MoHI: Boosting Motion Generation via Human Intention Understanding

ANONYMOUS AUTHOR(S)

We propose MoHI, a motion generation framework that explicitly models human intention as the underlying cause of motion. By explicitly disentangling intention prediction from motion synthesis during training and jointly optimizing the two objectives, MoHI captures the motivational logic underlying human actions and provides clearer semantic guidance for coherent motion generation. Experiments on HumanML3D demonstrate state-of-the-art performance, with +4.5% improvement in R-Precision Top-1 and 38.6% lower FID over the state-of-the-art method. Fine-tuned on motion captioning, MoHI also outperforms recent LLM-based approaches, highlighting its unified strength in both motion understanding and generation.

Additional Key Words and Phrases: Motion Generation, Human Intention Prediction, Motion Caption

ACM Reference Format:
Anonymous Author(s). 2018. MoHI: Boosting Motion Generation via Human Intention Understanding. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Human motion generation has become a central research focus in artificial intelligence, with applications in animation, robotics, and virtual reality [7]. Recent advances in generative modeling and large language models have enabled the synthesis of motions that appear plausible and semantically aligned with natural language descriptions [10, 13, 16]. Despite these advances, existing systems still face persistent challenges in achieving fine-grained spatiotemporal precision, ensuring controllability, and generalizing to diverse scenarios.

A fundamental limitation lies in the absence of explicit modeling of human intention. Current text-to-motion generation approaches including recent instruction-tuned large language models [5, 6, 8, 15] treat motion and language as a surface-level mapping during training. Despite leveraging large-scale linguistic priors and impressive generative capacity, these models fundamentally neglect the causal structure of intention that governs how and why actions unfold. As a result, while they can generate motions that appear lexically or semantically aligned with input descriptions, they often fail to capture the deeper motivational logic behind human behavior. This leads to sequences that, though superficially plausible, lack internal coherence, physical grounding, or purposeful progression—manifesting as rigid, unnatural, or contextually implausible motions that fall short of authentic human movement.

To address these challenges, we introduce MoHI, a motion generation framework that explicitly models human intention. In our design, fine-grained textual descriptions of incomplete motions are leveraged as explicit intent cues, capturing not only the observed action but also the latent goals that constrain what should follow. MoHI is trained to simultaneously predict human intention and generate motion, ensuring that synthesized sequences are both coherent and purpose-driven. Furthermore, by disentangling intent prediction from motion synthesis, the model reduces semantic entanglement and achieves substantial gains in generation quality.

Comprehensive experiments on HumanML3D demonstrate that MoHI outperforms state-of-the-art approaches on Text-to-Motion (T2M) tasks. By further fine-tuning MoHI on motion captioning, the model also surpasses recent LLM-based captioning methods. More importantly, MoHI highlights that understanding the core of motion from the perspective of intention significantly facilitates motion generation. This provides a unified perspective for advancing both motion understanding and motion generation.

2 Method

2.1 Model Architecture

We introduce OmniMoGen, a unified framework designed for generating human motion and intention. Each input modality is first mapped into a latent representation through dedicated encoders. To encourage generative modeling, random masking is applied to the motion tokens. These masked motion tokens are then modulated by a Conditional Masked Transformer, which integrates conditioning signals from other modalities at both the semantic-level and dynamic token-level. The resulting tokens act as a shared representation that is capable of producing both high-level intentions and complete motion sequences. A motion decoder finally reconstructs the output into the original motion domain.

2.1.1 *Modality-Specific Encoder.*

Motion Auto-Encoder. For a motion sequence $\mathbf{x}_m \in \mathbb{R}^{l_m \times c_m}$ with l_m frames and c_m feature dimensions, we employ a temporal convolutional auto-encoder to compress it into a latent representation $\mathbf{z}_m = f_m(\mathbf{x}_m) \in \mathbb{R}^{l'_m \times d_m}$, where d_m is the embedding dimension. A symmetric decoder reconstructs the motion as $\hat{\mathbf{x}}_m = g_m(\mathbf{z}_m)$. The reconstruction objective is defined by a smooth L1 loss.

Language Encoder. We adopt pretrained CLIP encoders [12], which remain frozen during training, to extract language features. For text input $\mathbf{x}_t$, the encoder produces contextual embeddings $\mathbf{z}_t = f_t(\mathbf{x}_t) \in \mathbb{R}^{l_t \times d_t}$, where the `[CLS]` token provides a global representation $\mathbf{z}_t^g$.

2.1.2 *Conditional Masked Transformer.*

The Conditional Masked Transformer fuses text signals into the motion tokens through two mechanisms: 1) semantic-level modulation, which injects global multimodal context into the motion representation using adaptive normalization; 2) mixture-of-attention with adaptive scope, which aligns motion tokens with the most relevant segments of text contexts through an adaptive Top-k attention strategy.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

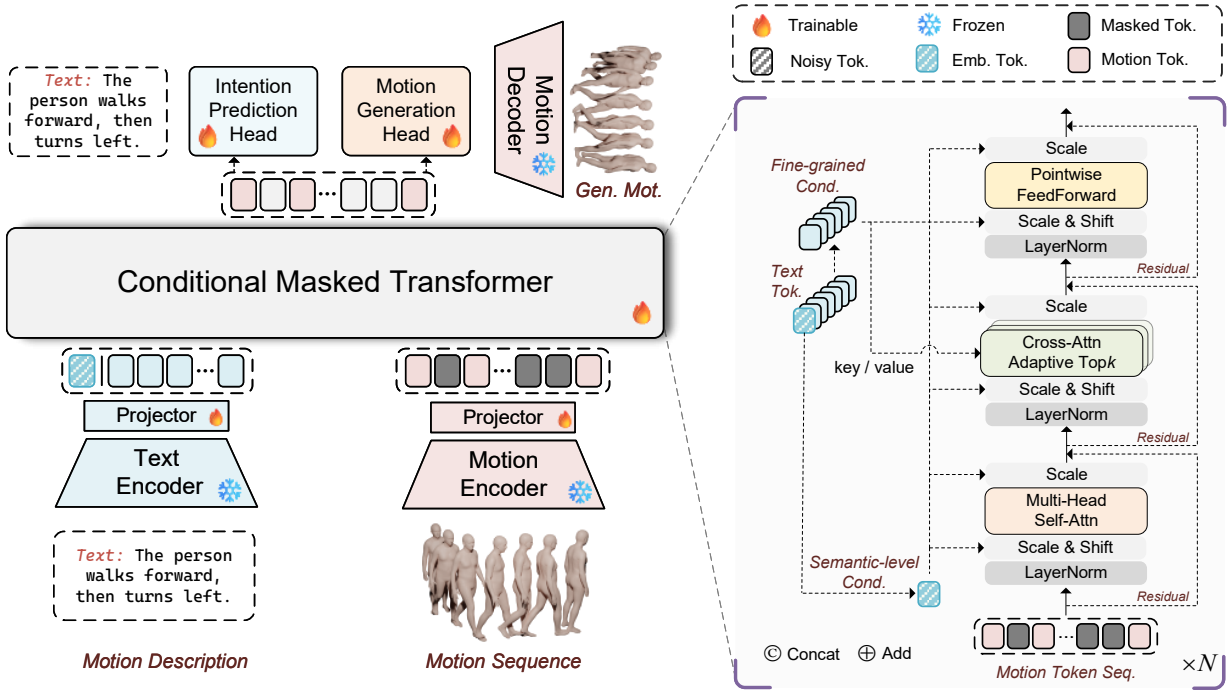

Fig. 1. Model architecture of MoHI. The framework first encodes motion and text inputs through modality-specific encoders, including a motion auto-encoder and frozen CLIP text encoder. Subsequently, textual conditioning signals are injected into masked motion tokens via a Conditional Masked Transformer with semantic modulation and adaptive Top-k cross-attention. The final outputs are generated through disentangled heads: a diffusion-based Motion Generation Head for human motion sequences and a T5-style Intention Prediction Head for predicting human intention.

Semantic-level Modulation. We compute a global context vector $\mathbf{c}^g = \mathbf{z}_t^g + \mathbf{z}_v^g$, which is mapped to modulation coefficients $(\alpha_c, \beta_c, \gamma_c)$ by a small MLP. For a normalized motion token $\bar{\mathbf{z}}_m = \mathrm{LN}(\mathbf{z}_m)$, modulation is applied as:

$$\mathbf{z}_m \leftarrow \mathbf{z}_m + \gamma_c \odot h(\alpha_c \odot \bar{\mathbf{z}}_m + \beta_c), \tag{1}$$

where $h(\cdot)$ denotes the sub-layer in the Transformer block, sequentially consisting of a self-attention module, a mixture-of-attention module with Adaptive Scope, and a bottleneck MLP layer.

Mixture-of-Attention with Adaptive Scope. We construct token-level embeddings $\mathbf{c}^{tok} = [\mathbf{z}_v; \mathbf{z}_t] \in \mathbb{R}^{(p+l)\times d}$, with missing modalities replaced by learnable embeddings. For query motion tokens $\mathbf{z}_m \in \mathbb{R}^{l'_m \times d}$, each expert computes queries, keys, and values, and produces an attention score matrix $\mathbf{A}^e$. To restrict the effective scope, we sort $\mathbf{A}^e$ in descending order and accumulate the attention weights until the cumulative mass exceeds a threshold τ. The attention is then normalized only over this selected subset of entries, whose size is adaptively determined by the threshold. Finally, contributions from all experts are summed.

2.1.3 Disentangled Generation Heads. Since motion and intention correspond to different semantic levels, we design two separate heads. The Intention Prediction Head (IPH) is responsible for producing textual descriptions of intentions, using a T5-style decoder conditioned on the output of the conditional masked transformer, denoted as $\mathbf{z}$. The Motion Generation Head (MGH) instead models continuous motion dynamics with a diffusion-based MLP model.

2.2 Training Strategy

We train the framework on two tasks: motion generation from full text using a diffusion-based velocity matching objective [9] with masked motion inputs, and intention prediction from partial motion using an autoregressive cross-entropy loss. After training, the model is further fine-tuned on the motion captioning task, where complete motion sequences are provided as input and the model generates descriptive text outputs.

3 Experiment

3.1 Dataset and Motion Representation

We conduct experiments on the HumanML3D dataset [3]. To ensure compact and effective motion encoding, we follow prior work [10] and remove redundant features (e.g., 6D rotations) to reduce distribution mismatch and generation errors. Each motion frame is represented as:

$$\mathbf{x}_m^i = [\dot{r}^a, \dot{r}^{xz}, \dot{r}^h, j^p] \tag{2}$$

where $\dot{r}^a$ denotes root angular velocity, $\dot{r}^{xz}$ the root linear velocities in the XZ-plane, $\dot{r}^h$ the root height, and $j^p \in \mathbb{R}^{3(N_j-1)}$ the local joint positions.

Table 1. The quantitative results of text-to-motion generation on the HumanML3D dataset. The best results are displayed in bold.

Methods	R Precision↑			FID↓	Matching↓	MModality↑	CLIP-score ↓
	Top 1	Top 2	Top 3				
T2M-GPT [17]	$0.470^{\pm.003}$	$0.659^{\pm.002}$	$0.758^{\pm.002}$	$0.335^{\pm.003}$	$3.505^{\pm.017}$	$2.018^{\pm.053}$	$0.607^{\pm.005}$
ReMoDiffuse [19]	$0.468^{\pm.003}$	$0.653^{\pm.003}$	$0.754^{\pm.005}$	$0.883^{\pm.021}$	$3.414^{\pm.020}$	$2.703^{\pm.154}$	$0.621^{\pm.003}$
MDM-50Step [14]	$0.440^{\pm.007}$	$0.636^{\pm.006}$	$0.742^{\pm.004}$	$0.518^{\pm.032}$	$3.640^{\pm.028}$	$\mathbf{3.604^{\pm.031}}$	$0.578^{\pm.003}$
MLD [1]	$0.461^{\pm.004}$	$0.651^{\pm.004}$	$0.750^{\pm.003}$	$0.431^{\pm.014}$	$3.445^{\pm.019}$	$3.506^{\pm.031}$	$0.610^{\pm.003}$
MMM [11]	$0.487^{\pm.003}$	$0.683^{\pm.002}$	$0.782^{\pm.001}$	$0.132^{\pm.004}$	$3.359^{\pm.009}$	$1.241^{\pm.073}$	$0.635^{\pm.003}$
MoMask [2]	$0.469^{\pm.004}$	$0.687^{\pm.003}$	$0.786^{\pm.003}$	$0.116^{\pm.006}$	$3.353^{\pm.010}$	$1.263^{\pm.079}$	$0.637^{\pm.003}$
MotionDiffuse [18]	$0.450^{\pm.006}$	$0.641^{\pm.005}$	$0.753^{\pm.005}$	$0.778^{\pm.005}$	$3.490^{\pm.023}$	$3.179^{\pm.046}$	$0.606^{\pm.004}$
MARDM-DDPM [10]	$0.492^{\pm.006}$	$0.690^{\pm.005}$	$0.790^{\pm.005}$	$0.116^{\pm.004}$	$3.349^{\pm.010}$	$2.470^{\pm.053}$	$0.637^{\pm.005}$
MARDM-SiT [10]	$0.500^{\pm.004}$	$0.695^{\pm.003}$	$0.795^{\pm.003}$	$0.114^{\pm.007}$	$3.270^{\pm.009}$	$2.231^{\pm.071}$	$0.642^{\pm.002}$
MotionAgent [16]	$0.485^{\pm.003}$	$0.680^{\pm.003}$	$0.780^{\pm.002}$	$0.202^{\pm.009}$	$3.327^{\pm.009}$	–	$0.634^{\pm.003}$
MoGIC w/o Int. (ours)	$0.533^{\pm0.012}$	$0.731^{\pm0.010}$	$0.826^{\pm0.010}$	$0.108^{\pm0.023}$	$3.078^{\pm0.037}$	$2.455^{\pm.062}$	$0.658^{\pm0.001}$
MoGIC (ours)	$\mathbf{0.545^{\pm0.003}}$	$\mathbf{0.741^{\pm0.003}}$	$\mathbf{0.835^{\pm0.002}}$	$\mathbf{0.070^{\pm0.004}}$	$\mathbf{2.999^{\pm0.011}}$	$2.448^{\pm.055}$	$\mathbf{0.669^{\pm0.001}}$

Table 2. Quantitative comparison with state-of-the-art methods on the motion captioning task

Methods	BLEU@1↑	BLEU@4↑	ROUGE↑	BERTScore↑
TM2T [4]	48.90	8.27	38.1	32.2
MotionGPT [5]	48.20	12.47	37.4	32.4
MotionChain [6]	48.10	**12.56**	33.9	36.9
MG-MotionLLM [15]	–	8.06	–	36.7
OmniMoGen	**53.13**	10.36	**40.6**	**40.7**

3.2 Experiment Settings

All experiments are conducted on an NVIDIA RTX 4090 GPU with a batch size of 64 using the Adam optimizer for 500 epochs. The motion generation loss is optimized every epoch, while the intention prediction loss is updated every 4 epochs. The conditional masked transformer consists of a single layer, with cross-attention implemented by two parallel modules: one with $k \in [1, 6]$ and threshold 0.8, and the other with $k \in [0, \infty]$ and threshold 1 (all condition tokens). The intention prediction head (IPH) adopts a 3-layer T5-style decoder, and the motion generation head (MGH) is a diffusion model built on a 10-layer MLP.

3.3 Comparisons on Text-to-Motion

We evaluate our proposed MoHI framework against a broad set of state-of-the-art (SOTA) methods on the HumanML3D benchmark, including diffusion-based models (e.g., MDM [14], MotionDiffuse [18]), autoregressive models (e.g., T2M-GPT [17]), and recent masked or multimodal approaches (e.g., MoMask [2], MARDM [10], MotionAgent [16]).

As shown in Table 1, MoHI achieves substantial improvements across all evaluation metrics. In particular, it sets new state-of-the-art performance in R-Precision, with notable gains of +4.5% Top-1 and +4.0% Top-3 over the strongest baseline. This indicates that motions generated by MoHI are more semantically aligned with the textual description. Additionally, MoHI attains the lowest FID score (0.070), demonstrating superior realism and distributional fidelity compared to reference motion data.

3.4 Comparisons on Motion Captioning

We further fine-tune MoHI on the motion captioning task to evaluate its ability in understanding and verbalizing human motion. Benefiting from the joint training of motion generation and intention prediction, MoHI learns to capture the latent causes driving actions, thereby producing captions that not only describe surface-level dynamics but also reflect the underlying intent and purpose.

Despite using only a 3-layer Transformer decoder as the Intention Prediction Head (IPH) and without relying on pretrained LLM weights, MoHI surpasses recent LLM-based approaches (Table 2).

3.5 Ablation Study on Intention Prediction

Results are summarized in Table 1. Compared to the variant without intention prediction (MoHI w/o Int.), the full MoHI model achieves consistent gains across most metrics, including a notable reduction in FID and improvements in R-Precision. This confirms that the disentangled optimization of intention prediction and motion generation is mutually reinforcing, enabling the model to synthesize motions that are both semantically aligned and contextually purposeful.

4 Conclusion

In this paper, we present MoHI, a framework that enhances human motion generation by explicitly modeling intention. Experiments on HumanML3D show that MoHI achieves state-of-the-art performance in the T2M task and, when fine-tuned, outperforms recent LLM-based models in motion captioning, demonstrating that intention-aware training benefits both motion synthesis and understanding.

References

[1] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. 2023. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18000–18010.

[2] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. 2024. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1900–1910.

[3] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating Diverse and Natural 3D Human Motions From Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5152–5161.

[4] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. 2022. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*. Springer, 580–597.

[5] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems* 36 (2023), 20067–20079.

[6] Biao Jiang, Xin Chen, Chi Zhang, Fukun Yin, Zhuoyuan Li, Gang YU, and Jiayuan Fan. 2024. MotionChain: Conversational Motion Controllers via Multimodal Prompts. In *European Conference on Computer Vision*.

[7] Peizhuo Li, Sebastian Starke, Yuting Ye, and Olga Sorkine-Hornung. 2024. Walkthedog: Cross-morphology motion alignment via phase manifolds. In *ACM SIGGRAPH 2024 Conference Papers*. 1–10.

[8] Mingshuang Luo, Ruibing Hou, Zhuo Li, Hong Chang, Zimo Liu, Yaowei Wang, and Shiguang Shan. 2024. M^3GPT: An Advanced Multimodal, Multitask Framework for Motion Comprehension and Generation. *Advances in Neural Information Processing Systems* 37 (2024), 28051–28077.

[9] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. 2024. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*. Springer, 23–40.

[10] Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. 2025. Rethinking Diffusion for Text-Driven Human Motion Generation: Redundant Representations, Evaluation, and Masked Autoregression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 27859–27871.

[11] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. 2024. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1546–1555.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[13] Junyu Shi, Lijiang Liu, Yong Sun, Zhiyuan Zhang, Jinni Zhou, and Qiang Nie. 2025. GenM^3: Generative Pretrained Multi-path Motion Model for Text Conditional Human Motion Generation. *arXiv preprint arXiv:2503.14919* (2025).

[14] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. 2023. Human Motion Diffusion Model. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=SJ1kSyO2jwu

[15] Bizhu Wu, Jinheng Xie, Keming Shen, Zhe Kong, Jianfeng Ren, Ruibin Bai, Rong Qu, and Linlin Shen. 2025. MG-MotionLLM: A unified framework for motion comprehension and generation across multiple granularities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 27849–27858.

[16] Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. 2025. Motion-Agent: A Conversational Framework for Human Motion Generation with LLMs. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=AvOhBgsE5R

[17] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. 2023. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14730–14740.

[18] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2024. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 6 (2024), 4115–4128.

[19] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. 2023. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 364–373.