

UniMotion: A Unified Framework for Motion-Text-Vision Understanding and Generation

Ziyi Wang[1], Xinshun Wang[1], Shuang Chen[2], Yang Cong[3], and Mengyuan Liu[1(✉)]

[1] State Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School, Shenzhen, China
liumengyuan@pku.edu.cn
[2] Donghua University, Shanghai, China
[3] School of Automation Science and Engineering, South China University of Technology, Guangzhou, China
https://wangzy01.github.io/UniMotion

Abstract. We present UniMotion, to our knowledge the first unified framework for simultaneous understanding and generation of human motion, natural language, and RGB images within a single architecture. Existing unified models handle only restricted modality subsets (e.g., Motion–Text or static Pose–Image) and predominantly rely on discrete tokenization, which introduces quantization errors and disrupts temporal continuity. UniMotion overcomes both limitations through a core principle: treating motion as a first-class continuous modality on equal footing with RGB. A novel Cross-Modal Aligned Motion VAE (CMA-VAE) and symmetric dual-path embedders construct parallel continuous pathways for Motion and RGB within a shared LLM backbone. To inject visual-semantic priors into motion representations without requiring images at inference, we propose Dual-Posterior KL Alignment (DPA), which distills a vision-fused encoder's richer posterior into the motion-only encoder. To address the cold-start problem—where text supervision alone is too sparse to calibrate the newly introduced motion pathway—we further propose Latent Reconstruction Alignment (LRA), a self-supervised pre-training strategy that uses dense motion latents as unambiguous conditions to co-calibrate the embedder, backbone, and flow head, establishing a stable motion-aware foundation for all downstream tasks. UniMotion achieves state-of-the-art performance across seven tasks spanning any-to-any understanding, generation, and editing among the three modalities, with especially strong advantages on cross-modal compositional tasks.

Keywords: Motion Generation · Unified Framework · MLLMs

Z. Wang, X. Wang and S. Chen—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-032-37092-1_37.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
P. Favaro et al. (Eds.): ECCV 2026, LNCS 17060, pp. 664–682, 2026.
https://doi.org/10.1007/978-3-032-37092-1_37

1 Introduction

Recent advances in unified multimodal large language models (MLLMs) have demonstrated strong and general capabilities in jointly understanding and generating text and images [24,29–32,41], exhibiting strong cross-modal reasoning within a shared semantic space. However, human motion—a critical dynamic modality—has not been systematically integrated into such unified frameworks. Motion sequences encode rich temporal dynamics and spatial structure indispensable for game animation, embodied intelligence, virtual reality, medical rehabilitation, and privacy-preserving action analysis. Constructing a single framework that unifies the Motion-Text-RGB tri-modality while supporting both understanding and generation remains an unresolved key problem.

Existing approaches have not truly addressed this challenge. One line of work, exemplified by MotionGPT [14], unifies Motion and Text through VQ-VAE tokenization [26], but lacks the ability to perceive or generate visual information. Another line, represented by UniPose [18], integrates human *pose* (single-frame body configuration) with vision and language, but is restricted to static estimation and image understanding, with no generation capability. Both rely on discrete tokenization, which inevitably introduces quantization errors [26] that disrupt temporal continuity and structural fidelity; moreover, the asymmetry between discrete tokens and the continuous RGB feature space complicates cross-modal alignment. In summary, prior methods handle only partial modality subsets, and none forms a unified generation-understanding system spanning continuous motion sequences, natural language, and real images.

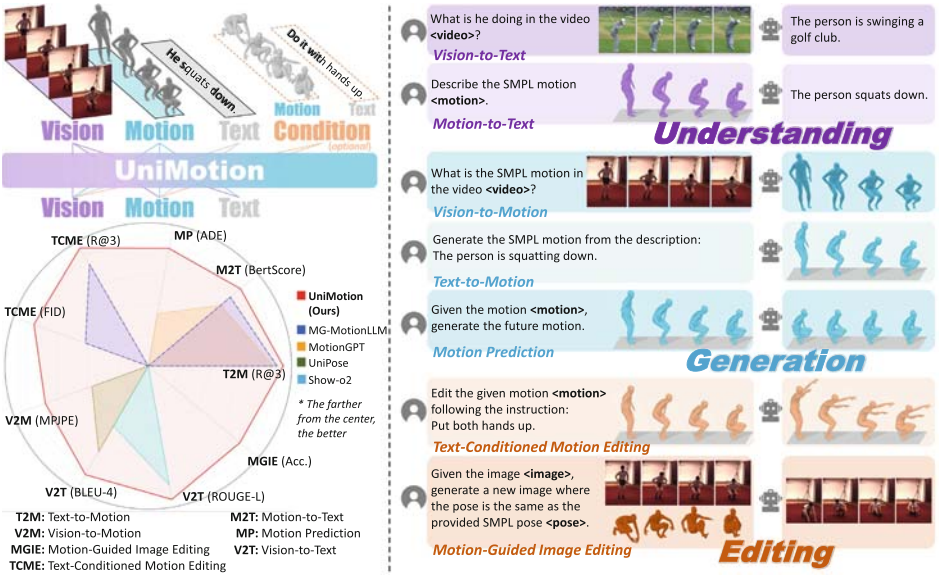


Fig. 1. Left: Overview and performance comparison of UniMotion, a unified framework for any-to-any Motion, Text, and Vision understanding, generation, and editing. UniMotion is the first model to support all seven tri-modal tasks and achieves consistent superiority over existing methods. **Right:** Representative task demonstrations.

To this end, we propose **UniMotion**: the first multimodal framework for unified understanding and generation across the Motion-Text-RGB tri-modality. As summarized in Fig. 1, UniMotion covers understanding, generation, and editing across the three modalities within one model, together with a unified performance comparison against prior partial solutions. The core design philosophy is to treat motion as a continuous modality on equal footing with RGB and construct symmetric continuous pathways for both. Unlike discrete-tokenization methods, we represent motion with a continuous **C**ross-**M**odal **A**ligned Motion VAE (**CMA-VAE**) and, through a dual-path embedder, provide complementary semantic-awareness and fine-grained generation channels that enable natural cross-modal alignment inside the shared LLM backbone. Within the backbone, we further employ lightweight modality-routed LoRA—routing each token to a modality-specific low-rank adapter so that motion and text/RGB modalities can adapt the shared parameters independently—and hybrid attention that reconciles motion's need for bidirectional temporal interaction with text's autoregressive constraints, alongside modality-specific flow heads that predict velocity fields in the respective latent spaces. This symmetric continuous design eliminates quantization errors at the architecture level, naturally unifying the understanding and generation of motion and images.

Beyond the architecture, we propose two key innovations for cross-modal semantic alignment under heterogeneous supervision. First, we design **D**ual-**P**osterior KL **A**lignment (**DPA**) for CMA-VAE by jointly training a Vision-Fused Motion Encoder $q_\psi(z \mid \text{motion}, \text{image})$ and a Motion Encoder $q_\phi(z \mid \text{motion})$, minimizing the KL divergence between their posteriors. This allows the Motion Encoder to absorb image-provided semantic supervision during training while requiring only motion at inference. Datasets lacking paired images participate by omitting the DPA loss. Second, we identify a supervision mismatch: the motion the model must generate is dense and kinematically rich, yet the text it learns from is much sparser—capturing only coarse action semantics while omitting details such as stride length, limb coordination, and subtle temporal transitions. Training the generation pathway solely from such under-specified signals leads to ambiguous learning, instability, and degraded fidelity. Yet the CMA-VAE latent z—the model's own continuous motion encoding—already preserves the full kinematic structure faithfully in compact form. This raises a natural question: before learning from sparse text, can the model first learn to generate motion from its own most informative encoding? We answer this with **L**atent **R**econstruction **A**lignment (**LRA**), a simple yet effective self-supervised pre-training strategy that treats z embeddings as "dense motion prompts" and trains the model to reconstruct z from noise in latent space. This self-reconstruction provides precise, unambiguous geometric supervision that jointly calibrates the embedder, LLM backbone, and flow head, establishing a robust motion-aware pathway as the shared foundation for all downstream tasks.

Thanks to these designs, UniMotion achieves state-of-the-art results across virtually all downstream tasks, supporting true any-to-any understanding and

generation among Motion, Text, and RGB. It demonstrates especially prominent advantages on cross-modal compositional tasks. Our main contributions are:

1. We propose UniMotion, the first framework to unify Motion, Text, and RGB understanding and generation in a single architecture, overcoming the modality-coverage and task-direction limitations of prior methods. Beyond reusing a pretrained Text–RGB MLLM, the key challenge we solve is inserting a kinematically-constrained *continuous* motion modality—absent from such backbones—into the shared space.
2. We propose a fully continuous motion paradigm: CMA-VAE encodes motion into a visual-semantically enriched continuous latent space, and symmetric dual-path embedders with hybrid attention and modality-routed LoRA construct parallel continuous pathways for Motion and RGB, eliminating quantization bottlenecks at the architecture level.
3. We propose Dual-Posterior KL Alignment (DPA) and Latent Reconstruction Alignment (LRA) as complementary alignment strategies: DPA injects visual-semantic supervision into the motion encoder via posterior alignment, while LRA leverages dense motion latents for self-supervised pathway calibration—jointly constructing a well-aligned tri-modal space that improves training stability and performance across all tasks.

Table 1. Comparison of UniMotion with representative methods from a task perspective. ✓ indicates the method supports the task; ✗ indicates it does not. UniMotion uniquely unifies comprehensive understanding, generation, and editing across Motion, Text, and RGB modalities within a single continuous motion-aware MLLM. "Size" is the backbone parameter count; UniMotion (1.5B) is smaller than the 7B MLLM baselines yet covers all seven tasks.

Method	Venue	Size	Understanding		Generation			Editing		Motion Repr.
			M2T	V2T	T2M	V2M	Pred	Mot.	Edit MGIE	
MotionGPT [14]	NeurIPS'23	220M	✓	✗	✓	✗	✓	✗	✗	VQ-VAE (discrete)
MG-MotionLLM [28]	CVPR'25	220M	✓	✗	✓	✗	✗	✓	✗	VQ-VAE (discrete)
UniPose [18]	CVPR'25	7B	✗	✓	✗	✓	✗	✗	✗	VQ-VAE pose tokens (discrete)
HMVLM [12]	NeurIPS'25	7B	✗	✓	✓	✓	✗	✗	✗	VQ-VAE body-part (discrete)
Show-o2 [32]	NeurIPS'25	1.5/7B	✗	✓	✗	✗	✗	✗	✗	—
UniMotion (Ours)	—	**1.5B**	✓	✓	✓	✓	✓	✓	✓	**CMA-VAE (continuous)**

2 Related Work

Table 1 summarizes the key distinctions between UniMotion and representative prior methods. We now discuss each related area below.

2.1 Human Motion Generation and Understanding

Human motion modeling covers text-driven generation (Text-to-Motion) and semantic understanding (Motion-to-Text), evolving from task-specific models to unified frameworks.

Discrete Tokenization Paradigm. Discretizing motion into codebook indices via VQ-VAE [26] has become mainstream. MotionGPT [14] treated motion as a "foreign language" for unified generation and understanding; MG-MotionLLM [28] advanced multi-granularity motion modeling. However, discretization inevitably introduces quantization errors, causing temporal jitter and detail loss, while codebook collapse limits diversity.

Continuous Representation Paradigm. To overcome quantization limitations, another line models motion in continuous latent space [33,40]. MLD [33] performs diffusion in the VAE [16] latent space, balancing quality and efficiency. Recent approaches combine motion VAEs with diffusion-based heads for smoother, more realistic generation while preserving semantic control.

Despite progress on Motion-Text tasks (Table 1), these methods remain confined to the Motion-Text subspace, lacking visual perception and generation needed for Vision-to-Motion or motion-guided editing.

2.2 Vision-Pose Multimodal Large Language Models

With the rise of MLLMs such as LLaVA [21], researchers have begun integrating human pose and skeletons into these models to enhance fine-grained behavior understanding. UniPose [18] discretizes 3D pose into tokens via VQ-VAE for unified pose understanding and generation. HMVLM [12] introduces MoE LoRA-based instruction tuning, yet still relies on discrete body-part tokenization and pairwise modality connections rather than a unified latent space.

These methods share common limitations: (1) most handle only static pose or pairwise connections without modeling continuous motion; (2) discrete quantization introduces precision bottlenecks; (3) visual interaction is typically unidirectional, lacking image generation capability.

2.3 Unified Multimodal Understanding and Generation

Unifying understanding and generation of arbitrary modalities within a single architecture is a fundamental direction toward building general artificial intelligence. Show-o [31] pioneered a single Transformer fusing autoregressive and discrete diffusion for joint text-image understanding and generation. Janus-Pro [4] substantially improved performance through data and model scale expansion; Show-o2 [32] further advanced cross-modal capabilities, with recent work exploring personalization [17] and efficient diffusion decoding [35].

While these works have achieved significant advances in the Text-RGB domain, human motion—encoding human intent and behavioral logic—remains absent from the core of unified frameworks. Treating motion as video pixels is inefficient and ignores skeletal topology and kinematic constraints.

3 Method

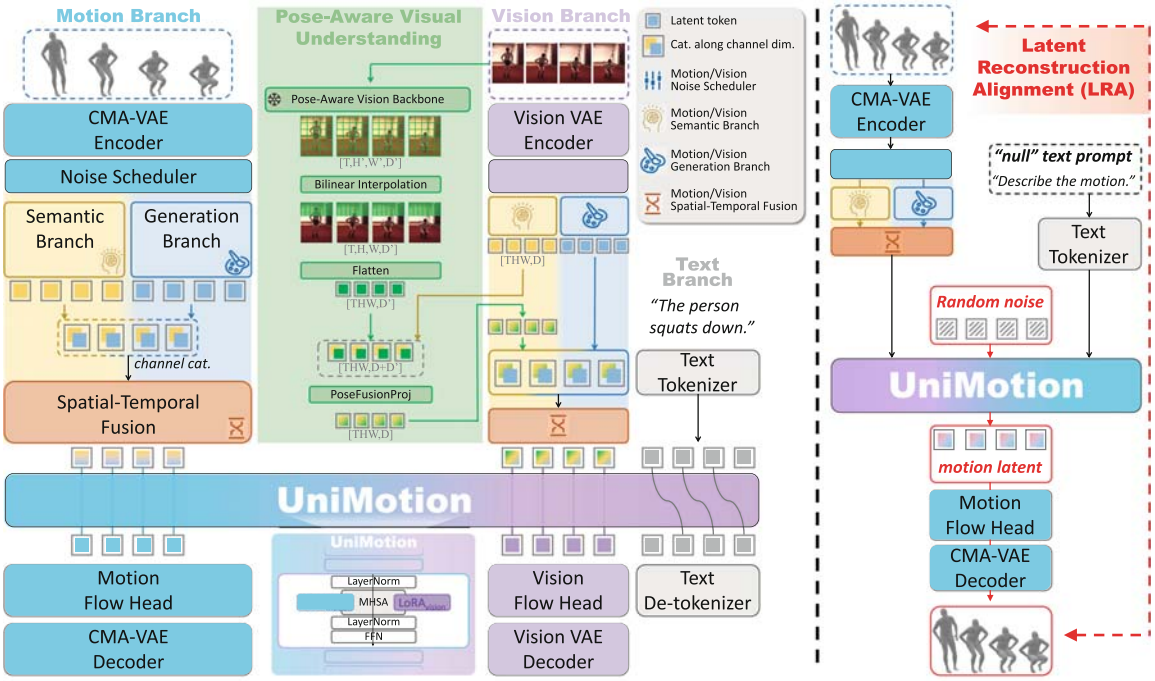


Fig. 2. Overview of UniMotion. **(Left)** UniMotion unifies motion, text, and RGB through symmetric continuous pathways: motion and images are encoded into continuous latents (via CMA-VAE and a vision VAE), mapped by a dual-path embedder that separates semantic abstraction from detail-preserving generation, and processed by a shared backbone for both multimodal understanding and modality-specific flow-based synthesis. **(Right)** Latent Reconstruction Alignment (LRA) pre-trains the motion pathway with a self-supervised Motion-to-Motion task, using motion latents as dense, unambiguous conditions to reconstruct motion from noise, thereby co-calibrating the embedder, backbone, and motion head before all downstream tri-modal learning.

3.1 Overview

UniMotion aims to unify modeling of Motion, Text, and RGB within a single framework, simultaneously supporting understanding and generation tasks (including T2M, M2T, Motion Prediction, Motion Editing, Vision-to-Motion, Vision-to-Text, and Motion-guided Image Editing (MGIE)). The core design philosophy is to treat motion as a continuous modality on equal footing with images and construct symmetric continuous pathways for both.

Unlike prior methods that rely on discrete VQ-VAE tokenization for representing motion dynamics (*e.g.*, MotionGPT, UniPose), we adopt continuous VAE representations for motion, which offers two key advantages: (1) it avoids irreversible information loss during quantization, preserving the temporal continuity and structural fidelity of motion; (2) continuous motion latents share a

symmetric representational form with the continuous RGB pathway, enabling cross-modal alignment naturally at the architecture level.

The overall framework, illustrated in Fig. 2, comprises three core components: (1) **CMA-VAE** (Sect. 3.2): encodes motion sequences into continuous latent representations with implicit visual-semantic supervision injected via Dual-Posterior KL Alignment (DPA); (2) **Unified Multimodal Architecture** (Sect. 3.3): built upon Show-o2 [32], with symmetric dual-path embedders for Motion and RGB, a shared LLM backbone, and modality-specific flow [20] heads; (3) **Latent Reconstruction Alignment (LRA)** (Sect. 3.4): using dense motion prompts to first warm up the motion pathway (the embedder, flow head, and motion-routed adaptation), followed by progressive multi-stage fine-tuning.

3.2 Cross-Modal Aligned Motion VAE (CMA-VAE)

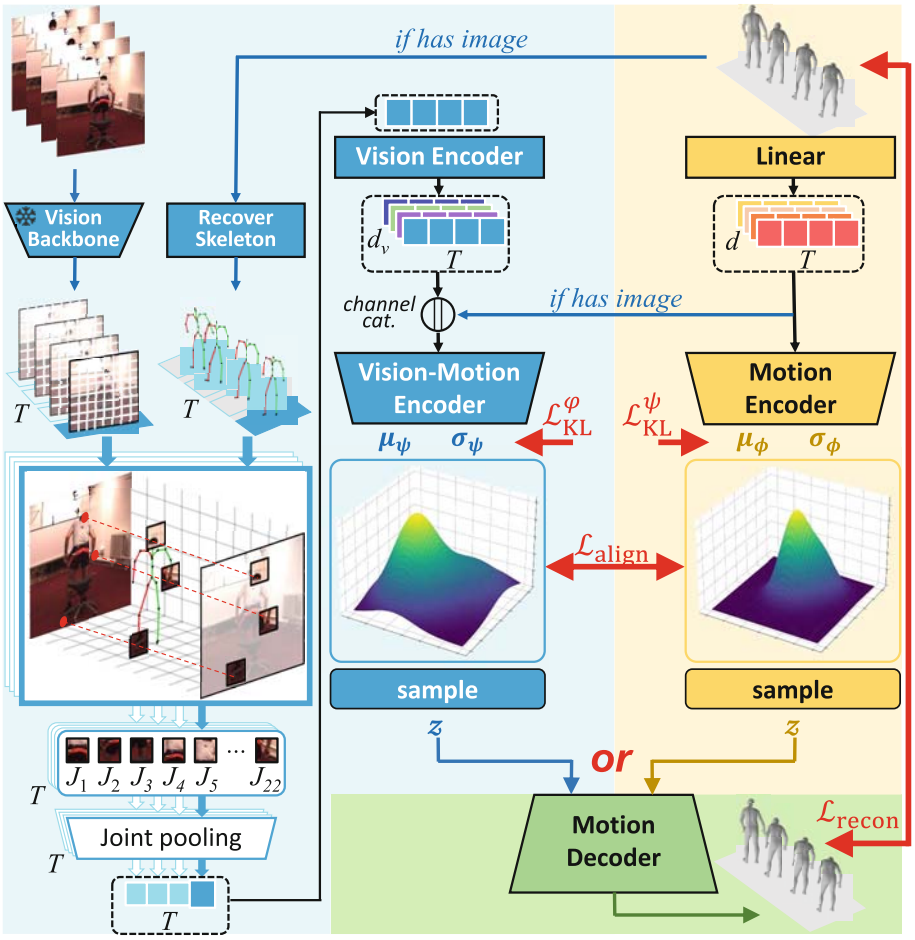


Fig. 3. CMA-VAE with DPA. CMA-VAE learns a continuous motion latent space using a motion-only encoder for inference and a vision-fused encoder for training-time visual supervision. When paired images are available, motion-guided visual features are fused with motion and distilled via DPA, enabling the shared decoder to learn visually informed motion latents without requiring images at inference.

CMA-VAE encodes variable-length motion sequences into continuous low-dimensional latent representations while injecting implicit visual-semantic supervision through Dual-Posterior KL Alignment (DPA). As shown in Fig. 3, its core consists of three components: Motion Encoder $q_\phi(z \mid \mathbf{m})$, Vision-Fused Motion Encoder $q_\psi(z \mid \mathbf{m}, \mathbf{v})$, and a shared Motion Decoder $p_\xi(\mathbf{m} \mid z)$. Unlike standard motion VAEs [33] that encode motion in isolation, CMA-VAE is cross-modal by construction: its motion-only inference encoder is coupled to a vision-fused training-time encoder via posterior alignment, so the latent space carries visual-semantic structure.

Motion Encoder. The Motion Encoder is used at inference. Given a motion sequence $\mathbf{m} \in \mathbb{R}^{T \times D_m}$, a linear layer maps each frame to the latent space, followed by learnable positional encodings and a SkipTransformer Encoder. A linear head predicts Gaussian parameters $(\mu_\phi, \log \sigma_\phi^2)$, and the latent code is obtained via reparameterization $z = \mu_\phi + \sigma_\phi \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, with $z \in \mathbb{R}^{T_z \times d}$.

Vision-Fused Motion Encoder. Used only during training, this encoder fuses motion and visual information. Its motion branch shares the front-end with the Motion Encoder, producing $\mathbf{h}_{\text{motion}} \in \mathbb{R}^{T_z \times d}$. The vision branch extracts spatial features from RGB image $\mathbf{v}$ via a frozen HRNet [23], then applies bilinear grid sampling at motion-guided 2D joint positions $\mathbf{j}_{2d}(\mathbf{m})$—skeleton projections derived from the motion sequence:

$$\mathbf{f}_{\text{vis}} = \text{VisionEnc}\big(\text{GridSample}(\text{HRNet}(\mathbf{v}), \mathbf{j}_{2d}(\mathbf{m}))\big) \in \mathbb{R}^{T_z \times d_v}. \tag{1}$$

After joint-dimension pooling, motion and visual features are concatenated and processed by an independent SkipTransformer Encoder to yield $(\mu_\psi, \log \sigma_\psi^2)$ and sample z_{fused}.

Motion Decoder. The decoder accepts latent z, passes it through positional encoding and SkipTransformer layers, and maps back to the D_m-dimensional motion space via a linear layer.

Dual-Posterior KL Alignment (DPA). The core idea of DPA is: constrain the Motion Encoder posterior $q_\phi(z \mid \mathbf{m})$ to approximate the Vision-Fused posterior $q_\psi(z \mid \mathbf{m}, \mathbf{v})$, so that the Motion Encoder implicitly absorbs visual-semantic supervision during training while requiring only motion input at inference. The total CMA-VAE training objective is:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{KL}} \left(\mathcal{L}_{\text{KL}}^{\phi} + \mathcal{L}_{\text{KL}}^{\psi}\right) + \lambda_{\text{align}} \cdot \mathcal{L}_{\text{align}}. \tag{2}$$

Reconstruction Loss. For samples with paired images, z_{fused} from the Vision-Fused Encoder is used for decoding; otherwise z_{motion} from the Motion Encoder is used:

$$\mathcal{L}_{\text{recon}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \text{SmoothL1}(\hat{m}_i, m_i). \tag{3}$$

KL Regularization. Both encoder posteriors are independently regularized toward the standard normal prior via the standard Gaussian KL $\mathcal{L}_{\mathrm{KL}}^{\star} = \frac{1}{2}\sum_{k=1}^{d}(\mu_{\star,k}^2 + \sigma_{\star,k}^2 - \log\sigma_{\star,k}^2 - 1)$, $\star \in \{\phi, \psi\}$.
DPA Alignment Loss. Computed only for samples with paired images, this term distills visual-semantic knowledge into the Motion Encoder:

$$\mathcal{L}_{\mathrm{align}} = D_{\mathrm{KL}}\big(q_\phi(z \mid \mathbf{m}) \,\|\, q_\psi(z \mid \mathbf{m}, \mathbf{v})\big). \tag{4}$$

For two diagonal Gaussians $\mathcal{N}(\mu_\phi, \sigma_\phi^2)$ and $\mathcal{N}(\mu_\psi, \sigma_\psi^2)$, the closed-form KL is:

$$D_{\mathrm{KL}} = \frac{1}{2}\sum_{k=1}^{d}\left(\log\frac{\sigma_{\psi,k}^2}{\sigma_{\phi,k}^2} + \frac{\sigma_{\phi,k}^2 + (\mu_{\phi,k} - \mu_{\psi,k})^2}{\sigma_{\psi,k}^2} - 1\right). \tag{5}$$

$\mathcal{L}_{\mathrm{align}}$ uses a linear warm-up schedule [8] to prevent strong alignment from disrupting reconstruction in early training. We adopt $D_{\mathrm{KL}}(q_\phi \| q_\psi)$ with the Vision-Fused posterior q_ψ detached as the alignment target. In the knowledge-distillation sense (student→teacher), this is the reverse KL, which is mode-seeking: it drives q_ϕ to concentrate on the most salient semantic modes of q_ψ, yielding compact, high-confidence motion representations that capture the core visual-semantic information while filtering out view-specific visual noise. The forward direction $D_{\mathrm{KL}}(q_\psi \| q_\phi)$ would be mode-covering, forcing q_ϕ to spread over all modes of q_ψ—including noisy or irrelevant visual modes—leading to an over-dispersed posterior that dilutes representational precision.
Flexible Data Utilization. For datasets without paired images (*e.g.*, HumanML3D), $\mathcal{L}_{\mathrm{align}}$ is simply dropped. For datasets with paired images (*e.g.*, Human3.6M), all three losses apply jointly. At inference, the Vision-Fused Encoder is entirely discarded, ensuring no overhead.

3.3 Unified Multimodal Architecture

UniMotion is built upon Show-o2 [32], extending it with a Motion modality pathway architecturally symmetric to RGB.

Dual-Path Embedder. Given the CMA-VAE latent $z \in \mathbb{R}^{T_z \times d}$, two parallel branches process z: a **Semantic branch** (MLP + Transformer Encoder layers) extracts high-level semantic features, mirroring SigLIP [36] on the vision side; a **Generation branch** (MLP + learnable positional encodings) maps z directly to the LLM hidden dimension, preserving fine-grained motion details and mirroring the vision PatchEmbed. The two outputs are concatenated and projected to the unified LLM hidden dimension via RMSNorm + MLP. All tasks uniformly use the fused embeddings. This dual-path design is particularly advantageous for motion-conditioned synthesis tasks (*e.g.*, Motion Editing and Prediction), where the two branches structurally decouple semantic comprehension from fine-grained detail preservation. To further strengthen visual-motion alignment, the RGB pathway is additionally equipped with a pose-aware vision backbone: initialized from a pretrained human body encoder [5] and kept frozen,

it extracts body-structure-aware features that complement SigLIP's global visual semantics. Both streams are fused into the same RGB token representation via projection, maintaining architectural symmetry with the Motion embedder.

Hybrid Attention and Modality-Routed LoRA. Hybrid attention maintains global causal ordering at the sequence level while enabling bidirectional full attention within each motion span, reconciling the flow matching objective—which requires simultaneous velocity field prediction across the entire motion latent—with text autoregressive generation. Modality-routed LoRA assigns separate low-rank adaptation branches for Text/RGB and Motion tokens in each attention layer, enabling modality-specific adaptation with only ∼2% additional parameters while preserving the LLM's existing capabilities.

Modality-Specific Flow Heads. For generation tasks, LLM hidden states are transformed back to the target modality latent space via modality-specific flow heads.

Motion Flow Head. A lightweight AdaLN-conditioned structure (Modulated Attention Blocks + MotionFinalLayer) maps backbone features to velocity predictions $\hat{v}_m \in \mathbb{R}^{T_z \times d_m}$ in the motion latent space. Timestep conditioning is injected via timestep embedding $\mathbf{c}_t$; the output layer is zero-initialized to stabilize early training.

Vision Flow Head. An isomorphic design whose output dimension aligns with the vision latent representation. This "shared backbone + modality-specific head" design balances parameter efficiency and cross-modal adaptation.

3.4 Latent Reconstruction Alignment (LRA)

The Cold-Start Problem and Dense Self-supervision. After DPA pre-training, the motion pathway is still uncalibrated: the embedder, motion flow head, and motion-routed adaptation have not yet been jointly aligned. Direct multi-task training from this state causes clear degradation (T2M R@3 only 0.801 vs. our final 0.841, see Sect. 4.3). While text descriptions are inherently sparse (one-to-many mapping), the CMA-VAE latent $z \in \mathbb{R}^{T_z \times d}$ is a dense, lossless encoding whose self-reconstruction constitutes an unambiguous one-to-one mapping—an ideal zero-cost pre-training signal for bootstrapping the motion pathway.

M2M Task Design. We instantiate this via a Motion-to-Motion (M2M) self-reconstruction task. The CMA-VAE Encoder produces z; the dual-path embedder projects z into the LLM, whose hidden states pass through the motion flow head to reconstruct z from noise:

$$\mathcal{L}_{\text{M2M}} = \mathbb{E}_{z_0 \sim \mathcal{N}(0,I),\, t \sim p(t)} \left[\| v_\theta(z_t,\, t \mid \text{Embed}_{\text{fused}}(z)) - (z - z_0) \|^2 \right], \tag{6}$$

where $z_t = t \cdot z + (1 - t) \cdot z_0$. Critically, the LLM receives $\text{Embed}_{\text{fused}}(z)$ as conditioning rather than the noised z_t (injected only into the flow head via

AdaLN), ensuring the motion pathway learns to structurally encode motion semantics rather than merely acting as a denoiser.

Co-calibration and Cross-Task Transfer. M2M simultaneously co-calibrates the embedder (compressing z into LLM-readable tokens), the motion-routed adaptation in the shared backbone (extracting structural cues), and the flow head (mastering latent-space geometry) in a tightly coupled manner—a calibration that sparse text supervision alone cannot provide due to ambiguous geometric feedback. Once calibrated, this pathway becomes the shared foundation for all downstream tasks: T2M benefits from the pre-calibrated flow head, M2T reuses the embedder's semantic compression, and Vision→M leverages the aligned pipeline so the LLM can focus purely on cross-modal mapping. Crucially, LRA does not degenerate into a trivial identity mapping: the dual-path embedder compresses z into LLM-compatible tokens via non-trivial projection and Transformer encoding, and the flow head must predict velocity fields from Gaussian noise conditioned on LLM hidden states—a probabilistic mapping that is architecturally incapable of bypassing the LLM.

4 Experiments

Our experiments span three dimensions: (1) **Unification**—systematic comparison with SOTA on tasks spanning Motion, Text, and RGB (Sect. 4.2); (2) **Continuity**—ablation validating the superiority of CMA-VAE (Sect. 4.3); (3) **Alignment**—ablation confirming the key contributions of DPA and LRA (Sect. 4.3).

4.1 Experimental Setup

Datasets. We evaluate UniMotion on HumanML3D [10] (Text-to-Motion, Motion-to-Text, Motion Prediction), MotionFix [1] (Motion Editing), Human3.6M [13] (Vision-to-Motion, Vision-to-Text), and MoVid [3] (Vision-to-Text). We also construct a triplet evaluation set from Human3.6M and 3DPW for Motion-guided Image Editing (MGIE). We adopt a unified 269-dimensional motion representation to support both generation and body recovery tasks.

Table 2. Unified multi-task comparison using one representative metric per task. "N/A" indicates the method does not support the task. UniMotion is the only method covering all seven tasks.

Method	T2M R@3↑	M2T BertScore↑	MotionPred ADE↓	MotionEdit R@3↑	V2M MPJPE↓	V2T BLEU-4↑	MGIE Mot.Acc↑
MotionGPT [14]	0.778	32.4	4.745	N/A	N/A	N/A	N/A
MG-MotionLLM [28]	0.802	36.7	N/A	73.23	N/A	N/A	N/A
UniPose [18]	N/A	N/A	N/A	N/A	81.8	17.3	N/A
Show-o2 [32]	N/A	N/A	N/A	N/A	N/A	12.1	N/A
UniMotion (Ours)	**0.841**	**41.2**	**3.172**	**84.94**	**75.0**	**21.9**	**0.67**

Implementation Details. The unified backbone is based on Show-o2 1.5B [32], equipped with a dual-path motion embedder and a motion flow head. Modality-routed LoRA (rank 32) is applied to all attention layers. We train with AdamW and use Euler ODE for generation inference. All training is conducted on 4×A6000 GPUs.

4.2 Comparison with State-of-the-Art

Table 3. Text-to-Motion generation on HumanML3D. → indicates closer to *Real* is better. UniMotion† is the single-task model; UniMotion is the unified multi-task model. **Bold**: best; underline: second best.

Type	Method	R@1↑	R@2↑	R@3↑	FID↓	MMDist↓	DIV→
—	Real	0.511	0.703	0.797	0.002	2.974	9.503
Gen. only	T2M-GPT [38]	0.491	0.680	0.775	0.116	3.118	9.761
	DiverseMotion [22]	0.515	0.706	0.802	0.072	2.941	9.683
	MoMask [9]	0.521	0.713	0.807	**0.045**	2.958	9.620
	UniMotion†	0.528	0.726	0.830	0.231	2.845	9.649
Gen. & Und.	TM2T [11]	0.424	0.618	0.729	1.501	3.467	8.589
	MotionGPT [14]	0.492	0.681	0.778	0.232	3.096	**9.528**
	HMVLM [12]	0.463	0.646	0.744	0.156	3.328	9.544
	MG-MotionLLM [28]	0.516	0.706	0.802	0.303	2.952	9.960
	UniMotion	**0.557**	**0.749**	**0.841**	0.194	**2.715**	9.583

Table 4. Motion-to-Text understanding on HumanML3D. Evaluation protocol follows [11]. UniMotion† is the single-task model. **Bold**: best; underline: second best.

Method	R@1↑	R@3↑	MMDist↓	Bleu@1↑	Bleu@4↑	Rouge↑	CIDEr↑	BertScore↑
Real	0.523	0.828	2.901	—	—	—	—	—
TM2T [11]	0.516	0.823	2.935	48.9	7.00	38.1	16.8	32.2
MotionGPT [14]	0.543	0.827	2.821	48.2	12.47	37.4	29.2	32.4
LaMPM2T [19]	0.547	0.831	2.808	47.8	13.04	37.1	28.9	32.7
MG-MotionLLM [28]	**0.592**	**0.866**	2.581	—	8.06	—	—	36.7
UniMotion†	0.547	0.843	2.562	56.9	18.6	44.7	35.4	38.5
UniMotion	0.562	0.857	**2.481**	**60.4**	**20.7**	**46.5**	**39.3**	**41.2**

Unified Multi-task Comparison. Table 2 provides a unified comparison across all seven tasks using one representative metric per task for transparent cross-task evaluation. UniMotion is the only method covering all tasks.

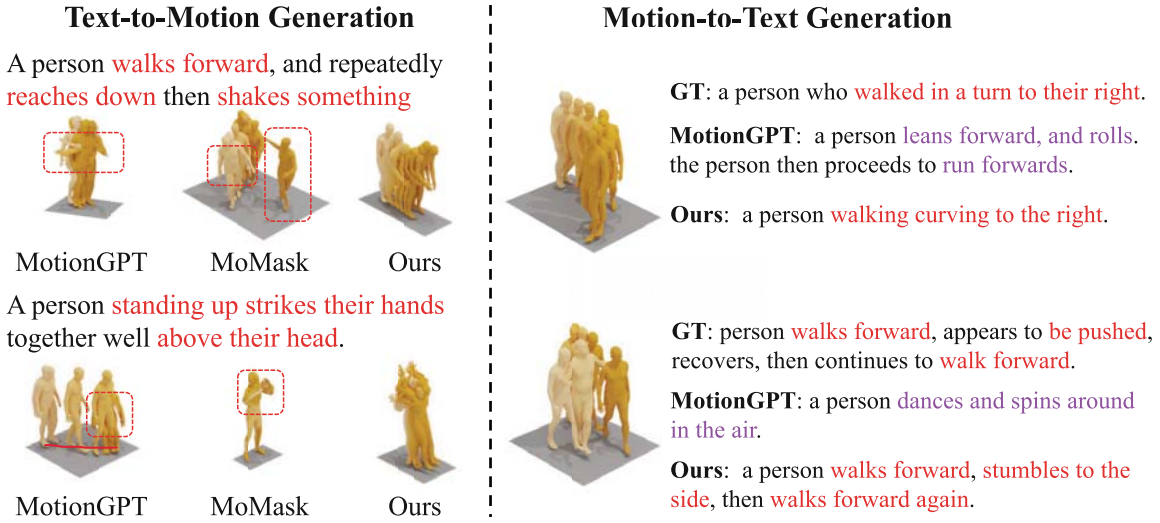


Fig. 4. Qualitative comparison on T2M and M2T.

Text-to-Motion Generation. Table 3 compares UniMotion with SOTA on HumanML3D. UniMotion† denotes the task-specific variant sharing the same architecture and training recipe, trained only on the evaluated task without cross-task supervision; UniMotion is the unified multi-task model.

UniMotion achieves the best R-Precision and MMDist by clear margins across compared methods, establishing strong semantic alignment between text and generated motion. Notably: (1) the substantial gains on these alignment metrics reflect CMA-VAE's continuous representation and DPA's effectiveness in binding motion-text semantics; (2) the unified multi-task model consistently outperforms the single-task variant (UniMotion†), demonstrating positive cross-modal transfer from RGB supervision; (3) while single-task discrete methods obtain lower FID (MoMask: 0.045)—reflecting their advantage in single-distribution fitting—UniMotion's leading semantic alignment highlights the cross-modal reasoning enabled by continuous representations within a unified framework. Furthermore, Diversity (9.583) closely matches the real data (9.503), confirming that alignment improvements preserve sample variability. As qualitatively confirmed in Fig. 4, MoMask's low FID does not prevent it from missing critical spatial constraints (*e.g.*, hands only reaching shoulder level instead of above the head), while UniMotion faithfully renders fine-grained modifiers and produces more accurate M2T captions than MotionGPT.

Motion-to-Text Understanding. As shown in Table 4, UniMotion leads all caption-quality metrics by substantial margins (Bert-Score: 41.2 vs. 36.7; CIDEr:

39.3 vs. 29.2; Bleu@4: 20.7 vs. 13.04), indicating semantically faithful and detail-rich descriptions rather than generic templates. We attribute this to CMA-VAE's continuous latents preserving kinematic nuances (*e.g.*, stride patterns, joint articulations) that discrete tokenization discards. MG-MotionLLM achieves higher retrieval R@k, reflecting stronger embedding-level matching, while UniMotion's advantage lies in generation quality. The unified model further surpasses the single-task baseline (BertScore: 41.2 vs. 38.5), confirming positive cross-modal transfer from DPA's visual-semantic supervision.

Motion Prediction and Text-Conditioned Motion Editing. UniMotion outperforms all baselines on both tasks (Table 5). Beyond CMA-VAE's continuous latents for preserving temporal dynamics, DPA and visual co-training inject visual-geometry priors (*e.g.*, plausible body configurations, global balance) that are especially helpful for forecasting future poses and editing local joints while maintaining whole-body coherence. For motion prediction, consistent gains in FID (0.871 vs. 0.905), ADE (3.172 vs. 4.745), and FDE (5.068 vs. 6.040) reflect these advantages. For motion editing, UniMotion substantially surpasses MG-MotionLLM (R@1: 63.81 vs. 47.96; R@3: 84.94 vs. 73.23; FID: 0.170 vs. 0.409), confirming that continuous representation combined with visual grounding enables precise text-guided modification at the joint level. Architecturally, the dual-path embedder is critical for these tasks: its semantic–generative decoupling enables the source motion to simultaneously provide high-level structural conditioning and fine-grained kinematic detail for the target.

Table 5. (a) Motion prediction on AMASS (HumanML3D subset), following MotionGPT [14]. (b) Text-conditioned motion editing on MotionFix; R@k is generated-to-target retrieval precision (%).

(a) Motion Prediction

Method	FID↓	ADE↓	FDE↓
Real	0.002	—	—
MDM [25]	6.031	5.446	8.561
MotionGPT [14]	0.905	4.745	6.040
UniMotion	**0.871**	**3.172**	**5.068**

(b) Motion Editing

Method	R@1↑	R@3↑	FID↓
GT	100.0	100.0	—
MDM [25]	39.10	54.84	0.917
MG-MotionLLM [28]	47.96	73.23	0.409
UniMotion	**63.81**	**84.94**	**0.170**

Vision-to-Motion and Motion-Guided Image Editing (MGIE). As reported in Table 7, among MLLM methods, UniMotion significantly surpasses the strong UniPose baseline on both MPJPE (75.0 vs. 81.8) and PA-MPJPE (46.1 vs. 50.9), benefiting from DPA's visual-geometry priors and CMA-VAE's precise continuous encoding. The remaining gap to specialist methods is expected for a general-purpose framework. For V2M we follow UniPose's single-reference-image protocol; the input image is encoded through the RGB pathway at inference, while the CMA-VAE vision-fused encoder used for DPA is discarded after

Table 6. Comparison of different motion representation methods. APE and AVE denote *mean-pose* Absolute Position Error and Average Velocity Error, respectively, in cm and cm/s. All variants were trained independently during the pure reconstruction phase, followed by multi-task fine-tuning under the same LLM backbone. **Bold**: best; underline: second best.

Motion Repr.	Reconstruction Quality			T2M	M2T	Pred	Edit
	APE↓	AVE↓	FID↓	R@3↑	BertScore↑	ADE↓	R@3↑
VQ-VAE (MotionGPT[14])	17.15	0.813	0.0674	0.771	35.0	4.713	69.71
MLD-VAE (MLD[33])	9.28	0.981	0.1283	0.810	37.2	3.784	77.46
CMA-VAE (Ours)	**3.53**	**0.428**	**0.0282**	**0.841**	**41.2**	**3.172**	**84.94**

Table 7. Vision-to-Motion on Human3.6M (H3.6M). **Bold**: best in category for each group and evaluation setting under the same protocol.

Method	Type	MPJPE↓	PA-MPJPE↓
Specialist Models			
HMR [15]	Spec.	100.7	67.7
PyMAF [37]	Spec.	64.2	44.9
SMPLer [34]	Spec.	**50.8**	37.3
HMR2.0 [7]	Spec.	52.2	37.1
TokenHMR [5]	Spec.	52.4	36.8
Zolly [27]	Spec.	55.0	**35.9**
MLLM-based Models			
ChatPose [6]	MLLM	146.7	92.4
UniPose [18]	MLLM	81.8	50.9
UniMotion	MLLM	**75.0**	**46.1**

Table 8. Motion-guided Image Editing (MGIE). UP+So2 is a two-stage text-mediated pipeline; OP+CN renders SMPL-derived skeletons as spatial conditions. Mot. Acc. is the hit rate based on HMR2.0-estimated pose, using PA-MPJPE $\leq$ 100.0 mm as success.

Method	FID↓	CLIP↑	Mot.Acc↑
UniPose+Show-o2 [18,32]	26.16	0.22	0.50
OpenPose+ControlNet [2,39]	22.34	0.29	0.59
UniMotion	**18.92**	**0.31**	**0.67**

Table 9. Ablation of DPA and LRA on T2M, M2T, Motion Prediction, Motion Editing, and Vision-to-Motion.

Setting	T2M R@3↑	M2T BertScore↑	MotionPred ADE↓	MotionEdit R@3↑	V2M MPJPE↓
w/o DPA	0.818	38.4	3.654	80.35	83.1
w/o LRA	0.801	38.1	3.777	78.72	84.3
Full UniMotion	**0.841**	**41.2**	**3.172**	**84.94**	**75.0**

training. Performance is driven mainly by our DPA/LRA alignment rather than the frozen pose-aware backbone (which alone reaches only 86.9 vs. our 75.0 MPJPE), and the model transfers zero-shot to in-the-wild 3DPW (93.6 MPJPE), indicating robustness beyond clean indoor imagery.

For MGIE, UniMotion is the first unified motion-aware MLLM supporting end-to-end motion-conditioned image generation in a shared latent space, without explicit skeleton rendering or text mediation. Against UniPose+Show-o2 [18,32] and OpenPose+ControlNet [2,39], it outperforms both on all metrics (Table 8), with the largest margin on Motion Accuracy (0.67 vs. 0.59), confirming that end-to-end latent-space reasoning avoids the information loss of staged pipelines.

4.3 Ablation Studies

CMA-VAE Motion Representation Comparison. To systematically evaluate CMA-VAE, we compare against VQ-VAE (MotionGPT [14]) and MLD-VAE (MLD [33]) in Table 6. VQ-VAE suffers the worst absolute position accuracy (APE=17.15) due to irreversible quantization errors, and transfers weakest to downstream tasks (T2M R@3=0.771, Edit R@3=69.71); its seemingly lower AVE (0.813) and FID (0.0674) compared to MLD-VAE reflect the bounded codebook artificially limiting deviations rather than faithful reconstruction. MLD-VAE improves global positioning (APE: 9.28) yet exhibits a position–velocity trade-off (AVE: 0.981, FID: 0.1283), confirming that a plain continuous VAE without cross-modal anchoring loosens temporal dynamics. Nevertheless, MLD-VAE substantially outperforms VQ-VAE on all downstream tasks, demonstrating that continuous latent spaces inherently facilitate LLM cross-modal alignment regardless of absolute reconstruction quality.

CMA-VAE, calibrated by DPA, resolves this trade-off and achieves the best results on all metrics—both reconstruction (APE=3.53, AVE=0.428, FID=0.0282) and downstream transfer (T2M R@3=0.841, Edit R@3=84.94). Adding DPA to the same architecture improves all tasks (T2M R@3: 0.818→0.841; Edit R@3: 80.35→84.94), confirming that implicit visual-semantic supervision is the core driver of CMA-VAE's superiority.

DPA and LRA Ablation. As shown in Table 9, removing either component causes consistent performance drops across all tasks. Without LRA, the motion pathway is insufficiently calibrated by sparse cross-modal supervision alone, leading to drops on T2M (R@3: 0.801 vs. 0.841), M2T (BertScore: 38.1 vs. 41.2), Motion Prediction (ADE: 3.777 vs. 3.172), Motion Editing (R@3: 78.72 vs. 84.94), and Vision→M (MPJPE: 84.3 vs. 75.0). Without DPA, the representation loses explicit visual-semantic alignment and likewise remains consistently below the full model across all tasks. Notably, DPA's gains extend to HumanML3D-only tasks (T2M, M2T) despite lacking paired images in that dataset, evidencing implicit knowledge transfer through shared encoder parameters. The two mechanisms are complementary: LRA calibrates the motion pathway's geometric capacity via dense self-supervision, while DPA enriches representational semantics through cross-modal posterior alignment.

5 Conclusion

We presented UniMotion, to our knowledge the first continuous Motion-Text-RGB unified framework in the motion-LLM setting for human-centric multimodal reasoning and unified understanding and generation within a single architecture. By treating motion as a continuous modality on equal footing with RGB, UniMotion constructs symmetric continuous pathways through CMA-VAE and a dual-path embedder, avoiding the quantization artifacts of discrete tokenization. Dual-Posterior KL Alignment and Latent Reconstruction Alignment jointly establish a well-aligned tri-modal latent space, yielding strong results across diverse tasks with clear advantages on cross-modal compositional tasks. As current tri-modal benchmarks are predominantly single-person, our evaluation focuses on single-person motion; the per-span hybrid attention is compatible with multi-person interaction, which we leave as future work.

Acknowledgement. This work was supported by National Natural Science Foundation of China (No. 62473007), Guangdong Outstanding Youth Fund (No. 2026B1515020015), Shenzhen Innovation in Science and Technology Foundation for The Excellent Youth Scholars (No. RCYX20231211090248064).

References

1. Athanasiou, N., Cseke, A., Diomataris, M., Black, M.J., Varol, G.: Motionfix: text-driven 3d human motion editing. In: SIGGRAPH Asia 2024 Conference Papers, pp. 1–11 (2024)
2. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. (2019)
3. Chen, L.H., et al.: Motionllm: understanding human behaviors from human motions and videos. arXiv preprint arXiv:2405.20340 (2024)

4. Chen, X., et al.: Janus-pro: unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
5. Dwivedi, S.K., Sun, Y., Patel, P., Feng, Y., Black, M.J.: Tokenhmr: advancing human mesh recovery with a tokenized pose representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1323–1333 (2024)
6. Feng, Y., Lin, J., Dwivedi, S.K., Sun, Y., Patel, P., Black, M.J.: Chatpose: chatting about 3d human pose. In: CVPR, pp. 2093–2103 (2024)
7. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4d: reconstructing and tracking humans with transformers. In: CVPR, pp. 14783–14794 (2023)
8. Goyal, P., et alcurate, large minibatch sgd: training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
9. 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)
10. 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 (CVPR), pp. 5152–5161 (2022)
11. 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)
12. Hu, L., Ye, Y., Xia, S.: Hmvlm: Human motion-vision-lanuage model via moe lora. arXiv preprint arXiv:2511.01463 (2025)
13. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6 m: large scale datasets and predictive methods for 3d human sensing in natural environments. TPAMI **36**(7), 1325–1339 (2013)
14. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: human motion as a foreign language. Adv. Neural. Inf. Process. Syst. **36**, 20067–20079 (2023)
15. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR, pp. 7122–7131 (2018)
16. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
17. Li, Y., et al.: Personalize your large vision-language models with in-context prompt tuning. arXiv preprint arXiv:2605.31513 (2026)
18. Li, Y., Hou, R., Chang, H., Shan, S., Chen, X.: Unipose: a unified multimodal framework for human pose comprehension, generation and editing. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 27805–27815 (2025)
19. Li, Z., et al.: Lamp: language-motion pretraining for motion generation, retrieval, and captioning. arXiv preprint arXiv:2410.07093 (2024)
20. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023). https://openreview.net/forum?id=PqvMRDCJT9t
21. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
22. Lou, Y., Zhu, L., Wang, Y., Wang, X., Yang, Y.: Diversemotion: towards diverse human motion generation via discrete diffusion. arXiv preprint arXiv:2309.01372 (2023)
23. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5693–5703 (2019)

24. Team, C.: Chameleon: mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
25. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Bermano, A.H., Cohen-Or, D.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
26. Van Den Oord, A., Vinyals, O.: Neural discrete representation learning. In: Advances in Neural Information Processing Systems, p. 30 (2017)
27. Wang, W., et al.: Zolly: zoom focal length correctly for perspective-distorted human mesh reconstruction. In: ICCV, pp. 3925–3935 (2023)
28. Wu, B., et al.: Mg-motionllm: a unified framework for motion comprehension and generation across multiple granularities. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 27849–27858 (2025)
29. Wu, C., et al.: Janus: decoupling visual encoding for unified multimodal understanding and generation. arXiv preprint arXiv:2410.13848 (2024)
30. Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.S.: Next-gpt: any-to-any multimodal llm. In: Forty-first International Conference on Machine Learning (2024)
31. Xie, J., et al.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv, (2024)
32. Xie, J., Yang, Z., Shou, M.Z.: Show-o2: Improved native unified multimodal models. arXiv preprint arXiv:2506.15564 (2025)
33. Xin, C., et al.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
34. Xu, X., Liu, L., Yan, S.: Smpler: taming transformers for monocular 3d human shape and pose estimation. In: TPAMI (2023)
35. Yuan, Y., Zhao, H., Meng, X.: Visual-redundancy-controlled parallel decoding for diffusion-based multimodal large language models. arXiv preprint arXiv:2605.25820 (2026)
36. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11975–11986 (2023)
37. Zhang, H., et al.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In: ICCV (2021)
38. Zhang, J., et al.: T2m-gpt: generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
39. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3836–3847 (2023)
40. Zhang, M., et al.: Motiondiffuse: text-driven human motion generation with diffusion model. IEEE Trans. Pattern Anal. Mach. Intell. **46**(6), 4115–4128 (2024)
41. Zhou, C., et al.: Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv preprint arXiv:2408.11039 (2024)