LaxMotion: Rethinking Supervision Granularity for 3D Human Motion Generation

Sheng Liu[1], Yuanzhi Liang[2], and Sidan Du[3⋆]

[1] Nanjing University, Nanjing, China
anderliu@smail.nju.edu.cn
[2] TeleAI, Shanghai, China
liangyzh18@outlook.com
[3] Nanjing University, Nanjing, China
coff128@nju.edu.cn

Abstract. Recent 3D human motion generation models demonstrate remarkable reconstruction accuracy yet struggle to generalize beyond training distributions. This limitation arises partly from the use of precise 3D supervision, which encourages models to fit fixed coordinate patterns instead of learning the essential 3D structure and motion–semantic cues required for robust generalization. To overcome this limitation, we propose LaxMotion, a framework that synthesizes realistic 3D motions without direct 3D pose supervision. Instead of regressing toward exact coordinates, LaxMotion learns 3D motion as a consistent explanation of global trajectories and monocular 2D kinematic cues. We introduce a structured motion factorization together with a reformulated training paradigm under relaxed observability. This design is further supported by relaxed regularization objectives that enforce view-consistent alignment, orientation coherence, and structural stability. Under this relaxed supervision paradigm, LaxMotion generates diverse, temporally coherent, and semantically aligned 3D motions, achieving performance comparable to or surpassing fully 3D-supervised methods. These results indicate that shifting supervision from exact coordinate matching to structural consistency promotes stronger reasoning and improved generalization, offering a scalable and data-efficient paradigm for 3D motion generation.

Keywords: 3D Motion Generation · Relaxed Supervision · Text-to-Motion

1 Introduction

Text-driven 3D human motion generation has progressed rapidly in recent years. Modern diffusion and token-based models can synthesize temporally coherent sequences with strong frame-level fidelity and low reconstruction error on standard benchmarks [5, 7, 24, 35, 42, 44]. However, high numerical accuracy does not always translate to robust motion understanding. Many systems generalize poorly

⋆ Corresponding author.

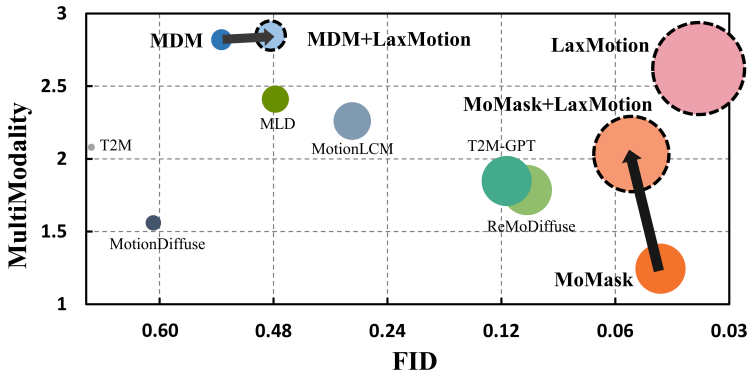

Fig. 1: MultiModality-FID comparisons of different methods on HumanML3D dataset. 3D-supervised methods, such as MDM [35], tend to achieve low FID but relatively high MultiModality, whereas MoMask [7] exhibits the opposite trend. LaxMotion, trained under a reformulated supervision scheme, relaxes the constraints of precise 3D signals. This encourages the model to learn true motion structure and semantic patterns, enabling it to achieve both high MultiModality and low FID simultaneously. The point size of each method represents its Quality–Multimodality Score (QM), which jointly reflects FID and MultiModality through an integrated measure detailed in the experimental section.

beyond the training distribution. They often struggle with unseen actions, new subjects, or compositional variations. They can also produce limited diversity for the same input. For instance, masked-token generators such as MoMask [7] may return highly similar samples across repeated generations [31]. This gap raises a practical question: are we optimizing the right supervision signal for generative motion modeling?

A common design choice in existing methods is to learn a direct mapping from text to 3D joint coordinates [5, 7, 35, 42, 44]. 3D motion capture provides precise annotations, but it is expensive and often limited in coverage and diversity [25, 32]. More importantly, coordinate-level supervision is highly overdetermined. It encourages exact matching to dataset-specific realizations of motion, including low-level details that are not essential to semantics. As a result, models can achieve strong reconstruction-oriented metrics by fitting the training distribution, while under-learning the invariances that matter for generalization and multi-modality. Recent analyses of VQ-based motion generators support this view. Discrete representations combined with reconstruction-driven objectives tend to reduce diversity and bias generation toward memorized patterns (Fig. 1) [21, 26].

These considerations suggest that the bottleneck is not only model capacity, but also the granularity of supervision. Text-to-motion generation is inherently one-to-many: a prompt admits multiple valid motions that differ in style, execution, viewpoint, and minor kinematic details. However, coordinate-level 3D supervision converts this one-to-many problem into a point-matching objective.

Every training example becomes a single privileged target in $\mathbb{R}^{3J\times T}$, and deviations are penalized even when they remain semantically correct. This mismatch encourages models to fit dataset-specific realizations and low-level coordinate patterns, which can inflate reconstruction-oriented metrics while suppressing diversity and weakening out-of-distribution generalization.

A natural way to better align the learning signal with the generative nature of the task is to supervise motion in a less over-determined space. Single-view 2D kinematics preserves articulation, relative limb dynamics, and temporal ordering, which are central to motion semantics. At the same time, 2D projections discard part of the depth- and camera-dependent details that can dominate coordinate regression. Crucially, a 2D motion sequence does not specify a unique 3D pose sequence; it corresponds to a set of plausible 3D explanations. Training against such a set-valued target relaxes the objective from exact coordinate reproduction to structural consistency. This creates room for multi-modality, while still constraining the motion to remain coherent under time and geometry.

Based on this principle, we train a 3D motion generator without applying losses on ground-truth 3D joint coordinates. We emphasize that this is not a claim that 3D data are unnecessary. On mocap benchmarks, we obtain the required relaxed supervision by projecting available 3D sequences, enabling controlled evaluation. In deployment, the same relaxed signals can be extracted from monocular videos at scale, offering substantially broader coverage. Our supervision combines single-view 2D relative motion with a global trajectory signal, and the model still outputs complete 3D motions. The role of training is to recover 3D as a consistent explanation of these 2D cues under geometric and physical regularization, rather than to memorize 3D coordinates.

To realize this idea, we introduce **LaxMotion**, a framework that shifts the paradigm of text-to-motion synthesis by rethinking the underlying supervision granularity. Instead of relying on rigid 3D pose-matching, LaxMotion induces high-fidelity 3D structures from global trajectories and monocular 2D kinematic cues through three synergistic strategies. First, we establish a representation reformulation of human motion by decoupling sequences into global trajectories and relative limb vectors, providing a consistent structural basis for coordinate-free pose alignment. Second, we reformulate the training paradigm by relaxing the information source available during learning. Instead of feeding fully specified 3D motion into the generator, we provide only partially observed monocular cues at training time, while still requiring the model to recover complete 3D motion. Finally, we move beyond traditional point-matching by Relaxation Regularizations. They include view-consistent projection losses, cross-view priors induced by random rotations, orientation constraints, and feature-level consistency. Together, these objectives stabilize geometry, improve temporal coherence, and preserve semantic alignment under diverse virtual viewpoints.

Extensive experiments demonstrate that LaxMotion generates natural and diverse 3D motions while remaining faithful to text. Despite using no direct 3D pose-level supervision, it achieves performance competitive with, and in several settings surpassing, strong 3D-supervised baselines on standard benchmarks.

These results suggest that relaxing coordinate-level supervision, when paired with appropriate constraints, can improve generalization and diversity in text-to-motion generation.

Our main contributions are:

- We identify a limitation of prevailing coordinate-level 3D supervision for generative motion modeling: it can favor dataset-specific fitting and reduce diversity, despite strong reconstruction scores.
- We propose **LaxMotion**, a framework which rethinks supervision granularity by learning from 2D kinematic cues and structural constraints rather than relying on dense 3D pose-level labels.
- We introduce a structured motion factorization, a reformulated training paradigm under relaxed observability, and the Relaxation Regularizations that enforce multi-view geometric stability and temporal coherence.
- We show that relaxed supervision can yield competitive or superior results to 3D-supervised training, offering a scalable and generalizable alternative supervision strategy.

2 Related Work

Text-Driven Motion Generation. Generating 3D human motion from textual descriptions has become an important research direction, aiming to synthesize semantically consistent and temporally coherent motions from language inputs [3,5,7–9,15,22,28,35,39,43,44]. Existing approaches can be broadly categorized into diffusion-based and VAE-based methods. Diffusion models [14, 33, 34] generate motions by progressively denoising a latent sequence under a controlled diffusion process. MDM [35] first proposed predicting samples instead of noise, enabling geometric constraints during training. ReMoDiffuse [44] adds a retrieval mechanism to refine denoising, while MotionDiffuse [43], MotionLCM [5], and Fg-t2m++ [39] improve semantic alignment and temporal smoothness through enhanced latent diffusion strategies. VAE-based frameworks such as T2M [8] model the probabilistic relationship between text and motion sequences via temporal VAEs [27]. MotionGPT [15] and T2M-GPT [42] combine VAE-based latent motion representations with GPT-like architectures [1, 17] to achieve strong semantic consistency. MoMask [7] further introduces hierarchical quantization to represent motion as multi-level discrete codes [19]. Despite their success, existing methods predominantly depend on dense 3D motion annotations and rigid coordinate-level supervision. Beyond the high cost of acquiring accurate 3D labels, such tightly constrained supervision can over-determine the learning objective, restricting the model's flexibility and limiting motion diversity. This motivates the exploration of alternative frameworks that relax supervision while preserving structural coherence, enabling high-quality and semantically faithful motion generation without relying on precise 3D pose annotations.

Weakly-Supervised Motion Generation. To alleviate the reliance on dense 3D annotations, weakly-supervised approaches leverage unpaired data or 2D-to-3D lifting priors [4,6,11,12,18,23,37,38,41]. Subsequent works further reduce explicit 3D

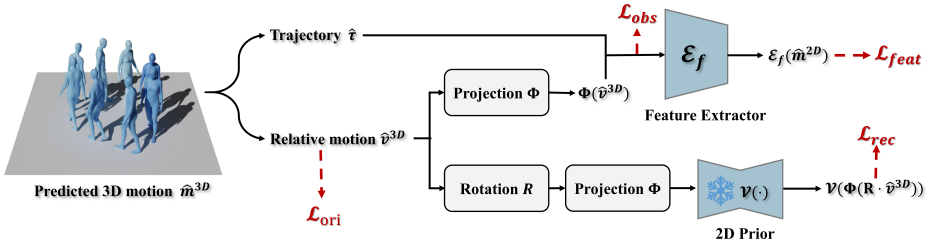

Fig. 2: Relaxed supervision for LaxMotion. Instead of rigid coordinate-level point matching, the generator is trained under relaxed supervision using partial 2D motion cues and reprojection consistency. This encourages structurally coherent, view-invariant 3D motion synthesis without explicit 3D pose annotations.

supervision by enforcing geometric consistency or ambiguity-aware constraints from 2D observations, including adversarial consistency [2], ambiguity mitigation [40], probabilistic flow modeling [36], and occlusion-robust extensions [13]. More recently, diffusion-based models have been introduced to capture complex motion distributions from in-the-wild videos. For example, MAS [16] adopts multi-view ancestral sampling for cross-view consistency, while MVLift [20] progressively reconstructs global 3D trajectories from 2D sequences. Motion-2-to-3 [29] pre-trains a 2D motion generator and subsequently fine-tunes with 3D data to disentangle local and global dynamics. Despite reducing reliance on explicit 3D annotations, existing methods often retain rigid coordinate-level objectives or require partial 3D fine-tuning, limiting motion diversity. Moreover, many approaches remain focused on point-wise alignment or rely on multi-view consistency to enforce constraints, rather than leveraging 'relaxed supervision' during training to induce structural consistency. In contrast, LaxMotion revisits supervision granularity and reformulates relaxed constraints, enabling coherent 3D motion synthesis without coordinate-level memorization or 3D pose fine-tuning.

3 Rethinking Supervision Granularity: LaxMotion

Our goal is to shift the paradigm of 3D motion generation away from rigid, over-determined coordinate point-matching. To this end, we introduce LaxMotion, a framework that rethinks supervision granularity by learning from 2D kinematic cues and structural constraints rather than relying on dense 3D pose labels. Our approach systematically relaxes the coordinate-level objective through three key components: First, a representation reformulation (Sec. 3.1) that transitions the target space from raw 3D points to decoupled kinematic structures; second, a reformulated training paradigm (Sec. 3.2) that relaxes the information source during learning; and finally, a relaxed regularization module (Sec. 3.3) that goes beyond point-matching by enforcing view-invariance and structural coherence, effectively supervising the 3D generation without 3D pose-level annotations.

3.1 Representation Reformulation: From Points to Structures

Instead of treating human motion as a set of over-determined point-level targets, we reformulate it into a structured representation. We decouple motion into a global trajectory and local limb articulations, which naturally preserves kinematic consistency across spatial dimensions. Specifically, let v^{3D} represents a set of 3D limb vectors defined by the skeletal topology:

$$v^{3D} = \{j^{parent(n)} - j^{child(n)}\}_{n=1}^{K} \tag{1}$$

where K is the number of limbs. $j^{parent(n)}$ and $j^{child(n)}$ denote the parent and child joints of n-th limb, respectively. A complete 3D motion sequence $m_{1:T}^{3D}$ over T frames is then factorized into a global trajectory sequence $\tau_{1:T}$ and a relative motion sequence $v_{1:T}^{3D}$:

$$m_{1:T}^{3D} = \{\tau_{1:T}, v_{1:T}^{3D}\} \tag{2}$$

This decoupling explicitly isolates the root translation from the internal body kinematics. Crucially, this structural factorization bridges the gap between 3D space and 2D projections. By defining motion through relative limb configurations rather than absolute points, we establish a representation that remains mathematically consistent under perspective or orthographic projection. This consistency allows us to fundamentally relax the need for dense 3D coordinate supervision. Instead of penalizing deviations from the exact 3D sequence $m_{1:T}^{3D}$, we drive the learning process using a hybrid kinematic observation, denoted as $m_{1:T}^{obs}$. Symmetrically to the 3D formulation, we substitute the over-determined 3D relative pose with its single-view 2D projection $v_{1:T}^{2D}$, while preserving the global trajectory $\tau_{1:T}$:

$$m_{1:T}^{obs} = \{\tau_{1:T}, v_{1:T}^{2D}\} \tag{3}$$

By structurally aligning the generated $\hat{m}_{1:T}^{3D}$ with its observable cues $m_{1:T}^{obs}$, the model is forced to learn the intrinsic 2D-to-3D geometric correspondence. It recovers 3D realizations as consistent explanations of the observed physical structures, rather than merely memorizing dataset-specific 3D point configurations.

3.2 Learning from Relaxed Observability

LaxMotion reformulates the training paradigm by relaxing the information source available during learning. Instead of training the generator with fully specified 3D motion m^{3D} as input, we provide only a partial observation m^{obs} at training time, while still requiring the model to recover complete 3D motion:

$$\hat{m}^{3D} = \mathcal{G}_\theta(m^{obs}) \tag{4}$$

Importantly, this observation signal is used only in training and is not required at inference. By reducing the observability of the training input, the learning process no longer depends on explicitly provided dense 3D pose annotations. More fundamentally, by replacing fully determined 3D inputs with incomplete

cues, the model is prevented from overfitting to exact coordinate patterns. Instead, it is encouraged to infer coherent 3D structure and motion semantics from limited information. The overall training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{relax} + \alpha \cdot \mathcal{L}_{prior} \tag{5}$$

where $\mathcal{L}_{relax}$ denotes Relaxation Regularization (detailed in Sec. 3.3), which encourages the model to generate 3D motions that are consistent with partial 2D observations. As the core supervisory signal, $\mathcal{L}_{relax}$ guides the model to infer coherent and semantically meaningful 3D motions from limited input information.

The term $\mathcal{L}_{prior}$ serves as a structural regularizer that constrains the solution space of the generator. In token-based (e.g., VQ-style) architectures, $\mathcal{L}_{prior}$ is instantiated as a commitment loss, which enforces alignment between continuous latent encodings and a predefined discrete codebook distribution, thereby stabilizing the structured latent space. In diffusion-based architectures, however, the latent distribution is already governed by a fixed Gaussian prior, and structural bias is implicitly encoded through the forward diffusion process and its noise scheduling. Consequently, no additional explicit structural regularization is required (i.e., $\alpha = 0$), and the optimization objective naturally reduces to minimizing the Relaxation Regularization alone.

3.3 Beyond Point-Matching: Relaxation Regularization

In the absence of explicit 3D coordinate ground truth, we relax the supervision granularity through a set of consistency-driven constraints. Instead of performing rigid point-matching, our framework optimizes the model by enforcing structural, view-invariant, and physical coherence.

View-Consistent Structural Regularization. To anchor the generated 3D motion $\hat{m}^{3D} = \{\hat{\tau}, \hat{v}^{3D}\}$ to the observed reality, we first enforce a structural alignment regularization. Specifically, we utilize a weak-perspective projection operator $\Phi(\cdot)$ to map the generated 3D relative limb vectors $\hat{v}^{3D}$ back to the 2D observation space. The structural loss is formulated as:

$$\mathcal{L}_{obs} = \|\Phi(\hat{v}^{3D}) - v^{2D}\|_2^2 + \|\hat{\tau} - \tau\|_2^2 \tag{6}$$

where τ and v^{2D} are the observed trajectory and 2D relative pose cues from m^{obs}, respectively. This term ensures that the generated 3D structure remains a mathematically valid explanation of the original observed evidence.

Cross-View Plausibility Regularization. While $\mathcal{L}_{obs}$ ensures 2D alignment from a single viewpoint, it cannot resolve depth ambiguity. We therefore propose a Cross-View Plausibility constraint to induce 3D consistency. We hypothesize that a physically valid 3D motion must yield 'natural' 2D projections under any arbitrary rotation R. To quantify this 'naturalness', we leverage a pretrained and frozen reconstruction-based 2D motion prior $\mathcal{V}(\cdot)$ (e.g., a 2D relative motion

VQ-VAE). The 3D motion is encouraged to maintain high reconstruction fidelity under random projections:

$$\mathcal{L}_{rec} = \|\mathcal{V}(\Phi(R \cdot \hat{v}^{3D})) - \Phi(R \cdot \hat{v}^{3D})\|_2^2 \tag{7}$$

Unlike traditional multi-view supervision methods that necessitate synchronized multi-camera arrays or calibrated pose labels from multiple perspectives, our Cross-View Plausibility constraint operates under a strictly single-view setting. By shifting the requirement from spatial correspondence across physical cameras to distributional consistency across virtual rotations, our approach significantly lowers the barrier for training 3D motion models, enabling high-fidelity synthesis even from monocular 'in-the-wild' 2D datasets where multi-view annotations are fundamentally unavailable.

Orientation Regularization. To enhance physical plausibility, we introduce an Orientation Regularization grounded in the geometric prior that global body orientation and foot direction are intrinsically coupled. Let $\hat{v}_{ori}$ denote the body orientation vector derived from the normalized cross product of hip vectors. Specifically, the forward projection of the foot direction $\hat{v}_{foot}$ onto the body orientation is constrained to be non-negative:

$$\mathcal{L}_{ori} = \max(0, -\hat{v}_{foot} \cdot \hat{v}_{ori}) \tag{8}$$

where $\hat{v}_{ori}$ can be computed as the normalized cross product of the left and right hip vectors:

$$\hat{v}_{ori} = \frac{\hat{v}_{rhip} \times \hat{v}_{lhip}}{\|\hat{v}_{rhip} \times \hat{v}_{lhip}\|_2} \tag{9}$$

Feature Consistency Regularization. To ensure that the reformulation holds at both the data and feature levels, we introduce Feature Consistency Regularization. The projected motion $\hat{m}^{obs} = \{\hat{\tau}, \Phi(\hat{v}^{3D})\}$ is passed back through one encoder $\mathcal{E}_f(\cdot)$ to ensure alignment between its latent representation and that of the original observation:

$$\mathcal{L}_{feat} = \|\mathcal{E}_f(\hat{m}^{obs}) - \mathcal{E}_f(m^{obs})\|_2^2 \tag{10}$$

The LaxMotion Regularization. The final regularization for our supervision is a weighted combination of these structural and prior-based constraints:

$$\mathcal{L}_{relax} = \mathcal{L}_{obs} + \lambda_1 \cdot \mathcal{L}_{rec} + \lambda_2 \cdot \mathcal{L}_{ori} + \lambda_3 \cdot \mathcal{L}_{feat} \tag{11}$$

where λ are hyperparameters balancing the influence of each geometric constraint.

4 Experiments

Datasets. We evaluate our approach on two major motion-language benchmarks: HumanML3D [8] and KIT-ML [30]. The **HumanML3D** dataset contains 14,616

Table 1: Quantitative evaluation on the HumanML3D test set. ± indicates a 95% confidence interval. '3D Finetune' indicates that LaxMotion is first trained with relaxed supervision to extract features, which are then fused into a 3D-supervised model and fine-tuned.

Supervision	Methods	R Precision↑			FID↓	Diversity→	MModality↑	QM Score↑
		Top 1	Top 2	Top 3				
	Real	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$9.503^{\pm.065}$	-	-
3D Motion	T2M [8]	$0.457^{\pm.002}$	$0.639^{\pm.003}$	$0.740^{\pm.003}$	$1.067^{\pm.002}$	$9.188^{\pm.002}$	$2.090^{\pm.083}$	2.023
	MDM [35]	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$9.559^{\pm.086}$	$2.799^{\pm.072}$	3.795
	MLD [3]	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$9.724^{\pm.082}$	$2.413^{\pm.079}$	3.509
	MotionDiffuse [43]	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$0.630^{\pm.001}$	$9.410^{\pm.049}$	$1.553^{\pm.042}$	1.957
	ReMoDiffuse [44]	$0.510^{\pm.005}$	$0.698^{\pm.006}$	$0.795^{\pm.004}$	$0.103^{\pm.004}$	$9.018^{\pm.075}$	$1.795^{\pm.043}$	5.593
	T2M-GPT [42]	$0.491^{\pm.003}$	$0.680^{\pm.003}$	$0.775^{\pm.002}$	$0.116^{\pm.004}$	$9.761^{\pm.081}$	$1.856^{\pm.011}$	5.449
	MotionLCM [5]	$0.502^{\pm.003}$	$0.698^{\pm.002}$	$0.798^{\pm.002}$	$0.304^{\pm.012}$	$9.607^{\pm.066}$	$2.259^{\pm.092}$	4.097
	MoMask [7]	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$0.045^{\pm.002}$	-	$1.241^{\pm.040}$	5.850
	Fg-T2M++ [39]	$0.513^{\pm.002}$	$0.702^{\pm.002}$	$0.801^{\pm.003}$	$0.089^{\pm.004}$	$9.223^{\pm.114}$	$2.625^{\pm.084}$	8.799
Relaxed Motion	LaxMotion (3D Finetune)	$\mathbf{0.526}^{\pm.003}$	$\mathbf{0.721}^{\pm.004}$	$\mathbf{0.812}^{\pm.004}$	$\mathbf{0.034}^{\pm.002}$	$\mathbf{9.519}^{\pm.079}$	$2.529^{\pm.081}$	**13.715**
	LaxMotion (MDM-Based)	$0.431^{\pm.004}$	$0.624^{\pm.005}$	$0.749^{\pm.006}$	$0.475^{\pm.041}$	$9.523^{\pm.078}$	$\mathbf{2.844}^{\pm.083}$	4.127
	LaxMotion (MoMask-Based)	$0.487^{\pm.002}$	$0.681^{\pm.002}$	$0.780^{\pm.002}$	$0.054^{\pm.003}$	$9.486^{\pm.048}$	$2.046^{\pm.068}$	8.805

Table 2: Quantitative evaluation on the KIT-ML test set. ± indicates a 95% confidence interval. '3D Finetune' indicates that LaxMotion is first trained with relaxed supervision to extract features, which are then fused into a 3D-supervised model and fine-tuned.

Supervision	Methods	R Precision↑			FID↓	Diversity→	MModality↑	QM Score↑
		Top 1	Top 2	Top 3				
	Real	$0.424^{\pm.005}$	$0.649^{\pm.006}$	$0.779^{\pm.006}$	$0.031^{\pm.004}$	$11.08^{\pm.097}$	-	-
3D Motion	T2M [8]	$0.370^{\pm.005}$	$0.569^{\pm.007}$	$0.693^{\pm.007}$	$2.770^{\pm.109}$	$10.91^{\pm.119}$	$1.482^{\pm.065}$	0.890
	MDM [35]	$0.164^{\pm.004}$	$0.291^{\pm.004}$	$0.396^{\pm.004}$	$0.497^{\pm.021}$	$10.85^{\pm.109}$	$1.907^{\pm.214}$	2.705
	MLD [3]	$0.390^{\pm.008}$	$0.609^{\pm.008}$	$0.734^{\pm.007}$	$0.404^{\pm.027}$	$10.80^{\pm.117}$	$2.192^{\pm.071}$	3.449
	MotionDiffuse [43]	$0.417^{\pm.004}$	$0.621^{\pm.004}$	$0.739^{\pm.004}$	$1.954^{\pm.062}$	$\mathbf{11.10}^{\pm.143}$	$0.730^{\pm.013}$	0.522
	ReMoDiffuse [44]	$0.427^{\pm.014}$	$0.641^{\pm.004}$	$0.765^{\pm.055}$	$0.155^{\pm.006}$	$10.80^{\pm.105}$	$1.239^{\pm.028}$	3.147
	T2M-GPT [42]	$0.416^{\pm.006}$	$0.627^{\pm.006}$	$0.745^{\pm.006}$	$0.514^{\pm.029}$	$10.92^{\pm.108}$	$1.570^{\pm.039}$	2.190
	MoMask [7]	$0.433^{\pm.007}$	$0.656^{\pm.005}$	$0.781^{\pm.005}$	$0.204^{\pm.011}$	-	$1.131^{\pm.043}$	2.504
Relaxed Motion	LaxMotion (3D Finetune)	$\mathbf{0.445}^{\pm.005}$	$\mathbf{0.672}^{\pm.005}$	$\mathbf{0.801}^{\pm.006}$	$\mathbf{0.135}^{\pm.010}$	$11.01^{\pm.097}$	$2.540^{\pm.080}$	**6.913**
	LaxMotion (MDM-Based)	$0.436^{\pm.004}$	$0.644^{\pm.004}$	$0.757^{\pm.005}$	$0.294^{\pm.019}$	$10.49^{\pm.112}$	$\mathbf{2.698}^{\pm.084}$	4.976
	LaxMotion (MoMask-Based)	$0.413^{\pm.008}$	$0.634^{\pm.007}$	$0.763^{\pm.007}$	$0.248^{\pm.016}$	$11.25^{\pm.109}$	$2.481^{\pm.074}$	4.982

motion samples from the AMASS [25] and HumanAct12 [10] datasets, each paired with three unique text descriptions, totaling 44,970 descriptions. The **KIT-ML** [27] dataset integrates data from multiple motion capture sources into a unified format, with 3,911 motion sequences and 6,278 text descriptions, providing a smaller benchmark for evaluation.

Evaluation Metrics. We adopt several standard metrics to evaluate motion quality, diversity, and text–motion alignment. FID measures the distance between generated and real motion feature distributions, where lower values indicate higher realism. R-Precision assesses semantic alignment by checking whether the correct text ranks within the top-k retrieved ($k = 1, 2, 3$). MMDist is the average Euclidean distance between motion and text features, reflecting cross-modal consistency. Diversity quantifies variability across different text inputs, while MultiModality (MModality) measures variance among motions generated

Table 3: Comparison with 2D-supervised methods on the HumanML3D test set under global and relative motion settings. 'Relative' denotes the setting without trajectory.

Setting	Method	R-Precision↑	FID↓	MMDist↓	Diversity→	MModality↑	QM Score↑
Global	MAS [16]	0.418	11.893	5.606	6.413	–	–
	Motion-2-to-3 [29]	0.697	0.321	3.579	9.286	1.892	3.339
	LaxMotion (MoMask-Based)	**0.780**	**0.054**	**3.155**	**9.486**	**2.046**	**8.805**
Relative	Real	0.713	–	3.791	7.731	–	–
	Motion-2-to-3 [29]	0.602	0.779	4.534	8.450	1.148	1.301
	MoMask [7]	0.680	0.154	4.168	8.247	0.996	2.538
	LaxMotion (MoMask-Based)	**0.683**	**0.137**	**4.056**	**8.190**	**1.311**	**3.542**

from the same text, capturing conditional diversity. To jointly evaluate overall generation quality and diversity under identical conditions, we introduce the Quality–MModeality Score (QM), defined as:

$$\text{QM Score} = \frac{\text{MModality}}{\sqrt{\text{FID}}} \tag{12}$$

A higher QM indicates a better balance between motion quality and diversity, providing a more holistic evaluation of generative performance.

4.1 Implementation Details.

Model Details. For the token-based pipeline, we adopt MoMask [7] as the baseline architecture. For the diffusion-based pipeline, we adopt MDM [35] as the baseline framework. In both cases, all architectural designs strictly follow their respective baselines to ensure fair comparison. For the VQ-VAE that learns the 2D motion distribution, we use a single codebook while keeping the other configurations identical.

Training Details. In our experiments, LaxMotion is implemented under two representative generation paradigms. For both paradigms, we set $\lambda_1 = \lambda_2 = \lambda_3 = 1$ during training. For the token-based pipeline, we follow MoMask [7] in terms of training protocol with the sole exception of setting $\alpha = 2$. For the diffusion-based pipeline, we adopt MDM [35] as the baseline and keep all training hyperparameters consistent with its original implementation. All experiments are conducted on a single RTX 3090 GPU.

4.2 Quantitative Results.

Comparison on HumanML3D. As the first 3D motion framework trained primarily under 2D pose cues and 3D trajectories, LaxMotion challenges the convention that coordinate-level 3D supervision is indispensable. As shown in Tab. 1, LaxMotion achieves a competitive FID of 0.054, performing on par with fully 3D-supervised SOTA methods despite avoiding the over-determined point-matching

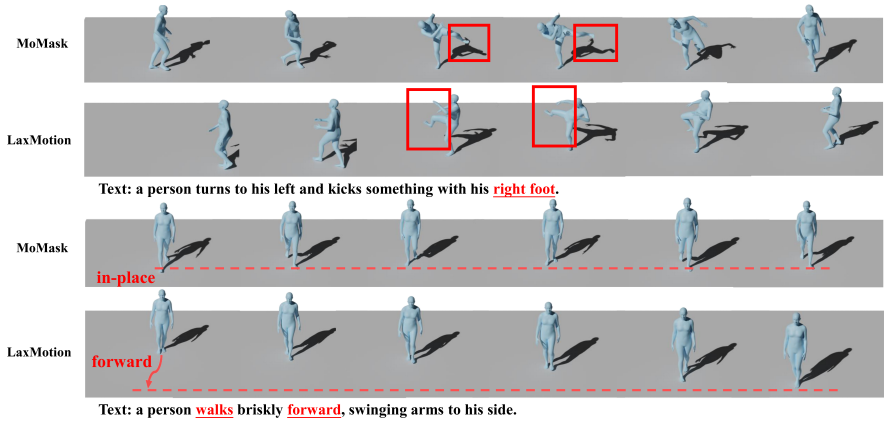

Fig. 3: Qualitative comparison on HumanML3D dataset. Compared with MoMask [7], our method demonstrates better generalization and a more accurate understanding of textual semantics, producing motions that more closely align with the given text descriptions.

objectives typical of coordinate regression. Notably, LaxMotion attains the highest QM score, reflecting its superior ability to navigate the one-to-many nature of text-to-motion generation by balancing fidelity with multimodality. When fused and fine-tuned with 3D-supervised representations [7], the FID further improves to 0.038, establishing a new SOTA while yielding substantial gains in Diversity. Detailed descriptions of the fusion strategy are provided in the supplementary material. This synergy confirms that 2D-derived representations encode essential structural invariances that effectively complement and regularize traditional 3D modeling. Beyond comparison with fully supervised models, LaxMotion significantly outperforms prior 2D-based approaches across all metrics under the global motion setting (Tab. 3). This demonstrates that by explicitly supervising the 3D trajectory alongside 2D kinematics, our framework generates globally coherent trajectories rather than merely lifting isolated poses. Under the more controlled relative motion setting, LaxMotion remains highly competitive, even surpassing the 3D-supervised MoMask [7]. This suggests that shifting from exact coordinate reproduction to structural consistency induces a more generalized and expressive motion representation.

Comparison on KIT-ML. As shown in Tab. 2, we further evaluate LaxMotion on KIT-ML. The results demonstrate that LaxMotion, supervised only by 3D trajectories and 2D pose cues, maintains performance competitive with fully 3D-supervised SOTA methods in FID and R-Precision. Most notably, it achieves the top QM score, reinforcing its superior capacity to balance motion realism with generative diversity—a direct benefit of our relaxed supervision strategy. Furthermore, fusing 2D-derived features with 3D-supervised models yields a new SOTA, confirming that 2D-learned representations capture rich structural and

Fig. 4: Comparison results on in-the-wild text prompts.

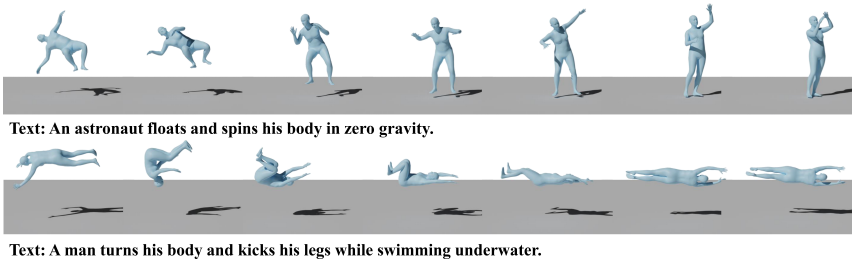

Fig. 5: Qualitative Results on In-the-Wild Motions. LaxMotion is capable of learning from and generating motions extracted from in-the-wild 2D videos. We selected several actions that cannot be physically performed or captured with 3D motion sensors under existing conditions, extracted their relaxed motion sequences, and used them for supervision. LaxMotion successfully generates high-quality 3D motions for these challenging scenarios, despite the absence of 3D pose annotations.

Table 4: Ablation Study on Relaxation.

Input	Supervision	R Precision↑ Top 1	Top 2	Top 3	FID↓	Diversity→	MModality↑	QM Score↑
-	-	0.511	0.703	0.797	0.002	9.503	-	-
Full	Relax	0.511	0.706	0.806	0.040	9.684	2.240	11.200
Relax	Full	0.507	0.688	0.791	0.048	9.805	1.288	5.879
Relax	Relax	**0.526**	**0.721**	**0.812**	**0.034**	**9.519**	**2.529**	**13.715**

semantic invariants that are often bypassed by traditional 3D coordinate regression. As hypothesized in our Introduction, these findings suggest that while precise 3D pose labels provide strong numerical anchors, they can overconstrain the learning process, leading to a 'point-matching' trap. In contrast, the structural abstraction inherent in 2D supervision compels the model to infer underlying kinematic relationships rather than memorizing dataset-specific coordinates. This shift from rote memorization to structural reasoning ultimately results in more robust and diverse motion generation.

Table 5: Ablation Study on Relaxation Regularization.

Config.	R Precision↑			FID↓	MMDist↓	Diversity→	MModality↑	QM Score↑
	Top 1	Top 2	Top 3					
Real	0.511	0.703	0.797	0.002	2.974	9.503	-	-
w.o. $\mathcal{L}_{rec}$	0.368	0.551	0.678	2.588	3.912	8.449	2.127	1.322
w.o. $\mathcal{L}_{ori}$	0.465	0.656	0.756	0.124	3.282	9.299	**2.225**	6.319
w.o. $\mathcal{L}_{feat}$	**0.501**	**0.686**	0.780	0.088	**3.125**	9.450	2.086	7.032
Ours	0.487	0.681	**0.780**	**0.054**	3.155	**9.486**	2.046	**8.805**

4.3 Qualitative Results.

Qualitative comparisons with SOTA 3D-supervised methods are shown in Fig. 3 and Fig. 4. Compared to MoMask [7], LaxMotion generates motions more aligned with textual semantics, demonstrating that relaxed supervision improves generalization. Since LaxMotion requires no 3D pose-level annotations, it can be directly applied to in-the-wild video data, learning motions difficult or impossible to capture in 3D. As shown in Fig. 5, it synthesizes realistic microgravity and underwater motions. Additional results are provided in the supplementary material, highlighting LaxMotion's potential for scalable and diverse 3D motion generation.

4.4 Ablation Studies.

Ablation Study on Relaxation. Tab. 4 reports the performance of the final model after fine-tuning with features extracted from models trained under varying granularities. The results show that features derived from more relaxed supervision significantly improve performance, indicating that relaxing the supervision granularity enables the model to capture more generalizable motion representations rather than overfitting to exact coordinates. When combined with partial observation during training, this relaxation further enhances the generalization and diversity of motions produced by the feature-fused model, demonstrating the practical benefits of loosening input constraints during training.

Ablation Study on Regularization. Since LaxMotion relies on 3D trajectories and 2D pose cues, the role of our Relaxation Regularization is pivotal for bridging the 2D-to-3D structural gap. The results in Tab. 5 verify the necessity of all three components. Among them, the 2D distribution reconstruction plays the most critical role in improving generation quality. By reconstructing 2D motions of different views, it effectively constrains the multi-view consistency of the generated results and helps the model learn a more accurate 3D spatial structure from purely 2D cues. In addition, both feature consistency regularization and orientation regularization further enhance motion generation quality. The orientation regularization encourages the model to learn consistent and physically plausible body orientations across frames, thereby improving the temporal coherence and realism of generated motions. Meanwhile, feature consistency regularization acts

Table 6: Ablation Study on 2D Distribution Learning.

Config.	R Precision↑			FID↓	MMDist↓	Diversity→	MModality↑	QM Score↑
	Top 1	Top 2	Top 3					
Real	0.511	0.703	0.797	0.002	2.974	9.503	-	-
with VAE	0.467	0.657	0.753	0.121	3.247	9.218	**2.168**	6.233
with AE	0.450	0.648	0.753	0.258	3.295	9.198	2.040	4.016
Ours	**0.487**	**0.681**	**0.780**	**0.054**	**3.155**	**9.486**	2.046	**8.805**

Table 7: Comparison of token-based reconstructions on various representations.

Config.	MPJPE↓ (mm)	R Precision↑			FID↓	MMDist↓	Diversity→
		Top 1	Top 2	Top 3			
Real	0	0.511	0.703	0.797	0.002	2.974	9.503
with joint coordinates	106.9	0.390	0.574	0.688	3.902	3.843	8.303
Ours	**56.4**	**0.487**	**0.681**	**0.780**	**0.054**	**3.155**	**9.486**

as a latent-space constraint that regularizes the distribution of encoded features, stabilizing the representation learning process and helping the model better disentangle motion semantics from spatial geometry. Collectively, these components enable LaxMotion to learn 3D-aware representations from 2D pose cues.

Ablation Study on 2D Prior Distribution Learning. To learn the prior distribution of relative 2D motions, LaxMotion adopts a VQ-VAE architecture for modeling. To verify its necessity, we compare the performance of using VAE and plain AE for distribution learning. As shown in Tab. 6, the results indicate that the performance of VAE is significantly weaker than that of VQ-VAE, while AE performs even worse. This performance gap arises because VQ-VAE, through vector quantization, introduces a discrete representation in the latent space, enabling the model to learn more stable and semantically consistent motion patterns, thus better capturing the structural characteristics of 2D motion distributions. In contrast, the VAE relies on continuous latent variables that often lead to blurred distribution boundaries, while the AE lacks explicit distribution modeling capability and can only perform point-to-point reconstruction, failing to capture the global statistical regularities of motion data. Therefore, VQ-VAE demonstrates clear superiority in learning a high-quality prior of 2D motion distributions, serving as a crucial component that supports ours training paradigm.

Ablation Study on LaxMotion Representation. In our experiments, LaxMotion adopts a limb vector–based representation. This choice is motivated by the fact that limb vectors naturally encode the relative structural relationships of the human skeleton, making it easier for the model to capture geometric constraints and temporal coherence of motions. Additionally, this representation reduces the influence of scale differences between joints on feature learning, thereby improv-

ing the stability and accuracy of 3D motion generation. In Tab. 7, we evaluate the impact of different kinematic representations by conducting reconstruction experiments on a token-based VQ-VAE pipeline, including an experiment with joint coordinates directly as the representation. However, this approach ignores the internal skeletal topology, making the model more dependent on absolute positions and more sensitive to noise and scale variations. As a result, the reconstruction quality drops significantly, leading to a substantial performance gap compared to LaxMotion using limb vectors.

5 Conclusion

We introduced LaxMotion, a framework that rethinks supervision granularity in text-to-3D motion generation. Instead of regressing toward exact 3D coordinates, LaxMotion learns 3D motion as a consistent explanation of global trajectories and monocular 2D kinematic cues. By shifting from point-level pose matching to relaxed supervision with consistency-driven regularization, our approach mitigates the over-determined nature of dense 3D labels and better aligns the objective with the one-to-many nature of generative motion. Despite not applying direct 3D pose losses, LaxMotion produces natural, diverse, and semantically faithful 3D motions, achieving competitive or superior results compared to fully supervised baselines. These results suggest that structural consistency, rather than exact coordinate memorization, is a more scalable and generalizable principle for 3D motion generation.

References

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
2. Chen, C.H., Tyagi, A., Agrawal, A., Drover, D., Mv, R., Stojanov, S., Rehg, J.M.: Unsupervised 3d pose estimation with geometric self-supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5714–5724 (2019)
3. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18000–18010 (2023)
4. Chen, Z., Liu, X., Sheng, B., Li, P.: Garnet: Graph attention residual networks based on adversarial learning for 3d human pose estimation. In: Advances in Computer Graphics: 37th Computer Graphics International Conference, CGI 2020, Geneva, Switzerland, October 20–23, 2020, Proceedings. pp. 276–287. Springer (2020)
5. Dai, W., Chen, L.H., Wang, J., Liu, J., Dai, B., Tang, Y.: Motionlcm: Real-time controllable motion generation via latent consistency model. In: European Conference on Computer Vision. pp. 390–408. Springer (2024)
6. Drover, D., MV, R., Chen, C.H., Agrawal, A., Tyagi, A., Phuoc Huynh, C.: Can 3d pose be learned from 2d projections alone? In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)

7. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1900–1910 (2024)
8. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5152–5161 (2022)
9. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: European Conference on Computer Vision. pp. 580–597. Springer (2022)
10. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2021–2029 (2020)
11. Habekost, J., Shiratori, T., Ye, Y., Komura, T.: Learning 3d global human motion estimation from unpaired, disjoint datasets. In: BMVC (2020)
12. Habibie, I., Xu, W., Mehta, D., Pons-Moll, G., Theobalt, C.: In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10905–10914 (2019)
13. Hardy, P., Kim, H.: Links" lifting independent keypoints"-partial pose lifting for occlusion handling with improved accuracy in 2d-3d human pose estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3426–3435 (2024)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
15. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. Advances in Neural Information Processing Systems **36**, 20067–20079 (2023)
16. Kapon, R., Tevet, G., Cohen-Or, D., Bermano, A.H.: Mas: Multi-view ancestral sampling for 3d motion generation using 2d diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1965–1974 (2024)
17. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022)
18. Kundu, J.N., Seth, S., Jampani, V., Rakesh, M., Babu, R.V., Chakraborty, A.: Self-supervised 3d human pose estimation via part guided novel image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6152–6162 (2020)
19. Lee, D., Kim, C., Kim, S., Cho, M., Han, W.S.: Autoregressive image generation using residual quantization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11523–11532 (2022)
20. Li, J., Liu, C.K., Wu, J.: Lifting motion to the 3d world via 2d diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17518–17528 (2025)
21. Li, W., Dai, B., Zhou, Z., Yao, Q., Wang, B.: Controlling character motions without observable driving source. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6194–6203 (2024)
22. Liang, H., Bao, J., Zhang, R., Ren, S., Xu, Y., Yang, S., Chen, X., Yu, J., Xu, L.: Omg: Towards open-vocabulary motion generation via mixture of controllers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 482–493 (2024)

23. Liu, S., Li, Y., Du, S.: Urpose: The model with unbiased rectified projection and reconstruction error for monocular unsupervised 3d human pose estimation. Neurocomputing p. 133036 (2026)
24. Liu, S., Liang, Y., Wang, J., Du, S., Zhang, C., Li, X.: Uni-inter: Unifying 3d human motion synthesis across diverse interaction contexts. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–11 (2025)
25. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5442–5451 (2019)
26. Meng, Z., Xie, Y., Peng, X., Han, Z., Jiang, H.: Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27859–27871 (2025)
27. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10985–10995 (2021)
28. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision. pp. 480–497. Springer (2022)
29. Pi, H., Guo, R., Shen, Z., Shuai, Q., Hu, Z., Wang, Z., Dong, Y., Hu, R., Komura, T., Peng, S., Zhou, X.: Motion-2-to-3: Leveraging 2d motion data for 3d motion generations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14305–14316 (October 2025)
30. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big data **4**(4), 236–252 (2016)
31. Qin, Z., Wang, Y., Yang, M., Zhou, S., Yang, M., Wang, L.: Embracing aleatoric uncertainty: Generating diverse 3d human motion. arXiv preprint arXiv:2508.20604 (2025)
32. Ren, Z., Huang, S., Li, X.: Realistic human motion generation with cross-diffusion models. In: European Conference on Computer Vision. pp. 345–362. Springer (2024)
33. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
34. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: Proceedings of the 40th International Conference on Machine Learning. pp. 32211–32252 (2023)
35. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=SJ1kSyO2jwu
36. Wandt, B., Little, J.J., Rhodin, H.: Elepose: Unsupervised 3d human pose estimation by predicting camera elevation and learning normalizing flows on 2d poses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6635–6645 (2022)
37. Wandt, B., Rosenhahn, B.: Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7782–7791 (2019)
38. Wang, C., Kong, C., Lucey, S.: Distill knowledge from nrsfm for weakly supervised 3d pose learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 743–752 (2019)

39. Wang, Y., Li, M., Liu, J., Leng, Z., Li, F.W., Zhang, Z., Liang, X.: Fg-t2m++: Llms-augmented fine-grained text driven human motion generation. International Journal of Computer Vision pp. 1–17 (2025)
40. Yu, Z., Ni, B., Xu, J., Wang, J., Zhao, C., Zhang, W.: Towards alleviating the modeling ambiguity of unsupervised monocular 3d human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8651–8660 (2021)
41. Zanfir, A., Bazavan, E.G., Xu, H., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. pp. 465–481. Springer (2020)
42. Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., Shen, X., Shan, Y.: Generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14730–14740 (2023)
43. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE transactions on pattern analysis and machine intelligence **46**(6), 4115–4128 (2024)
44. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 364–373 (2023)