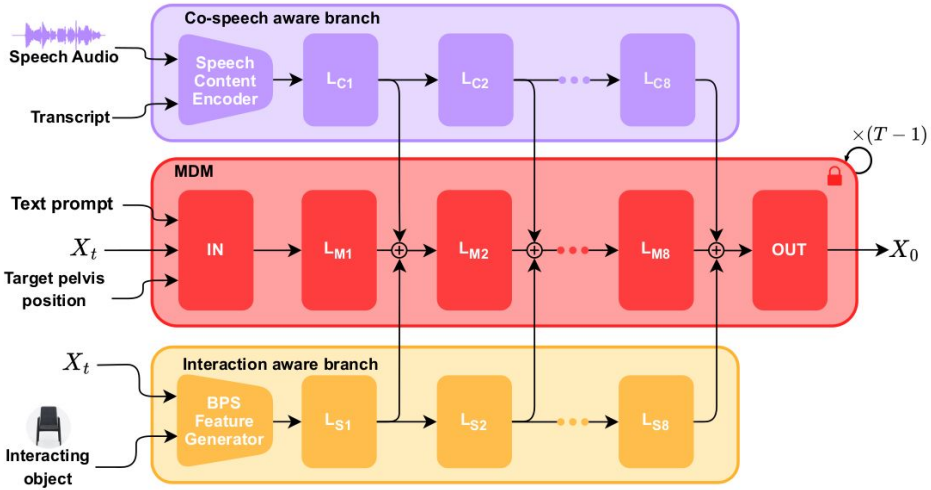
We condition a frozen Motion Diffusion Model on:

- Audio
- Text Prompt
- Target Object

Using modular adaptation branches.



InteracTalker Architecture



Core components:

1. Motion Diffusion Model
2. Interaction-Aware Branch
3. Co-Speech Branch

Adaptive Fusion Strategy

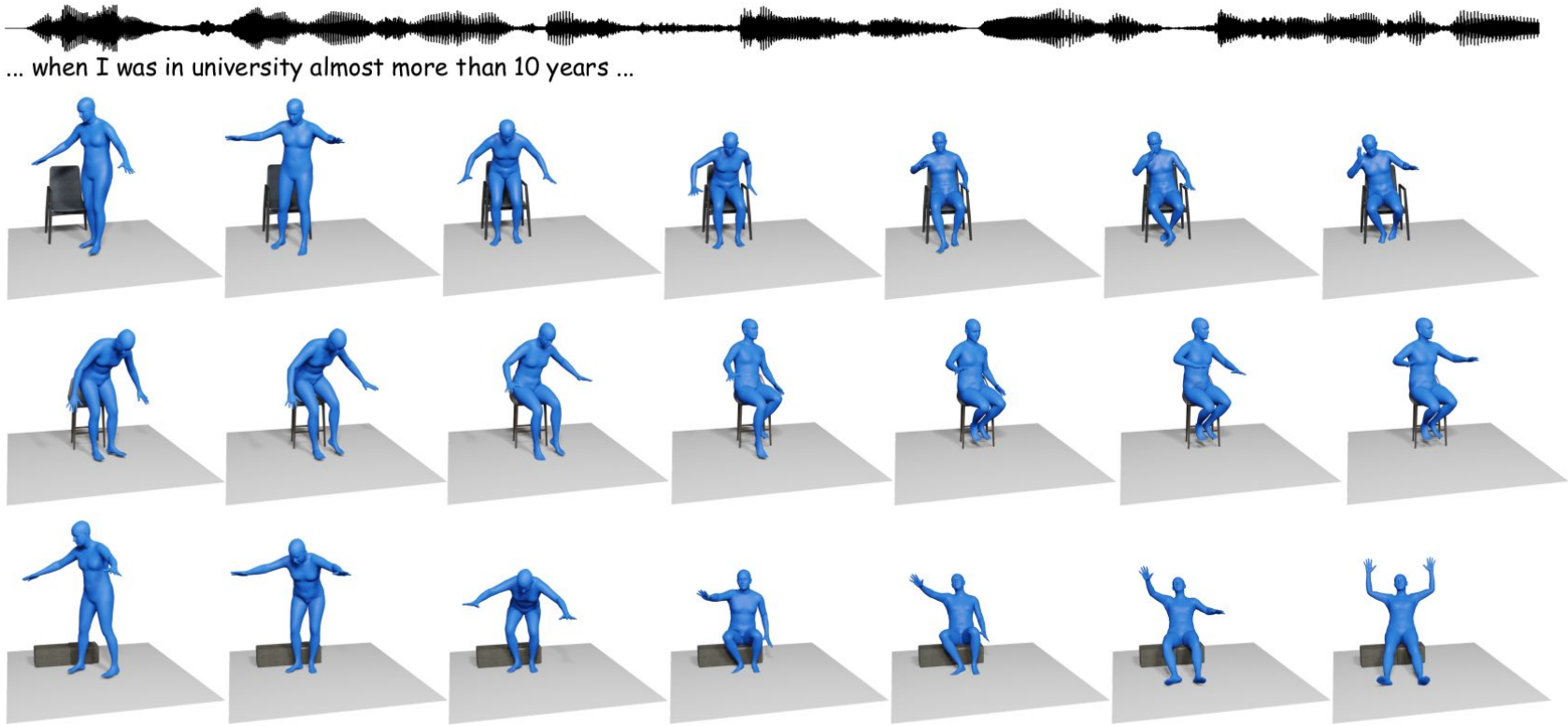
$$\mathcal{X}_t = \mathcal{X}_{uncond} + \sum_{i=1}^{n} \lambda_{cond_i} \mathcal{X}_{cond_i}$$

Weights dynamically updated via:

- Anchor supervision
- Gradient-based updates
- No test-time optimization

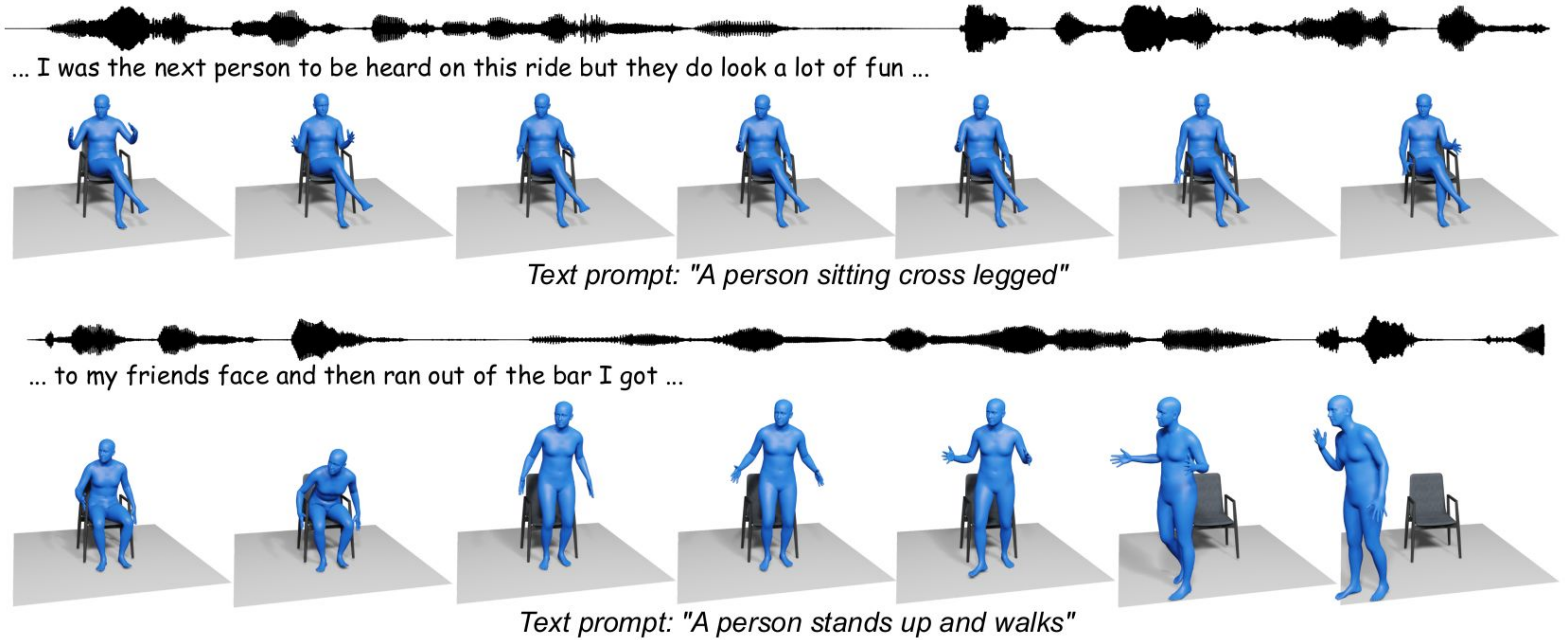
Balanced joint conditioning during diffusion.

Robust Across Object Variations



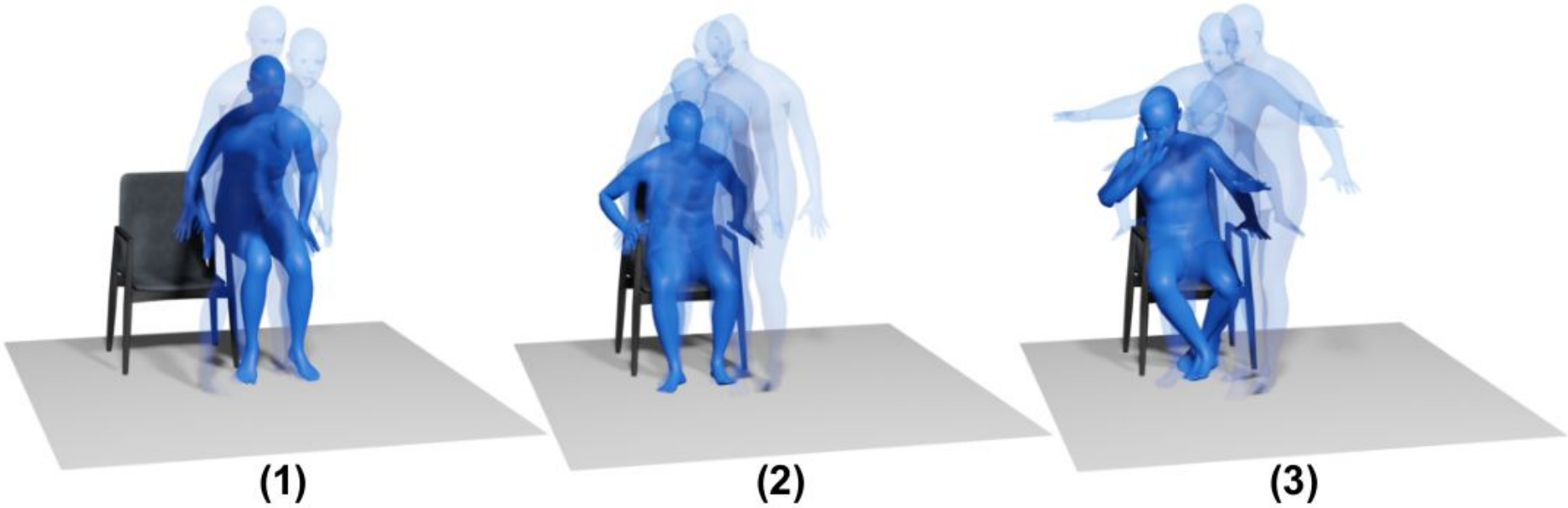
Text prompt: "A person sits down"

Semantic Control via Text Prompt



Same object, different prompts → distinct behaviors.

Why Unified Conditioning Matters



Incremental realism with unified conditioning.

User Study (64 Participants)

Method	Motion Realism (1 – 5 ↑)	Natural Motion (% ↑)	Object Interaction (% ↑)	Gesture Quality (% ↑)	Fewer Penetrations (% ↑)
Concat	2.77	29.7%	20.3%	26.6%	20.3%
InteracTalker	**3.77**	**70.3%**	**79.7%**	**73.4%**	**79.7%**

75% Preferred Ours

State-of-the-Art Performance

Co-Speech:
FGD ↓ **1.017** **(Best)**

Interaction:
Lower penetration
Lower goal error
No test-time optimization

Contributions

- ✔ First unified co-speech + object-aware diffusion
- ✔ Modular adaptation modules
- ✔ Adaptive fusion strategy
- ✔ SOTA in both subtasks
- ✔ Physically plausible without optimization

Unified diffusion enables realistic, controllable, physically plausible motion.

Thank You

Questions?

Project Page

Personal Website

LinkedIn

Applying for PhD positions in 3D Computer Vision and Digital Humans (Fall 2026)