Multimodal Context-Aware Human Motion Generation with Language, Vision, and Object

Junyu Shi Yong Sun Zhiyuan Zhang Lijiang Liu Yuxin He Zhengjie Zhang Qiang Nie*
The Hong Kong University of Science and Technology (Guangzhou)
{jshi890, ysun691, zzhang400, lliu135, yhe469, zzhang666}@connect.hkust-gz.edu.cn
qiangnie@hkust-gz.edu.cn

Abstract

Motion generation has made substantial progress in synthesizing human motion from language, yet text alone remains ambiguous for specifying fine-grained spatial, temporal, and interaction details. Visual observations offer a powerful source of complementary context: even sparse images can reveal intermediate body states, object affordances, and interaction relations that are difficult to specify precisely in language. Moreover, contextual constraints like visual, object, and partial-motion conditions can ground motion generation with higher physical fidelity and controllability. We therefore study a unified multimodal context-aware framework for motion generation under diverse conditions. The key challenge is that heterogeneous conditions impose constraints at different spatiotemporal granularities and exhibit token- and phase-dependent relevance. Moreover, incorporating multiple modalities without a robust motion prior may entangle modality-specific semantics and cause negative transfer. To this end, we combine global multimodal modulation with fine-grained motion token-level context selection, enabling the model to adaptively exploit relevant signals during generation. We further adopt progressive training that first learns a strong language-conditioned motion prior, then extends to multimodal context with different modality combinations. To support multimodal training, we introduce Mo900H, a large-scale benchmark integrating 21 motion datasets with over 900 hours of human motion. Our method reduces FID by 21.1% and 48.3% compared with SOTA methods on the HumanML3D and Mo900H datasets, while also improving motion captioning and enabling diverse multimodal-conditioned generation. Code is available at https://github.com/SorenShi/MC-MoGen.

1 Introduction

Language-to-motion generation provides a natural interface for human motion synthesis [12, 36, 27, 44], but a text prompt rarely specifies the full spatial and temporal constraints of the target motion. Consider the instruction "Lift the box". The prompt identifies the action and the target object, but it does not specify where the box is, how the person approaches it, or when contact and lifting occur. In applications such as character animation, embodied simulation, and human-object interaction, these missing details are often supplied more naturally by visual observations, object states, or partial motion evidence than by longer text.

Existing motion-generation approaches only partially address this problem. First, language-to-motion models typically rely on detailed annotations or GPT-augmented text as the motion specification [54, 3], yet language remains an incomplete control signal for spatial and fine-grained temporal constraints. Second, existing image-conditioned motion studies mainly use visual inputs for pose

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

estimation and reconstruction [23], leaving sparse visual signals underexplored as controllable prompts that specify intermediate action states. Although videos can be converted into 3D poses via off-the-shelf estimators, this incurs extra computation and produces poses in estimator-specific coordinate systems misaligned with the generator's canonical representation. Third, while recent methods incorporate conditions beyond language [23, 25, 65], they still lack a unified framework that jointly leverages multimodal spatial constraints, especially visual sequences and object conditions.

In this paper, we study multimodal context-aware motion generation. We treat language as high-level intent, while optional contextual signals serve as constraints that narrow the space of plausible motions. Such context includes sparse visual frames, object geometry and trajectories, and visible motion fragments. Rather than deterministically specifying a single motion, these signals provide complementary spatial, geometric, temporal, and kinematic constraints that shape the distribution of plausible outputs. This formulation enables a unified model to generate motion from text alone when no additional context is available, while leveraging extra evidence when the prompt is underspecified.

This setting raises three challenges. First, multimodal conditions are heterogeneous: language provides high-level intent, while visual observations, object states, and partial motions impose spatiotemporal and interaction constraints at different granularities. Second, the relevance between motion tokens and contextual conditions is phase-dependent, as different motion tokens correspond to different action phases and require distinct contextual cues during generation. Third, directly incorporating multiple modalities before learning a robust motion prior may entangle modality-specific semantics and cause negative transfer, especially when multimodal supervision is incomplete across datasets.

To address these challenges, we propose a unified multimodal context-aware motion generation framework. First, a Multimodal Context Transformer (MCT) integrates any available subset of language, visual, and object conditions at both the global and fine-grained levels. Second, a Context-Selective Cross-Attention module enables each motion token to adaptively attend to the most relevant condition tokens across modalities rather than uniformly aggregating all inputs. Third, we adopt a progressive training strategy that first learns a strong language-conditioned motion prior and then extends the model to multimodal conditioning under available modality combinations, using modality-specific null tokens for missing or dropped conditions. During the multimodal stage, we formulate sparse visual frames as temporally localized key-state observations for motion infilling rather than estimated 3D key-pose references; by providing body-state, scene, and interaction evidence, these frames condition completion and enrich the generator with transferable visual-motion priors that further benefit Language-to-Motion (L2M) generation after L2M fine-tuning. Finally, a disentangled Motion Understanding Head predicts discrete motion descriptions as an auxiliary captioning task, regularizing the shared motion representation with semantic supervision while enabling strong motion captioning capability.

To support multimodal learning and evaluation, we integrate and automatically annotate 21 high-quality motion datasets into a large-scale benchmark, Mo900H, spanning over 900 hours of human motion, 216.4k language annotations, 43.1k videos, and 7.8k object motions. Extensive experiments on HumanML3D and Mo900H demonstrate the versatility of our framework: in language-to-motion generation, our method outperforms SOTA methods by 21.1% and 48.3% on HumanML3D and Mo900H, respectively, and improves motion captioning without caption-specific fine-tuning. Moreover, our framework turns visual observations and object states into effective grounding signals within a unified generative model, enabling intuitive motion specification while strengthening shared representations and standard language-to-motion tasks.

2 Related Work

Language-to-motion generation Language-conditioned human motion synthesis has progressed from contrastive and variational formulations to diffusion, discrete-token, and masked or autoregressive transformer models. Early datasets and methods established paired motion-language generation as a standard benchmark [10, 39, 48], while diffusion models improved diversity and temporal realism [49, 5, 64]. Meanwhile, motion captioning and alignment studies have shown that motion can be modeled as a language-like modality, enabling tokenized text-motion modeling and motion-language models for generation, captioning, retrieval, and understanding [11, 18, 32, 51, 53, 71], as well as motion-aware embeddings and shared multimodal spaces [24, 4, 21]. Building on these

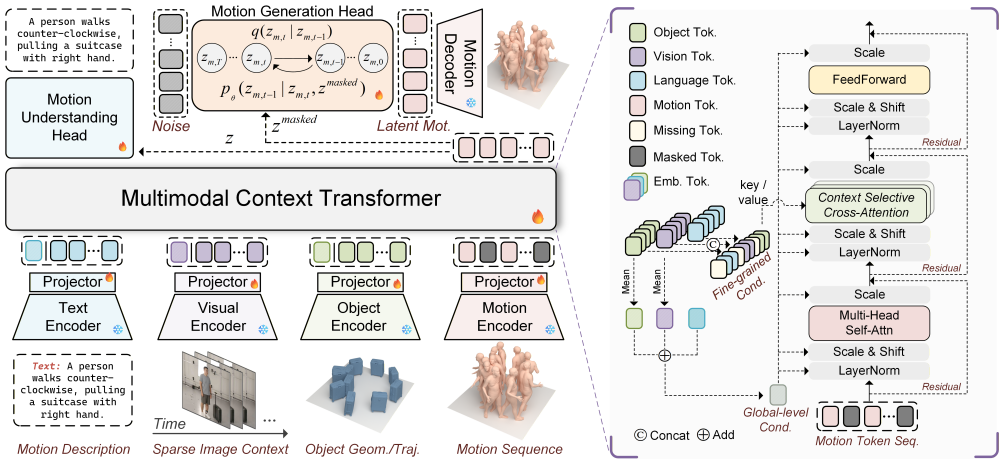


Figure 1: Overview of the proposed framework. Each available modality is first encoded into latent tokens, the Multimodal Context Transformer integrates global and fine-grained conditions, the Motion Generation Head reconstructs the motion sequence, and the Motion Understanding Head predicts a language description from the shared representation.

insights, recent generators use discrete codebooks or masked token prediction to improve efficiency, editability, and alignment [60, 12, 40, 50, 17, 36, 44], while scaling and pretraining strengthen open-vocabulary priors [24, 31, 7, 69]. Yet these methods rely on language as the primary specification, leaving spatiotemporal human and interaction configurations largely implicit; our work retains language-to-motion as a special case while leveraging additional context to resolve ambiguity.

Multimodal and object-grounded motion generation A complementary line of work augments motion generation with non-language constraints such as sparse tracking signals, controllable joints, trajectories, or keyframes [6, 55, 43, 68]. Image- or scene-conditioned methods use visual context to adapt motion to an environment [16, 37], while human-scene and human-object interaction models condition on 3D scenes, object geometry, contact, or trajectories [15, 52, 47, 20, 22, 62, 33, 30, 2]. Related motion forecasting studies model multi-scale correlations and past-movement representations [70, 45] or aggregate multi-granular context: MGTR [9] predicts traffic trajectories from agent histories, maps, and LiDAR via proximity search, while HUMOF [46] and SAST [38] forecast 3D human motion from observed histories and scene or social context. Recent unified frameworks further connect motion generation, editing, estimation, and completion in a shared model [14, 23, 25, 57], with OmniMotion-X [57] unifying multimodal controls via condition-token prefixing. Building on these advances, our framework accommodates flexible combinations of optional sparse RGB key states and 3D object geometry/trajectories under incomplete supervision, unifying language-only generation, vision-guided completion, object-conditioned generation, and motion in-betweening by decoupling global modulation from query-specific variable-budget context selection.

3 Method

We address multimodal context-aware motion generation, where motion is synthesized from language and an optional subset of context signals, including sparse visual observations, objects, and visible motion fragments. Figure 1 summarizes the framework: each modality is encoded into latent tokens, the Multimodal Context Transformer integrates global and fine-grained conditions, and two task-specific heads decode the shared representation for motion generation and understanding.

3.1 Modality-Specific Encoder

Given a motion sequence $X_m \in \mathbb{R}^{l_m \times c_m}$, we encode it into a compact latent representation $Z_m = f_m(X_m) \in \mathbb{R}^{l'_m \times d_m}$ using a temporal convolutional auto-encoder, where $l'_m \leq l_m$. A symmetric decoder reconstructs the motion as $\hat{X}_m = g_m(Z_m)$, and the auto-encoder is trained with a smooth L1 loss, $\mathcal{L}_{\text{rec}} = \text{SmoothL1}(\hat{X}_m, X_m)$.

For language, a frozen CLIP text encoder produces token embeddings $Z_t = f_t(X_t) \in \mathbb{R}^{l_t \times d_t}$, with the [CLS] token Z_t^g used as the global language feature. For vision, frames sampled at 1 fps are encoded independently as $z_v^p = f_v(x_v^p) \in \mathbb{R}^{d_v}$, and mean pooled to obtain the global visual representation Z_v^g. For object context, we use geometry, initial position, and trajectory information. The object geometry Z_o^{geo} is encoded from its point cloud by PointNet++ [42]. For each frame, the 3D translation and axis-angle rotation are concatenated into a 6D object trajectory $X_o^{\text{traj}} \in \mathbb{R}^{l_o \times 6}$. The first-frame translation is taken as the object initial position $x_o^{\text{init}} \in \mathbb{R}^3$ and projected by a linear layer to obtain Z_o^{init}, while the full trajectory is encoded by a 1D convolutional encoder as $Z_o^{\text{traj}} = f_o^{\text{traj}}(X_o^{\text{traj}})$.

3.2 Multimodal Context Transformer

The Multimodal Context Transformer addresses two requirements of context-aware generation: 1) integrating conditions with different granularities; and 2) letting each motion token focus on condition tokens relevant to its current action phase. Each MCT layer thus applies global modulation, self-attention over motion tokens, and Context-Selective Cross-Attention.

Global-level Modulation We adopt Adaptive LayerNorm modulation (AdaLN). The global condition features of each modality, Z_t^g, Z_v^g, Z_o^{geo}, and Z_o^{init}, are mapped by modality-specific projections into a common d_c-dimensional conditioning embedding and summed to obtain a conditioning vector C^g. This additive form preserves the condition dimension, allows missing modalities to be zero-filled, and satisfies the single-vector input requirement of AdaLN. The aggregated global context is mapped to modulation coefficients $(\alpha_c, \beta_c, \gamma_c) = W_{\text{ada}}(C^g), \alpha_c, \beta_c, \gamma_c \in \mathbb{R}^{d_m}$. Given normalized motion tokens $\bar{Z}_m = \text{LN}(Z_m)$, modulation and gated residual connections are applied as

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

where $h(\cdot)$ denotes the corresponding MCT sub-layer (Self-Attention, Context-Selective Cross-Attention, or Feed-Forward).

Context-Selective Cross-Attention. To enable fine-grained alignment between motion tokens and multimodal conditions, we perform context-selective cross-attention over concatenated conditions $C^{\text{tok}} = [Z_v; Z_t; Z_o^{\text{traj}}] \in \mathbb{R}^{L_c \times d}$, where features from all modalities are projected to dimension d, and missing modalities are represented by learnable missing tokens. Given motion tokens $Z_m \in \mathbb{R}^{l_m \times d}$, motion queries attend to condition tokens through multi-head cross-attention.

The module contains multiple context-selection branches with different sparsity levels. For branch b, let τ_b, $k_{\min}^b$, and $k_{\max}^b$ denote the retained attention-mass threshold and the minimum/maximum number of selected condition tokens. For head h, let A_h be the masked attention distribution over valid condition tokens, and let $A_{h,i,(r)}$ be the r-th largest attention weight for motion token i. The dynamic context size is computed as

$$k_{b,h,i}^{\text{dyn}} = \text{clip}\Big(\min\big\{k \mid \textstyle\sum_{r=1}^{k} A_{h,i,(r)} \geq \tau_b\big\},\ k_{\min}^b,\ \min(k_{\max}^b, L_c)\Big). \tag{2}$$

Let $\Omega_{b,h,i}$ denote the corresponding top-$k_{b,h,i}^{\text{dyn}}$ condition-token indices. Attention is then restricted to this subset and re-normalized:

$$\widetilde{A}_{b,h,i,j} = \frac{A_{h,i,j}\mathbf{1}\{j \in \Omega_{b,h,i}\}}{\sum_{j' \in \Omega_{b,h,i}} A_{h,i,j'} + \epsilon}, \qquad Y_b = \text{Concat}_{h=1}^{H}(\widetilde{A}_{b,h} V_h) W_b^O. \tag{3}$$

Finally, branch outputs are aggregated through condition-dependent residual gates to obtain the context-enhanced motion representation. By assigning branches to different attention-mass thresholds and context-size ranges, the module captures both compact and broad multimodal dependencies while dynamically adapting the amount of contextual information used by each motion token.

3.3 Disentangled Generation Heads

Motion understanding and motion generation operate in different output spaces: the former predicts a discrete natural-language description, whereas the latter synthesizes a continuous motion trajectory. We therefore decode the shared representation with two separate heads rather than a single shared decoder.

Motion Understanding Head (MUH) The MUH is a lightweight T5-style decoder that autoregressively generates a motion description conditioned on the shared MCT representation, with all language-condition tokens dropped and replaced by the language null token. It both produces captions at inference time and regularizes the latent space to preserve high-level action semantics.

Table 1: Quantitative results on Mo900H. The best results are in bold, and the second-best results are underlined. * indicates the default setting. All compared methods are retrained on Mo900H with 100 epochs. The evaluator is also trained on Mo900H following the MARDM protocol [36].

Methods	R-Precision Top-1 ↑	FID ↓	Matching ↓	Diversity ↑	Jitter Degree ↓	Dynamic Degree ↑	Foot Floating ↓	Foot Sliding ↓
MMM [40]	$0.545^{\pm.001}$	$0.169^{\pm.006}$	$3.964^{\pm.014}$	$17.201^{\pm.128}$	0.0089	<u>0.0251</u>	0.1916	0.0109
MoMask [12]	$0.549^{\pm.001}$	$0.151^{\pm.005}$	$3.845^{\pm.012}$	$17.284^{\pm.121}$	0.0104	0.0222	0.2073	0.0085
MARDM-DDPM [36]	$0.512^{\pm.001}$	$0.212^{\pm.008}$	$3.793^{\pm.011}$	$17.337^{\pm.134}$	0.0316	0.0193	0.2898	0.0135
MARDM-SiT [36]	$0.543^{\pm.001}$	$0.161^{\pm.005}$	$3.928^{\pm.013}$	$\mathbf{17.718^{\pm.127}}$	0.0093	0.0218	0.1861	0.0064
MG-MotionLLM [53]	$0.498^{\pm.002}$	$0.276^{\pm.010}$	$4.123^{\pm.015}$	$17.092^{\pm.139}$	0.0115	0.0187	0.2018	0.0178
MoScale [69]	$0.570^{\pm.001}$	$0.155^{\pm.007}$	$3.806^{\pm.010}$	$17.011^{\pm.151}$	0.0097	**0.0283**	0.1714	0.0058
MoMask++ [13]	$0.565^{\pm.001}$	$0.395^{\pm.013}$	$3.817^{\pm.010}$	$16.861^{\pm.196}$	0.0102	0.0223	0.1822	0.0064
Ours (1-stage)	$0.584^{\pm.002}$	$0.165^{\pm.006}$	$3.803^{\pm.013}$	<u>$17.510^{\pm.126}$</u>	0.0079	0.0221	0.1509	0.0053
Ours (1-stage w/o V & O)	$0.604^{\pm.002}$	$0.105^{\pm.005}$	$3.528^{\pm.011}$	$17.266^{\pm.130}$	<u>0.0069</u>	0.0230	0.1671	0.0057
Ours* (2-stage)	<u>$0.615^{\pm.002}$</u>	<u>$0.078^{\pm.004}$</u>	<u>$3.457^{\pm.013}$</u>	$17.254^{\pm.132}$	**0.0064**	0.0227	<u>0.1492</u>	**0.0046**
Ours (2-stage w/o und.)	$0.595^{\pm.002}$	$0.099^{\pm.005}$	$3.649^{\pm.014}$	$17.056^{\pm.171}$	0.0071	0.0240	0.1837	0.0062
Ours (ft. on L2M)	$\mathbf{0.620^{\pm.002}}$	$\mathbf{0.067^{\pm.003}}$	$\mathbf{3.400^{\pm.010}}$	$17.495^{\pm.116}$	**0.0064**	0.0229	**0.1471**	<u>0.0047</u>
Ours (ft. on L2M, w/o V)	$0.590^{\pm.002}$	$0.111^{\pm.005}$	$3.703^{\pm.012}$	$17.229^{\pm.130}$	<u>0.0069</u>	0.0234	0.1522	0.0051

Motion Generation Head (MGH) Since motion is continuous, we adopt a continuous-time interpolant model in the Motion Generation Head (MGH) following SiT [34, 36]. Given the masked embedding H^{mask}, MGH learns a conditional velocity field $\mathbf{v}_\theta(\cdot, t, H^{\text{mask}})$ mapping a Gaussian noise latent to the ground-truth motion latent $Z_{m,0}$. At inference, we sample $Z_{m,T} \sim \mathcal{N}(0, I)$ and integrate the velocity field via an ODE or SDE solver to obtain $Z_{m,0}$, which the motion decoder maps to $\hat{X}_m$.

4 Cross-Modal Generative Training

Since the available modality set varies across datasets, we train the model progressively with modality dropout for robustness to missing inputs, replacing unavailable or dropped conditions with modality-specific null tokens under incomplete supervision.

Stage I: Language-to-Motion Pretraining. We first train a language-conditioned motion generator using all motion-language pairs. Motion tokens are randomly masked, and the model reconstructs the complete motion latent sequence from visible motion tokens and the language description, establishing a stable language-to-motion prior before multimodal training.

Stage II: Multimodal Generative Training. We then activate the additional conditioning signals available in the data and continue training under unimodal and multimodal settings. Because vision and object conditions do not co-occur in Mo900H, Stage II uses all observed modality combinations except the unavailable vision-object pair, handling flexible multimodal combinations within a unified framework. For samples with visual input, we construct a vision-guided completion task by masking a contiguous motion segment and reconstructing it from aligned visual frames (optionally with language), encouraging the model to use visual frames as temporally localized and spatially controlled key states rather than generic scene features.

Auxiliary Motion Understanding Objective. In Stage II, the Motion Understanding Head predicts a language description from the shared motion-conditioned representation, encouraging semantic consistency between the latent motion and language spaces within a unified model.

Training Objective. The overall objective combines a continuous motion generation loss and an autoregressive Motion Understanding Head loss:

$$\mathcal{L} = \lambda_{\text{motion}} \mathbb{E}_{t,\boldsymbol{\epsilon}} \Big[\big\| \mathbf{v}_\theta(Z_{m,t}, t, H^{\text{mask}}) - (\dot{\alpha}_t Z_{m,0} + \dot{\sigma}_t \boldsymbol{\epsilon}) \big\|_2^2 \Big] + \lambda_{\text{und}} \mathbb{E}_{(X_m, Y)} \Big[-\textstyle\sum_{i=1}^{T_y} \log P_\phi(y_i \mid y_{<i}, H_m^{\text{cap}}) \Big], \tag{4}$$

where $Z_{m,0}$ is the clean motion latent and $Z_{m,t} = \alpha_t Z_{m,0} + \sigma_t \boldsymbol{\epsilon}$ is its noisy interpolation at timestep t. H^{mask} denotes the MCT representation conditioned on visible motion tokens and available modalities, and H_m^{cap} is the captioning representation with all language-condition tokens dropped. The first term trains MGH to predict the velocity target, while the second trains MUH to generate the description $Y = \{y_i\}_{i=1}^{T_y}$.

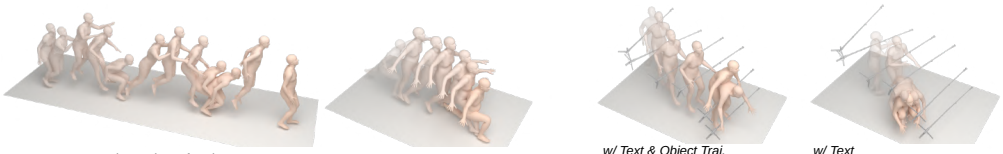


Figure 2: Visualization of language-to-motion and interaction-aware generation. Object geometry and trajectory provide spatiotemporal constraints for interaction motions.

5 Experiments

We evaluated the framework on HumanML3D and Mo900H from four aspects: language-to-motion generation and captioning, zero-shot motion in-betweening, whether sparse visual frames improve ambiguous motion completion, and whether object context enhances interaction-centric generation.

5.1 Integrated Motion Dataset

Motion Dataset We curated and standardized 21 high-quality motion datasets (spanning single-person activities and human-object interactions) into a 22-joint, 30 fps format capped at 20s. For visually paired datasets lacking language annotations, we generated and manually filtered captions using GPT-5-mini, substituting absent RGB videos with rendered mesh sequences (all downsampled to 1 fps). The resulting Mo900H comprises 900 hours of motion, 216.4k descriptions, 43.1k videos, and 7.8k object motions. It is organized by modality into Mo900H-M (all motions), Mo900H-MV (motion-vision), Mo900H-MLV (motion-language-vision), Mo900H-ML (motion-language), and Mo900H-MLO (motion-language-object), where M, L, V, and O denote motion, language, vision, and object. We also evaluate on HumanML3D [10] following [36] (Sec. A.1).

Table 2: Quantitative results on HumanML3D. All the results are obtained following MARDM [36].

Methods	R-Prec. ↑	FID ↓	Match. ↓	CLIP ↑
T2M-GPT [60]	$0.470^{\pm.003}$	$0.335^{\pm.003}$	$3.505^{\pm.017}$	$0.607^{\pm.005}$
ReMoDiffuse [63]	$0.468^{\pm.003}$	$0.883^{\pm.021}$	$3.414^{\pm.020}$	$0.621^{\pm.003}$
MDM-50Step [49]	$0.440^{\pm.007}$	$0.518^{\pm.032}$	$3.640^{\pm.028}$	$0.578^{\pm.003}$
MLD [5]	$0.461^{\pm.004}$	$0.431^{\pm.014}$	$3.445^{\pm.019}$	$0.610^{\pm.003}$
MMM [40]	$0.487^{\pm.003}$	$0.132^{\pm.004}$	$3.359^{\pm.009}$	$0.635^{\pm.003}$
MoMask [12]	$0.469^{\pm.004}$	$0.116^{\pm.006}$	$3.353^{\pm.010}$	$0.637^{\pm.003}$
MotionDiffuse [64]	$0.450^{\pm.006}$	$0.778^{\pm.005}$	$3.490^{\pm.023}$	$0.606^{\pm.004}$
MARDM-DDPM [36]	$0.492^{\pm.006}$	$0.116^{\pm.004}$	$3.349^{\pm.010}$	$0.637^{\pm.005}$
MARDM-SiT [36]	$0.500^{\pm.004}$	$0.114^{\pm.007}$	$3.270^{\pm.009}$	$0.642^{\pm.002}$
MotionAgent [54]	$0.485^{\pm.003}$	$0.202^{\pm.009}$	$3.327^{\pm.009}$	$0.634^{\pm.003}$
MoScale [69]	$0.520^{\pm.003}$	$\underline{0.090}^{\pm.003}$	$3.251^{\pm.008}$	$0.644^{\pm.001}$
MoMask++ [13]	$0.490^{\pm.003}$	$0.109^{\pm.005}$	$3.350^{\pm.008}$	$0.636^{\pm.002}$
Ours *w/o Und.*	$\underline{0.533}^{\pm.012}$	$0.108^{\pm.023}$	$\underline{3.078}^{\pm.037}$	$\underline{0.658}^{\pm.001}$
Ours	$\mathbf{0.544}^{\pm.002}$	$\mathbf{0.071}^{\pm.004}$	$\mathbf{2.995}^{\pm.010}$	$\mathbf{0.669}^{\pm.001}$

Motion Representation We adopt a compact motion representation [36] that removes redundant features such as 6D rotations to mitigate distribution mismatch. At time step i, the motion is represented as $\mathbf{x}^i_{\text{motion}} = [\dot{r}_{\text{angular}}, \dot{\mathbf{r}}^{xz}_{\text{linear}}, r_{\text{height}}, \mathbf{j}_{\text{position}}]$, where $\dot{r}_{\text{angular}}$ is the root angular velocity, $\dot{\mathbf{r}}^{xz}_{\text{linear}}$ is the root linear velocity in the XZ-plane, r_{height} is the root height, and $\mathbf{j}_{\text{position}} \in \mathbb{R}^{3(N_{\text{joint}}-1)}$ denotes local joint positions, capturing essential kinematic information for motion generation.

5.2 Experiment Settings

All experiments are conducted on NVIDIA RTX 4090 and RTX PRO 6000 GPUs with a batch size of 64 using the Adam optimizer (lr 2×10^{-4}, 2000-step warm-up). We report Physics-based Motion Quality, R-Precision, FID, Matching, Diversity, CLIP, BLEU, ROUGE, BERTScore, and task-specific metrics. For HumanML3D (motion-language pairs only), we use single-stage training for 500 epochs. For Mo900H, we use two-stage training: Stage I trains language-to-motion generation for 60 epochs, and Stage II unlocks all multimodal conditions and adds the Motion Understanding Head objective for 20 epochs, while the Motion Generation Head is optimized throughout both stages. The MCT uses 384 hidden channels (1 layer on HumanML3D, 2 layers on Mo900H). Context-Selective Cross-Attention uses two branches: an adaptive sparse branch with $(\tau, k_{\min}, k_{\max}) = (0.8, 1, 6)$ and a full-attention branch with $(\tau, k_{\min}, k_{\max}) = (1.0, 0, L_c)$. The Motion Understanding Head is a 3-layer T5-style decoder, and the Motion Generation Head is a 10-layer MLP diffusion model with 1280 hidden channels.

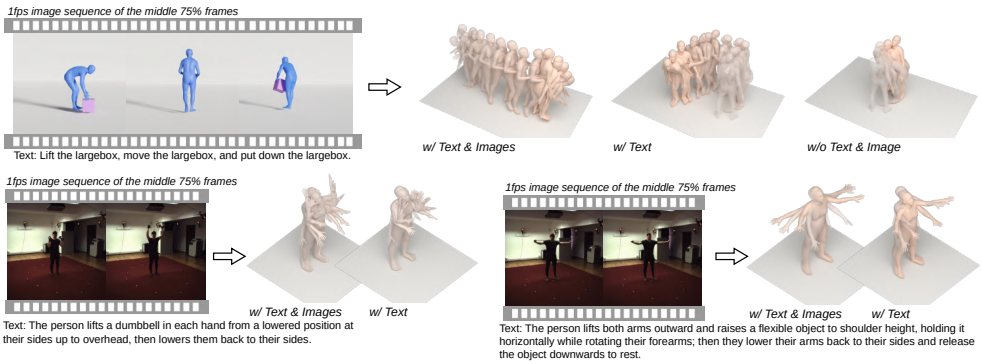


Figure 3: Effect of the visual modality on motion generation. We mask the middle 75% of the motion sequence and perform completion conditioned on 1 fps visual observations with text, or text alone.

5.3 Evaluation of Our Methods

Language-to-Motion Generation & Motion Captioning We first evaluate language-to-motion generation on Mo900H (Table 1). Here, 1-stage denotes direct multimodal generative training without the progressive two-stage schedule, while *1-stage w/o V & O* removes visual and object conditions and trains only on motion-language pairs. Direct one-stage multimodal training underperforms the language-only variant, suggesting negative transfer across heterogeneous context tasks with different supervision densities and mask patterns. *Ours* (2-stage)* is our default setting, which first stabilizes the language-motion prior and then introduces multimodal conditions, avoiding this degradation and improving FID from 0.165 to 0.078 and Top-1 R-Precision from 0.584 to 0.615. This confirms that multimodal grounding is most effective after learning a stable language-conditioned motion prior.

Removing the Motion Understanding Head objective in Stage II (*Ours (2-stage w/o und.)*) degrades performance, increasing FID from 0.078 to 0.099 and Matching from 3.457 to 3.649, verifying that the Motion Understanding Head provides semantic supervision that regularizes the motion latent space for both captioning and generation. Figure 2 (left) shows qualitative language-to-motion results, generating plausible and semantically consistent motions even for less common prompts such as long jumps and lunges. On HumanML3D (motion-language pairs only), we jointly optimize the Motion Generation and Understanding Head losses in a single stage, achieving clear improvements over state-of-the-art methods in Table 2 (full table in Sec. A.5).

Table 3: Comparisons of motion captioning results on HumanML3D and Mo900H-ML.

Dataset	Method	BLEU@1↑	BLEU@4↑	ROUGE↑	BERTScore↑
HumanML3D	TM2T [11]	48.90	8.27	38.1	32.2
	MotionGPT [18]	48.20	12.47	37.4	32.4
	MotionChain [19]	48.10	**12.56**	33.9	36.9
	MotionGPT3 [71]	51.06	8.43	38.7	32.0
	MG-MotionLLM [53]	–	8.06	–	36.7
	Ours	**53.44**	10.09	**41.3**	**41.2**
Mo900H	TM2T [11]	31.49	16.71	44.77	37.39
	MG-MotionLLM [53]	37.67	19.34	50.18	40.05
	Ours	**45.00**	**22.54**	**54.73**	**41.32**

Table 4: Motion in-between results on Mo900H-ML with and without language conditions.

Setting	Method	Top-1↑	FID↓	Matching↓	Diversity↑
w/o lang.	MARDM [36]	0.636	0.157	3.783	16.991
	Ours	**0.720**	**0.074**	**2.576**	**17.806**
w/ lang.	MARDM [36]	0.677	0.053	3.243	16.920
	Ours	**0.743**	**0.012**	**2.293**	**17.545**

For motion captioning, the Motion Understanding Head is jointly trained in Stage II to generate natural-language descriptions from motion-only representations without caption-specific fine-tuning. As shown in Table 3, our method achieves the best BLEU@1, ROUGE, and BERTScore on HumanML3D and outperforms TM2T and MG-MotionLLM across all metrics on Mo900H, showing that the auxiliary head provides strong captioning ability while regularizing the shared motion representation.

Motion In-between We evaluate our method on the motion in-between task, generating plausible infix motion from partial motion contexts on HumanML3D and Mo900H compared with MARDM [36] under two settings: (i) in-between with language and visible motion fragments, and (ii) in-between

with motion fragments alone. Without task-specific fine-tuning, our method consistently outperforms baselines in Table 4.

Vision-Conditioned Motion Completion. We evaluate whether sparse visual observations provide controllable conditions for motion completion (Table 5). Vision conditioning consistently improves completion quality, especially under ambiguous random masking where motion tokens are removed at arbitrary positions and lengths: on the MV split, adding vision reduces FID from 0.180 to 0.066 for in-between completion and from 1.649 to 0.203 for random completion, showing that visual cues supply pose and scene information difficult to infer from partial motion alone.

On the MLV split, language and vision are complementary: language provides semantic intent, while vision reduces spatiotemporal ambiguity. For random completion, language alone reduces FID from 1.620 to 0.116, and adding vision further improves FID to 0.085 and Top-1 to 0.466; for in-between completion, the full language-vision condition likewise achieves the lowest FID. Figure 3 shows that visual key frames guide more faithful completion than text alone: when prompted to lift, move, and put down a box, visual context supplies object scale, initial position, and movement direction that language alone leaves underspecified.

Table 5: Vision-augmented motion completion results.

Set	Condition	Pattern	FID↓	MM.↑	Top-1↑	Match.↓
MV	Real (GT)	–	–	–	0.359	7.282
	w/o vision	In-between	0.180	0.800	0.342	7.529
	w/ vision	In-between	0.066	1.237	0.349	7.312
	w/o vision	Random	1.649	7.691	0.256	9.624
	w/ vision	Random	0.203	4.005	0.302	7.713
	w/ vision (0.5 fps)	Random	0.293	5.331	0.300	7.727
	w/ vision (2 fps)	Random	0.173	3.730	0.311	7.710
MLV	Real (GT)	–	–	–	0.472	2.341
	w/o language+vision	In-between	0.061	0.487	0.446	2.562
	w/ language	In-between	0.039	0.446	0.471	2.312
	w/ vision	In-between	0.037	0.502	0.459	2.460
	w/ language+vision	In-between	0.036	0.423	0.474	2.306
	w/o language+vision	Random	1.620	7.751	0.358	5.655
	w/ language	Random	0.116	1.952	0.453	2.407
	w/ vision	Random	0.290	1.997	0.396	3.430
	w/ language+vision	Random	0.085	1.874	0.466	2.392
	w/ vision (0.5 fps)	Random	0.359	2.429	0.373	3.269
	w/ vision (2 fps)	Random	0.241	1.709	0.382	3.255

Unlike approaches that condition on a single static scene image [16] or canonicalized 3D key poses with planned trajectories [68], our framework completes partial motion directly from temporally aligned sparse RGB frames. To examine whether end-to-end visual conditioning provides benefits beyond explicit per-frame pose recovery, we build a pose-estimation baseline on Mo900H-MLV using SAM 3D Body [59] at the same frames and timestamps. Since monocularly estimated 3D poses reside in camera-specific coordinate systems rather than the canonical motion space, we encode them as structured 3D motion conditions at the corresponding timestamps in place of visual tokens. As shown in Table 6, while estimated 3D poses improve FID over language alone (0.116 → 0.093), direct visual conditioning achieves superior FID (0.085), R-Precision across all top-k, Matching (2.392 vs. 2.459), Diversity (16.939 vs. 16.817), and Multimodality (1.874 vs. 1.580), demonstrating that raw visual frames provide richer scene and interaction cues than isolated 3D body pose estimates.

Table 6: Comparison with estimated 3D body conditioning on Mo900H-MLV (random completion).

Condition	FID↓	Top-1↑	Top-2↑	Top-3↑	Match.↓	Div.↑	MM.↑
w/ lang.	0.116	0.453	0.675	0.802	2.407	16.913	**1.952**
w/ lang.+3D body [59]	0.093	0.455	0.681	0.812	2.459	16.817	1.580
w/ lang.+visual (Ours)	**0.085**	**0.466**	**0.692**	**0.827**	**2.392**	**16.939**	1.874

Table 7: Generation with object contexts.

Condition	FID↓	MM.↑	Top-1↑	Match.↓	Contact F1↑
Real (GT)	–	–	0.345	1.876	–
Language	1.570	**2.436**	**0.340**	2.334	0.145
Language + Obj. Init.	0.168	2.245	0.329	2.217	0.173
Language + Obj. Traj.	**0.151**	2.360	0.329	**2.039**	**0.182**

Object-Conditioned Motion Generation. We further evaluate object-conditioned generation on Mo900H-MLO to assess interaction motion synthesis. As shown in Table 7, language-only generation obtains a high FID of 1.570, indicating that language alone is insufficient to determine precise spatial configurations of the human and object. Adding object initialization reduces FID to 0.168, while using the full object trajectory further improves FID to 0.151, Matching to 2.039, and Contact F1 to 0.182. Figure 2 shows that object geometry and trajectory provide spatiotemporal constraints for plausible interaction motions. Additional spatial interaction metrics (MPJPE, foot sliding, and contact precision, recall, and accuracy) and a comparison with InterAct [58] are provided in Sec. A.6.

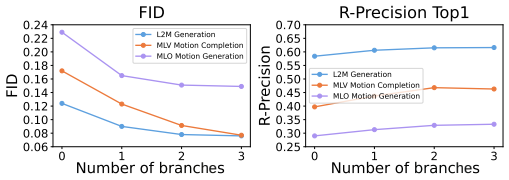


Figure 4: Ablation study of context-selective branches.

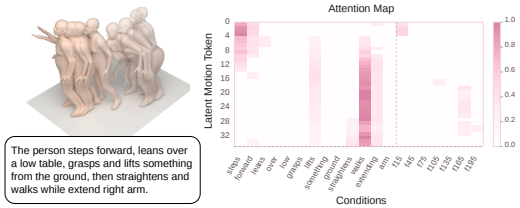


Figure 5: Visualization of context-selective attention map. `f*` denotes the image at the *th frame.

Table 8: Ablation of global modulation and Context-Selective Cross-Attention (CSCA) control variants on L2M generation and vision-guided completion (LV2M). All CSCA variants retain global modulation.

Conditioning Variant	L2M Generation						LV2M Completion					
	FID↓	Top-1↑	Top-2↑	Top-3↑	Match.↓	Div.↑	FID↓	Top-1↑	Top-2↑	Top-3↑	Match.↓	Div.↑
w/o Global Modulation	0.161	0.545	0.694	0.779	4.557	16.994	0.167	0.375	0.514	0.649	3.736	16.422
No CA	0.124	0.584	0.748	0.811	3.932	17.145	0.172	0.397	0.559	0.694	3.420	**17.005**
1× Dense CA	0.090	0.606	0.782	0.844	3.610	**17.301**	0.123	0.436	0.638	0.775	2.729	16.591
2× Dense CAs	0.103	0.605	0.780	0.833	3.618	17.206	0.144	0.431	0.631	0.768	2.841	16.832
Dense + Fixed Top-k	0.087	0.610	0.793	0.859	3.507	17.055	0.099	0.456	0.675	0.803	2.442	16.910
Dense + Global-Shared Top-k	0.095	0.605	0.774	0.836	3.801	16.993	0.119	0.440	0.643	0.787	2.601	16.787
Dense + Adaptive Sparse (Ours)	**0.078**	**0.615**	**0.802**	**0.879**	**3.457**	17.254	**0.085**	**0.466**	**0.692**	**0.827**	**2.392**	16.939

Table 9: Counterfactual context perturbations and partial corruption robustness on visual-guided completion (Mo900H-MLV, left) and object-conditioned generation (Mo900H-MLO, right). KS-MPJPE measures pose error at visual key-state timestamps, and HandObjDis measures wrist-to-object distance.

(a) Visual Context Perturbations (Mo900H-MLV)

Visual Perturbation	FID↓	Top-1↑	KS-MPJPE↓
None (unperturbed)	**0.036**	0.474	**0.216**
Cross-sample image swap	0.133	0.421	0.371
Shift by 1 visual-token pos.	0.045	0.473	0.237
Shift by 2 visual-token pos.	0.048	0.472	0.240
Noise on 20% image tokens	0.060	0.465	0.248
Drop 20% image tokens	0.037	**0.475**	0.229

(b) Object Context Perturbations (Mo900H-MLO)

Object Perturbation	FID↓	Top-1↑	HandObjDis↓
GT motion/object reference	–	–	0.248
None (unperturbed)	**0.151**	**0.329**	**0.453**
Object-feature swap	0.584	0.305	0.778
Object-trajectory swap	0.204	0.314	0.820
Noise on 20% traj. tokens	0.256	0.322	0.681
Drop 20% traj. tokens	0.179	0.327	0.650

5.4 Ablation Study

Effect of Vision-Conditioned Generative Training and FPS of Vision Input. Table 1 shows that including visual modality in multimodal generative training improves subsequent L2M fine-tuning: compared with the variant without vision, *Ours (ft. on L2M)* improves Top-1 from 0.590 to 0.620 and reduces FID from 0.111 to 0.067, indicating transferable fine-grained motion and interaction priors. Table 5 also shows that increasing visual sampling from 0.5 fps to 2 fps reduces random-completion FID on both MV (0.293 → 0.173) and MLV (0.359 → 0.241) while lowering MM, as denser visual key states impose stronger spatiotemporal constraints.

Effectiveness of Global Modulation and Context-Selective Cross-Attention. We vary the number of context-selection branches from 0 to 3: no cross-attention, a full-attention branch, our default two-branch design adding an adaptive sparse branch with $(\tau, k_{\min}, k_{\max}) = (0.8, 1, 6)$, and a three-branch variant adding a medium-range branch ($k_{\min} = 6, k_{\max} = 10$). As shown in Fig. 4, context-selective branches consistently improve FID and R-Precision across L2M generation, MLV completion, and MLO generation, with two branches capturing most of the gain. The attention map of the adaptive sparse branch in Fig. 5 further shows phase-dependent selection: motion tokens attend to different text phrases for semantic guidance and image frames as sparse visual body-state anchors as the action progresses.

We further analyze the interplay between sequence-level global modulation and token-level context selection on L2M generation and vision-guided completion (LV2M) in Table 8. Removing the

mean-pooled AdaLN pathway (*w/o Global Modulation*) while retaining CSCA substantially degrades FID on both L2M ($0.078 \rightarrow 0.161$) and LV2M ($0.085 \rightarrow 0.167$), confirming that global modulation and token-level cross-attention are complementary. We also evaluate three two-branch controls covering branch capacity ($2\times$ *Dense CAs*), fixed sparsity (*Dense + Fixed Top-k*), and query-shared selection (*Dense + Global-Shared Top-k*). Duplicating dense cross-attention ($2\times$ *Dense CAs*) underperforms $1\times$ *Dense CA*, indicating that the gain stems from selective sparsity rather than extra branch parameters. Fixed top-k improves over dense attention by filtering low-relevance tokens, whereas sharing a global top-k subset across queries (*Dense + Global-Shared Top-k*) degrades LV2M FID (0.119 vs. 0.085), underscoring query-specific selection across action phases. Finally, our adaptive sparse branch (*Dense + Adaptive Sparse*) outperforms fixed top-k across all fidelity and alignment metrics by dynamically adjusting the active token budget per query.

Counterfactual Grounding and Modality Robustness. To evaluate how the generator utilizes spatiotemporal cues from visual and object contexts and assess its sensitivity to partial corruption, we conduct counterfactual perturbation experiments in Table 9. For vision-guided completion (Table 9a), keeping language and visible motion fixed while replacing images with frames from another sample (*Cross-sample image swap*) degrades all metrics (FID $0.036 \rightarrow 0.133$, KS-MPJPE $0.216 \rightarrow 0.371$). Conversely, small temporal shifts (1 or 2 positions) and corrupting or dropping 20% of image tokens cause minimal change (e.g., FID 0.037, KS-MPJPE 0.229 under 20% drop), showing strong resilience to partial visual noise and missing frames.

For object-conditioned generation (Table 9b), swapping object geometry (*Object-feature swap*) strongly degrades FID ($0.151 \rightarrow 0.584$), whereas swapping the object trajectory (*Object-trajectory swap*) causes the largest increase in HandObjDis ($0.453 \rightarrow 0.820$). These distinct failure modes verify that our model leverages both 3D object geometry and trajectory to satisfy spatial interaction constraints, while showing much smaller degradation under 20% trajectory noise or token dropout.

6 Conclusion

We present a unified framework for multimodal context-aware human motion generation, using language for high-level intent and visual, object, and partial-motion cues as complementary constraints. With global multimodal modulation, context-selective cross-attention, progressive training, and an auxiliary Motion Understanding Head, our model effectively integrates heterogeneous conditions while preserving a strong language-to-motion prior. Experiments on HumanML3D and Mo900H show improved performance in motion generation, captioning, vision-conditioned completion, and object-conditioned interaction synthesis, demonstrating the value of multimodal context in reducing ambiguity for motion generation.

Acknowledgments and Disclosure of Funding

This work is partially supported by the Guangzhou-HKUST(GZ) Joint Funding Program (2025A03J3656) and partially supported by the Guangdong Basic and Applied Basic Research Foundation (2025A1515110155).

Competing Interests. The authors declare that they have no competing financial interests relevant to this work.

References

[1] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. Circle: Capture in rich contextual environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21211–21221, 2023.

[2] Xinhao Cai, Minghang Zheng, Xin Jin, and Yang Liu. InteractMove: Text-controlled human-object interaction generation in 3d scenes with movable objects. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9491–9499, 2025.

[3] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei

Yang, and Ziwei Liu. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *17th European Conference on Computer Vision*, pages 557–577. Springer, 2022.

[4] Changan Chen, Juze Zhang, Shrinidhi K Lakshmikanth, Yusu Fang, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, and Ehsan Adeli. The language of motion: Unifying verbal and non-verbal language of 3d human motion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6200–6211, 2025.

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

[6] Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 481–490, 2023.

[7] Ke Fan, Shunlin Lu, Minyue Dai, Runyi Yu, Lixing Xiao, Zhiyang Dou, Junting Dong, Lizhuang Ma, and Jingbo Wang. Go to zero: Towards zero-shot motion generation with million-scale data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13336–13348, 2025.

[8] Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9919–9928, 2021.

[9] Yiqian Gan, Hao Xiao, Yizhe Zhao, Ethan Zhang, Zhe Huang, Xin Ye, and Lingting Ge. MGTR: Multi-granular transformer for motion prediction with LiDAR. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16230–16236. IEEE, 2024.

[10] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, June 2022.

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

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

[13] Chuan Guo, Inwoo Hwang, Jian Wang, and Bing Zhou. Snapmogen: Human motion generation from expressive texts. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[14] Ziyan Guo, Zeyu Hu, De Wen Soh, and Na Zhao. Motionlab: Unified human motion generation and editing via the motion-condition-motion paradigm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13869–13879, 2025.

[15] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019.

[16] Hsin-Ping Huang, Yang Zhou, Jui-Hsien Wang, Difan Liu, Feng Liu, Ming-Hsuan Yang, and Zhan Xu. Move-in-2d: 2d-conditioned human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22766–22775, 2025.

[17] Minjae Jeong, Yechan Hwang, Jaejin Lee, Sungyoon Jung, and Won Hwa Kim. HGM³: Hierarchical generative masked motion modeling with hard token mining. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IEul1M5pyk.

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

[19] Biao Jiang, Xin Chen, Chi Zhang, Fukun Yin, Zhuoyuan Li, Gang Yu, and Jiayuan Fan. Motionchain: Conversational motion controllers via multimodal prompts. In *European Conference on Computer Vision*, pages 54–74. Springer, 2024.

[20] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9365–9376, 2023.

[21] Kaleab A. Kinfu and Rene Vidal. Motionbind: Multi-modal human motion alignment for retrieval, recognition, and generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=sUjwDdyspc.

[22] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023.

[23] Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: A generalist model for human motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11766–11776, 2025.

[24] Zhe Li, Weihao Yuan, Yisheng HE, Lingteng Qiu, Shenhao Zhu, Xiaodong Gu, Weichao Shen, Yuan Dong, Zilong Dong, and Laurence Tianruo Yang. LaMP: Language-motion pretraining for motion generation, retrieval, and captioning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LYawG8YkPa.

[25] Zhe Li, Weihao Yuan, Weichao Shen, Siyu Zhu, Zilong Dong, and Chang Xu. Omnimotion: Multimodal motion generation with continuous masked autoregression. *arXiv preprint arXiv:2510.14954*, 2025.

[26] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36, 2023.

[27] Jing Lin, Ruisi Wang, Junzhe Lu, Ziqi Huang, Guorui Song, Ailing Zeng, Xian Liu, Chen Wei, Wanqi Yin, Qingping SUN, Zhongang Cai, Lei Yang, and Ziwei Liu. The quest for generalizable motion generation: Data, model, and evaluation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=KNke6Pkq4o.

[28] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European conference on computer vision*, pages 612–630. Springer, 2022.

[29] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1144–1154, 2024.

[30] Yun Liu, Chengwen Zhang, Ruofan Xing, Bingda Tang, Bowen Yang, and Li Yi. CORE4D: A 4d human-object-human interaction dataset for collaborative object rearrangement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1769–1782, 2025.

[31] Shunlin Lu, Jingbo Wang, Zeyu Lu, Ling-Hao Chen, Wenxun Dai, Junting Dong, Zhiyang Dou, Bo Dai, and Ruimao Zhang. Scamo: Exploring the scaling law in autoregressive motion generation model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27872–27882, 2025.

[32] Mingshuang Luo, Ruibing Hou, Zhuo Li, Hong Chang, Zimo Liu, Yaowei Wang, and Shiguang Shan. M ^{3} gpt: An advanced multimodal, multitask framework for motion comprehension and generation. *Advances in Neural Information Processing Systems*, 37:28051–28077, 2024.

[33] Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, et al. Himo: A new benchmark for full-body human interacting with multiple objects. In *European Conference on Computer Vision*, pages 300–318. Springer, 2024.

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

[35] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017.

[36] Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. 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)*, pages 27859–27871, June 2025.

[37] Zoltán Á. Milacski, Koichiro Niinuma, Ryosuke Kawamura, Fernando de la Torre, and László A. Jeni. Ghost: Grounded human motion generation with open vocabulary scene-and-text contexts. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 4108–4118, 2025.

[38] Felix B. Mueller, Julian Tanke, and Juergen Gall. Massively multi-person 3d human motion forecasting with scene context. In *European Conference on Computer Vision Workshops*, pages 154–170. Springer, 2024.

[39] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022.

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

[41] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 722–731, June 2021.

[42] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[43] Sigal Raab, Inbal Leibovitch, Guy Tevet, Moab Arar, Amit Haim Bermano, and Daniel Cohen-Or. Single motion diffusion. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=DrhZneqz4n.

[44] Junyu Shi, Lijiang Liu, Yong Sun, Zhiyuan Zhang, Jinni Zhou, and Qiang Nie. Genm3: Generative pretrained multi-path motion model for text conditional human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13129–13139, 2025.

[45] Junyu Shi, Haoting Wu, Zhiyuan Zhang, Lijiang Liu, Yong Sun, and Qiang Nie. MoReFun: Past-movement guided motion representation learning for future motion prediction and understanding. In *Proceedings of the 34th ACM International Conference on Multimedia*, 2026. doi: 10.1145/3767308.3836437.

[46] Caiyi Sun, Yujing Sun, Xiao Han, Zemin Yang, Jiawei Liu, Xinge Zhu, Siu-Ming Yiu, and Yuexin Ma. HUMOF: Human motion forecasting in interactive social scenes. In *The Fourteenth International Conference on Learning Representations*, 2026.

[47] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 581–600. Springer, 2020.

[48] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022.

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

[50] Yilin Wang, chuan guo, Yuxuan Mu, Muhammad Gohar Javed, Xinxin Zuo, Juwei Lu, Hai Jiang, and Li cheng. Motiondreamer: One-to-many motion synthesis with localized generative masked transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=d23EVDRJ6g.

[51] Yuan Wang, Di Huang, Yaqi Zhang, Wanli Ouyang, Jile Jiao, Xuetao Feng, Yan Zhou, Pengfei Wan, Shixiang Tang, and Dan Xu. Motiongpt-2: A general-purpose motion-language model for motion generation and understanding. *arXiv preprint arXiv:2410.21747*, 2024.

[52] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems*, 35:14959–14971, 2022.

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

[54] Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Motion-agent: A conversational framework for human motion generation with LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=AvOhBgsE5R.

[55] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gd0lAEtWso.

[56] Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, et al. Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19801–19811, 2024.

[57] Guowei Xu, Yuxuan Bian, Ailing Zeng, Zhuo Chen, Mingyi Shi, Shaoli Huang, Wen Li, Lixin Duan, and Qiang Xu. OmniMotion-X: Versatile multimodal whole-body motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pages 3641–3652, 2026.

[58] Sirui Xu, Dongting Li, Yucheng Zhang, Xiyan Xu, Qi Long, Ziyin Wang, Yunzhi Lu, Shuchang Dong, Hezi Jiang, Akshat Gupta, et al. Interact: Advancing large-scale versatile 3d human-object interaction generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7048–7060, 2025.

[59] Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jia-Wei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollár, and Kris Kitani. SAM 3D Body: Robust full-body human mesh recovery. *arXiv preprint arXiv:2602.15989*, 2026.

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

[61] Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8834–8845, 2023.

[62] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-mˆ 3: Capture multiple humans and objects interaction within contextual environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 516–526, 2024.

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

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

[65] Mingyuan Zhang, Daisheng Jin, Chenyang Gu, Fangzhou Hong, Zhongang Cai, Jingfang Huang, Chongzhi Zhang, Xinying Guo, Lei Yang, Ying He, and Ziwei Liu. Large motion model for unified multi-modal motion generation. In *Computer Vision – ECCV 2024*, pages 397–421, 2024.

[66] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European conference on computer vision*, pages 180–200. Springer, 2022.

[67] Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. I'm hoi: Inertia-aware monocular capture of 3d human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 729–741, June 2024.

[68] Yang Zhao, Yan Zhang, and Xubo Yang. IKMo: Image-keyframed motion generation with trajectory-pose conditioned motion diffusion model. *arXiv preprint arXiv:2505.21146*, 2025.

[69] Zhiwei Zheng, Shibo Jin, Lingjie Liu, and Mingmin Zhao. Next-scale autoregressive models for text-to-motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16376–16386, 2026.

[70] Jianqi Zhong, Junyu Shi, and Wenming Cao. Multiscale graph redefining: Correlation-based multiscale graph clustering network for human motion prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 37(10):5075–5089, 2026.

[71] Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. MotionGPT3: Human motion as a second modality. In *The Fourteenth International Conference on Learning Representations*, 2026.

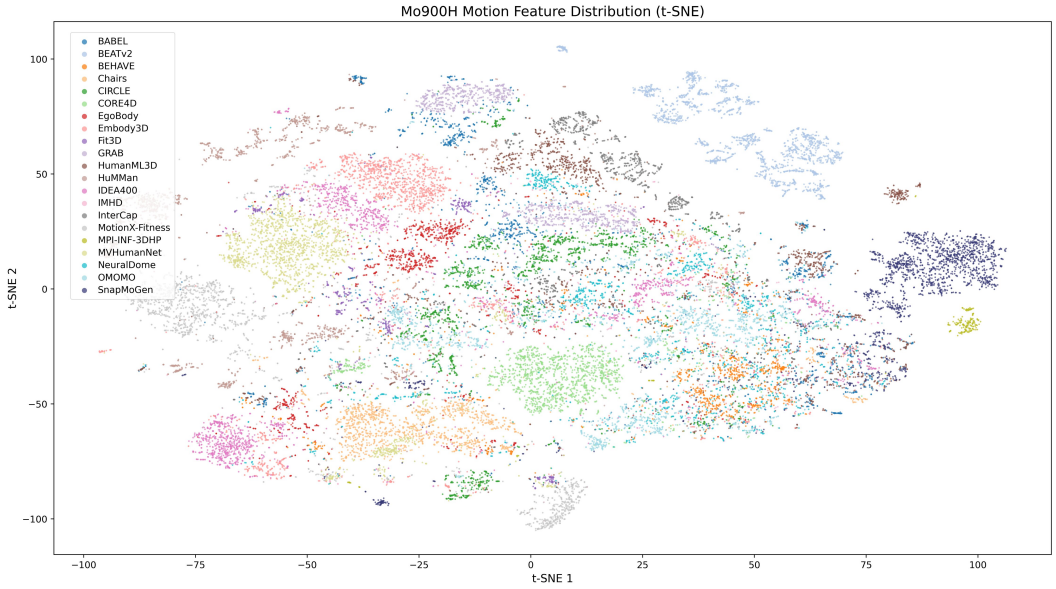


Figure A1: Visualization of data distributions. For each dataset, we randomly sample 2,000 motion sequences. Each sequence is temporally averaged to obtain a compact feature representation, which is then reduced in dimensionality using t-SNE.

A Technical appendices and supplementary material

A.1 More Details of Mo900H

A.1.1 Integration and Annotation of Mo900H Dataset

To support four-modality training over motion, language, vision, and object conditions, we integrate and process 20 motion-capture-based motion datasets and one additional dataset whose motions are estimated from in-the-wild videos. The motion-capture datasets provide high-quality motion annotations, while the in-the-wild video-based dataset is included to enrich the visual modality and improve the diversity of visual-conditioned motion contexts.

The original datasets differ substantially in motion representation, skeleton template, coordinate convention, facing direction, frame rate, and modality availability. To obtain a unified training representation, we canonicalize all motions into a common 22-joint skeleton, resample all sequences to 30 FPS, and limit the maximum sequence length to 20 seconds. Sequences longer than 20 seconds are temporally cropped, while sequences shorter than 2 seconds are discarded. We also align the initial facing direction of each character to the positive z axis to reduce orientation inconsistency across datasets. All motion sequences are retargeted to a standard skeleton template following HumanML3D [10].

Following the compact motion representation used in prior motion generation work [36], each motion sequence is represented by root angular velocity, root linear velocity on the ground plane, root height, and local joint positions. This produces a 67-dimensional frame-wise motion representation. This unified representation reduces redundant global information while preserving the local pose dynamics required for diffusion-based motion generation. We also provide a 263-dimensional HumanML3D-style motion representation.

Figure A1 visualizes the distribution of the integrated motion data. For each dataset, we randomly sample 2,000 motion sequences, temporally average each sequence to obtain a compact motion-level feature, and project the features into two dimensions using t-SNE.

Table A1: Statistics of Mo900H subsets used in this work. † indicates videos are rendered motion visualizations. *Auto text* indicates whether the text annotations were generated or re-generated by the vision-language annotation pipeline. Subsets with the suffix `_InterAct` follow the data split and text annotation protocol of InterAct. Specifically, OMOMO uses the original OMOMO text annotations and data split, while OMOMO_InterAct uses the text annotations and data split defined in InterAct.

Dataset	Motions	Duration (h)	Texts	Videos	Obj. motions	Auto text
BABEL [41]	15,096	41.6	–	15,096†	–	–
BEATv2 [28, 29]	10,867	60.3	–	–	–	–
BEHAVE	988	3.9	–	988	–	–
CIRCLE [1]	5,743	9.4	–	–	–	–
CORE4D [30]	1,692	3.8	–	–	–	–
Chairs_InterAct [20, 58]	1,493	3.7	4,478	–	1,493	No
EgoBody [66]	1,257	3.3	–	–	–	–
Embody3D [27]	98,591	499.9	–	–	–	–
Fit3D [8]	564	2.5	564	564	–	Yes
GRAB_InterAct [47, 58]	1,333	3.7	1,333	–	–	No
HuMMan [3]	4,036	3.6	4,036	–	–	No
HumanML3D [10]	29,226	56.4	87,310	–	–	No
IDEA400 [27]	11,609	23.1	11,653	11,609	–	Yes
IMHD [67]	204	0.2	204	204	–	Yes
IMHD_InterAct [67, 58]	562	1.5	1,677	–	562	No
InterCap	1,374	2.9	1,374	1,374	–	Yes
MPI-INF-3DHP [35]	209	1.1	209	209	–	Yes
MVHumanNet [56]	25,694	71.1	–	–	–	–
MotionX-Fitness [26]	8,211	34.3	8,211	8,211	–	Yes
NeuralDome_InterAct [61, 58]	1,486	3.8	4,441	–	1,486	No
OMOMO [22]	4,838	8.2	4,838	4,838†	–	No
OMOMO_InterAct [22, 58]	4,294	7.4	4,294	–	4,294	No
SnapMoGen [13]	40,896	87.7	81,754	–	–	No
Total	270,263	933.3	216,376	43,093	7,835	–

A.1.2 Vision and Language Annotation

Not all source datasets provide all modalities. For datasets without RGB videos, such as OMOMO and BABEL, we render virtual human videos from their motion-capture SMPL body sequences. These rendered visualizations provide accurate 3D pose and trajectory observations, enabling vision-conditioned training when raw RGB footage is unavailable.

For datasets with real RGB videos that lack language annotations or require finer action descriptions, specifically Fit3D, IDEA400 (re-annotated for more detailed descriptions), IMHD, MotionX-Fitness (re-annotated for more detailed descriptions), InterCap, and MPI-INF-3DHP, we first split the motion sequences into clips of at most 20 seconds, and then use GPT-5-mini to generate action-centered descriptions from video frames sampled at 3 FPS and resized to 540p. The prompt instructs the model to describe only observable human motion, physical human-object interaction, object state changes, and movement trajectories, while avoiding appearance, identity, clothing, camera properties, scene aesthetics, intentions, or other non-motion-related information. Among the generated fields, we use the `Detail Movement` output for training because it provides a continuous, low-level description of body motion over time.

The captioning prompt is shown below:

```
Task: Generate objective, action-centered captions for a human motion.

1. Only describe:
- Human body motion (torso, arms, hands, legs)
- Human-object or human-scene interaction only if physical contact occurs
- Observable physical effects (object movement, rotation, lifting,
  placement, support change)
- Moving direction or trajectory (forward, backward, left, right,
```

```
  in circle)

2. Do NOT describe:
- Appearance, identity, clothing, age, gender
- Object color, texture, brand, or semantics beyond function
- Camera, viewpoint, lighting, video quality
- Emotions, intentions, or inferred goals

3. Output Structure (Do NOT output irrelevant content)
<Action Summary>
Describe the motion in 1-3 concise sentence
<Detail Movement>
Describe the motion in detail (Do not divide into multiple paragraphs).
<Movement Stages>
#Stage 1:
Human: objective body movement
Interaction: contact and object state change
#Stage 2:
...

4. Action Summary
- Describe the main motion in one/two concise sentence
- If applicable, include the core object interaction outcome only

5. Detail Movement
- Describe the motion in detail in a single continuous paragraph
- Focus on a continuous, low-level physical description of the motion
  sequence, emphasizing how the body moves over time

6. Movement Stages
- Split motion into ordered stages by visible motion changes
- Each stage MUST contain:
-- Human: Objective body movement
-- Interaction (if has): Describe contact and object state change
   (on ground / held / on surface). If no interaction at this stage,
   write None

7. Language
- Physical, neutral, factual
- No interpretation or speculation
- No redundancy across stages

The following images are chronologically ordered video frames sampled from
one motion clip at 3 FPS. Use the full sequence jointly.
```

A.1.3 Caption Filtering and Dataset-Level Rejection

Since automatically generated captions can be affected by unreliable visual inputs, occlusions, ambiguous motion, or insufficient temporal evidence, we adopt a three-stage filtering strategy operating at complementary source, model, and sample levels to remove distinct types of annotation noise that cannot be reliably detected by a single criterion. First, we perform dataset-level rejection to remove sources with systematic visual or motion ambiguities that make reliable captioning difficult. Second, we conduct model-based rejection using an early motion generator to identify annotations or data sources that introduce negative transfer to motion generation. Third, we manually inspect the remaining VLM-annotated captions to verify their motion relevance and temporal consistency.

As a result, we remove the text annotations for the following datasets: The text modality of MVHumanNet is excluded because many of its images contain large black regions introduced by the original storage format, which makes visual captioning unreliable. Text annotations for BEHAVE are also removed because its motions often contain repeated direction switches, and the VLM-generated

Table A2: Modality composition and sequence-level alignment breakdown of Mo900H.

Component / Aligned Modality Subset	Count	Proportion / Status
Real RGB visual observations	23,159	53.7% of 43,093 videos
Rendered motion visualizations	19,934	46.3% of 43,093 videos
Source-provided text annotations	194,161	89.7% of 216,376 annotations
GPT-generated / re-generated annotations	22,215	10.3% of 216,376 annotations
Aligned M-V pairs (`Mo900H-MV`)	43,093	Sequence-level
Aligned M-L-V triplets (`Mo900H-MLV`)	27,009	Sequence-level
↪ Real RGB + GPT-generated/re-generated text	22,171	82.1% of `Mo900H-MLV`
↪ Rendered vision + source-provided text	4,838	17.9% of `Mo900H-MLV` (OMOMO)
Aligned M-L-O triplets (`Mo900H-MLO`)	7,835	Sequence-level
Aligned M-L-V-O quadruplets	0	Disjoint visual and object sources

captions are not sufficiently precise in describing these directional changes. Captions for rendered videos such as those from BABEL are likewise excluded, ensuring that all retained VLM-generated descriptions are paired exclusively with real RGB observations.

A.1.4 Object Modality and Modality Composition Summary

For motion sequences with object modality, we use the motion-object interaction data integrated by InterAct and adopt the text annotations provided by InterAct. These sequences provide paired human motion and object motion, allowing the model to learn object-conditioned motion generation.

Per-subset statistics of Mo900H are summarized in Table A1, and Table A2 details the overall breakdown of real versus rendered visual observations, source-provided versus GPT-generated text annotations, and sequence-level multimodal alignments. As visual and 3D object annotations originate from complementary source datasets, Mo900H is organized into aligned modality subsets (`Mo900H-M`, `Mo900H-MV`, `Mo900H-ML`, `Mo900H-MLV`, and `Mo900H-MLO`) to support progressive cross-modal training.

A.1.5 Why Not Directly Use Large-Scale In-the-Wild Motion Datasets?

Large-scale in-the-wild motion datasets, such as MotionMillion, are constructed by crawling human motion videos from the web and extracting 3D human poses using monocular pose estimation algorithms. Such datasets are attractive because they cover broad motion categories and diverse human activities. However, they are less suitable for our multimodal setting for two main reasons.

First, these datasets generally do not release the original visual modality. Therefore, they cannot directly support training with visual conditions, which is essential for our four-modality setting. Second, motions estimated from web videos often suffer from motion degradation caused by low image quality, occlusion, fast camera motion, frequent viewpoint changes, and tracking instability. These issues can introduce jitter, high-frequency artifacts, foot-ground inconsistency, abrupt joint jumps, and physically implausible accelerations. Moreover, because web videos often contain frequent shot changes, the effective duration of each continuous motion segment is usually short. As a result, although such datasets may provide broad coverage of the motion space, their motion annotations can be less stable and less physically reliable for high-quality motion generation.

A.1.6 Motion Annotation Quality Analysis

We compare Mo900H with MotionMillion (GV part[2]) to analyze the quality of motion annotations from temporal, physical, and sequence-length perspectives. All motions are evaluated after being converted to a unified 22-joint representation at 30 FPS. Let a motion sequence be denoted as $\mathbf{X} = \{\mathbf{x}_t^j\}_{t=1,j=1}^{T,J}$, where $\mathbf{x}_t^j \in \mathbb{R}^3$ is the 3D position of joint j at frame t. The frame interval is

[2] https://huggingface.co/datasets/InternRobotics/MotionMillion/tree/main/motion_272rpr/MotionGV

$\Delta t = 1/\text{fps}$. We compute velocity and acceleration as

$$\mathbf{v}_t^j = \frac{\mathbf{x}_{t+1}^j - \mathbf{x}_t^j}{\Delta t}, \qquad \mathbf{a}_t^j = \frac{\mathbf{v}_{t+1}^j - \mathbf{v}_t^j}{\Delta t}. \tag{5}$$

Jitter. Jitter measures the average local temporal irregularity of joint trajectories:

$$\text{Jitter} = \frac{1}{(T-2)J} \sum_{t=1}^{T-2} \sum_{j=1}^{J} \left\| \mathbf{a}_t^j \right\|_2. \tag{6}$$

Lower jitter indicates smoother motion trajectories.

Foot Sliding. Foot sliding measures the horizontal displacement of foot joints during estimated ground contact. Let $\mathcal{F}$ be the set of foot joints, $c_t^f \in \{0, 1\}$ be the binary contact state of foot f, and $\mathbf{p}_t^f$ be its ground-plane position. We compute

$$\text{FootSlide} = \frac{\sum_{t=1}^{T-1} \sum_{f \in \mathcal{F}} c_t^f \left\| \mathbf{p}_{t+1}^f - \mathbf{p}_t^f \right\|_2}{\sum_{t=1}^{T-1} \sum_{f \in \mathcal{F}} c_t^f + \epsilon}. \tag{7}$$

Lower values indicate more stable foot-ground contact.

Foot Floating. Foot floating measures whether contact feet are unrealistically above the ground. Let h_t^f denote the height of foot f above the ground, and τ_h be a small tolerance threshold. We define

$$\text{FootFloat} = \frac{\sum_{t=1}^{T} \sum_{f \in \mathcal{F}} c_t^f \max(h_t^f - \tau_h, 0)}{\sum_{t=1}^{T} \sum_{f \in \mathcal{F}} c_t^f + \epsilon}. \tag{8}$$

Lower values indicate that contact feet remain closer to the ground.

Acceleration Spike. Acceleration spike measures the proportion of frames where joint acceleration exceeds a plausible range. For each joint j, we estimate a reference threshold

$$\tau_a^j = Q_{0.99}\left(\left\{ \left\| \mathbf{a}_t^j \right\|_2 \right\}_{\text{ref}} \right), \tag{9}$$

and compute

$$\text{AccSpike} = \frac{1}{(T-2)J} \sum_{t=1}^{T-2} \sum_{j=1}^{J} \mathbb{I}\left[\left\| \mathbf{a}_t^j \right\|_2 > \tau_a^j \right]. \tag{10}$$

This metric captures rare but severe abrupt changes in local motion dynamics.

High-Frequency Noise Ratio. High-frequency noise ratio measures the proportion of trajectory energy lying in high-frequency bands. For each joint trajectory, we compute its temporal Fourier transform $\mathcal{F}(\mathbf{x}^j)(\omega)$. Given a cutoff frequency ω_c,

$$\text{HFNoise} = \frac{\sum_{j=1}^{J} \sum_{\omega > \omega_c} \left\| \mathcal{F}(\mathbf{x}^j)(\omega) \right\|_2^2}{\sum_{j=1}^{J} \sum_{\omega} \left\| \mathcal{F}(\mathbf{x}^j)(\omega) \right\|_2^2 + \epsilon}. \tag{11}$$

Lower values indicate less high-frequency pose noise and better temporal stability.

As shown in Table A3, Mo900H provides cleaner and more temporally stable motion annotations than MotionMillion. These results support the use of curated motion-capture-based data as the main source for high-quality multimodal motion generation.

Table A3: Comparison of motion annotation quality between Mo900H and MotionMillion. Lower values indicate better motion quality for all metrics except average length.

Dataset	Jitter↓	Foot Sliding↓	Foot Floating↓	Acceleration Spike↓	HFNoise Ratio↓	Avg. Length Seconds↑
MotionMillion	0.0250	0.0049	0.0401	0.1408	0.0352	1.933
Mo900H	**0.0046**	**0.0017**	**0.0282**	**0.0359**	**0.0046**	**12.724**

A.2 More Implementation Details

We use a Motion Autoencoder, consisting of a Motion Encoder and a Motion Decoder, to map raw motion sequences into compact latent motion representations and reconstruct them back to the motion space. For experiments on HumanML3D, we directly adopt the Motion AE pretrained by MARDM [36]. Since HumanML3D contains only paired text-motion data and follows the standard motion representation used in prior language-to-motion generation benchmarks, we keep the pretrained Motion AE unchanged and use it for all HumanML3D experiments.

For experiments on Mo900H, we train a Motion AE using the same codebase and a comparable model configuration to the MARDM Motion AE. Considering that Mo900H contains more diverse modalities and significantly longer motion sequences, we adjust the preprocessing and training settings accordingly. Specifically, we set the maximum motion length to 20 and apply a 6× temporal downsampling to the motion sequences. The Motion AE is trained for 6 epochs on Mo900H before being used in the subsequent generative model training. After training, the Motion Encoder and Motion Decoder are fixed and shared across the two-stage training pipeline of the generative model.

Due to the increased modality diversity and longer motion duration in Mo900H, training the Mo900H generation model requires substantially more GPU memory than HumanML3D. In our implementation, the reported training configuration requires approximately 55 GB of GPU memory. We train the model on either a single NVIDIA RTX Pro 6000 GPU or two NVIDIA RTX 4090 GPUs. On a workstation equipped with one NVIDIA RTX Pro 6000 GPU and an AMD Ryzen 9 9950X CPU, training takes approximately 16 minutes per epoch under the reported configuration.

A.3 Evaluation Metrics

For motion generation, we use several common quantitative metrics to evaluate different aspects of model performance. FID (Fréchet Inception Distance) measures how close the distribution of generated motions is to real motions, reflecting overall realism. R-Precision and Matching Score evaluate whether the generated motions match the given text descriptions, focusing on text-motion alignment. To check the diversity of outputs, we report Multimodality Diversity, which measures how much variation the model can produce under the same condition. We also include CLIP-Score following [36], which leverages pretrained vision-language models to further test semantic consistency between motion and text.

For the Language-to-Motion task on Mo900H, we further report physics-based motion quality metrics following MBench [27], including Jitter (Eq. 6), Foot Sliding (Eq. 7), Foot Floating (Eq. 8), and Dynamic. For human-object interaction motions, we additionally report Mean Per Joint Position Error (MPJPE), Foot Sliding (FS), and contact-state metrics, including Contact Precision (C-Prec.), Contact Recall (C-Rec.), Contact Accuracy (C-Acc.), and Contact F1, to measure whether the generated motion reproduces the expected spatial configuration and hand-object contact state. Following the contact-state matching protocol of InterAct [58], we first compute, at each frame, the minimum distance between the wrist/hand joints and the posed object point cloud. A frame is labeled as contact if this distance is below a threshold τ. We obtain binary contact sequences for both generated and ground-truth motions, and compute precision, recall, accuracy, and F1 score over all frames. Since Mo900H represents the body using a sparse 22-joint skeleton rather than the dense body markers used in InterAct, the original 0.05m threshold is overly strict: the hand joint does not necessarily lie on the actual hand-object contact surface. We therefore set $\tau = 0.10$m, which better reflects the spatial resolution of the 22-joint representation while keeping the same contact-consistency criterion. In our counterfactual perturbation experiments, we also report KS-MPJPE (MPJPE evaluated at the

Table A4: Ablation study of iteration number in masked modeling and sampling steps in diffusion.

Method	Time steps	ODE steps	Time (s) (*bs=32*)	R-Precision Top-1 ↑	FID ↓	Diversity ↑	Matching ↓
MARDM [36]	18	adaptive	13.709	0.543	0.161	**17.718**	3.928
MMM [40]	10	–	0.692	0.545	0.169	17.201	3.964
Ours	18	1	**0.114**	0.102	174.217	3.253	14.995
Ours	18	6	0.769	0.615	0.078	17.254	3.457
Ours	18	10	1.430	0.617	0.080	17.499	3.383
Ours	18	20	2.907	0.620	**0.069**	17.525	3.387
Ours	5	6	0.211	**0.624**	0.176	17.575	**3.343**
Ours	10	6	0.411	0.614	0.109	17.459	3.392
Ours	25	6	1.020	0.616	0.104	17.477	3.419

Table A5: Full table of quantitative results on HumanML3D (as a supplement to Table 2). The best results are in **bold**. Metrics for some baseline methods are cited from MARDM [36].

Methods	R-Precision ↑			FID ↓	Matching ↓	CLIP-score ↑
	Top-1	Top-2	Top-3			
T2M-GPT [60]	$0.470^{\pm.003}$	$0.659^{\pm.002}$	$0.758^{\pm.002}$	$0.335^{\pm.003}$	$3.505^{\pm.017}$	$0.607^{\pm.005}$
ReMoDiffuse [63]	$0.468^{\pm.003}$	$0.653^{\pm.003}$	$0.754^{\pm.005}$	$0.883^{\pm.021}$	$3.414^{\pm.020}$	$0.621^{\pm.003}$
MDM-50Step [49]	$0.440^{\pm.007}$	$0.636^{\pm.006}$	$0.742^{\pm.004}$	$0.518^{\pm.032}$	$3.640^{\pm.028}$	$0.578^{\pm.003}$
MLD [5]	$0.461^{\pm.004}$	$0.651^{\pm.004}$	$0.750^{\pm.003}$	$0.431^{\pm.014}$	$3.445^{\pm.019}$	$0.610^{\pm.003}$
MMM [40]	$0.487^{\pm.003}$	$0.683^{\pm.002}$	$0.782^{\pm.001}$	$0.132^{\pm.004}$	$3.359^{\pm.009}$	$0.635^{\pm.003}$
MoMask [12]	$0.469^{\pm.004}$	$0.687^{\pm.003}$	$0.786^{\pm.003}$	$0.116^{\pm.006}$	$3.353^{\pm.010}$	$0.637^{\pm.003}$
MotionDiffuse [64]	$0.450^{\pm.006}$	$0.641^{\pm.005}$	$0.753^{\pm.005}$	$0.778^{\pm.005}$	$3.490^{\pm.023}$	$0.606^{\pm.004}$
MARDM-DDPM [36]	$0.492^{\pm.006}$	$0.690^{\pm.005}$	$0.790^{\pm.005}$	$0.116^{\pm.004}$	$3.349^{\pm.010}$	$0.637^{\pm.005}$
MARDM-SiT [36]	$0.500^{\pm.004}$	$0.695^{\pm.003}$	$0.795^{\pm.003}$	$0.114^{\pm.007}$	$3.270^{\pm.009}$	$0.642^{\pm.002}$
MotionAgent [54]	$0.485^{\pm.003}$	$0.680^{\pm.003}$	$0.780^{\pm.002}$	$0.202^{\pm.009}$	$3.327^{\pm.009}$	$0.634^{\pm.003}$
MoScale [69]	$0.520^{\pm.003}$	$0.713^{\pm.002}$	$0.808^{\pm.002}$	$0.090^{\pm.003}$	$3.251^{\pm.008}$	$0.644^{\pm.001}$
MoMask++ [13]	$0.490^{\pm.003}$	$0.684^{\pm.003}$	$0.782^{\pm.002}$	$0.109^{\pm.005}$	$3.350^{\pm.008}$	$0.636^{\pm.002}$
Ours *w/o Und.*	$0.533^{\pm.012}$	$0.731^{\pm.010}$	$0.826^{\pm.010}$	$0.108^{\pm.023}$	$3.078^{\pm.037}$	$0.658^{\pm.001}$
Ours	$\mathbf{0.544^{\pm.002}}$	$\mathbf{0.740^{\pm.003}}$	$\mathbf{0.834^{\pm.002}}$	$\mathbf{0.071^{\pm.004}}$	$\mathbf{2.995^{\pm0.010}}$	$\mathbf{0.669^{\pm0.001}}$

timestamps of visual key-state frames) for visual completion and HandObjDis (mean distance between the generated wrist joints and the ground-truth target object) for object-conditioned generation.

For text generation, we evaluate the captions generated for motions using both n-gram and semantic metrics. BLEU@4 and BLEU@1 measure overlap at different n-gram levels, while ROUGE emphasizes recall of important phrases. To capture similarity beyond exact wording, we report BERTScore, which uses contextual embeddings to measure semantic closeness to reference captions.

A.4 Inference Speed

All inference experiments are conducted on a workstation equipped with an AMD 9950X CPU and an NVIDIA RTX 6000 PRO GPU. We measure inference efficiency as the wall-clock latency for generating 32 samples in parallel. We analyze two major factors that affect inference speed and performance:

Number of Iterations. Our model consists of an MCT backbone and a diffusion head. Inference starts from a fully masked latent variable, and the model reconstructs the masked regions iteratively. We set the default number of iterations to 18, and further vary this value to examine its effect on inference speed and performance. Results are reported in Table A4.

Number of Diffusion Sampling Steps. The diffusion module adopts Euler sampling with 10 default steps on HumanML3D and 6 default steps on Mo900H. Unlike MARDM [36], which relies on adaptive solvers (dopri5), our approach generates realistic motions with only a small number of steps. We also compare different numbers of sampling steps and report the corresponding runtime and performance on Mo900H in Table A4.

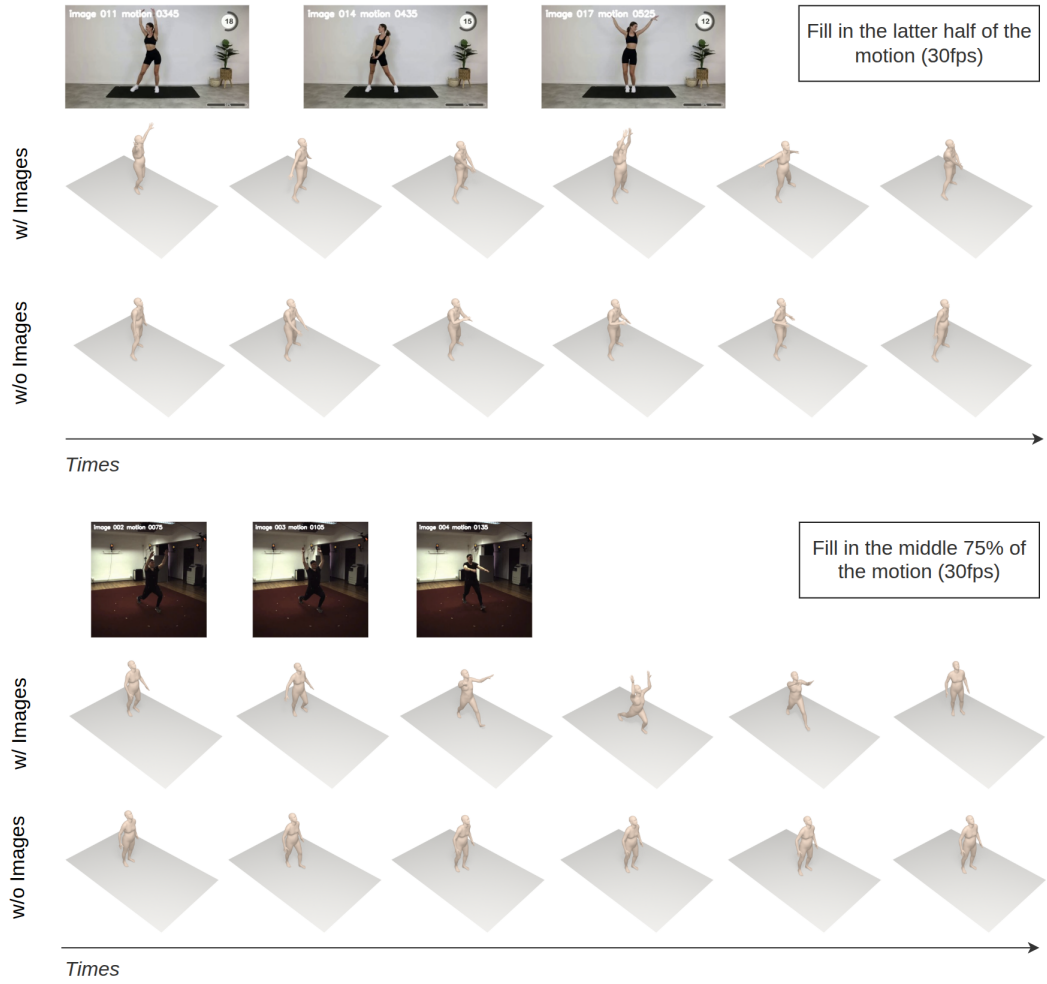


Figure A2: Qualitative examples of image-guided motion completion. Given partial motion observations (30 FPS) and visual reference frames (1 FPS), our model generates complete motions (30 FPS) that are coherent with the observed motion and well aligned with the reference image sequence.

These experiments demonstrate that, compared with other diffusion-based methods, our model achieves a clear advantage in real-time inference, maintaining competitive performance even with 6-10 sampling steps.

A.5 Full Results on HumanML3D

Table A5 reports the full set of metrics on HumanML3D as a supplement to Table 2 in the main paper.

A.6 Additional Quantitative Results on Object-Conditioned Generation

As a supplement to Table 7 in the main paper, Table A6 reports extended spatial interaction metrics on Mo900H-MLO, including Mean Per Joint Position Error (MPJPE), Foot Sliding (FS), Contact Precision (C-Prec.), Contact Recall (C-Rec.), Contact Accuracy (C-Acc.), and Contact F1, alongside a reference comparison with InterAct [58]. InterAct jointly synthesizes human and object motion from text alone without observed object context, whereas our framework conditions on text together with either the initial object configuration or the full object trajectory. To ensure consistent metric computation, we convert InterAct's marker-based outputs to our 22-joint skeleton representation and evaluate wrist-object contact against the ground-truth object trajectory under the same $\tau = 0.10\,\mathrm{m}$ threshold. Providing explicit object context yields lower FID (0.151 vs. 0.598), stronger text-motion alignment across all R-Precision and Matching metrics, and higher contact precision (0.549 vs.

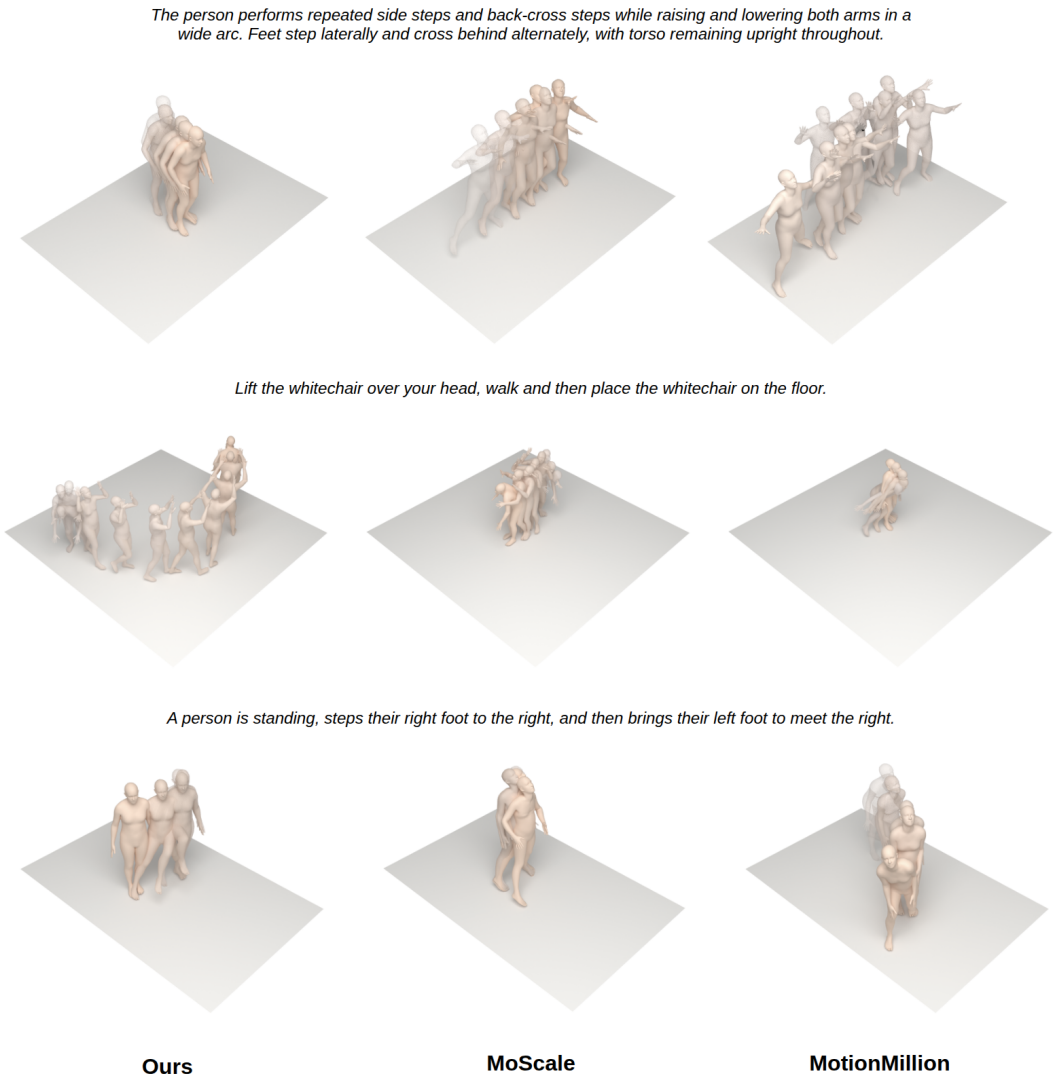


Figure A3: Qualitative comparison for language-to-motion generation. MoScale is trained on Mo900H, and MotionMillion uses the official pretrained weights released with the MotionMillion dataset. Although MotionMillion uses a larger 3B model and a stronger FLAN-T5-XL text encoder, our method achieves better text responsiveness and more realistic motion generation.

0.512) with comparable contact accuracy (0.635) and foot sliding (0.005), while joint human-object synthesis in InterAct achieves lower MPJPE (0.214) and higher contact recall (0.140) and F1 (0.220). Conditioning on the full object trajectory also consistently improves over object initialization alone across FID, Top-2/Top-3 R-Precision, Matching, MPJPE, Foot Sliding, and all contact metrics.

A.7 Qualitative Results on Image-Guided Motion Completion

We provide qualitative examples of image-guided motion completion in Fig. A2. The figure shows two settings. In the first case, the model observes the first half of the motion, while the second half is masked and guided by visual reference frames. In the second case, the model is given the first and last 25% of the motion and completes the missing middle part.

The results show that incorporating image guidance significantly improves the controllability of motion generation. The generated motions are temporally coherent with the given motion context and follow the visual reference sequence well.

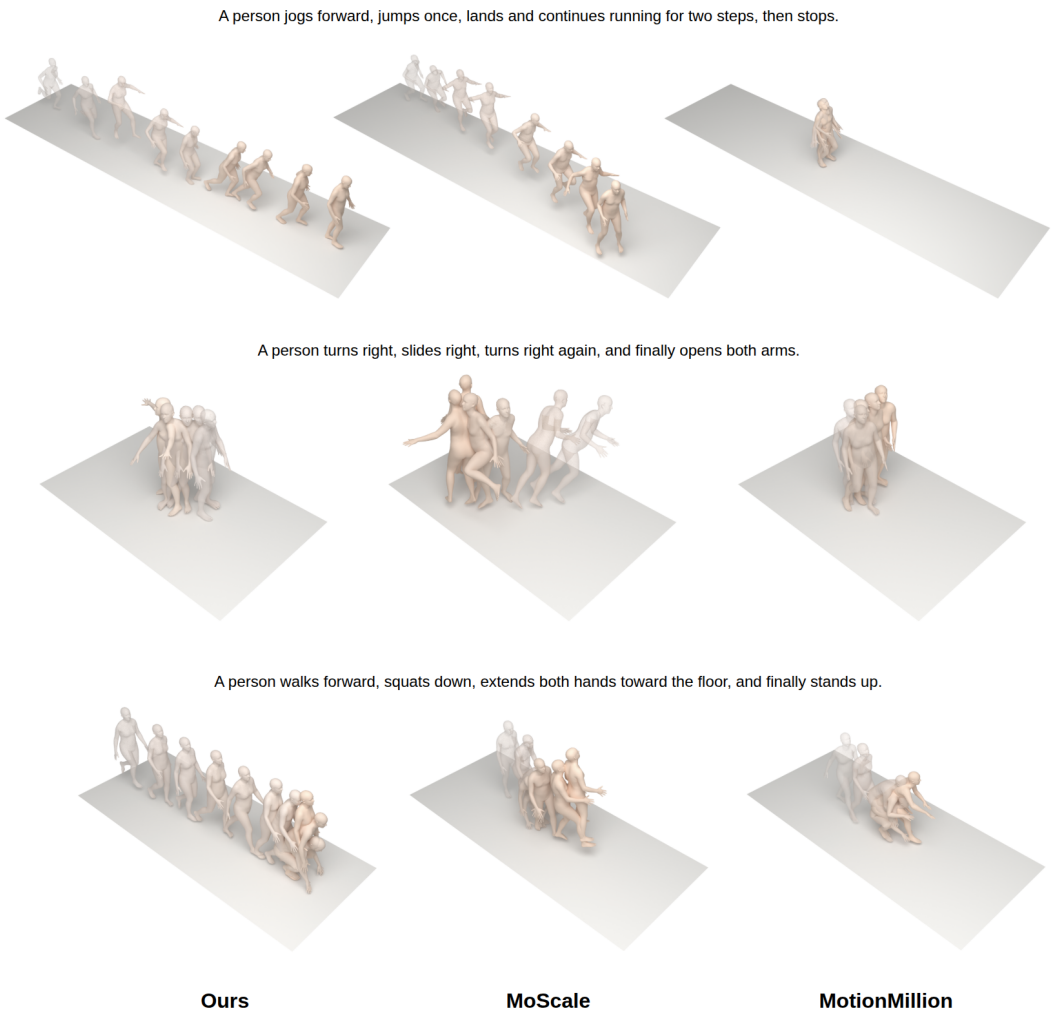


Figure A3: Qualitative comparison for language-to-motion generation, continued.

Table A6: Extended evaluation of object-conditioned motion generation on Mo900H-MLO (as a supplement to Table 7). †InterAct [58] jointly generates human and object motion from text alone without observed object context. FS: Foot Sliding; C-Prec., C-Rec., C-Acc., F1: wrist-object contact metrics.

Method / Condition	FID↓	R-Precision↑			Matching↓	MPJPE↓	FS↓	Contact Quality↑			
		Top-1	Top-2	Top-3				C-Prec.	C-Rec.	C-Acc.	F1
InterAct† [58]	0.598	0.286	0.487	0.622	2.581	**0.214**	**0.005**	0.512	**0.140**	**0.635**	**0.220**
Ours (Language + Obj. Init.)	0.168	**0.329**	0.521	0.676	2.217	0.225	0.006	0.524	0.104	0.632	0.173
Ours (Language + Obj. Traj.)	**0.151**	**0.329**	**0.523**	**0.685**	**2.039**	0.223	**0.005**	**0.549**	0.109	**0.635**	0.182

A.8 Qualitative Comparison on Language-to-Motion Generation

We further provide qualitative comparisons for language-to-motion generation in Fig. A3. We compare our method with MoScale [69] and MotionMillion [7]. For MoScale, we use the model trained on Mo900H. For MotionMillion, we use the official pretrained weights released with the MotionMillion dataset.

It is worth noting that MotionMillion is trained on the MotionMillion dataset with a substantially larger model and a stronger text encoder: the evaluated MotionMillion model has 3B parameters and uses FLAN-T5-XL, while our model uses CLIP-ViT-B/32 as the text encoder. Despite this stronger

setup, our method produces motions that better follow the input text and exhibit more realistic and natural dynamics. In particular, our results show clearer semantic alignment with action descriptions and fewer implausible body poses or motion artifacts.

A.9 Limitations

Although our framework supports flexible multimodal context-aware motion generation, it is still limited by the availability of high-quality multimodal motion data. In particular, existing datasets that simultaneously provide visual observations, interacting objects, object trajectories, and text annotations remain limited. As a result, some important modality combinations, such as jointly observed visual and object contexts, are not available or very limited, which restricts the model's ability to fully learn fine-grained correlations among scene appearance, object geometry, object trajectory, and human motion. Another limitation is that our model mainly learns contextual constraints from data rather than explicitly modeling physical feasibility and contact dynamics. Therefore, although visual and object conditions improve spatial grounding, the generated motions may still suffer from inaccurate contact timing or physically implausible transitions in complex interactions. Future work could address these limitations by collecting richer paired multimodal datasets and incorporating contact-aware or physics-based objectives for more realistic human-object motion generation.

B Broader Impacts

This work advances multimodal context-aware human motion generation, offering significant positive potential for embodied AI simulation, character animation, and human-robot interaction by allowing intuitive, fine-grained motion synthesis. However, as a generative model, it carries inherent dual-use risks. High-fidelity motion generation could potentially be misused to synthesize misleading media or "deepfakes," misrepresenting an individual's actions in surveillance or video evidence when combined with photorealistic rendering pipelines. Furthermore, because our model is trained on diverse human motion datasets, it may inadvertently learn and perpetuate demographic, cultural, or physical biases present in the training data, potentially leading to stereotypical or unrepresentative motion generation for certain demographic groups.

C Safeguards

To mitigate the potential risks of misuse associated with high-fidelity human motion generation, we are taking a cautious approach to the release of our resources. The Mo900H benchmark and our pre-trained generative models will be released under a custom, non-commercial research license that explicitly prohibits malicious applications, including the generation of deceptive media. Access will be gated and restricted to verified researchers who agree to these terms. As a technical safeguard for future deployment, we also encourage the community to develop motion-trajectory watermarking techniques to ensure that synthesized interactions can be reliably distinguished from authentic human capture data.

D Dataset Links and Licenses

We summarize the official access links and license terms of the datasets used to construct Mo900H in Table A7. For assets governed by custom terms, registration agreements, or source-specific restrictions, we used the data only under the corresponding official terms.

Table A7: Official links and license terms of datasets used in Mo900H. "Custom" denotes an official dataset-specific license or registration agreement rather than a standard SPDX license. Entries marked as "not explicitly specified" are used conservatively under their official data-access terms.

Dataset	Official link	License / terms
BABEL	https://babel.is.tue.mpg.de/	Custom non-commercial scientific research / education / artistic-use license.
BEATv2	https://pantomatrix.github.io/BEAT/	Non-commercial license / official dataset terms.

Continued on next page

Dataset	Official link	License / terms
BEHAVE	https://virtualhumans.mpi-inf.mpg.de/behave/	Custom non-commercial scientific research license.
CIRCLE	https://tml.stanford.edu/circle_dataset/	CC BY-NC 4.0 for data.
CORE4D	https://core4d.github.io/	CC BY 4.0 for dataset.
Chairs_InterAct	https://github.com/wzyabcas/InterAct	InterAct access terms; follows original source licenses. Chairs license is not explicitly specified by the source.
EgoBody	https://egobody.ethz.ch/	CC BY-NC-SA 4.0 with additional no-copy / no-share / no-distribution terms.
Embody3D	https://www.meta.com/emerging-tech/codec-avatars/embody-3d/	Meta XRCIA Noncommercial Research License Agreement for dataset.
Fit3D	https://fit3d.imar.ro/	Custom non-commercial scientific research / education / artistic-use license; no redistribution without permission.
GRAB_InterAct	https://grab.is.tue.mpg.de/	GRAB custom non-commercial scientific research / education / artistic-use license; InterAct conversion follows original terms.
HuMMan	https://caizhongang.com/projects/HuMMan/	S-Lab License v1.0.
HumanML3D	https://github.com/EricGuo5513/HumanML3D	Repository under MIT license; underlying motion assets follow AMASS / HumanAct12 original data terms.
IDEA400	https://github.com/IDEA-Research/Motion-X	Motion-X annotations under CC BY-NC-SA 4.0; IDEA400 sub-dataset listed under CC BY-SA 4.0 in Motion-X license notes.
IMHD	https://afterjourney00.github.io/IM-HOI.github.io/	CC BY-NC-SA 4.0 / non-commercial scientific research dataset terms.
IMHD_InterAct	https://github.com/wzyabcas/InterAct	InterAct access terms; source IMHD uses CC BY-NC-SA 4.0 / non-commercial scientific research terms.
InterCap	https://intercap.is.tue.mpg.de/	Custom non-commercial scientific research / education / artistic-use license.
MPI-INF-3DHP	https://vcai.mpi-inf.mpg.de/3dhp-dataset/	Custom non-commercial research license / official dataset terms.
MVHumanNet	https://x-zhangyang.github.io/MVHumanNet/	MVHumanNet data Terms of Use; code under CC BY-NC 4.0.
MotionX-Fitness	https://motion-x-dataset.github.io/	Motion-X annotations under CC BY-NC-SA 4.0; source-specific terms also apply.
NeuralDome_InterAct	https://github.com/wzyabcas/InterAct	InterAct access terms; HODome / NeuralDome source reported under Apache License in InterAct documentation.
OMOMO	https://lijiaman.github.io/projects/omomo/	Official code under MIT license; dataset license not explicitly specified on the project / code page.
OMOMO_InterAct	https://github.com/wzyabcas/InterAct	InterAct access terms; OMOMO source license not explicitly specified.
SnapMoGen	https://huggingface.co/datasets/Ericguo5513/SnapMoGen	Snap Inc. Non-Commercial License.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's main contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We mention our limitations in Appendix A.9.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not present formal theoretical results, theorems, or proofs. The equations are used to define the model components, attention mechanism, and training objectives.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the model architecture, modality-specific encoders, Multimodal Context Transformer, generation and understanding heads, training objectives, two-stage training procedure, dataset construction, preprocessing, evaluation metrics, baselines, and inference settings. See Sec. 5 and Appendix A.1. These details gives a reasonable methodological path to reproduce the main experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and instructions are publicly available at https://github.com/SorenShi/MC-MoGen.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides sufficient training and test details to support the reported results. See Sec. 5 and Appendix A.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports error bars indicating 95% confidence intervals for key metrics such as FID, R-Precision, and MM Dist across multiple benchmarks. See Table 2 and Table 1

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Sec. 5 and Appendix A.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research aligns with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: We discuss the positive applications in embodied AI and animation, as well as potential negative risks such as deepfake generation and training data demographic biases, in Appendix B.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: We detail our mitigation strategies in Appendix C, which include gated access, non-commercial licensing for models and datasets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All source datasets are cited and used according to their released terms. We summarize the official links and license terms of the datasets used in Mo900H in Appendix D.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce Mo900H and document its construction, preprocessing, modality availability, and statistics in Appendix A.1, with documentation and access instructions provided at `https://github.com/SorenShi/MC-MoGen`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The paper does not involve research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: See Appendix A.1. LLM (GPT-5-mini) is used for data annotation.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.