LaMoGen: Language to Motion Generation Through LLM-Guided Symbolic Inference

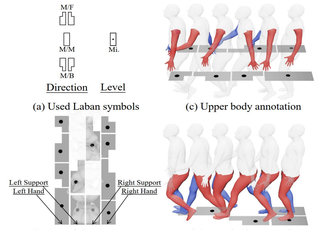
Junkun Jiang, Ho Yin Au, Jingyu Xiang, Jie Chen*

Department of Computer Science, Hong Kong Baptist University

CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivations

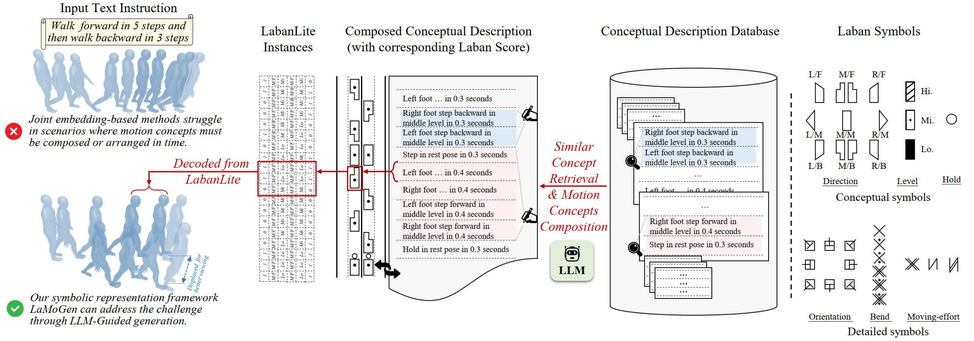
- **Human motions encode rich semantics that often correspond to intentions and instructions described in natural language.** Precisely mapping between language and motion is essential for computational understanding and modeling of human behavior.
- **Current text-to-motion methods struggle with fine-grained semantic fidelity**, especially in temporal and spatial coordination specified by instructions (e.g., "Walk forward in 5 steps and then walk backward in 3 steps"), which leads to generic, imprecise motion generation.
- **The traditional symbolic system Labanotation offers a compact and interpretable way to specify complex, compositional movements**, but its abstraction makes it challenging to directly connect symbolic representations with continuous motion data via machine learning.



Key Technology

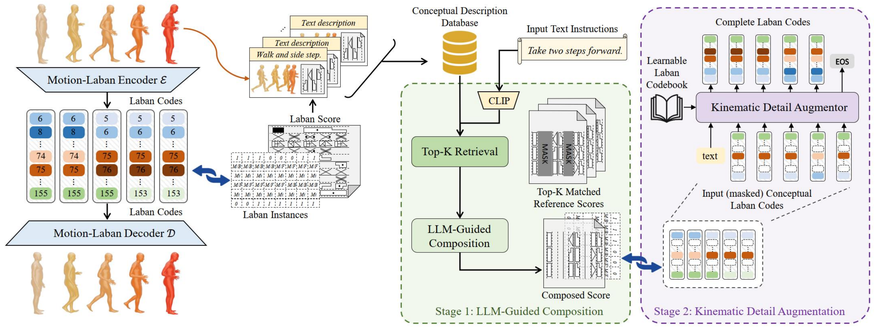
- **LabanLite:** A frame- and body-part-wise motion representation based on Labanotation, enabling explicit and interpretable encoding of temporal and structural attributes. *Bridges symbolic abstraction with machine learning for LLM-driven, interpretable motion planning.*
- **LaMoGen Framework:** An end-to-end text instructions to LabanLite symbols via LLM-based planning, then generates fine-grained motion sequences.
- **Labanotation-based Benchmark:** A evaluation benchmark for evaluating motion-text alignment in symbolic, temporal, and harmony aspects., enabling transparent and interpretable assessment of generated motions.

Overview of LaMoGen



- (Left) Traditional joint embeddings often fail at semantic consistency.
- (Middle) LLMs compose motion plans via **retrieval-augmented prompting**:
  - A **Conceptual Description Database** maps motion captions to **Laban Scores** and **Conceptual Descriptions**.
  - User queries are matched to captions using semantic similarity (CLIP); top-K Scores and Descriptions are retrieved.
  - LLMs edit or compose new conceptual plans (Scores + Descriptions) from retrieved data.
- (Right) Laban symbols adopted in LaMoGen.

Text-Labanotation-Motion Network



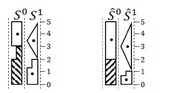
(a) Laban-Motion Encoder-Decoder

(b) LLM-Guided Text-Laban-Motion Generator

- Enables conversions between motions and Laban instances.
- LLMs perform high-level planning via retrieval-augmented prompting.
- The Kinematic Detail Augmentor adds fine-grained details through autoregressive generation.

Laban Benchmark

- Prediction and Ground truth



- Laban Metrics

Semantic Alignment

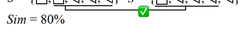
Harmonious Alignment

(S^0, S^1)={ }, { }

$(S^0, \hat{S}^1)$={ }, { }

$Sim = 50\%$

Temporal Alignment



- Motion-Laban-Text Paired Dataset

Examples of textual description

- Take a stride to the right — this is step two, following the sequence: right, left.
- Slide over to the right—step two in the progression goes: right, left.
- Proceed by stepping rightward for step two, continuing the series: right, left.

WordCloud

Experiment Results

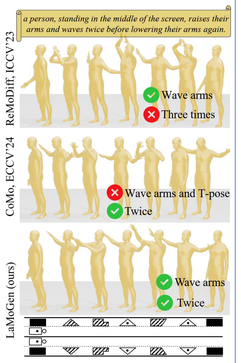
Method	SMT ↑				TMP ↑				HMN ↑			R@3 ↑	FID ↓
	supL	supR	armL	armR	supL	supR	armL	armR	arm-arm	arm-sup	sup-sup		
Real data	-	-	-	-	-	-	-	-	-	-	-	0.216	0.001
MDM [22]	0.380	0.380	0.335	0.257	0.316	0.316	0.329	0.231	0.119	0.226	0.258	0.180	22.81
ReMoDiff [29]	0.470	0.470	0.427	0.395	0.377	0.377	0.385	0.322	0.179	0.264	0.351	0.192	7.121
MoDiff [30]	0.491	0.491	0.470	0.411	0.362	0.362	0.412	0.328	0.180	0.281	0.361	0.196	5.701
CoMo [7]	0.358	0.358	0.474	0.382	0.211	0.211	0.284	0.250	0.203	0.298	0.252	0.176	21.94
MotionGPT [9]	0.488	0.488	0.466	0.400	0.353	0.353	0.343	0.340	0.244	0.300	0.378	0.195	2.072
None	0.523	0.523	0.430	0.392	0.337	0.337	0.361	0.385	0.215	0.356	0.393	0.199	5.562
Qwen3	0.571	0.571	0.478	0.495	0.401	0.401	0.456	0.448	0.369	0.334	0.450	**0.212**	1.903
DeepSeekV3	0.552	0.552	0.496	0.500	0.475	0.475	0.486	0.486	0.370	0.326	0.463	0.206	1.859
GPT4.1	0.583	0.583	0.493	0.476	0.507	0.507	0.501	0.492	0.303	0.367	0.508	0.208	1.861
Human	**0.648**	**0.648**	**0.592**	**0.616**	**0.619**	**0.619**	**0.643**	**0.632**	**0.379**	**0.449**	**0.558**	0.211	**1.769**

LaMoGen Composer

Method	R-precision ↑			FID ↓	MM-Dist ↓	Diversity →	Multi-Mod. ↑
	Top-1	Top-2	Top-3				
Real data	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.085}$	-
Guo et al. [5]	$0.457^{\pm.002}$	$0.639^{\pm.003}$	$0.740^{\pm.003}$	$1.067^{\pm.002}$	$3.340^{\pm.008}$	$9.188^{\pm.002}$	$2.090^{\pm.083}$
MDM [22]	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$\mathbf{9.559^{\pm.086}}$	$\mathbf{2.799^{\pm.072}}$
ReMoDiff [29]	$0.510^{\pm.005}$	$0.698^{\pm.006}$	$0.795^{\pm.004}$	$\mathbf{0.103^{\pm.004}}$	$\mathbf{2.974^{\pm.016}}$	$9.018^{\pm.075}$	$1.795^{\pm.043}$
MoDiff [30]	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$0.630^{\pm.001}$	$3.113^{\pm.001}$	$9.410^{\pm.049}$	$1.533^{\pm.042}$
CoMo [7]	$0.502^{\pm.002}$	$0.692^{\pm.007}$	$0.790^{\pm.002}$	$0.262^{\pm.004}$	$3.032^{\pm.015}$	$9.936^{\pm.066}$	$1.013^{\pm.046}$
KP [18]	$0.496^{\pm.000}$	-	-	$0.275^{\pm.000}$	-	$9.975^{\pm.000}$	$2.218^{\pm.000}$
MotionGPT [9]	$0.492^{\pm.003}$	$0.681^{\pm.003}$	$0.775^{\pm.002}$	$0.232^{\pm.008}$	$3.096^{\pm.008}$	$9.528^{\pm.071}$	$2.008^{\pm.084}$
None	$0.438^{\pm.005}$	$0.591^{\pm.004}$	$0.755^{\pm.003}$	$1.091^{\pm.013}$	$3.999^{\pm.020}$	$8.555^{\pm.097}$	$1.421^{\pm.089}$
GPT4.1mini	$0.453^{\pm.003}$	$0.679^{\pm.003}$	$0.779^{\pm.004}$	$0.561^{\pm.008}$	$3.717^{\pm.011}$	$8.434^{\pm.115}$	$1.263^{\pm.034}$
GPT4.1	$0.491^{\pm.002}$	$0.694^{\pm.002}$	$0.796^{\pm.003}$	$0.252^{\pm.003}$	$3.087^{\pm.003}$	$9.124^{\pm.058}$	$1.131^{\pm.027}$
Human	$\mathbf{0.513^{\pm.003}}$	$\mathbf{0.704^{\pm.002}}$	$\mathbf{0.813^{\pm.006}}$	$0.206^{\pm.003}$	$2.993^{\pm.009}$	$9.635^{\pm.109}$	$0.973^{\pm.024}$

HumanML3D

LaMoGen Composer



Conclusion and Takeaway

- Propose LabanLite, a symbolic motion representation aligning textual instructions with semantically and temporally matched sub-actions.
- Bridge high-level language instructions with low-level motion execution, enabling language-driven agents to generate interpretable and modular motion plans.
- Advance expressive and controllable robotics with improved text-to-motion alignment.

Project Page

My CV
Open to academic positions