VersatileMotion: A Unified Framework for Motion Synthesis and Comprehension

Zeyu Ling[1], Bo Han[2], Shiyang Li[1], Jikang Cheng[3], Hongdeng Shen[3], and Changqing Zou[1,4⋆]

[1] Zhejiang University, Hangzhou, China
{zeyuling, 12521030}@zju.edu.cn
[2] ByteDance, Hangzhou, China
borishan.815@bytedance.com
[3] University of Chinese Academy of Sciences, Hangzhou, China
{chengjikang23, shenhongdeng23}@mails.ucas.ac.cn
[4] Zhejiang Lab, Hangzhou, China
changqing.zou@zju.edu.cn

Fig. 1: VersatileMotion supports both single-agent and multi-agent motion synthesis and understanding, enabling seamless cross-modal conversion among text, audio, and motion (both single and multi-agent).

Abstract. Human motion modeling must support generation from text, music, and speech, comprehension through captioning, temporal completion, and multi-person coordination—yet existing methods typically tackle each scenario with a separate, task-specific system. We argue that discrete tokenization plus autoregressive (AR) modeling possesses unique *compositional* advantages for building a unified motion model: all modalities share a single vocabulary and training objective, tasks are defined by token arrangement rather than architectural changes, and the same model naturally supports both generation and understanding by simply reordering the sequence. However, realizing this potential

⋆ Corresponding author.

has been blocked by two obstacles. First, current motion tokenizers use root-relative, non-causal, frame-level encodings that discard world-space information, break consistency with forward-only AR decoding, and scale poorly—confining existing discrete AR models to text–motion pairs in single-human settings. Second, multi-modal motion data with paired text and audio annotations remains scarce. We address both obstacles. We introduce VerMoVQ, a causal spatio-temporal VQ-VAE that factorizes each frame into low-dimensional *body tokens* while preserving world-space root trajectories, naturally enabling multi-person tokenization via per-agent concatenation and arbitrary-length encoding through strict temporal causality. We curate MotionHub, a ∼380K-clip corpus in unified SMPL representation with large-scale multi-modal annotations spanning nine motion tasks. Built on these foundations, we train VersatileMotion, a single decoder-only AR Transformer that covers text↔motion, music↔dance, speech→gesture, motion prediction, in-betweening, and multi-person interaction—all within one model. Across nine benchmarks, VersatileMotion surpasses strong task-specific baselines on seven and enables cross-modal compositions (e.g., text+audio→dance) through its compositional token structure without task-specific architectural changes. Code and data are available at ZeyuLing/VersatileMotion.

Keywords: Human motion synthesis · Motion understanding · Unified multimodal model

1 Introduction

Human motion modeling spans a diverse set of scenarios: generating motion from text, music, or speech; understanding motion through captioning; predicting or in-filling temporal segments; and coordinating multiple interacting agents (Fig 1). A truly *unified* system that handles all of these within a single model would not only eliminate per-task engineering but also enable cross-task knowledge transfer—learning richer motion priors from the combined data of all tasks.

Two dominant paradigms have emerged. *Diffusion-based* methods operate in pose or latent space and deliver high-fidelity results on individual generation tasks [10, 46, 71]. However, each task typically requires its own conditioning architecture (cross-attention for text, rhythm alignment for music, prosody encoders for speech), and diffusion models cannot natively perform *understanding* tasks such as motion-to-text captioning—they only map noise to data, not motion to language. Multi-person settings further demand specialized architectures (e.g., ComMDM [66], InterGen [40]). As a result, diffusion-based methods are fundamentally task-specific and difficult to unify.

Discrete approaches first tokenize motion via a VQ-VAE and then apply an autoregressive (AR) Transformer. This paradigm offers a distinct structural advantage: once all modalities—text, audio, and motion—are mapped into a shared discrete vocabulary, tasks reduce to different *arrangements* of the same token types, and the model simply performs next-token prediction regardless of the specific task. The same formulation naturally supports both generation

(condition→motion) and understanding (motion→text) by reordering the token sequence. Adding a new modality means extending the vocabulary; adding a new task means defining a new token arrangement—neither requires architectural changes. This *compositionality* is the same property that makes large language models effective across diverse NLP tasks [5] and has been demonstrated in vision via unified token vocabularies [70].

Early motion instantiations of this paradigm—TM2T [24], MotionGPT [31], M^3GPT [56]—validated the basic pipeline: VQ-VAE tokenization followed by a GPT-style decoder can handle text↔motion and even incorporate music. Yet these systems remain far from the compositional potential of the discrete AR framework: they are largely confined to text–motion pairs in single-human settings. We identify two concrete obstacles. **(1) Tokenizer limitations.** Current motion VQ-VAEs use coordinate-centric, root-relative, non-causal, frame-level encodings [23,40]: they normalize away world-space trajectories (precluding multi-person spatial reasoning), compress each frame into a single high-dimensional vector (entangling body structure), and use bidirectional convolutions (creating a mismatch with forward-only AR decoding). **(2) Data coverage.** Most training corpora provide only text–motion pairs; audio-conditioned (music–dance, speech–gesture) and multi-person clips are scarce and lack text annotations.

In this work, we remove both obstacles and demonstrate, for the first time, that a single discrete AR model can competitively handle nine diverse motion generation and understanding tasks—including multi-person scenarios and cross-modal compositions—within a unified framework.

Technical contributions. To unlock the AR model's compositional potential, we introduce VerMoVQ, a causal spatio-temporal motion VQ-VAE. VerMoVQ factorizes each frame into low-dimensional *body tokens*—one per joint group—in SMPL-style parameterization that retains world-space root trajectories. Causal temporal convolutions ensure forward-only consistency with AR generation, while spatial body-attention blocks capture inter-joint couplings. This design naturally extends to multi-person sequences: agents share the same codebook and their token streams are simply concatenated.

For data coverage, we curate **MotionHub**, a unified corpus of ∼380K single- and multi-human motion clips standardized to SMPL representation. For domains lacking text—especially music–dance and speech–gesture—we perform large-scale human and VLM-assisted captioning, producing paired multi-modal examples rarely available in public motion corpora. We establish **MotionBench**, a standardized evaluation suite covering nine motion tasks.

Built on these foundations, we train VersatileMotion, a single decoder-only AR Transformer over a unified motion–text–audio vocabulary, using two-stage generalist training that balances breadth across all tasks with depth in data-scarce domains. In evaluations across nine benchmarks, VersatileMotion exceeds prior state-of-the-art on seven while enabling flexible cross-modal behaviors (e.g., text+audio→dance) through its compositional token structure without task-specific architectural changes.

In summary, our contributions are:

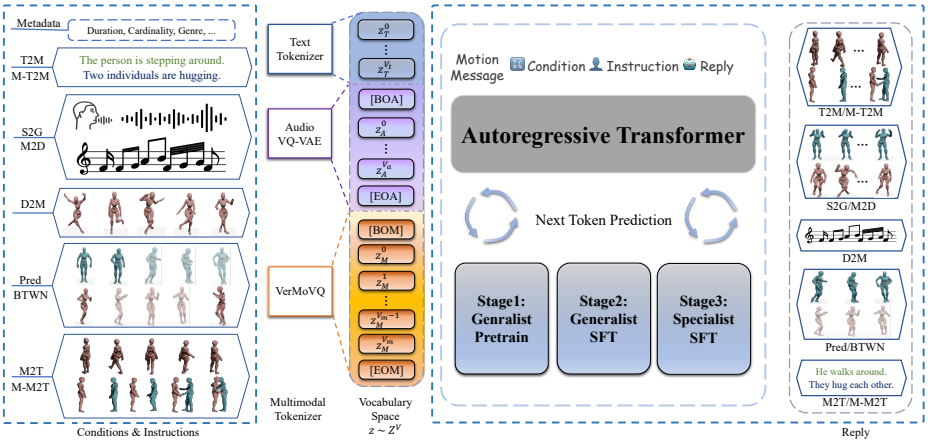

Fig. 2: Overview of VersatileMotion. All modalities (text, audio, motion) are tokenized into a shared vocabulary. Different tasks correspond to different *arrangements* of the same token types, processed by a single AR Transformer with one training objective (next-token prediction). Multi-person motion is handled by per-agent token concatenation with a shared codebook, requiring no architectural changes.

– We demonstrate that the discrete AR paradigm possesses unique compositional advantages for unified motion modeling, and present VersatileMotion, the first single model that competitively handles nine motion generation and understanding tasks spanning text, audio, single-person, and multi-person scenarios.
– To unlock this compositional potential, we design VerMoVQ, a causal spatio-temporal VQ-VAE with per-joint body tokens and world-space root preservation that removes the tokenizer-level barriers confining prior discrete AR models to narrow task scopes.
– We curate MotionHub (∼380K clips, unified SMPL, large-scale multi-modal annotations) and establish MotionBench, a standardized nine-task evaluation suite for motion generation and understanding.

2 Related Work

2.1 Motion Generation and Comprehension

Motion tasks are commonly split into *generation* (text-to-motion, music-to-dance, speech-to-gesture) and *comprehension* (motion-to-text, dance-to-music, motion-to-music), with most methods following either diffusion or discrete autoregressive (AR) routes. For **text** $\leftrightarrow$ **motion**, diffusion models operate in pose or latent space and yield strong realism and diversity [10, 40, 45, 71]; in parallel, AR approaches tokenize motion via a VQ-VAE and train Transformers on token

Table 1: Unified multitasking comparison. *T2M*: text-to-motion, *M2D*: music-to-dance, *S2G*: speech-to-gesture, *M-T2M*: multi-agent T2M, *M2M*: motion prediction/in-between, *M2T*: motion-to-text, *D2M*: dance-to-music, *M-M2T*: multi-agent M2T.

Methods	Synthesis					Comprehension		
	T2M	M2D	S2G	M-T2M	M2M	M2T	D2M	M-M2T
MDM [71]	✓	✗	✗	✓	✗	✗	✗	✗
TM2T [24]	✓	✗	✗	✗	✗	✓	✗	✗
TM2D [20]	✓	✓	✗	✗	✗	✗	✗	✗
UDE [94]	✓	✓	✗	✗	✗	✗	✗	✗
MCM [45]	✓	✓	✓	✗	✗	✗	✗	✗
MotionGPT [31]	✓	✗	✗	✗	✓	✓	✗	✗
LMM [89]	✓	✓	✗	✗	✓	✗	✗	✗
UDE2 [93]	✓	✓	✓	✗	✗	✗	✗	✗
UniMuMo [78]	✗	✓	✗	✗	✗	✗	✓	✗
LoM [8]	✓	✗	✓	✗	✗	✗	✗	✗
M^3GPT [56]	✓	✓	✗	✗	✓	✓	✓	✗
MotionAgent [75]	✓	✗	✗	✓	✗	✓	✗	✗
VersatileMotion	✓	✓	✓	✓	✓	✓	✓	✓

sequences [22, 24, 54, 64, 86]. TM2T and T2M-GPT showed that such tokens support reciprocal text–motion generation [24, 86], inspiring a wave of discrete-token frameworks. However, most target single-person motion and text-only conditioning, and are trained on coordinate-based, root-relative formats (e.g., HumanML3D [23]) in which global trajectories are normalized to emphasize *relative* rather than absolute world-space motion [40].

Music ↔ dance extends these ideas to audio: generation uses diffusion [1, 37, 72] or AR Transformers [26, 35, 38, 68], while the reverse direction treats music as tokens and learns sequence-to-sequence mappings from motion [13, 18, 25, 96]. **Speech-to-gesture** focuses on expressive upper-body motion synchronized with speech and often decouples hands/face from the torso [48, 49, 80].

Despite rich cross-modal correspondences, these systems are typically specialized to a direction or modality pair and do not share a common tokenizer across tasks (Tab 1).

2.2 Discrete Motion Tokenizers and Motion LLMs

The *motion tokenizer* underpins the motion-as-language paradigm. Early systems adopt HumanML3D-style, coordinate-centric representations [23, 40, 43] mixing root-relative joint positions, velocities, contacts, and sometimes rotations; AR models are then trained on fixed-length windows of discrete codes [22, 24, 54, 64, 86]. Normalizing global root motion simplifies single-person learning but complicates multi-human interaction, where absolute spatial relations are essential.

To raise quantization fidelity, MoMask introduces hierarchical residual tokens for high-frequency detail [22]; MMM couples masked modeling with a tokenizer for editing and inpainting [64]; ScaMo and Go-To-Zero use FSQ-based motion VAEs and large-scale AR training [14, 55]; and DisCoRD adds a rectified-flow decoder in latent space to reduce artifacts [11]. These advances lower tokenization noise on single-person text-to-motion, yet most tokenizers remain non-causal in time and rooted in single-human, root-relative coordinates.

Built atop these tokenizers, **motion LLMs** apply sequence models to the resulting motion tokens. T2M-GPT and TM2T use GPT-style decoders [24, 86]; MotionGPT/MotionGPT-2 adopt T5 to unify text–motion tasks [31, 73]; LMM builds a MoE over FineMoGen [89, 90]. Language of Motion, M^{3}GPT, and UDE-2 add music or multi-task conditioning [8, 56, 93]; UniMuMo targets dance and music–text conversion [78]; and MotionLLM / Motion-Agent enable conversational control [9, 75]. Nevertheless, they largely inherit tokenizer constraints: non-causality, single-human orientation, root-relative motion, and a narrow task set centered on text–motion.

3 VersatileMotion Framework

Given text, audio, and/or past-motion conditions, our goal is a single model that can generate or understand 3D human motion for one or more agents across diverse tasks. The discrete AR paradigm offers a natural solution: once all modalities are tokenized into a shared vocabulary, tasks become different arrangements of the same token types, and the model simply performs next-token prediction regardless of the specific task. Realizing this potential, however, requires a motion tokenizer that does not artificially restrict the AR model's versatility. We first describe the motion representation (Sec 3.1) and the motion tokenizer VerMoVQ (Sec 3.2), then present the unified AR Transformer VersatileMotion (Sec 3.3). The overall framework is illustrated in Fig 2.

3.1 Motion Representation

For the AR model to reason about multi-person interactions and produce deployment-ready outputs, the motion representation must preserve absolute world-space positions—root-relative formats eliminate this information by design. Yet most motion pipelines [23, 40] use a coordinate-centric, HumanML3D-style scheme that normalizes the pelvis to the origin and expresses all joints root-relatively, discarding global trajectories. While convenient for single-person learning, this format requires post-hoc IK/retargeting at inference and, critically, cannot represent the spatial relations between interacting agents.

We instead encode motion directly in a SMPL-style state that preserves global motion. For each frame t,

$$x_t = \big(p_t,\, v_t,\, R_t\big), \tag{1}$$

where $p_t, v_t \in \mathbb{R}^3$ are world-space root position and velocity, and $R_t \in \mathbb{R}^{J\times 6}$ stacks continuous 6D rotations for all J joints. Retaining p_t produces deployment-ready

SMPL parameters without post-processing and enables multi-human tokenization as simple per-agent concatenation in a shared code space.

3.2 VerMoVQ: A Causal Spatio-Temporal Motion VQ-VAE

For the AR model to exploit its compositional potential across multi-person, multi-modal, and temporal scenarios, the motion tokenizer must satisfy three requirements: (1) preserve world-space positions for multi-agent reasoning, (2) produce tokens compatible with forward-only generation, and (3) scale gracefully to large and diverse corpora. Current motion tokenizers violate all three: they normalize away global trajectories, use bidirectional convolutions, and compress each frame into a single high-dimensional vector. We address these with VerMoVQ.

We represent each clip as a spatio-temporal tensor

$$X \in \mathbb{R}^{T \times K \times 6}, \tag{2}$$

where T is the number of frames and K the number of body tokens per frame. Each token is 6D: one token packs root position and velocity $[p_t, v_t] \in \mathbb{R}^6$, and the remaining $K-1$ tokens hold 6D joint rotations $r_{t,j} \in \mathbb{R}^6$. Unlike *frame tokens* that compress an entire frame into a single high-dimensional vector, we use low-dimensional *body tokens* to factor motion across joints while keeping the embedding size fixed—a design choice whose superiority we verify in ablations (Sec 4.6). Formally, VerMoVQ comprises a spatio-temporal encoder E, a vector quantizer Q with codebook $\mathcal{C}$, and a spatio-temporal decoder $\mathcal{D}$:

$$Z = E(X) \in \mathbb{R}^{L \times K \times d}, \qquad \hat{X} = \mathcal{D}\big(Q(Z)\big), \tag{3}$$

where $L \leq T$ is the downsampled temporal length and d the latent channel width. Quantization is applied *per* time step and body token: for each $z_{\ell,k} \in \mathbb{R}^d$, Q selects its nearest codeword in $\mathcal{C}$, yielding a discrete tensor of indices $Q(Z) \in \{1, \ldots, K_c\}^{L \times K}$. For the AR Transformer, we flatten (ℓ, k) into a 1D token stream.

Causal temporal convolutions. A non-causal tokenizer that conditions on future frames during encoding creates a systematic mismatch with the AR Transformer, which generates tokens in a forward-only manner. To eliminate this discrepancy, all temporal convolutions in VerMoVQ are strictly *causal*: when encoding or decoding frame t, the receptive field includes only frames $\leq t$, never future frames. This ensures forward-only consistency and makes the tokenizer suitable for prediction, streaming, and autoregressive chaining of variable-length segments.

Spatial body-attention blocks. While temporal convolutions capture dynamics across time, they process each body token independently along the spatial axis, ignoring the physical couplings between joints. To model these inter-joint dependencies, we process the K body tokens $\{X_{t,k}\}_{k=1}^{K}$ at each time step t with self-attention along the body dimension, capturing interactions between global motion and articulated rotations.

Quantization. We adopt finite scalar quantization (FSQ) [60], which quantizes each latent dimension using fixed scalar levels and requires no codebook or commitment losses. This produces a token tensor $I \in \{0, \ldots, K_c{-}1\}^{L \times K}$, flattened over (ℓ, k) for AR training.

Training objective. Let $x_t = (p_t, v_t, R_t)$ and $\hat{x}_t = (\hat{p}_t, \hat{v}_t, \hat{R}_t)$ be the ground-truth and reconstructed motion at frame t. The total loss combines an L1 parameter loss on translation, velocity, and rotations ($\mathcal{L}_{\text{param}}$) and a forward-kinematics joint loss ($\mathcal{L}_{\text{fk}}$) computed on root-centered 3D joint positions via FK:

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{param}} + \lambda_{\text{fk}} \mathcal{L}_{\text{fk}}. \tag{4}$$

3.3 Unified Autoregressive Transformer

The decoder-only AR formulation offers three structural advantages for unified motion modeling: (1) all modalities share one vocabulary and one training objective (next-token prediction), eliminating per-task architectural engineering; (2) any combination of modalities can serve as a prefix, enabling flexible task composition at inference without task-specific architectural changes; and (3) the same model naturally supports both generation (condition→motion) and understanding (motion→text) by reordering the token sequence.

Multimodal tokenization. All modalities are mapped to a shared vocabulary Z: text via a subword tokenizer; audio (speech/music/ambient) via a VQ-VAE into tokens `<AUD_x>`; motion via VerMoVQ into tokens `<MOT_y>`. Sequences use modality-specific `[BOS]/[EOS]` markers; multi-agent motion wraps each agent with `[MOT_BOS]/[MOT_EOS]` and separates agents by `[AGENT_SEP]`. Because motion tokens retain world-space root trajectories, inter-agent spatial relations are implicit, so arbitrary group sizes are supported without architectural changes.

Motion message formulation. Each task is encoded as one token stream with three parts: **Instruction** (natural language), **Condition** (any subset of modalities and optional metadata, each enclosed by BOS/EOS markers), and **Reply** (target tokens with BOS/EOS). During training, optional condition fields are stochastically included to support composite tasks; at inference, the model autoregressively generates the reply given the preceding segments. Representative formats appear in Tab 14.

We illustrate a single motion message that simultaneously conditions on multi-agent motion, audio, and text as follows:

Instruction $\mathcal{I}$ `[INSTR_BOS]` Generate motion w.r.t. the given music and caption. `[INSTR_EOS]`
Condition $\mathcal{C}$ `[COND_BOS]` `[TEXT_BOS]` Two dancers perform complementary sequences: one spins clockwise, the other leaps upward. `[TEXT_EOS]` `[AUDIO_BOS]` <AUD_23> <AUD_98> <AUD_188> ... <AUD_47> `[AUDIO_EOS]` [NUM_PERSON] [DURATION] `[COND_EOS]`
Reply $\mathcal{R}$ `[REPLY_BOS]` `[MOT_BOS]` <MOT_12> <MOT_100> ... <MOT_240> [AGENT_SEP] <MOT_26> <MOT_38> ... <MOT_35> `[MOT_EOS]` `[REPLY_EOS]`

Generalist pretraining and fine-tuning. Training proceeds in two stages: (i) **Generalist pretraining**—sample diverse subtasks from MotionHub and optimize next-token likelihood on full instruction–condition–reply sequences to learn a unified prior across all modalities and tasks; (ii) **Generalist SFT**—fine-tune on benchmark (condition, instruction)→reply pairs with reply-only supervision to adapt to task constraints while keeping cross-modal competence.

Dual-pathway task sampling. The tasks in MotionHub are highly imbalanced: single-person text-to-motion dominates, whereas audio-conditioned tasks are scarce (e.g., music-conditioned samples are under 1% of the eligible data). Drawing each training example uniformly from the corpus—a *data-proportional* pathway, in which a task's exposure scales with its supporting data—thus lets such rare tasks be sampled almost never, so the model never learns to attend to the audio condition. We therefore sample along two pathways: (a) the data-proportional pathway above, which favors abundant skills and stabilizes the shared prior; and (b) a *task-uniform* pathway that first selects a task uniformly from the eligible task pool and then draws examples for that task, giving every task equal expected exposure regardless of its data volume. We use pathway (a) during generalist pretraining to build broad cross-modal priors, and pathway (b) during generalist SFT to lift the under-exposed audio tasks while rehearsing well-learned skills so they are not forgotten.

4 Experiments

4.1 Experimental Setup

Data. All training and evaluation use subsets of **MotionHub**, our large-scale motion–language corpus that covers single- and multi-person settings across diverse tasks. We standardize all sources to a unified SMPL representation at 30 fps. Full data composition, pre-processing details, and the motion/evaluator feature extractors are described in Appendix Sec C.

Evaluation protocol. We evaluate tokenizer reconstruction plus nine downstream tasks, grouped as: *(i)* motion tokenizer reconstruction (Sec 4.2), *(ii)* language-conditioned tasks—text-to-motion (Sec 4.3), multi-person text-to-motion (Sec 4.4), motion-to-text, motion prediction and in-betweening, long-term sequential generation, and *(iii)* audio-conditioned tasks—music-to-dance / dance-to-music and speech-to-gesture. Ablation studies on the training strategy and VerMoVQ architecture are presented in Sec 4.6.

Baselines. We compare against strong recent methods spanning three paradigms: (a) *diffusion-based*: MDM [71], MotionDiffuse [87], MLD [10], ReMoDiffuse [88], HY-Motion [46]; (b) *autoregressive & masked*: T2M-GPT [86], MotionGPT [31], MoMask [22], Go-To-Zero [14], MotionStreamer [76]; and (c) *unified multi-task models*: M^{3}GPT [56], LMM [89], MotionLLM [9], LoM [8]. For multi-person interaction we additionally compare with ComMDM [66], InterGen [40], InterMask [28], and InterControl [74].

Table 2: Motion tokenizer reconstruction quality. All tokenizers are evaluated on the MotionHub test set. MPJPE / PA-MPJPE are in **mm**, MPJRE in **degrees**, and "CB Util." denotes codebook utilization (%). **Bold** / underline mark the best / second best within each #P group.

#P	Method	Repr.	Quant.	|CB|	MPJPE↓	PA-MPJPE↓	MPJRE↓	CB Util.↑
1P	*Prior tokenizers*							
	MLD [10]	HML3D	VAE	–	72.52	48.13	15.74	–
	MotionStreamer [76]	SMPL	TAE	–	36.42	27.21	7.76	–
	TM2T [24]	HML3D	VQ	1024	86.28	57.63	19.41	52.0
	MotionGPT [31]	HML3D	VQ	512	71.47	51.49	18.93	99.9
	MoMask [22]	HML3D	RVQ	512×6	47.90	30.02	15.84	**100.0**
	Go-To-Zero [14]	SMPL	FSQ	64k	27.19	21.19	6.35	83.7
	LoM [8]	SMPL	VQ	512×4	48.97	37.11	7.97	**100.0**
	VerMoVQ (Ours)							
	1D	SMPL	RVQ	1k×6	**12.34**	9.49	3.43	**100.0**
	2D	SMPL	FSQ	1k	24.42	14.77	2.71	**100.0**
	2D	SMPL	FSQ	4k	19.81	11.67	2.15	99.2
	2D	SMPL	FSQ	16k	18.13	10.51	1.93	97.3
	2D	SMPL	FSQ	64k	13.54	**8.38**	**1.50**	59.2
2P	*Prior tokenizers*							
	Go-To-Zero [14]	SMPL	FSQ	64k	38.80	28.89	7.60	60.9
	MotionStreamer [76]	SMPL	TAE	–	78.04	57.53	16.03	–
	InterMask [28]	InterX	VQ	1024	67.10	51.14	10.61	**99.5**
	VerMoVQ (Ours)							
	2D	SMPL	FSQ	16k	**18.52**	**10.73**	**2.07**	97.8

4.2 Motion Tokenizer Reconstruction

Reconstruction fidelity sets the quality ceiling of any discrete AR system, since information lost at tokenization cannot be recovered downstream. We evaluate on both single- and two-person motion, reporting MPJPE and PA-MPJPE (mm), MPJRE (degrees), and codebook utilization. As shown in Tab 2, the 2D spatio-temporal factorization of VerMoVQ yields substantially lower joint errors than all baselines: in the 1P setting, VerMoVQ (2D, FSQ 16k) achieves 18.13 mm MPJPE, outperforming the best prior VQ-based tokenizer Go-To-Zero (27.19 mm) by ∼33%, while the continuous-output MotionStreamer (36.42 mm) also trails significantly. The improvement stems from low-dimensional body tokens that reduce per-codebook quantization pressure compared to single high-dimensional frame vectors. While our 1D RVQ variant attains an even lower reconstruction error (12.34 mm), we adopt the 2D tokenizer for downstream AR modeling because it delivers substantially better *generation* quality: as shown in Tab 12, the 2D design consistently outperforms 1D on gFID across all codebook sizes. In the 2P setting, the same tokenizer applied per-person maintains strong quality (18.52 mm), demonstrating that world-space root preservation enables simple per-agent tokenization with a shared codebook.

4.3 Text-to-Motion (Single-Person)

Table 3: Text-to-motion (1P) on HumanML3D and MotionHub. All metrics are computed using our trained TMR features. "∗" denotes models retrained on the corresponding dataset. **Bold**: best; underline: second best.

Method	HumanML3D [23]					MotionHub				
	R-P T1↑	R-P T3↑	FID↓	MM-D↓	Div.→	R-P T1↑	R-P T3↑	FID↓	MM-D↓	Div.→
Real	0.776	0.902	0.000	0.906	21.63	0.664	0.845	0.000	0.988	22.98
Diffusion / flow-based										
MDM [71]	0.223	0.415	0.362	1.173	21.46	0.115	0.244	0.445	1.211	21.19
MLD [10]	0.232	0.418	0.373	1.170	21.30	0.127	0.256	0.400	1.209	21.32
ViMoGen [42]	0.155	0.297	0.400	1.205	19.78	0.158	0.304	0.533	1.213	20.05
MotionStreamer [76]	0.460	0.715	0.065	**1.024**	21.89	0.190	0.367	0.411	1.176	21.38
HY-Motion [46]	**0.561**	**0.797**	0.185	1.063	22.36	**0.417**	**0.621**	0.363	1.148	22.87
Discrete / autoregressive / masked										
TM2T [24]	0.130	0.288	0.491	1.213	19.35	0.070	0.167	0.658	1.253	18.79
TM2D [20]	0.075	0.168	0.446	1.248	19.88	0.041	0.115	0.545	1.275	20.10
T2M-GPT* [86]	0.369	0.589	0.338	1.120	21.43	0.142	0.284	0.463	1.217	21.06
MoMask* [22]	0.494	0.749	0.346	1.194	23.23	0.169	0.347	0.434	1.198	21.68
MotionGPT [31]	0.313	0.502	0.324	1.155	21.29	0.129	0.243	0.434	1.237	21.10
MotionGPT3 [95]	0.414	0.639	0.372	1.113	20.82	0.206	0.364	0.441	1.181	21.08
LoM [8]	0.170	0.334	0.303	1.148	18.72	0.106	0.221	0.623	1.210	18.85
Go-To-Zero [14]	0.443	0.653	0.075	1.052	**21.64**	0.291	0.464	0.101	1.130	**22.96**
VersatileMotion	0.457	0.692	**0.062**	1.039	21.74	0.408	0.612	**0.098**	**1.122**	22.82

Tab 3 reports R-Precision, FID, MM Dist and Diversity on HumanML3D and MotionHub (TMR features). On HumanML3D, where all baselines are trained, VersatileMotion achieves the lowest FID (0.062), ahead of MotionStreamer (0.065), with competitive R-Precision. On MotionHub, where most baselines were not trained, the gap widens: VersatileMotion attains the best FID (0.098) and MM-Dist, trailing HY-Motion only on R-Precision. This confirms the VQ+AR paradigm can rival the best diffusion models given a well-designed tokenizer and large-scale data.

Table 4: Multi-person text-to-motion (2P) on MotionHub. "∗" denotes retrained models. **Bold**: best; underline: second best.

Method	R-P↑	FID↓	MM-D↓	Div→
Real	0.851	–	0.946	25.25
ComMDM* [66]	0.141	1.01	1.357	17.80
InterGen [40]	0.367	0.77	1.265	22.51
InterCtrl [74]	0.413	0.66	1.211	22.16
InterMask [28]	**0.508**	**0.43**	**1.147**	**23.44**
VersatileMotion	0.473	0.51	1.165	23.11

4.4 Text-to-Motion (Multi-Person)

We evaluate multi-person text-to-motion (2P) in Tab 4 on the MotionHub multi-person test set, using a multi-person TMR encoder for feature extraction. Because most baselines discard absolute root position/orientation, their two-person results are computed in a first-frame root-relative frame and post-processed by re-injecting the ground-truth root; VersatileMotion is evaluated directly in world

space. VersatileMotion achieves the second-best results across all metrics, narrowly behind InterMask, while substantially outperforming the diffusion-based ComMDM. Notably, this is accomplished by simple per-agent token concatenation in the shared codebook, without any interaction-specific architecture.

4.5 Additional Generation and Understanding Tasks

Beyond text-to-motion, VersatileMotion is evaluated on seven additional tasks spanning language-conditioned and audio-conditioned domains, with results below and in the supplementary material.

Motion-to-text (Tab 5). VersatileMotion attains the best retrieval accuracy on HumanML3D (R-Prec = 0.686, MM-Dist = 1.044 vs. 0.536 / 1.062 for TM2T), while TM2T stays competitive only on surface-fluency scores (BERT/METEOR/Sim). The gap widens sharply on MotionHub, where VersatileMotion leads every metric for both single-person (R-Prec = 0.509 vs. 0.186) and two-person (R-Prec = 0.778 vs. 0.691) captioning.

Table 5: Motion-to-text on HumanML3D and MotionHub. R-Prec (Top-3, 64 cands.) and MM Dist use TMR features; BERT uses RoBERTa-large; Sim uses Sentence-BERT. "∗" = retrained on MotionHub. MotionGPT3 is omitted: its released weights yield a degenerate caption decoder (motion understanding works, but text decoding is not reproducible). Best in **bold**, second underlined.

Method	R-P↑	MM-D↓	BERT↑	MET↑	Sim↑
HumanML3D (1P)					
Real	0.830	0.945	1.003	1.003	1.003
MotionGPT [31]	0.398	1.123	0.895	0.367	0.598
TM2T [24]	0.536	1.062	**0.909**	**0.403**	**0.670**
VersatileMotion	**0.686**	**1.044**	**0.909**	0.370	0.642
MotionHub (1P)					
Real	0.930	0.835	1.006	1.008	1.009
MotionGPT [31]	0.153	1.260	0.884	0.192	0.484
TM2T [24]	0.186	1.257	0.884	0.190	0.539
VersatileMotion	**0.509**	**1.115**	**0.901**	**0.422**	**0.694**
MotionHub (2P)					
Real	0.922	0.904	1.005	1.004	1.005
TM2T* [24]	0.623	1.089	0.882	0.378	0.615
MotionGPT* [31]	0.691	1.045	0.907	0.420	0.662
VersatileMotion	**0.778**	**0.994**	**0.929**	**0.466**	**0.728**

Motion prediction and in-betweening (Tab 6). Across all three conditioning regimes and both tasks, VersatileMotion is best on R-Precision, FID, ADE, and FDE. In the hardest single-frame regime it reaches FID = 0.166 (in-betweening) against 0.276 for the strongest baseline MotionGPT3, whereas MotionGPT cannot operate from a single frame at all; the lead persists with no text (40%/20% context), where VersatileMotion attains the lowest FID by a wide margin (0.147 vs. 0.240), handling arbitrary masking without task-specific architecture or retraining.

Long-term sequential generation (Tab 9). On BABEL [65], VersatileMotion achieves the best subsequence quality (R@3 = 0.489, FID = 0.106), slightly surpassing MotionStreamer (0.469, 0.121) while matching transition smoothness; the causal tokenizer naturally supports variable-length chaining without explicit transition modeling.

Table 6: Motion prediction (Pred) and in-betweening (Btw) on MotionHub. Three conditioning regimes: (a) 1 frame + text; (b) 10%/5% frames + text; (c) 40%/20% frames, no text.

		(a) 1 frame + text				*(b) 10%/5% + text*				*(c) 40%/20%, no text*			
Method	Task	R-P↑	FID↓	ADE↓	FDE↓	R-P↑	FID↓	ADE↓	FDE↓	FID↓	Div→	ADE↓	FDE↓
Real	–	0.845	0.000	0.000	0.000	0.846	0.000	0.000	0.000	0.000	22.91	0.000	0.000
MDM [71]	Pred	0.351	0.360	0.951	1.819	0.349	0.379	0.829	1.671	0.307	21.74	0.424	1.203
	Btw	0.399	0.356	0.950	1.792	0.412	0.345	0.887	1.675	0.277	21.90	0.662	1.334
MotionGPT [31]	Pred	–	–	–	–	0.047	0.329	1.238	2.145	0.321	21.00	0.519	1.169
	Btw	–	–	–	–	0.045	0.327	1.184	1.934	0.359	20.84	1.067	1.811
MotionGPT3 [95]	Pred	0.285	0.293	1.041	1.929	0.311	0.272	0.865	1.680	0.267	21.36	0.398	1.088
	Btw	0.312	0.276	1.021	1.884	0.341	0.268	0.843	1.650	0.240	21.55	0.570	1.246
VersatileMotion	Pred	**0.423**	**0.184**	**0.816**	**1.542**	**0.439**	**0.175**	**0.681**	**1.320**	**0.157**	21.69	**0.303**	**0.876**
	Btw	**0.453**	**0.166**	**0.794**	**1.482**	**0.475**	**0.159**	**0.661**	**1.290**	**0.147**	21.87	**0.423**	**1.016**

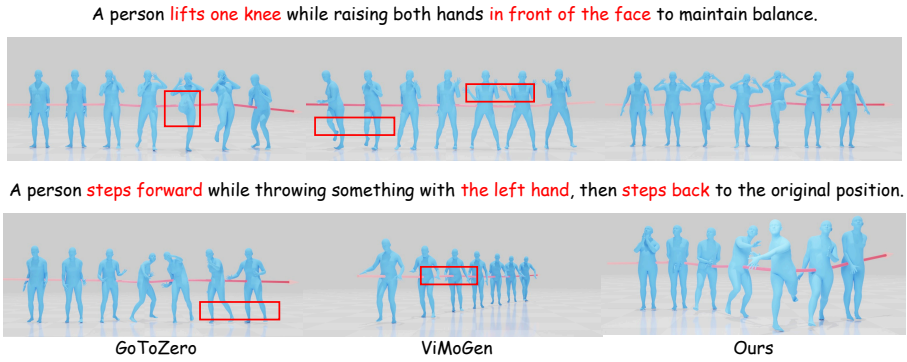

Fig. 3: Qualitative comparison on single-person text-to-motion. VersatileMotion produces coherent full-body motion with correct semantics, while Go-To-Zero [14] exhibits unnatural root trajectories and ViMoGen [42] shows temporal inconsistencies.

Music-to-dance and dance-to-music (Tab 10 and Tab 11). For music→dance, VersatileMotion attains the best FID and kinetic FID on both AIST++ (FID = 0.152, $\text{FID}_k = 24.85$) and FineDance (0.185, 32.48), surpassing dedicated dance generators; adding text guidance (music+text→dance) further improves fidelity (FID = 0.138 on AIST++). For dance→music (Tab 11), VersatileMotion attains the strongest beat alignment (BCS/BHS/F1) with the lowest deviation while remaining competitive on FAD, outperforming dedicated dance-to-music systems.

Speech-to-gesture (Tab 15). We compare against CaMN [49], DiffStyGest [79], EMAGE [48], TalkSHOW [80], MCM [45], and LoM [8]. Full results are in the supplementary material.

Table 7: Multi-task co-training ablation (VersatileMotion-1B). Jointly training all tasks vs. single-task specialists, across text-to-motion (T2M), motion-to-text (M2T), music+text→dance (M+T→D), and text+speech→gesture (T+S→G).

Setting	T2M			M2T		M+T→D		T+S→G	
	FID↓	R-T3↑	MMD↓	R-T3↑	MMD↓	FID_k↓	BAS↑	FID↓	BA→
Single-task specialist	0.075	0.680	1.041	0.666	1.073	32.5	0.215	0.399	0.183
All tasks mixed (Ours)	**0.062**	**0.692**	**1.039**	**0.686**	**1.044**	**23.52**	**0.238**	**0.346**	**0.219**

Table 8: Training-strategy ablations (VersatileMotion-1B). (i) generation↔understanding; (ii) LLM backbone; (iii) pseudo multi-person pretraining.

Setting	T2M			M2T	
	FID↓	R-T3↑	MMD↓	R-T3↑	MMD↓
(i) Generation ↔ understanding					
Generation only (T2M)	0.074	0.683	1.046	–	–
Understanding only (M2T)	–	–	–	0.658	1.067
Joint gen. + understand.	**0.062**	**0.692**	**1.039**	**0.686**	**1.044**
(ii) LLM backbone					
Random Init – 1B	0.147	0.549	1.122	0.484	1.160
Qwen3 – 0.6B	0.093	0.655	1.066	0.625	1.045
LLaMA 3.2 – 1B	0.062	**0.692**	**1.039**	**0.686**	**1.044**
Qwen3 – 1.7B	**0.060**	0.690	1.045	0.673	1.053
(iii) Pseudo multi-person pretrain[†]					
w/o pseudo multi-person	0.60	0.427	1.218	0.544	1.098
w/ pseudo multi-person	**0.51**	**0.473**	**1.165**	**0.778**	**0.994**

[†]Evaluated on two-person (2P) tasks.

4.6 Ablation Study

Multi-task co-training (Tab 7). Joint multi-task training improves both music+text→dance and text+speech→gesture quality, benefiting from richer cross-domain motion priors and stronger semantic alignment from large-scale text-motion pairs.

Generation ↔ understanding (Tab 8 (i)). Joint training benefits both T2M and M2T directions, indicating that generation and comprehension provide complementary learning signals.

LLM backbone (Tab 8 (ii)). Pre-trained initialization clearly beats random; among Qwen3-0.6B, LLaMA 3.2-1B, and Qwen3-1.7B, we use LLaMA 3.2-1B by default, as it gives the best R-Precision, MM-Dist, and M2T quality at lower cost than the larger Qwen3-1.7B, which only marginally lowers FID.

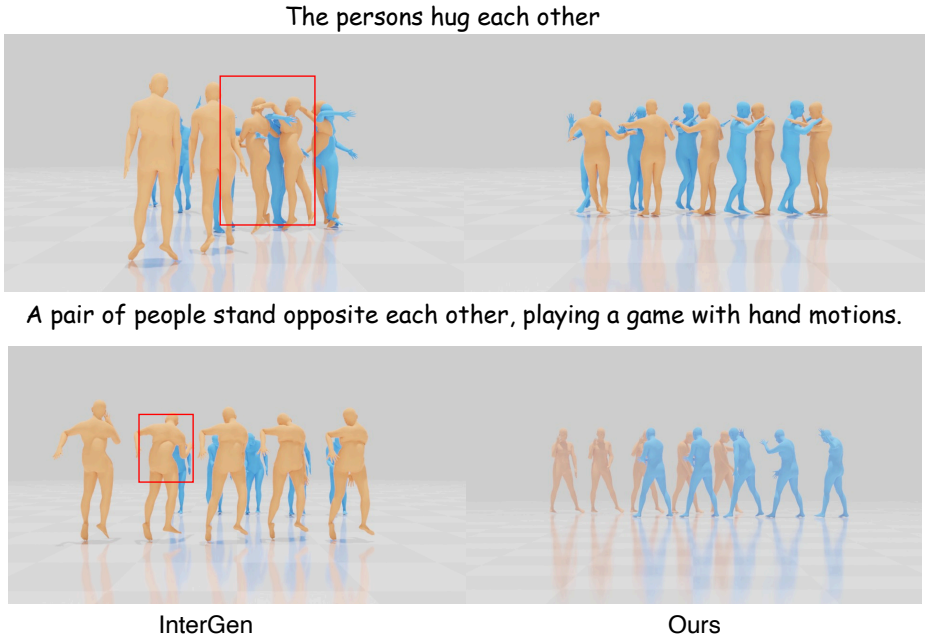

Fig. 4: Qualitative comparison on multi-person text-to-motion between VersatileMotion and InterGen [40]. VersatileMotion generates plausible two-person interactions via per-agent token concatenation, while InterGen struggles with spatial coordination.

Pseudo multi-person pretraining (Tab 8 (iii)). Composing single-person clips into pseudo two-person data during pretraining markedly improves two-person T2M and M2T.

Tokenizer design (Tab 12). In the supplementary, VerMoVQ's 2D design consistently beats 1D, and FK loss with body-attention further improve joint accuracy.

Qualitative results. Fig 3 and Fig 4 show coherent single-person motion and plausible two-person interactions from VersatileMotion, while baselines exhibit unnatural roots, jitter, or weak coordination.

5 Conclusion

We presented VersatileMotion, a unified autoregressive framework that casts diverse human-motion tasks as next-token prediction over a shared discrete vocabulary, spanning text-, music-, and speech-conditioned generation and comprehension. Powered by VerMoVQ, a world-space motion tokenizer, and the ∼380K-clip MotionHub corpus, a single model rivals or surpasses task-specific baselines across nine tasks, showing that earlier discrete methods were limited by their tokenizers and data, not the autoregressive paradigm.

Acknowledgments

This research was also supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F020007.

References

1. Alexanderson, S., Nagy, R., Beskow, J., Henter, G.E.: Listen, denoise, action! audio-driven motion synthesis with diffusion models. ACM Transactions on Graphics **42**(4), 1–20 (2023)
2. Aristidou, A., Shamir, A., Chrysanthou, Y.: Digital dance ethnography: Organizing large dance collections. J. Comput. Cult. Herit. **12**(4) (Nov 2019). https://doi.org/10.1145/3344383, https://doi.org/10.1145/3344383
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Barquero, G., Escalera, S., Palmero, C.: Seamless human motion composition with blended positional encodings. In: CVPR (2024)
5. Brown, T., Mann, B., Ryder, N., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems **33**, 1877–1901 (2020)
6. Cai, Z., Yin, W., Zeng, A., Wei, C., Sun, Q., Yanjun, W., Pang, H.E., Mei, H., Zhang, M., Zhang, L., et al.: Smpler-x: Scaling up expressive human pose and shape estimation. Advances in Neural Information Processing Systems **36**, 11454–11468 (2023)
7. Chatzitofis, A., Saroglou, L., Boutis, P., Drakoulis, P., Zioulis, N., Subramanyam, S., Kevelham, B., Charbonnier, C., Cesar, P., Zarpalas, D., et al.: Human4d: A human-centric multimodal dataset for motions and immersive media. IEEE Access **8**, 176241–176262 (2020)
8. Chen, C., Zhang, J., Lakshmikanth, S.K., et al.: The language of motion: Unifying verbal and non-verbal language of 3d human motion. In: CVPR (2025)
9. Chen, L.H., Lu, S., Zeng, A., et al.: Motionllm: Understanding human behaviors from human motions and videos. arXiv preprint arXiv:2405.20340 (2024)
10. Chen, X., Jiang, B., Liu, W., et al.: Executing your commands via motion diffusion in latent space. In: CVPR (2023)
11. Cho, J., Kim, J., Kim, J., Kim, M., Kang, M., Hong, S., Oh, T.H., Yu, Y.: Discord: Discrete tokens to continuous motion via rectified flow decoding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14602–14612 (October 2025)
12. Contributors, M.: MMEngine: Openmmlab foundational library for training deep learning models. https://github.com/open-mmlab/mmengine (2022)
13. Di, S., Jiang, Z., Liu, S., et al.: Video background music generation with controllable music transformer. In: ACM Multimedia (2021)
14. Fan, K., Lu, S., Dai, M., Yu, R., Xiao, L., Dou, Z., Dong, J., Ma, L., Wang, J.: Go to zero: Towards zero-shot motion generation with million-scale data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13336–13348 (2025)
15. Fieraru, M., Zanfir, M., Oneata, E., Popa, A.I., Olaru, V., Sminchisescu, C.: Learning complex 3d human self-contact. In: AAAI. vol. 35, pp. 1343–1351 (2021)

16. Fieraru, M., Zanfir, M., Oneata, E., Popa, A.I., Olaru, V., Sminchisescu, C.: Reconstructing three-dimensional models of interacting humans. arXiv preprint arXiv:2308.01854 (2023)
17. Fieraru, M., Zanfir, M., Pirlea, S.C., Olaru, V., Sminchisescu, C.: Aifit: Automatic 3d human-interpretable feedback models for fitness training. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2021)
18. Gan, C., Huang, D., Chen, P., et al.: Foley music: Learning to generate music from videos. In: ECCV (2020)
19. Ghorbani, N., Black, M.J.: SOMA: Solving optical marker-based mocap automatically. In: Proc. International Conference on Computer Vision (ICCV). pp. 11117–11126 (Oct 2021)
20. Gong, K., Lian, D., Chang, H., Guo, C., Jiang, Z., Zuo, X., Mi, M.B., Wang, X.: Tm2d: Bimodality driven 3d dance generation via music-text integration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9942–9952 (2023)
21. Gopinath, D., Won, J.: Fairmotion-tools to load, process and visualize motion capture data. `https://github.com/facebookresearch/fairmotion` (2020)
22. Guo, C., Mu, Y., Javed, M.G., et al.: Momask: Generative masked modeling of 3d human motions. In: CVPR (2024)
23. Guo, C., Zou, S., Zuo, X., et al.: Generating diverse and natural 3d human motions from text. In: CVPR (2022)
24. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: ECCV (2022)
25. Han, B., Li, Y., Shen, Y., et al.: Dance2midi: Dance-driven multi-instrument music generation. Computational Visual Media **10**(4), 791–802 (2024)
26. Han, B., Peng, H., Dong, M., et al.: Amd: Autoregressive motion diffusion. In: AAAI (2024)
27. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence **36**(7), 1325–1339 (2013)
28. Javed, M.G., Guo, C., Cheng, L., Li, X.: Intermask: 3d human interaction generation via collaborative masked modeling. arXiv preprint arXiv:2410.10010 (2024)
29. Ji, S., Jiang, Z., Cheng, X., et al.: Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. arXiv preprint arXiv:2408.16532 (2024)
30. Ji, Y., Xu, F., Yang, Y., Shen, F., Shen, H.T., Zheng, W.S.: A large-scale varying-view rgb-d action dataset for arbitrary-view human action recognition. arXiv preprint arXiv:1904.10681 (2019)
31. Jiang, B., Chen, X., Liu, W., et al.: Motiongpt: Human motion as a foreign language. In: NeurIPS (2023)
32. Jiang, N., Zhang, Z., Li, H., Ma, X., Wang, Z., Chen, Y., Liu, T., Zhu, Y., Huang, S.: Scaling up dynamic human-scene interaction modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1737–1747 (2024)
33. Kim, B., Jeong, H.I., Sung, J., Cheng, Y., Lee, J., Chang, J.Y., Choi, S.I., Choi, Y., Shin, S., Kim, J., et al.: Personabooth: Personalized text-to-motion generation. arXiv preprint arXiv:2503.07390 (2025)
34. Krebs, F., Meixner, A., Patzer, I., Asfour, T.: The KIT bimanual manipulation dataset. In: IEEE/RAS International Conference on Humanoid Robots (Humanoids). pp. 499–506 (2021)

35. Li, B., Zhao, Y., Zhelun, S., Sheng, L.: Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In: AAAI (2022)
36. Li, R., Zhang, Y., Zhang, Y., Zhang, H., Guo, J., Zhang, Y., Liu, Y., Li, X.: Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1524–1534 (2024)
37. Li, R., Zhao, J., Zhang, Y., et al.: Finedance: A fine-grained choreography dataset for 3d full body dance generation. In: ICCV (2023)
38. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: ICCV (2021)
39. Li, S., Dong, W., Zhang, Y., Tang, F., Ma, C., Deussen, O., Lee, T.Y., Xu, C.: Dance-to-music generation with encoder-based textual inversion. In: SIGGRAPH Asia (2024)
40. Liang, H., Zhang, W., Li, W., et al.: Intergen: Diffusion-based multi-human motion generation under complex interactions. IJCV **1**, 1–21 (2024)
41. Liao, Y., Fu, Y., Cheng, Z., Wang, J.: Animationgpt: An aigc tool for generating game combat motion assets. `https://github.com/fyyakaxyy/AnimationGPT` (2024), combatMotion dataset and code
42. Lin, J., Wang, R., Lu, J., Huang, Z., Song, G., Zeng, A., Liu, X., Wei, C., Yin, W., Sun, Q., Cai, Z., Yang, L., Liu, Z.: The quest for generalizable motion generation: Data, model, and evaluation. arXiv preprint arXiv:2510.26794 (2025)
43. Lin, J., Zeng, A., Lu, S., et al.: Motion-x: A large-scale 3d expressive whole-body human motion dataset. Advances in Neural Information Processing Systems **36**, 25268–25280 (2023)
44. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. pp. 740–755. Springer (2014)
45. Ling, Z., Han, B., Wong, Y., et al.: Mcm: Multi-condition motion synthesis framework. In: IJCAI (2024)
46. Ling, Z., et al.: Hy-motion: A hybrid diffusion framework for versatile motion generation. Techinal Report (2025)
47. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
48. Liu, H., Zhu, Z., Becherini, G., et al.: Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In: CVPR (2024)
49. Liu, H., Zhu, Z., Iwamoto, N., et al.: Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In: ECCV (2022)
50. Liu, H., Zhu, Z., Iwamoto, N., Peng, Y., Li, Z., Zhou, Y., Bozkurt, E., Zheng, B.: Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In: European conference on computer vision. pp. 612–630. Springer (2022)
51. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2684–2701 (2019)
52. Liu, X., Wu, Q., Zhou, H., Xu, Y., Qian, R., Lin, X., Zhou, X., Wu, W., Dai, B., Zhou, B.: Learning hierarchical cross-modal association for co-speech gesture generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10462–10472 (2022)

53. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM Transactions on Graphics **34**(6) (2015)
54. Lu, S., Chen, L.H., Zeng, A., et al.: Humantomato: Text-aligned whole-body motion generation. In: ICML (2024)
55. Lu, S., Wang, J., Lu, Z., et al.: Scamo: Exploring the scaling law in autoregressive motion generation model. arXiv preprint arXiv:2412.14559 (2024)
56. Luo, M., Hou, R., Chang, H., et al.: An advanced multimodal, multitask framework for motion comprehension and generation. arXiv preprint arXiv:2405.16273 (2024)
57. Mandery, C., Terlemez, O., Do, M., Vahrenkamp, N., Asfour, T.: The KIT whole-body human motion database. In: International Conference on Advanced Robotics (ICAR). pp. 329–336 (2015)
58. Mandery, C., Terlemez, O., Do, M., Vahrenkamp, N., Asfour, T.: Unifying representations and large-scale whole-body motion databases for studying human motion. IEEE Transactions on Robotics **32**(4), 796–809 (2016)
59. McFee, B., Raffel, C., Liang, D., et al.: librosa: Audio and music signal analysis in python. In: SciPy (2015)
60. Mentzer, F., Minnen, D., Agustsson, E., Tschannen, M.: Finite scalar quantization: Vq-vae made simple. In: ICLR (2023)
61. Paszke, A.: Pytorch: An imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703 (2019)
62. 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)
63. Petrovich, M., Black, M.J., Varol, G.: Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In: ICCV (2023)
64. Pinyoanuntapong, E., Wang, P., Lee, M., Chen, C.: Mmm: Generative masked motion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1546–1555 (June 2024)
65. Punnakkal, A.R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., Black, M.J.: BABEL: Bodies, action and behavior with english labels. In: CVPR (2021)
66. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. arXiv preprint arXiv:2303.01418 (2023)
67. Shen, Z., Pi, H., Xia, Y., Cen, Z., Peng, S., Hu, Z., Bao, H., Hu, R., Zhou, X.: World-grounded human motion recovery via gravity-view coordinates. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
68. Siyao, L., Yu, W., Gu, T., et al.: Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In: CVPR (2022)
69. Sun, Q., Wang, Y., Zeng, A., Yin, W., Wei, C., Wang, W., Mei, H., Leung, C.S., Liu, Z., Yang, L., et al.: Aios: All-in-one-stage expressive human pose and shape estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1834–1843 (2024)
70. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
71. Tevet, G., Raab, S., Gordon, B., et al.: Human motion diffusion model. In: ICLR (2023)
72. Tseng, J., Castellon, R., Liu, K.: EDGE: Editable dance generation from music. In: CVPR (2023)
73. Wang, Y., Huang, D., Zhang, Y., et al.: Motiongpt-2: A general-purpose motion-language model for motion generation and understanding. arXiv preprint arXiv:2410.21747 (2024)

74. Wang, Z., Wang, J., Li, Y., Lin, D., Dai, B.: Intercontrol: Zero-shot human interaction generation by controlling every joint. In: NeurIPS (2024)
75. Wu, Q., Zhao, Y., Wang, Y., Liu, X., Tai, Y.W., Tang, C.K.: Motion-agent: A conversational framework for human motion generation with llms. arXiv preprint arXiv:2405.17013 (2024)
76. Xiao, L., Lu, S., Pi, H., Fan, K., Pan, L., Zhou, Y., Feng, Z., Zhou, X., Peng, S., Wang, J.: Motionstreamer: Streaming motion generation via diffusion-based autoregressive model in causal latent space. arXiv preprint arXiv:2503.15451 (2025)
77. Xu, L., Lv, X., Yan, Y., et al.: Inter-x: Towards versatile human-human interaction analysis. In: CVPR (2024)
78. Yang, H., Su, K., Zhang, Y., et al.: Unimumo: Unified text, music and motion generation. arXiv preprint arXiv:2410.04534 (2024)
79. Yang, S., Wu, Z., Li, M., Zhang, Z., Hao, L., Bao, W., Cheng, M., Xiao, L.: Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models. arXiv preprint arXiv:2305.04919 (2023)
80. Yi, H., Liang, H., Liu, Y., et al.: Generating holistic 3d human motion from speech. In: CVPR (2023)
81. Yin, W., Cai, Z., Wang, R., Zeng, A., Wei, C., Sun, Q., Mei, H., Wang, Y., Pang, H.E., Zhang, M., et al.: Smplest-x: Ultimate scaling for expressive human pose and shape estimation. arXiv preprint arXiv:2501.09782 (2025)
82. Yin, Y., Guo, C., Kaufmann, M., Zarate, J.J., Song, J., Hilliges, O.: Hi4d: 4d instance segmentation of close human interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17016–17027 (2023)
83. Yoon, Y., Cha, B., Lee, J.H., Jang, M., Lee, J., Kim, J., Lee, G.: Speech gesture generation from the trimodal context of text, audio, and speaker identity. ACM Transactions on Graphics (TOG) **39**(6), 1–16 (2020)
84. Yu, J., Wang, Y., Chen, X., et al.: Long-term rhythmic video soundtracker. In: ICML (2023)
85. Zeng, A., Yang, L., Ju, X., Li, J., Wang, J., Xu, Q.: Smoothnet: A plug-and-play network for refining human poses in videos. In: European Conference on Computer Vision. pp. 625–642. Springer (2022)
86. Zhang, J., Zhang, Y., Cun, X., et al.: Generating human motion from textual descriptions with discrete representations. In: CVPR (2023)
87. Zhang, M., Cai, Z., Pan, L., et al.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(6), 4115–4128 (2024)
88. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. In: ICCV (2023)
89. Zhang, M., Jin, D., Gu, C., et al.: Large motion model for unified multi-modal motion generation. arXiv preprint arXiv:2404.01284 (2024)
90. Zhang, M., Li, H., Cai, Z., et al.: Finemogen: Fine-grained spatio-temporal motion generation and editing. In: NeurIPS (2023)
91. Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: ICLR (2020)
92. Zhang, Y., Lin, J., Zeng, A., et al.: Motion-x++: A large-scale multimodal 3d whole-body human motion dataset. arXiv preprint arXiv:2501.05098 (2025)
93. Zhou, Z., Wan, Y., Wang, B.: A unified framework for multimodal, multi-part human motion synthesis. arXiv preprint arXiv:2311.16471 (2023)
94. Zhou, Z., Wang, B.: Ude: A unified driving engine for human motion generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5632–5641 (2023)

95. Zhu, B., Jiang, B., Wang, S., Tang, S., Chen, T., Luo, L., Zheng, Y., Chen, X.: Motiongpt3: Human motion as a second modality. arXiv preprint arXiv:2506.24086 (2025)
96. Zhu, Y., Olszewski, K., Wu, Y., et al.: Quantized gan for complex music generation from dance videos. In: ECCV (2022)
97. Zhu, Y., Wu, Y., Olszewski, K., et al.: Discrete contrastive diffusion for cross-modal music and image generation. In: ICLR (2023)
98. Zou, S., Zuo, X., Qian, Y., Wang, S., Guo, C., Xu, C., Gong, M., Cheng, L.: Polarization human shape and pose dataset. arXiv preprint arXiv:2004.14899 (2020)