COLLAGE: Collaborative Human-Agent Interaction Generation using Hierarchical Latent Diffusion and Language Models

Divyanshu Daiya[1], **Damon Conover**[2], **Aniket Bera**[1]

[1] Purdue University, [2] DEVCOM Army Research Laboratory

ATLANTA 2025 ICRA IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION

Motivation

Challenges. Generating collaborative (Human–Object–Human) motions is difficult because of:

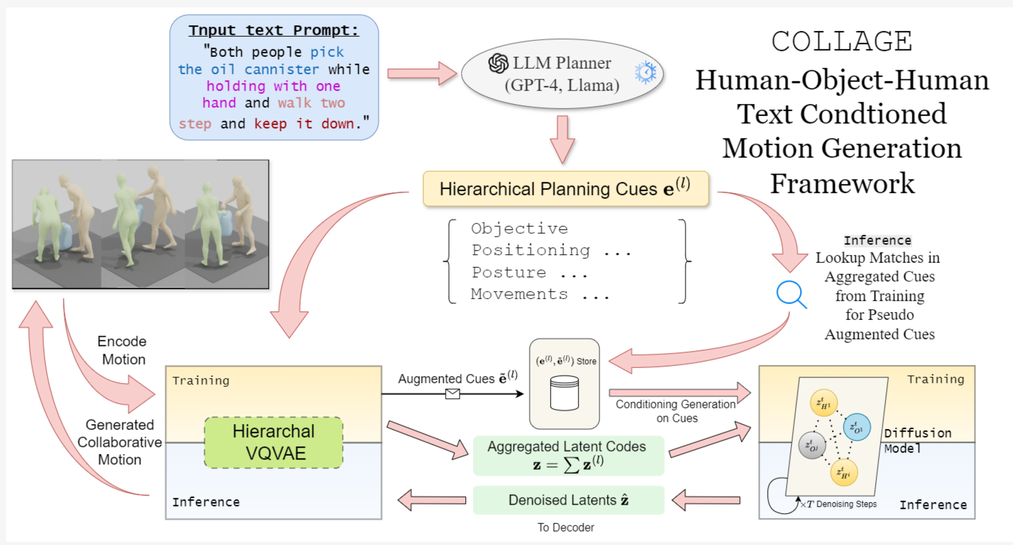
- Complex multi-agent planning,
- Limited **datasets that capture fine-grained synergy**,
- Need for fine control over interactions.

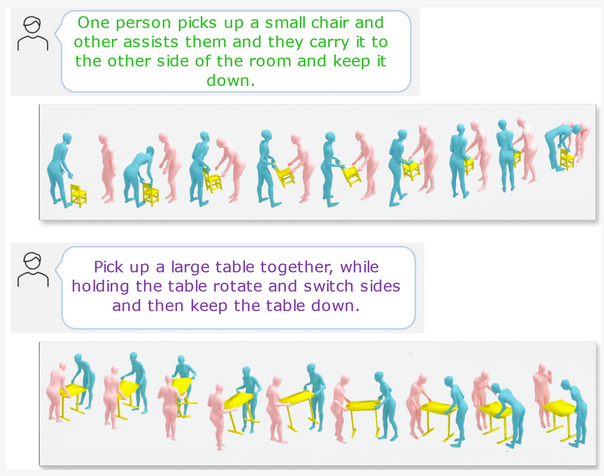
Key Contributions.

- **LLM-guided hierarchical VQ-VAE** supplies a structured, coarse-to-fine action script, directly taming the complexity of multi-agent planning.
- **Text ↔ code semantic anchors** learned via contrastive training enrich sparse data with strong language priors, compensating for limited fine-grained synergy examples.
- **Anchor-conditioned diffusion** refines motions in codebook space within **5–55 steps**, delivering precise, high-fidelity control and achieving SOTA generative performance on CORE4D and InterHuman datasets.

Overview

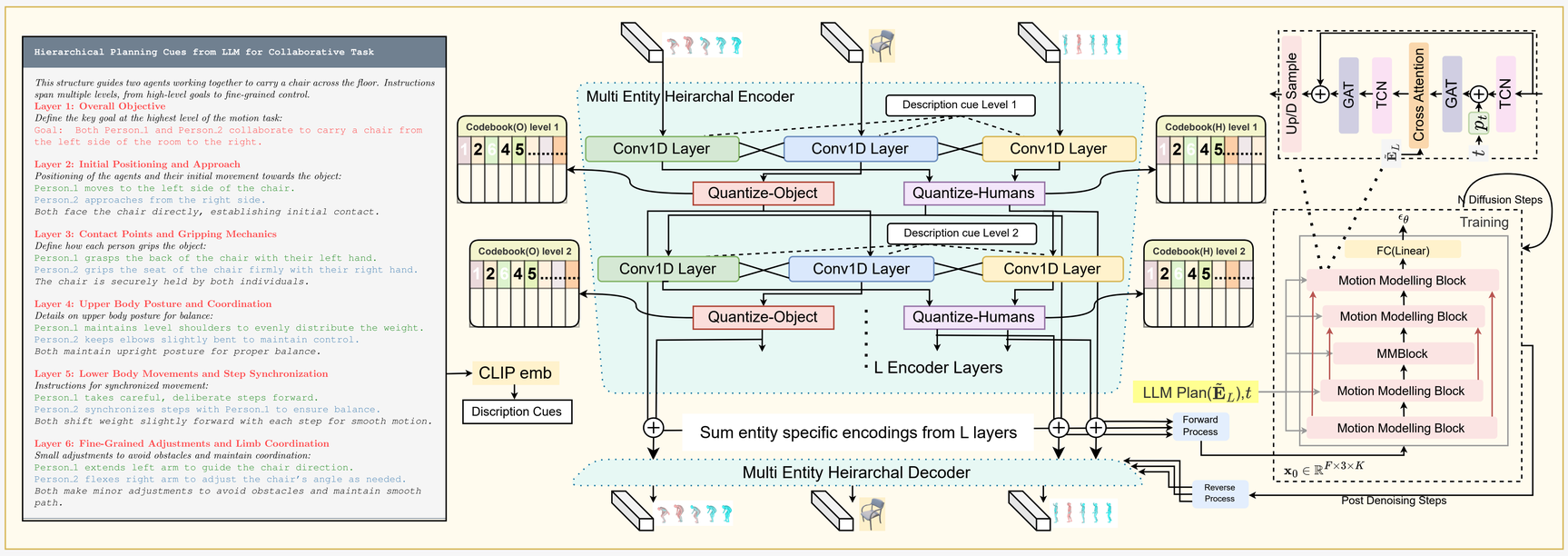
Goal: Generate realistic multi-actor + object collaboration from text or geometry.





High-Level Insight: LLM-based semantics + hierarchical latents yield coherent, controllable multi-agent interactions.

Methodology



Hierarchical VQ-VAE.

- Each layer encodes a different motion scale, and **LLM text cues** are injected at every level to steer semantics in real time.
- **Cross-code / text alignment** binds discrete latent codes to sentence embeddings, giving robust semantic anchors.
- Those anchors guide **anchor-conditioned diffusion** at test-time, so motions stay text-consistent while requiring markedly fewer denoise steps.
- A **compact loss mix**—reconstruction, contact, penetration and alignment—jointly enforces physical plausibility and semantic fidelity.
- The resulting latents are **disentangled and stable**, enabling smooth multi-agent synergy.

Latent Generation & Planning.

- A **Graph U-Net** built from TCN + GAT blocks models multi-actor synergy.
- Training minimises the **noise-prediction loss** $\|\epsilon - \epsilon_\theta(\mathbf{x}_t, \mathbf{E}_L)\|^2$.
- **Time-modulated weights** emphasise broad constraints early and fine detail late.
- Converges in just **5–55 DDIM steps**.

Incorporating Language.

- **Contrastive links** tie textual prompts to the learned codebook, guaranteeing semantic relevance.
- **Cross-attention** fuses text and motion signals inside every TCN/GAT layer.
- **Hierarchical GPT cues** such as "Approach → Grasp → Lift" guide generation step-by-step along the diffusion timeline.

Iterative Refinement.

- **LLM sub-actions** are injected at each partial denoise step, acting as intermediate waypoints.
- Codes **progressively align** to those sub-actions while Gaussian noise shrinks, tightening the motion trajectory.
- The procedure preserves consistent agent – agent - object collaboration even on long, complex tasks.

Experimental Results

Datasets: CORE-4D & InterHuman; **Metrics:** R-Precision (Top-1/2/3), FID, Diversity, Multimodality, MM-Distance, Joint Position Error (RR.J$_e$), Vertex Position Error (RR.V$_e$), Contact Accuracy (C$_{acc}$).

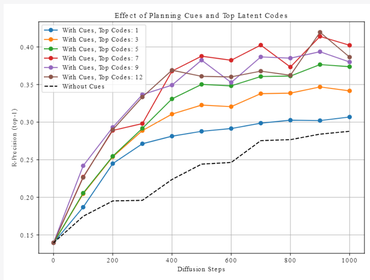
Text-Conditioned Generation on CORE4D and InterHuman

Methods(CORE-4D)	R Precision↑			FID↓	MM Dist↓	Diversity→	MModality↑
	Top 1	Top 2	Top 3				
Real	$0.312^{\pm0.007}$	$0.587^{\pm0.006}$	$0.673^{\pm0.006}$	$0.005^{\pm0.0005}$	$4.124^{\pm0.019}$	$8.151^{\pm0.091}$	-
TEMOS	$0.065^{\pm0.006}$	$0.179^{\pm0.006}$	$0.211^{\pm0.005}$	$9.214^{\pm0.0758}$	$8.536^{\pm0.019}$	$4.671^{\pm0.091}$	$0.510^{\pm0.052}$
T2M	$0.195^{\pm0.003}$	$0.141^{\pm0.003}$	$0.267^{\pm0.002}$	$11.258^{\pm0.0694}$	$5.867^{\pm0.013}$	$2.738^{\pm0.078}$	$1.672^{\pm0.041}$
MDM	$0.163^{\pm0.013}$	$0.257^{\pm0.010}$	$0.348^{\pm0.008}$	$9.671^{\pm0.0629}$	$10.219^{\pm0.020}$	$\mathbf{7.395^{\pm0.090}}$	$\mathbf{3.526^{\pm0.074}}$
MDM(GRU)	$0.168^{\pm0.009}$	$0.279^{\pm0.008}$	$0.361^{\pm0.010}$	$9.587^{\pm0.1382}$	$10.228^{\pm0.025}$	$6.951^{\pm0.151}$	$3.170^{\pm0.046}$
ComMDM	$0.187^{\pm0.005}$	$0.256^{\pm0.007}$	$0.301^{\pm0.007}$	$9.217^{\pm0.0727}$	$7.541^{\pm0.023}$	$5.367^{\pm0.080}$	$0.721^{\pm0.065}$
InterGen	$0.206^{\pm0.007}$	$0.312^{\pm0.008}$	$0.401^{\pm0.008}$	$7.217^{\pm0.2321}$	$10.251^{\pm0.017}$	$6.162^{\pm0.225}$	$3.402^{\pm0.063}$
COLLAGE	$\mathbf{0.229^{\pm0.008}}$	$\mathbf{0.332^{\pm0.009}}$	$\mathbf{0.435^{\pm0.009}}$	$\mathbf{6.890^{\pm0.2198}}$	$\mathbf{5.526^{\pm0.016}}$	$7.373^{\pm0.237}$	$3.589^{\pm0.066}$
w/o Hierarchy	$0.201^{\pm0.007}$	$0.309^{\pm0.008}$	$0.411^{\pm0.008}$	$7.452^{\pm0.2381}$	$5.582^{\pm0.018}$	$6.995^{\pm0.224}$	$3.209^{\pm0.058}$
w/o LLM	$0.208^{\pm0.007}$	$0.315^{\pm0.008}$	$0.419^{\pm0.008}$	$7.235^{\pm0.2305}$	$5.561^{\pm0.017}$	$6.549^{\pm0.230}$	$3.152^{\pm0.063}$
w/o Time Modulation	$0.218^{\pm0.008}$	$0.317^{\pm0.009}$	$0.420^{\pm0.009}$	$7.071^{\pm0.2251}$	$5.556^{\pm0.017}$	$7.263^{\pm0.234}$	$3.474^{\pm0.065}$

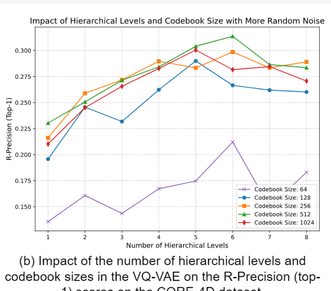
Methods (InterHuman)	R Precision↑			FID↓	MM Dist↓	Diversity→	MModality↑
	Top 1	Top 2	Top 3				
Real	$0.452^{\pm0.008}$	$0.610^{\pm0.009}$	$0.701^{\pm0.008}$	$0.273^{\pm0.007}$	$3.755^{\pm0.008}$	$7.948^{\pm0.064}$	-
TEMOS	$0.224^{\pm0.010}$	$0.316^{\pm0.013}$	$0.450^{\pm0.018}$	$17.375^{\pm0.043}$	$6.342^{\pm0.015}$	$6.939^{\pm0.071}$	$0.535^{\pm0.014}$
T2M	$0.238^{\pm0.012}$	$0.325^{\pm0.012}$	$0.464^{\pm0.014}$	$13.769^{\pm0.072}$	$5.731^{\pm0.018}$	$7.046^{\pm0.022}$	$1.387^{\pm0.076}$
MDM	$0.153^{\pm0.009}$	$0.260^{\pm0.011}$	$0.339^{\pm0.012}$	$9.167^{\pm0.056}$	$7.125^{\pm0.018}$	$\mathbf{7.602^{\pm0.045}}$	$2.355^{\pm0.080}$
MDM(GRU)	$0.179^{\pm0.006}$	$0.299^{\pm0.005}$	$0.387^{\pm0.007}$	$32.617^{\pm0.1221}$	$9.557^{\pm0.019}$	$7.003^{\pm0.134}$	$\mathbf{3.430^{\pm0.035}}$
ComMDM	$0.223^{\pm0.010}$	$0.334^{\pm0.008}$	$0.466^{\pm0.012}$	$7.069^{\pm0.054}$	$6.212^{\pm0.021}$	$7.244^{\pm0.038}$	$1.822^{\pm0.052}$
InterGen	$0.371^{\pm0.010}$	$0.515^{\pm0.012}$	$0.624^{\pm0.010}$	$5.918^{\pm0.079}$	$5.108^{\pm0.014}$	$7.387^{\pm0.029}$	$2.141^{\pm0.063}$
COLLAGE	$\mathbf{0.383^{\pm0.005}}$	$\mathbf{0.547^{\pm0.006}}$	$\mathbf{0.657^{\pm0.006}}$	$\mathbf{4.987^{\pm0.2061}}$	$\mathbf{4.992^{\pm0.012}}$	$7.515^{\pm0.214}$	$2.872^{\pm0.057}$
w/o Hierarchy	$0.355^{\pm0.009}$	$0.521^{\pm0.010}$	$0.632^{\pm0.009}$	$5.543^{\pm0.2154}$	$5.048^{\pm0.015}$	$7.137^{\pm0.201}$	$2.492^{\pm0.061}$
w/o LLM	$0.362^{\pm0.008}$	$0.528^{\pm0.009}$	$0.639^{\pm0.008}$	$5.326^{\pm0.2087}$	$5.027^{\pm0.014}$	$6.691^{\pm0.207}$	$2.435^{\pm0.059}$
w/o Time Modulation	$0.372^{\pm0.007}$	$0.536^{\pm0.008}$	$0.647^{\pm0.007}$	$5.162^{\pm0.2129}$	$5.021^{\pm0.013}$	$7.405^{\pm0.211}$	$2.767^{\pm0.062}$

Object-Conditioned

Test Set	Method	$RR.J_e$ (mm, ↓)	$RR.V_e$ (mm, ↓)	C_{acc} (%, ↑)	FID (↓)
S1	MDM	138.0 (± 0.3)	194.6 (± 0.2)	76.9 (± 0.5)	7.7 (± 0.2)
	OMOMO	137.8 (± 0.2)	196.7 (± 0.3)	78.2 (± 0.5)	8.3 (± 0.6)
	COLLAGE	**131.2 (± 0.2)**	**185.1 (± 0.2)**	**80.5 (± 0.4)**	**7.2 (± 0.2)**
S2	MDM	145.9 (± 0.2)	208.2 (± 0.2)	76.7 (± 0.1)	7.7 (± 0.2)
	OMOMO	145.2 (± 0.6)	209.9 (± 1.0)	77.8 (± 0.3)	8.3 (± 1.0)
	COLLAGE	**138.5 (± 0.5)**	**198.7 (± 0.8)**	**79.9 (± 0.2)**	**7.3 (± 0.8)**



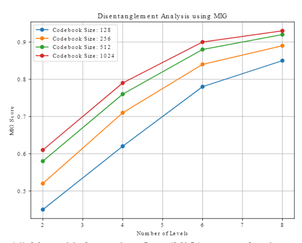
(a) Effect of LLM-guided planning cues and the number of top latent codes on the diffusion speed and denoising process.



(b) Impact of the number of hierarchical levels and codebook sizes in the VQ-VAE on the R-Precision (top-1) scores on the CORE-4D dataset.

Method	Inference Time (s)
TEMOS	1.24 (± 0.08)
T2M	0.96 (± 0.05)
MDM	0.85 (± 0.06)
ComMDM	1.12 (± 0.09)
InterGen	0.79 (± 0.04)
OurModel	**0.58 (± 0.02)**

(c)Average inference time per sample on the CORE-4D dataset.



(d) Mutual Information Gap (MIG) scores for the hierarchical VQ-VAE model across different numbers of levels and codebook sizes.

Observations

- **Hierarchy Matters**: Removing levels harms synergy and quality, drops R-Precision by **12 %** and worsens FID by **8 %**
- **Language Guidance is crucial** Excluding LLM cues lowers R-Precision by **9 %** and raises FID by **5 %**.
- With only 15 steps, FID is **4 %** better than InterGen while **65 %** faster; 55 steps yields peak quality.
- **Diversity & Task Conformity**: Hierarchical latent diffusion explores a **broader codebook space** for varied interactions, while LLM-guided anchors ensure **prompt consistency**.
- **SOTA Performance**: Against InterGen, R-Precision improves by **+11 %** (CORE-4D) and **+3 %** (InterHuman) and FID drops by **–5 %** and **–16 %**.

Discussion

- We leverage hierarchical cues to accelerate alignment and generation, and our **coarse-to-fine latents reduce collisions** while preserving long-range coordination in dense scenes.
- We envision that our pipeline can support **humanoid robots or VR avatars**, synthesizing contact-aware trajectories from high-level goals (e.g., "hand over the wrench").
- We plan to extend our system with **interactive motion editing** via user-adjustable handles and to integrate **scene-based cues** for more refined motion trajectories.

For more information, and visualizations, *Please visit / Scan* →