IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 May 2026, accepted 26 May 2026, date of publication 2 June 2026, date of current version 8 June 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3699430

RESEARCH ARTICLE

Trimodal Disentangled Representation Learning for Motion Understanding

MYUNGIN KIM[1] **AND RI YU**[1,2]
[1]Department of Artificial Intelligence, Ajou University, Suwon-si 16499, South Korea
[2]Department of Software and Computer Engineering, Ajou University, Suwon-si 16499, South Korea

Corresponding author: Ri Yu (riyu@ajou.ac.kr)

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) through the Artificial Intelligence Convergence Innovation Human Resources Development grant funded by Korean Government [Ministry of Science and ICT (MSIT)] under Grant IITP-2026-RS-2023-00255968.

ABSTRACT Trimodal motion understanding systems that integrate motion, video, and text face a fundamental challenge: existing approaches typically excel at either generation or retrieval, but rarely both. This trade-off arises from conflicting optimization objectives—generation requires modality-independent representations to preserve diversity, whereas retrieval benefits from tightly aligned cross-modal embeddings for accurate matching. To address this challenge, we introduce a TRImodAl representation Disentanglement framework (TRIAD) that decomposes trimodal embeddings into task-specific components: a generative subspace for generation and a discriminative subspace for retrieval. A shallow dual-router Mixture-of-Experts architecture enables independent optimization pathways for stable task-specific disentanglement. Our results demonstrate that explicit task-specific decomposition allows a single unified model to perform both generation and retrieval. TRIAD provides empirical analysis on jointly supporting generation and retrieval within a single trimodal framework.

INDEX TERMS Cross-modal retrieval, human motion generation, mixture of experts, multimodal learning, representation disentanglement, trimodal learning.

I. INTRODUCTION

Human motion understanding has become essential for character animation, video game development, and robotics [1], [2], [3]. Recent advances in generative models [4], [5], [6] have significantly improved motion synthesis capabilities, enabling more realistic and controllable motion generation. Motion can be understood through multiple modalities: natural language descriptions that capture semantic intent, video sequences that provide visual context, and motion data itself. Dual-modality approaches have shown success in text-to-motion generation [7], [8] and video-based motion analysis [9], [10]. However, text descriptions alone cannot fully convey spatial nuances such as movement style or speed variations, while video references provide rich visual context but lack explicit semantic abstraction. Integrating all three modalities—text, video, and motion—within a unified framework enables complementary supervision: text provides high-level semantic guidance, video supplies fine-grained visual cues, and their combination offers richer training signals for motion representation learning. Despite these potential benefits, trimodal learning remains largely unexplored in motion understanding.

Current trimodal motion systems face a fundamental architectural limitation: they excel at either generation or retrieval, but not both. MotionCLIP [11], as the pioneering work in trimodal motion understanding, was designed primarily for generative tasks. Its VAE-based architecture successfully enables text-to-motion generation and motion editing by learning a continuous latent space aligned with CLIP embeddings. However, this generative design inherently struggles with discriminative tasks—when evaluated on motion retrieval benchmarks, MotionCLIP's performance is significantly inferior to specialized retrieval methods.

LAVIMO [12] represents a different approach, focusing on cross-modal retrieval through contrastive learning with

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

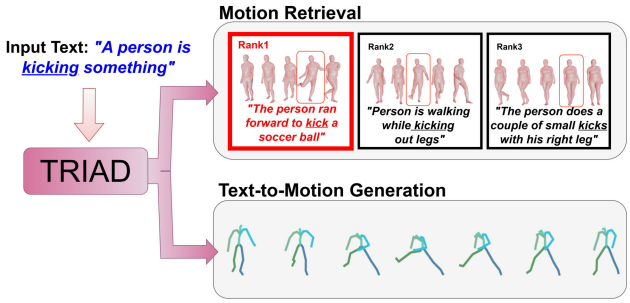

FIGURE 1. Dual-capability motion understanding through disentanglement. TRIAD enables both motion retrieval and generation from a single text query. Given "a person is kicking something," the discriminative branch retrieves semantically similar motions (top, Rank1-3), while the generative branch synthesizes novel kicking motions (bottom). This unified framework addresses the fundamental trade-off between retrieval and generation in existing approaches.

attention-based feature fusion. By fusing features from different modalities early in the pipeline, LAVIMO achieves strong performance in text-to-motion and video-to-motion retrieval tasks. However, this early fusion strategy comes with a trade-off: the tight coupling of modalities makes it difficult to use these representations for generation tasks, where maintaining modality independence is beneficial for producing diverse outputs.

This fundamental trade-off stems from competing optimization objectives. When a single representation space must serve both generative modeling (requiring diverse, expressive features) and discriminative tasks (requiring tightly aligned cross-modal features), the resulting compromise satisfies neither requirement optimally. Real-world applications, however, demand both capabilities—animators need to generate novel motions and search existing databases efficiently.

To this end, we propose a representational disentanglement strategy to address this trade-off (Fig. 1). Our method decomposes trimodal representations into specialized components for motion generation (z^{gen}) and motion discrimination (z^{disc}). This disentanglement enables independent optimization pathways for different downstream tasks while maintaining semantic coherence across modalities. While disentanglement has proven successful in other multimodal domains—particularly in sentiment analysis where studies decompose text, audio, and visual information—adapting these principles to spatiotemporal motion data presents unique challenges due to the temporal dependencies and kinematic constraints inherent in human motion. To the best of our knowledge, this is the first attempt to explicitly decompose trimodal motion representations into task-specific subspaces for generation and retrieval. Table 1 summarizes this contrast: prior trimodal works (MotionCLIP, LAVIMO) sit on opposite ends of the generation–retrieval trade-off, while our task-specific disentanglement is the first to overcome it within a single trimodal framework. We employ a shallow Mixture of Experts (MoE) architecture [13], [14] with dual routing shared across modalities to facilitate stable representation disentanglement.

To validate our disentanglement approach, we carefully select evaluation tasks that test discriminative and generative capabilities independently. For the discriminative branch, we employ cross-modal retrieval, which has been established as a fundamental benchmark in vision-language models such as CoCa [15] and BLIP [16]. Retrieval tasks effectively evaluate the quality of cross-modal alignment by measuring how well embeddings from different modalities correspond to each other—a core requirement for discriminative understanding. For the generative branch, we employ text-to-motion synthesis, following pioneering multimodal motion works including MotionCLIP [11] and TEMOS [7]. While specialized text-to-motion methods such as T2M-GPT [17] and MoMask [8] demonstrate high-fidelity generation, trimodal approaches offer distinct advantages for downstream applications. The learned embedding space enables semantic compositionality—arithmetic operations on embeddings naturally support motion editing and style transfer, such as combining ''running'' motion with ''boxing'' text to generate boxing-running animations, as demonstrated in MotionCLIP [11].

Our contributions are three-fold:

- We introduce the first application of representation disentanglement to motion understanding, enabling simultaneous optimization for both generative and discriminative tasks.
- We demonstrate that disentangled discriminative representations (z^{disc}) achieve effective cross-modal motion retrieval, improving upon generative baselines without requiring a separate retrieval model.
- We show that disentangled generative representations (z^{gen}) improve text-to-motion generation quality over MotionCLIP [11], validating that task-specific decomposition enhances generation capability.

II. RELATED WORK

A. TEXT-TO-MOTION GENERATION

Text-driven motion synthesis has progressed from early RNN-based approaches [18] to transformer architectures including ACTOR [19] and TEMOS [7]. Recent methods employ diffusion models (MDM [20], MotionDiffuse [21]), trimodal embeddings (MotionCLIP [11]), reciprocal modeling (TM2T [22]), and discrete representations (T2M-GPT [17]). These approaches primarily optimize for generation quality, focusing on motion realism, diversity, and text alignment.

B. CROSS-MODAL MOTION UNDERSTANDING

1) MOTION RETRIEVAL

Motion retrieval has gained prominence with the growth of motion capture databases. Early motion retrieval relied on low-level similarity measures such as motion graphs [23] and pose-sequence matching [24], which required handcrafted representations and lacked semantic querying. TMR [25] pioneered text-to-motion retrieval by creating cross-modal

TABLE 1. Conceptual comparison with prior trimodal motion frameworks. TRIAD is the first to support both generation and retrieval through explicit task-specific disentanglement.

Method	Generation	Retrieval	Disentangle	Architecture
MotionCLIP [11]	✓ (strong)	× (weak)	×	VAE
LAVIMO [12]	×	✓ (strong)	×	Early fusion
TRIAD (Ours)	✓	✓	✓ **(task-specific)**	**Dual-MoE**

embedding spaces through contrastive learning, establishing the foundation for language-based motion search. Recent advances in video-motion retrieval include MotionSet [26], which converts video clips to binary silhouette representations. Motion retrieval remains a standard component in animation pipelines, still used by many recent studies [27] as a baseline.

2) TRIMODAL MOTION UNDERSTANDING

MotionCLIP [11] represents the pioneering work in trimodal motion understanding, aligning motion representations with CLIP's text and image embeddings. This approach enabled remarkable capabilities including motion generation, semantic interpolation, and motion editing through natural language. However, MotionCLIP's architecture, while effective for generative tasks, inherently struggles with discriminative applications such as motion retrieval.

LAVIMO [12] addressed these retrieval limitations by designing attention-based fusion mechanisms specifically for cross-modal retrieval. Through specialized video encoders and contrastive learning, LAVIMO achieved state-of-the-art performance in text-to-motion and video-to-motion retrieval. However, its early fusion strategy that tightly couples modalities during training may be suboptimal for generation tasks, where maintaining modality independence is beneficial for producing diverse outputs. This reflects a natural trade-off in trimodal motion understanding, where architectures are typically specialized for either generation or retrieval tasks. Practical systems, however, benefit from supporting both capabilities within a unified architecture.

More recently, MotionBind [28] proposed a multi-modal alignment framework for motion understanding that supports retrieval, recognition, and generation through a shared representation space. While MotionBind demonstrates strong multi-task capabilities, it employs a unified shared representation without explicit task-specific decomposition. Our approach differs by explicitly factorizing representations into task-specific subspaces (z^{gen} and z^{disc}) through a dedicated dual-router MoE mechanism, providing a more structured exploration of the generation-retrieval trade-off.

C. REPRESENTATION DISENTANGLEMENT

Representation disentanglement aims to factorize learned representations into interpretable components. Early theoretical work on disentanglement in VAEs [29], [30] established foundational principles. In sentiment analysis, MISA [31] pioneered binary disentanglement into modality-invariant and modality-specific subspaces. TAILOR [32] employed BERT-like transformers for granularity-descent fusion. TriDIRA [33] advanced through triple disentanglement into task-relevant (r^*), effective modality-specific ($r \cap u$), and ineffective (u^*) components. More recent approaches incorporate language-focused attractors, as in DLF [34], to stabilize cross-modal factorization. However, applying these methods to motion understanding introduces unique challenges from temporal dependencies and kinematic constraints. Our work adapts these principles to the trimodal motion domain while addressing these motion-specific issues.

III. METHOD

A. OVERVIEW

Our framework addresses the fundamental trade-off between generation and retrieval in trimodal motion understanding through representation disentanglement. As illustrated in Fig. 2, each modality (motion, video, text) is first encoded into 512-dimensional representations. These representations are then processed through our MoE-based disentanglement module, which decomposes them into task-specific components: z^{gen} for generation tasks and z^{disc} for retrieval tasks.

The architecture branches into two specialized pathways: (i) a retrieval branch that employs feature fusion with z^{disc} representations of three modalities and utilizes triplet loss for hard negative mining, and (ii) a generation branch that maintains modality independence using z^{gen} representations with contrastive loss for cross-modal alignment.

B. MODALITY-SPECIFIC ENCODERS AND MOTION DECODER

Each of the three modalities captures a complementary aspect of human motion: text supplies high-level semantic intent (*what* the person is doing), video supplies fine-grained visual context such as gait or stylistic nuance that text cannot easily express, and motion data itself provides the ground-truth kinematics. Because these modalities live in fundamentally different signal spaces—tokens, pixels, and joint rotations—a single network cannot consume them directly. We therefore introduce one dedicated encoder per modality (E_{vi}, E_{te}, E_{mo}, defined in Eqs. (1), (5), and (6)), each projecting its input into a common 512-dimensional embedding space, plus a motion decoder D_{mo} (Eq. (7)) that inverts the motion latent back to executable trajectories so that the same shared space can support both retrieval and generation.

Video Encoder. The video modality grounds motion in observable visual evidence, such as gait, posture, and stylistic

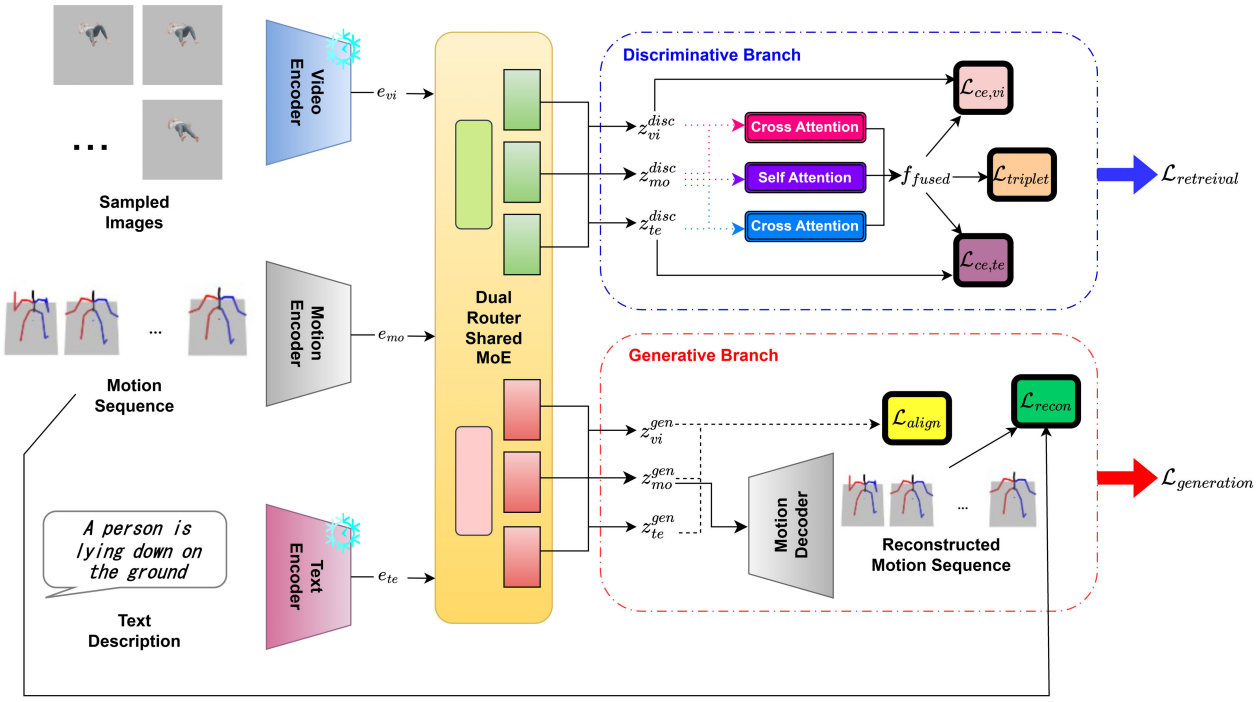

FIGURE 2. Overview of our trimodal disentanglement framework. Three modalities (video, motion, text) are encoded into 512-dimensional embeddings. Video and text encoders are frozen to preserve T2L's pretrained capabilities. Our modality-shared dual-router MoE module decomposes embeddings into task-specific representations: z^{disc} for discriminative tasks and z^{gen} for generative tasks. The discriminative branch fuses features via attention mechanisms and optimizes $\mathcal{L}_{retrieval} = \mathcal{L}_{ce} + \lambda_{triplet}\mathcal{L}_{triplet}$ for cross-modal retrieval. The generative branch maintains modality independence, reconstructing motions through the decoder while optimizing $\mathcal{L}_{generation} = \mathcal{L}_{align} + \lambda_{recon}\mathcal{L}_{recon}$ for text-to-motion generation.

execution that a sentence alone cannot fully convey (e.g., the difference between "walking" and "walking with a limp"). To exploit this evidence we introduce a temporally-aware video encoder, denoted E_{vi}, that consumes a frame sequence and emits a single 512-dimensional embedding aligned with the text and motion encoders. Concretely, E_{vi} aggregates 8 frames sampled uniformly across the video into one such embedding. This sampling strategy balances computational efficiency with temporal coverage, following standard practices in action recognition [35], [36], [37] where 8-16 frames are commonly used. Given frame sequences of a video $V = [I_1, \ldots, I_{l_{vi}}]$ where $I_i \in \mathbb{R}^{H\times W\times 3}$ is an RGB image frame and l_{vi} is the number of video frames, the video encoder produces:

$$e_{vi} = E_{vi}(V) \in \mathbb{R}^{512} \tag{1}$$

Following MotionCLIP [11], we use the CLIP [38] ViT-B/16 model as the per-frame visual backbone of E_{vi}. To obtain a multi-frame embedding rather than MotionCLIP's single-frame one, we adopt the temporal extension proposed by T2L [39]. T2L is an action-recognition model; its temporal transformer, pretrained on UCF-101 [40], aggregates eight uniformly sampled frames over the CLIP backbone into a single embedding. In short, the per-frame backbone is the standard CLIP ViT-B/16, while the multi-frame aggregation module is borrowed from T2L; this design preserves CLIP's strong vision-language alignment while recovering the kinematic dynamics that MotionCLIP's single-frame encoding fails to capture. Each frame I_i is first resized so that its shorter side is 256 pixels (preserving aspect ratio) and then center-cropped to 224×224 to match the ViT-B/16 input resolution, then normalized with the CLIP-standard mean [0.4815, 0.4578, 0.4082] and standard deviation [0.2686, 0.2613, 0.2758]. For videos with $l_{vi} > 8$ frames, we use centered uniform sampling:

$$\text{offset} = \frac{l_{vi}/8 - 1}{2} \tag{2}$$

$$\text{index}_i = \left\lfloor \frac{i \cdot l_{vi}}{8} + \text{offset} \right\rfloor \text{ for } i \in \{0, 1, \ldots, 7\} \tag{3}$$

For short sequences where $l_{vi} \leq 8$, we apply repeat padding:

$$\text{index}_i = \left\lfloor \frac{i \cdot l_{vi}}{8} \right\rfloor \text{ for } i \in \{0, 1, \ldots, 7\} \tag{4}$$

where index_i denotes the frame index selected from the original video sequence. This ensures consistent 8-frame input regardless of video length. The video encoder remains frozen during training to preserve pretrained representations. We apply no augmentation to the video frames themselves, as it would be redundant on both fronts: (i) E_{vi} is frozen and thus cannot be trained to be invariant to such perturbations, and (ii) visual diversity is already covered by

our rendering-time randomization of clothing textures, floor textures, and backgrounds. Full rendering details, including the texture and background randomization, are given in the ''Synthetic RGB Videos'' paragraph of Section IV.

Text Encoder. Natural language is the most accessible interface for users to specify or query motions, but as a symbolic signal a sentence cannot be directly compared with continuous video or motion features. Likewise, the role of the text encoder is to project a sentence into the same 512-dimensional space as the video and motion encoders. We instantiate this with E_{te}, the CLIP-based text encoder shipped with T2L [39], which inherits CLIP's vision-aligned language space [38]. Given a text description $T = [W_1, W_2, \ldots, W_{l_{te}}]$ where each W_i is a sub-word token produced by CLIP's byte-pair-encoding (BPE) tokenizer, the encoder processes the input through transformer layers to produce a text embedding:

$$e_{te} = E_{te}(T) \in \mathbb{R}^{512} \tag{5}$$

The resulting sequence length l_{te} is capped at 77 (CLIP's standard context window): longer descriptions are truncated and shorter ones are padded with the special end-of-text token. Like the video encoder E_{vi}, our text encoder E_{te} is kept frozen throughout training to preserve CLIP's pretrained vision-language alignment.

Motion Encoder and Decoder. To represent human motion, we adopt the SMPL parametric body model [41], [42], the de facto standard in 3D motion synthesis, in which a posed body is parameterized by a compact set of joint rotations. A motion sequence is then a variable-length stream of such per-frame pose vectors, written compactly as $P \in \mathbb{R}^{T\times 150}$. The 150 dimensions of each per-frame pose vector correspond to 6D rotations [43] for 24 body joints plus one root joint with translation (padded to 25×6). This raw representation cannot be directly compared with the fixed-size 512-dimensional text and video embeddings e_{te} and e_{vi} produced by our text and video encoders: P is a sequence whose length T differs from one motion to another, while e_{te} and e_{vi} are single 512-dimensional vectors of a fixed size. We therefore introduce two motion-side modules: a motion encoder E_{mo} that compresses an arbitrary-length raw SMPL pose sequence into a single 512-dimensional embedding of the same dimensionality as e_{vi} and e_{te}, and a motion decoder D_{mo} that maps a generative latent back to executable SMPL pose trajectories. D_{mo} is what makes generation possible by inverting the shared latent back into joint rotations. Following MotionCLIP [11], we employ a Transformer-based VAE architecture [19] for motion representation learning. The motion encoder E_{mo} learns continuous latent representations of human motion sequences. . The encoder consists of an 8-layer Transformer with 4 attention heads, hidden dimension of 512, and feed-forward dimension of 1024. The encoder maps P into a 512-dimensional latent vector:

$$e_{mo} = E_{mo}(P) \in \mathbb{R}^{512} \tag{6}$$

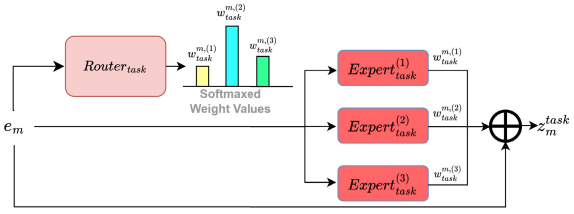

FIGURE 3. Architecture of a single router system in our dual-router MoE. The router computes softmax weights over three experts, which independently process input embedding e_m and combine via weighted sum to produce z_m^{task} where $task \in \{gen, disc\}$.

This latent captures spatial joint dependencies and temporal dynamics. The motion decoder D_{mo} reconstructs the motion sequence from the latent representation:

$$\hat{P} = D_{mo}(e_{mo}) \in \mathbb{R}^{T\times 150} \tag{7}$$

The decoder mirrors the encoder architecture (8)-layer Transformer decoder, 4 heads, hidden dimension 512) and reconstructs SMPL parameters $\hat{P} \in \mathbb{R}^{T\times 150}$ from the motion embedding. While the decoder is trained only with motion embeddings (e_{mo}), it is designed to operate in a shared embedding space where motion, text (e_{te}), and video (e_{vi}) embeddings are mutually compatible at inference time.

C. MOE-BASED DISENTANGLEMENT MODULE

Recall that the goal of disentanglement in our framework is to let one shared embedding space serve two conflicting objectives within a single model: retrieval and generation. Representation disentanglement has proven effective in multimodal sentiment analysis [33], [34], decomposing representations into task-specific components. However, when adapting such approaches to motion understanding directly, we observed that deep shared transformer architectures employed in [33] are prone to representation collapse [44], where embeddings converge to similar values regardless of input.

To address this issue, we employ a dual-router Mixture of Experts (MoE) architecture [13], [14] that maintains stable disentanglement without collapse. Our MoE module consists of two independent routers (for generation and retrieval tasks) and two groups of shallow expert networks that are shared across modalities. The architecture of a single router system is illustrated in Fig. 3. Each router is a linear layer mapping embeddings to expert weights: $\text{Router}_{task} : \mathbb{R}^{512} \to \mathbb{R}^3$.

Given an input embedding $e_m \in \mathbb{R}^{512}$ from modality $m \in$ {mo (motion), vi (video), te (text)}, each router computes routing weights over its expert set, where $task \in \{gen, disc\}$:

$$w^m_{task} = \text{softmax}(\text{Router}_{task}(e_m)). \tag{8}$$

The router maps each input embedding e_m to a soft distribution over the three experts; using a separate router

per task ($task \in \{gen, disc\}$) yields two distinct expert-weight patterns from the same input.

Each expert $\text{Expert}_{task}^{(k)}$ is a 2-layer MLP with ReLU activation and layer normalization, where $k \in \{1, 2, 3\}$. The routers independently select and combine experts:

$$z_m^{task} = \sum_{k=1}^{3} w_{task}^{m,(k)} \cdot \text{Expert}_{task}^{(k)}(e_m). \tag{9}$$

Combining the experts with these task-specific weights yields two distinct subspaces, z_m^{gen} and z_m^{disc}, that can be optimized independently without competing for the same parameters.

To prevent over-transformation, we apply residual connections with learnable weight α:

$$z_m^{task} \leftarrow \alpha e_m + (1 - \alpha) z_m^{task}. \tag{10}$$

The learnable residual lets the module fall back to identity when $\alpha \to 1$, providing a stable starting point for training.

The output is L2-normalized. The total MoE module contains 6 experts (3 per branch) with approximately 3.16M parameters. The residual weight α is initialized to 0.8 and learned during training. This shallow architecture with dual routing prevents representation collapse while enabling stable task-specific optimization.

D. DISCRIMINATIVE TASK BRANCH (MOTION RETRIEVAL)

Attention-based Feature Fusion. The retrieval branch processes z^{disc} representations through attention-based fusion, following the previous work in motion retrieval [12]. We compute three attention components:

$$f_{\text{self}} = \text{SelfAttn}(z_{mo}^{disc}, z_{mo}^{disc}, z_{mo}^{disc}), \tag{11}$$

$$f_{\text{text}} = \text{CrossAttn}(z_{mo}^{disc}, z_{te}^{disc}, z_{te}^{disc}), \tag{12}$$

$$f_{\text{video}} = \text{CrossAttn}(z_{mo}^{disc}, z_{vi}^{disc}, z_{vi}^{disc}), \tag{13}$$

where mo, te, vi denote motion, text, and video modalities. Intuitively, f_{self} retains the motion's own intrinsic information; f_{text} injects text-derived semantic context into the motion query; and f_{video} injects video-derived visual context into the motion query. The final fused representation combines these with modality weights:

$$f_{\text{fused}} = \frac{\lambda_{mo} f_{\text{self}} + \lambda_{te} f_{\text{text}} + \lambda_{vi} f_{\text{video}}}{\|\lambda_{mo} f_{\text{self}} + \lambda_{te} f_{\text{text}} + \lambda_{vi} f_{\text{video}}\|_2}. \tag{14}$$

The λ weights determine how much motion-self, text-derived, and video-derived information enter the fused representation; L2-normalization then yields a unit vector suitable for cosine-similarity comparison in the InfoNCE loss of Eq. (15).

To prevent over-reliance on auxiliary modalities (i.e., text and video) during inference, we employ curriculum learning [45] that gradually shifts from multi-modal fusion ($\lambda_{mo} = \lambda_{te} = \lambda_{vi} = 1.0$) to motion-dominant fusion by linearly decaying λ_{te} and λ_{vi} to 0.0 (further details in Section IV-B).

Optimization with Hard Negative Mining. The fused representations are optimized using InfoNCE-based contrastive loss [46]:

$$\mathcal{L}_{ce,m} = -\frac{1}{2B} \sum_{i=1}^{B} \Bigg(\log \frac{\exp(\cos(f_{\text{fused},i}, z_{m,i}^{disc})/\tau)}{\sum_j \exp(\cos(f_{\text{fused},i}, z_{m,j}^{disc})/\tau)} + \log \frac{\exp(\cos(z_{m,i}^{disc}, f_{\text{fused},i})/\tau)}{\sum_j \exp(\cos(z_{m,i}^{disc}, f_{\text{fused},j})/\tau)} \Bigg), \tag{15}$$

where τ is the temperature parameter and $m \in \{\text{vi}, \text{te}\}$. During training, we jointly optimize both video-motion and text-motion alignment by computing this loss for both video and text modalities. Here, B is the training mini-batch size, set to 20 in our implementation (see Section IV). Equation (15) is the InfoNCE loss for the discriminative branch: it pulls each positive cross-modal pair $(f_{\text{fused},i}, z_{m,i}^{disc})$ together while pushing all $B-1$ in-batch negatives apart in both directions (motion→auxiliary and auxiliary→motion). A larger B exposes each anchor to more negatives, making z^{disc} better at distinguishing similar samples. We additionally apply triplet margin loss [47] for hard negative mining:

$$\mathcal{L}_{\text{triplet}} = \max(0, \delta + \cos(a, n) - \cos(a, p)), \tag{16}$$

where a, p, n denote anchor, positive, and negative samples, with hard negatives $n = \arg\max_{i \neq j} \cos(a_j, p_i)$ selected within each batch, the margin parameter δ is set to 0.2. Equation (16) complements the soft InfoNCE signal of Eq. (15) with a hard-margin objective: rather than treating all negatives equally, it concentrates gradient on the single batch-hard distractor most likely to be confused with the anchor. This is particularly important on motion datasets such as HumanML3D [48], which contain many semantically similar samples (e.g., variations of "walking forward"), where soft contrast alone leaves these near-duplicates poorly separated in the embedding space. The total retrieval loss is:

$$\mathcal{L}_{retrieval} = \mathcal{L}_{ce,te} + \mathcal{L}_{ce,vi} + \lambda_{triplet} \mathcal{L}_{triplet}, \tag{17}$$

where we set $\lambda_{triplet} = 0.5$. Equation (17) sums two InfoNCE terms (one per auxiliary modality) with a single triplet term, weighted by $\lambda_{triplet}$=0.5. The choice $\lambda_{triplet} < 1$ keeps InfoNCE as the dominant learning signal while letting the triplet term sharpen decision boundaries for hard negatives.

E. GENERATIVE TASK BRANCH (TEXT-TO-MOTION GENERATION)

The generation branch processes z^{gen} representations without fusion to preserve modality independence. Fusing features would create dependency on auxiliary modalities during inference, preventing standalone generation. Our goal is to align the three modalities in a shared embedding space while maintaining their independence.

Contrastive Alignment. Following the same InfoNCE formulation as Eq. (15), we align z^{gen} representations

between motion and each auxiliary modality:

$$\mathcal{L}_{align,m} = -\frac{1}{2B}\sum_{i=1}^{B}\Bigg(\log\frac{\exp(\cos(z^{gen}_{mo,i}, z^{gen}_{m,i})/\tau)}{\sum_j \exp(\cos(z^{gen}_{mo,i}, z^{gen}_{m,j})/\tau)} + \log\frac{\exp(\cos(z^{gen}_{m,i}, z^{gen}_{mo,i})/\tau)}{\sum_j \exp(\cos(z^{gen}_{m,i}, z^{gen}_{mo,j})/\tau)}\Bigg) \tag{18}$$

where $m \in \{\text{vi}, \text{te}\}$. Equation (18) deliberately operates on the unfused z^{gen} representations rather than on the attention-fused features used in Eq. (15). This preserves modality independence: at inference, the generative branch must condition on text alone (without paired video or motion), so coupling modalities through fusion at training time would create a dependency that cannot be satisfied at deployment, preventing standalone text-to-motion generation. The total alignment loss is:

$$\mathcal{L}_{align} = \mathcal{L}_{align,vi} + \mathcal{L}_{align,te} \tag{19}$$

Equation (19) sums the text-motion and video-motion alignment terms with equal weight, treating both auxiliary modalities as equally important anchors for z^{gen}.

Reconstruction. The generative representations are trained to reconstruct motion sequences:

$$\mathcal{L}_{generation} = \mathcal{L}_{align} + \lambda_{recon}\mathcal{L}_{recon} \tag{20}$$

where $\lambda_{recon} = 100$ for scaling, and $\mathcal{L}_{recon} = \mathcal{L}^{rot}_{MSE} + \mathcal{L}^{pos}_{MSE} + \mathcal{L}^{vel}_{MSE}$ computes reconstruction errors in rotation space, 3D joint positions, and temporal velocities, respectively. The large weight λ_{recon}=100 in Eq. (20) compensates for a roughly 100-fold magnitude gap between the two loss families: the cosine-similarity-based alignment terms in Eq. (19) take values around 1, while the per-element MSE on 6D rotations and joint positions sits near 0.01.

F. TOTAL TRAINING OBJECTIVE

The total loss combines both discriminative and generative objectives:

$$\mathcal{L}_{total} = \lambda_{retrieval}\mathcal{L}_{retrieval} + \mathcal{L}_{generation} \tag{21}$$

where $\lambda_{retrieval}$ is set to 1.5. Equation (21) jointly optimizes the discriminative and generative branches that share the same modality encoders and MoE module. The weighting $\lambda_{retrieval}$=1.5 was chosen so that the smaller gradients from the retrieval loss (Eq. (17)) remain comparable to those of the reconstruction-dominated generation loss (Eq. (20)).

IV. EXPERIMENTS

A. DATASET AND IMPLEMENTATION DETAILS

Motion Datasets with Text Descriptions. We conduct experiments on HumanML3D [48], containing 14,616 motion sequences with text descriptions, and KIT-ML [49], containing 3,911 sequences. Following the standard split, HumanML3D and KIT-ML dataset are divided into 80% training, 5% validation, and 15% test sets. All motion sequences are represented using SMPL [41] pose parameters.

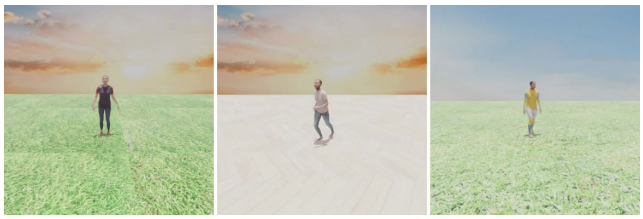

FIGURE 4. Representative samples of our synthetic video dataset. We render SMPL motions with diverse clothing textures (SMPLitex [54]) and randomized backgrounds to increase visual diversity.

For HumanML3D and KIT-ML, we extract SMPL parameters from SMPLH [42] format of the AMASS [50] dataset by converting to SMPL format (i.e., 24 body joints). For HumanAct12 [51], which provides only 3D joint positions, we fit SMPL parameters using SMPLify following the joints2smpl implementation [52].

Synthetic RGB Videos. To provide video modality, we render each motion sequence into human-centric videos at 384 × 384 resolution using SMPL-Blender-AddOn [53] at 20 FPS. To increase visual diversity, we randomize floor textures and background environments. We apply diverse clothing textures from the SMPLitex dataset [54], ensuring varied human appearances across sequences. Representative samples are shown in Fig. 4.

Implementation Details. We employ a two-stage training procedure. First, we pre-train the motion encoder for 200 epochs to align with frozen T2L [39] video and text encoders using contrastive learning. Second, we attach our dual-router MoE module and fine-tune the entire framework for 200 epochs using AdamW [55] optimizer with learning rate 8×10^{-5}, weight decay 0.01, and batch size 20. Additionally, we use pose keypoints extracted from videos [56], [57] as auxiliary data augmentation during training but not at inference. All experiments are conducted on NVIDIA GeForce RTX 4080 Super GPUs. The complete training procedure takes approximately 48 hours.

Evaluation Metrics. For retrieval, we report Recall at top-K (R@K, higher is better) and Median Rank (MedR, lower is better). For generation, we use Fréchet Inception Distance (FID) [17] measuring distributional distance between generated and real motion features (lower is better), and diversity computed as average pairwise Euclidean distance over 300 sampled pairs [17]. Following the standard evaluation protocol [48], we convert SMPL pose outputs to the 263-dimensional feature representation for FID computation.

From Method to Experiments. Each component of our framework introduced in Section III is exercised by a specific experiment in the rest of this section: the InfoNCE retrieval loss in Eq. (15) and its triplet companion in Eq. (16) are validated by the cross-modal retrieval results in Tables 2–5; the generative objective $\mathcal{L}_{generation}$ in Eq. (20) is validated by the FID/diversity numbers in Table 6; and the dual-router MoE design defined in Eq. (8)–(10) is validated by the architectural ablation in Table 7 and the cross-task validation in Table 8.

B. QUANTITATIVE RESULTS

We evaluate our method on four cross-modal retrieval tasks (text-to-motion, motion-to-text, video-to-motion, motion-to-video) and text-to-motion generation using HumanML3D and KIT-ML datasets. For retrieval, we report Recall at top-K (R@K, higher is better) and Median Rank (MedR, lower is better) under three retrieval protocols [25]: the **All** protocol evaluates the full test set without filtering; the **Dissimilar Subset** protocol evaluates a refined subset of 100 maximally-dissimilar pairs (a sanity check on discriminative ability); and the **Small Batches** protocol averages results over multiple random 32-pair batches [48].

Feature Fusion at Inference. As described in Section III-D, our discriminative branch employs curriculum learning during training with auxiliary modality weights defined as:

$$\lambda_{te} = \lambda_{vi} = \begin{cases} 1.0 & \text{if epoch} < 35 \\ 1.0 - \dfrac{\text{epoch} - 35}{70} & \text{if } 35 \leq \text{epoch} \leq 105 \\ 0.0 & \text{if epoch} > 105 \end{cases} \tag{22}$$

Early in training, the model relies on all three modalities (high $\lambda_{te}, \lambda_{vi}$) to bootstrap the motion embedding. The curriculum then linearly decays these auxiliary weights to zero so that, by the end of training, z_{mo}^{disc} alone carries the discriminative information—matching the inference setting where only the query modality is available. At inference, we report results under three configurations: $\lambda_{te} = \lambda_{vi} = 0$ (motion-only), $\lambda_{te} = \lambda_{vi} = 0.02$ (minimal auxiliary guidance), and $\lambda_{te} = \lambda_{vi} = 1$ (full fusion).

Important note on inference configurations: The $\lambda_{te} = \lambda_{vi} = 0$ setting represents the standard retrieval protocol where only the query modality and candidate modality are available. Configurations with $\lambda_{te} = \lambda_{vi} > 0$ assume access to paired auxiliary modalities at inference time and are reported as an *oracle analysis* to study the upper-bound potential of multi-modal fusion. Results with $\lambda_{te} = \lambda_{vi} > 0$ are marked with † in the tables and are not directly comparable to standard retrieval baselines.

Cross-Modal Retrieval Results. Tables 2 and 3 present comprehensive retrieval results on HumanML3D and KIT-ML datasets. Across both datasets, we observe consistent patterns validating our disentanglement approach: First, our z^{disc} representations achieve robust cross-modal alignment even without fusion ($\lambda_{te} = \lambda_{vi} = 0$). Compared to MotionCLIP's VAE-based approach, we demonstrate substantially stronger baseline performance across all retrieval tasks, validating that task-specific disentanglement (Eqs. (8)–(10)) is more effective for cross-modal alignment than asking a generative architecture (such as MotionCLIP's VAE) to also serve retrieval. Second, minimal auxiliary guidance ($\lambda_{te} = \lambda_{vi} = 0.02$, i.e., the curriculum-decayed weights of Eq. (22)) provides substantial gains, narrowing the gap with specialized retrieval methods like LAVIMO across multiple retrieval tasks. Third, full fusion ($\lambda_{te} = \lambda_{vi} = 1$) demonstrates the upper-bound potential of our feature fusion strategy (Eqs. (11)–(14)). We achieve strong performance across all retrieval tasks on both datasets, with particularly impressive results on video-to-motion and motion-to-video retrieval. The strong generalization from large (HumanML3D: 14,616 sequences) to small (KIT-ML: 3,911 sequences) datasets confirms the robustness and scalability of our approach across different dataset scales.

FIGURE 5. Text-to-motion generation results. Gallery showing 8 uniformly sampled frames for diverse action categories. Our disentangled z^{gen} representations generate semantically consistent motions with natural temporal transitions.

Beyond the **All** protocol summarized in Tables 2 and 3, we additionally report results under the **Dissimilar Subset** and **Small Batches** protocols [25] in Tables 4 and 5.

Text-to-Motion Generation Results. Table 6 presents generation quality comparison on both HumanML3D and KIT-ML datasets. We compare with MotionCLIP [11], which represents the baseline for trimodal embedding approaches. Our disentangled z^{gen} representations achieve lower FID across both datasets (HumanML3D: 5.164 vs. 6.74; KIT-ML: 6.15 vs. 6.97), validating that task-specific decomposition (Eqs. (8)–(10)), optimized under the $\mathcal{L}_{generation}$ objective of Eq. (20), preserves generation quality in the trimodal setting. Notably, our approach demonstrates substantially improved diversity on both datasets (HumanML3D: 0.936 vs. 0.162; KIT-ML: 0.570 vs. 0.062). We note that dedicated generation methods (e.g., MoMask [8]: 0.045, T2M-GPT [17]: 0.116, MDM [20]: 0.544 FID on HumanML3D) achieve substantially lower FID, as they are optimized solely for generation rather than jointly supporting retrieval.

C. QUALITATIVE RESULTS

Text-to-Motion Generation. Fig. 5 shows text-to-motion generation across diverse action categories. Our model successfully captures semantic distinctions between different motion types—combat actions (e.g., boxing, taekwondo, fencing) exhibit characteristic poses and movement dynamics distinct from locomotion patterns (e.g., running, climbing). The generated sequences demonstrate natural temporal progression, validating that our disentangled z^{gen} representations preserve generation quality. Fig. 6 further presents

TABLE 2. Cross-modal motion retrieval results on HumanML3D test set (All protocol). R@K (↑): Recall at top-K, higher is better. MedR (↓): Median rank, lower is better. Results marked with † assume access to paired auxiliary modalities at inference (oracle analysis).

Protocol	Methods	Relation between Text and Motion									
		text to motion (t2m)					motion to text (m2t)				
		R@1↑	R@3↑	R@5↑	R@10↑	MedR↓	R@1↑	R@3↑	R@5↑	R@10↑	MedR↓
All	TEMOS [7]	2.12	5.87	8.26	13.52	173	3.86	6.94	9.38	14.00	183.25
	MotionCLIP [11]	2.33	8.93	12.77	18.14	103	5.12	8.35	12.46	19.02	91.42
	TMR [25]	5.68	14.04	20.34	30.94	28	9.95	17.95	23.56	32.69	28.5
	LAVIMO [12]	6.37	15.60	21.95	33.67	24	9.72	18.73	25.00	36.55	22.5
	Yu et al. (4-modal) [58]	9.41	21.95	30.07	**43.83**	**14**	10.03	22.75	30.40	43.74	14
	Ours (λ_{te}=λ_{vi}=0)	2.42	5.93	8.74	15.33	79	2.67	6.80	10.45	17.59	74
	Ours (λ_{te}=λ_{vi}=0.02)†	4.74	10.10	14.03	22.40	47	5.13	11.50	16.42	25.05	43
	Ours (λ_{te}=λ_{vi}=1)†	**14.69**	**26.57**	**32.89**	42.63	16.5	**63.64**	**82.50**	**88.89**	**95.19**	**1**

Protocol	Methods	Relation between Video and Motion									
		video to motion (v2m)					motion to video (m2v)				
		R@1↑	R@3↑	R@5↑	R@10↑	MedR↓	R@1↑	R@3↑	R@5↑	R@10↑	MedR↓
All	MotionCLIP [11]	4.96	12.48	17.46	26.46	43	3.52	9.74	13.82	22.34	58
	MotionSet [26]	8.26	35.83	48.91	56.32	16	10.17	37.39	51.25	58.97	13
	LAVIMO [12]	21.25	50.51	63.40	77.50	3	25.76	54.71	66.87	79.73	3
	Yu et al. (4-modal) [58]	64.78	88.33	93.56	96.65	**1**	66.54	88.79	94.48	97.53	**1**
	Ours (λ_{te}=λ_{vi}=0)	9.40	20.46	28.33	40.08	19	8.58	20.35	28.15	40.58	18
	Ours (λ_{te}=λ_{vi}=0.02)†	20.55	38.12	46.65	57.98	7	17.06	37.29	47.65	60.72	6
	Ours (λ_{te}=λ_{vi}=1)†	**78.35**	**99.59**	**99.91**	**100.0**	**1**	**81.71**	**99.25**	**99.82**	**99.95**	**1**

TABLE 3. Cross-modal motion retrieval results on KIT-ML test set (All protocol). Evaluation protocols and configurations follow Table 2. Results marked with † assume access to paired auxiliary modalities at inference (oracle analysis).

Protocol	Methods	Relation between Text and Motion									
		text to motion (t2m)					motion to text (m2t)				
		R@1↑	R@3↑	R@5↑	R@10↑	MedR↓	R@1↑	R@3↑	R@5↑	R@10↑	MedR↓
All	TEMOS [7]	7.11	17.59	24.10	35.66	24	11.69	20.12	26.63	36.39	26.5
	MotionCLIP [11]	4.87	14.36	20.09	31.57	26	6.55	17.12	25.48	34.97	23
	TMR [25]	7.23	20.36	28.31	40.12	17	11.20	20.12	28.07	38.55	18
	LAVIMO [12]	10.16	24.61	34.57	49.80	11	15.43	26.95	34.57	53.32	10
	Yu (4-modal) [58]	13.44	34.88	44.14	60.22	7	16.08	36.24	46.87	58.04	6
	Ours (λ_{te}=λ_{vi}=0)	4.33	14.38	20.36	34.22	20	8.4	18.58	24.55	36.64	18.5
	Ours (λ_{te}=λ_{vi}=0.02)†	10.05	24.68	34.22	47.84	11	14.25	26.21	36.39	52.93	9.5
	Ours (λ_{te}=λ_{vi}=1)†	**40.97**	**76.46**	**84.73**	**93.13**	**2**	**72.39**	**90.84**	**96.44**	**99.36**	**1.5**

Protocol	Methods	Relation between Video and Motion									
		video to motion (v2m)					motion to video (m2v)				
		R@1↑	R@3↑	R@5↑	R@10↑	MedR↓	R@1↑	R@3↑	R@5↑	R@10↑	MedR↓
All	MotionCLIP [11]	3.71	8.79	11.52	17.97	52.5	4.49	9.96	12.11	21.68	51
	MotionSet [26]	9.15	35.27	46.18	61.31	13	11.08	35.03	45.29	60.53	13
	LAVIMO [12]	18.75	42.58	55.86	73.24	4	23.05	44.14	56.25	72.07	4
	Yu (4-modal) [58]	65.40	89.92	95.64	98.09	**1**	62.94	91.28	96.46	99.46	**1**
	Ours (λ_{te}=λ_{vi}=0)	9.54	24.94	35.24	48.85	11	8.91	24.94	34.73	49.87	11
	Ours (λ_{te}=λ_{vi}=0.02)†	27.48	54.71	63.36	76.72	3	24.05	51.53	64.5	79.26	3
	Ours (λ_{te}=λ_{vi}=1)†	**91.73**	**100**	**100**	**100**	**1**	**91.98**	**100**	**100**	**100**	**1**

generation results for more complex text prompts involving compositional actions and longer descriptions.

Motion Editing. Similar to MotionCLIP [11], our embedding-based approach enables compositional motion editing through latent arithmetic in z^{gen}: $z^{\text{gen}}_{\text{result}} = z^{\text{gen}}_{\text{base}} - z^{\text{gen}}_{\text{standing}} + z^{\text{gen}}_{\text{component}}$. Subtracting the neutral standing pose isolates motion-specific components, enabling meaningful composition as shown in Fig. 7: *walking* → *waving*, *walking* → *boxing*, and *running* → *throwing*. Additional motion editing examples are presented in Fig. 8.

D. ABLATION STUDIES

Model Selection Protocol. We select checkpoints based on validation set performance after epoch 105, when curriculum learning for retrieval task (i.e., decay of λ_{te}, λ_{vi}) becomes meaningful. For generation evaluation, we select the epoch with the lowest validation FID, following the standard protocol in motion generation literature [17], [48].

Effect of MoE Disentanglement. Table 7 demonstrates the critical role of MoE disentanglement. Without it (i.e., bypassing the dual-router routing of Eq. (8)–(10) and directly integrating trimodal encoders), the model suffers representation collapse (FID 9.85, diversity 0.030). Adding the MoE disentanglement defined in Eq. (8)–(10) recovers and improves performance (FID 5.16, diversity 0.936), demonstrating that MoE disentanglement is critical for stable trimodal learning.

Loss Function Analysis. Among MoE configurations, GRAM loss [59] alone achieves FID 7.34. Combining GRAM with contrastive loss (CE) yields FID 5.96 with higher diversity (1.150), but CE alone achieves better FID (5.16) with controlled diversity (0.936). We adopt CE-only training for our final model.

Cross-Task Validation. Table 8 shows cross-task evaluation results. z^{disc} achieves substantially better retrieval performance than z^{gen} (8.74% vs 3.10% R@5), confirming task-specific specialization. Conversely, z^{gen} demonstrates

TABLE 4. Cross-modal motion retrieval results on HumanML3D test set (Dissimilar Subset and Small Batches protocols). Results marked with † assume access to paired auxiliary modalities at inference (oracle analysis).

Protocol	Methods	Text to Motion (t2m)					Motion to Text (m2t)				
		R@1↑	R@3↑	R@5↑	R@10↑	MedR↓	R@1↑	R@3↑	R@5↑	R@10↑	MedR↓
Dissimilar Subset	TEMOS [7]	20.00	37.00	47.00	62.00	6	24.00	39.00	47.00	62.00	6.74
	MotionCLIP [11]	21.00	37.00	49.00	65.00	6	21.00	45.00	53.00	69.00	5.12
	TMR [25]	34.00	61.00	68.00	76.00	2	47.00	65.00	71.00	78.00	2.50
	LAVIMO [12]	**50.00**	**79.00**	86.00	90.00	**1**	56.00	81.00	86.00	92.00	**1**
	Ours (λ_{te}=λ_{vi}=0)	23.00	46.00	60.00	72.00	4	26.00	49.00	63.00	75.00	4
	Ours (λ_{te}=λ_{vi}=0.02)†	27.00	53.00	70.00	82.00	3	38.00	59.00	74.00	81.00	3
	Ours (λ_{te}=λ_{vi}=1)†	46.00	72.00	**87.00**	**95.00**	2	**81.00**	**99.00**	**99.00**	**100.0**	**1**
Small Batches	TEMOS [7]	40.49	61.14	70.96	84.15	2.33	39.96	61.79	72.40	85.89	2.33
	MotionCLIP [11]	46.24	68.93	80.47	91.35	1.88	44.76	65.22	77.83	90.19	2.03
	TMR [25]	67.16	86.81	91.43	95.36	1.04	67.97	86.35	91.70	95.27	1.03
	LAVIMO [12]	68.58	85.02	88.77	92.58	**1.01**	68.64	85.52	88.76	92.82	1.01
	Ours (λ_{te}=λ_{vi}=0)	51.28	78.22	86.84	95.38	1.48	52.22	78.00	86.78	95.47	1.38
	Ours (λ_{te}=λ_{vi}=0.02)†	60.50	85.00	92.09	97.56	1.12	61.69	84.91	91.78	97.50	1.08
	Ours (λ_{te}=λ_{vi}=1)†	**72.25**	**89.78**	**94.91**	**98.97**	**1.01**	**98.47**	**99.88**	**100.0**	**100.0**	**1**

Protocol	Methods	Video to Motion (v2m)					Motion to Video (m2v)				
		R@1↑	R@3↑	R@5↑	R@10↑	MedR↓	R@1↑	R@3↑	R@5↑	R@10↑	MedR↓
Dissimilar Subset	MotionCLIP [11]	42.00	72.00	80.00	88.00	2	43.00	69.00	76.00	86.00	2
	MotionSet [26]	58.00	81.00	89.00	97.00	2	59.00	86.00	90.00	98.00	2
	LAVIMO [12]	74.00	98.00	**100.0**	**100.0**	**1**	79.00	**100.0**	**100.0**	**100.0**	**1**
	Ours (λ_{te}=λ_{vi}=0)	46.00	73.00	80.00	89.00	2	45.00	77.00	83.00	89.00	2
	Ours (λ_{te}=λ_{vi}=0.02)†	59.00	84.00	92.00	95.00	1	61.00	86.00	94.00	97.00	1
	Ours (λ_{te}=λ_{vi}=1)†	**94.00**	**100.0**	**100.0**	**100.0**	**1**	**94.00**	**100.0**	**100.0**	**100.0**	**1**

TABLE 5. Cross-modal motion retrieval results on KIT-ML test set (Dissimilar Subset and Small Batches protocols). Results marked with † assume access to paired auxiliary modalities at inference (oracle analysis).

Protocol	Methods	Text to Motion (t2m)					Motion to Text (m2t)				
		R@1↑	R@3↑	R@5↑	R@10↑	MedR↓	R@1↑	R@3↑	R@5↑	R@10↑	MedR↓
Dissimilar Subset	TEMOS [7]	24.00	46.00	54.00	70.00	5	33.00	45.00	49.00	64.00	6.50
	MotionCLIP [11]	19.00	41.00	50.00	69.00	6	28.00	43.00	48.00	65.00	7
	TMR [25]	26.00	60.00	70.00	83.00	3	34.00	60.00	69.00	82.00	3.50
	LAVIMO [12]	30.00	63.00	73.00	84.00	3	48.00	66.00	76.00	82.00	2
	Ours (λ_{te}=λ_{vi}=0)	15.00	33.00	46.00	69.00	6	17.00	35.00	51.00	71.00	5.50
	Ours (λ_{te}=λ_{vi}=0.02)†	21.00	45.00	60.00	82.00	4	25.00	55.00	63.00	81.00	3.50
	Ours (λ_{te}=λ_{vi}=1)†	**49.00**	**81.00**	**89.00**	**95.00**	**2**	**80.00**	**94.00**	**100.0**	**100.0**	**1.50**
Small Batches	TEMOS [7]	43.88	67.00	74.00	84.75	2.06	41.88	65.62	75.25	85.75	2.25
	MotionCLIP [11]	41.29	69.50	78.83	90.12	1.73	39.55	68.13	77.94	90.85	2.16
	TMR [25]	49.25	78.25	87.88	95.00	1.50	50.12	76.88	88.88	94.75	1.53
	LAVIMO [12]	58.10	86.34	93.08	96.47	1.08	60.23	86.44	93.22	95.87	1.20
	Ours (λ_{te}=λ_{vi}=0)	47.12	79.03	89.09	95.56	1.70	50.88	80.59	89.81	96.28	1.50
	Ours (λ_{te}=λ_{vi}=0.02)†	59.72	88.03	94.53	97.94	1.13	63.84	89.28	95.03	98.31	1.09
	Ours (λ_{te}=λ_{vi}=1)†	**92.66**	**99.66**	**100.0**	**100.0**	**1**	**97.81**	**99.97**	**100.0**	**100.0**	**1**

Protocol	Methods	Video to Motion (v2m)					Motion to Video (m2v)				
		R@1↑	R@3↑	R@5↑	R@10↑	MedR↓	R@1↑	R@3↑	R@5↑	R@10↑	MedR↓
Dissimilar Subset	MotionCLIP [11]	15.00	31.00	40.00	59.00	8	13.00	25.00	41.00	63.00	6
	MotionSet [26]	40.00	73.00	85.00	89.00	2	40.00	73.00	82.00	87.00	2
	LAVIMO [12]	66.00	90.00	96.00	99.00	**1**	66.00	93.00	98.00	**100.0**	**1**
	Ours (λ_{te}=λ_{vi}=0)	19.00	48.00	67.00	79.00	4	17.00	47.00	61.00	77.00	4
	Ours (λ_{te}=λ_{vi}=0.02)†	37.00	77.00	84.00	96.00	2	32.00	69.00	85.00	98.00	2
	Ours (λ_{te}=λ_{vi}=1)†	**97.00**	**100.0**	**100.0**	**100.0**	**1**	**89.00**	**100.0**	**100.0**	**100.0**	**1**

TABLE 6. Text-to-motion generation results. Comparison with trimodal baseline (MotionCLIP) on HumanML3D and KIT-ML test sets.

Method	HumanML3D		KIT-ML	
	FID↓	Div↑	FID↓	Div↑
MotionCLIP [11]	6.74	0.162	6.97	0.062
Ours (z^{gen})	**5.164**	**0.936**	**6.15**	**0.570**

higher generation quality (FID 5.16 vs 5.63), confirming that each representation specializes in its designated task. This empirically validates the dual-router design of Eq. (8): assigning a separate router per task is what produces two genuinely task-specific subspaces, rather than a single shared one.

Effect of Number of Experts. Table 9 compares different numbers of experts in our DoubleRouterMoE architecture. The 6-expert configuration (3 per branch) achieves the best FID of 5.16. The 8-expert variant shows comparable performance (FID 5.27) but with increased parameter count, while 4 experts significantly underperform (FID 5.72), suggesting insufficient capacity for effective task-specific disentanglement.

E. INFERENCE COST COMPARISON

We measured per-sample inference time on the text-to-motion generation path on a single NVIDIA RTX 4080 SUPER GPU (16 GB; PyTorch 2.4.1, CUDA 12.4). Each model was timed over 30 forward passes after

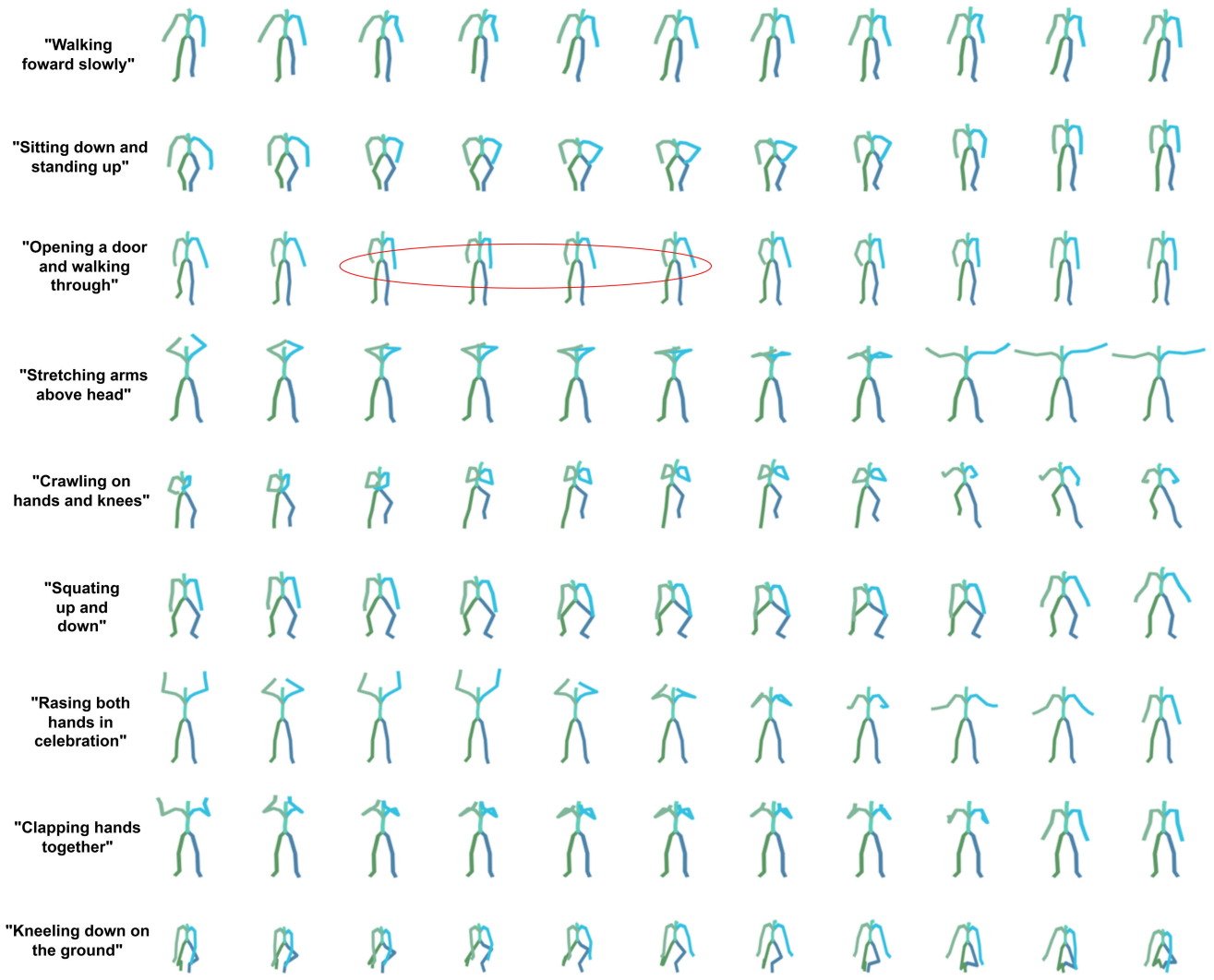

FIGURE 6. Generation results for complex text prompts. Key frames from generated motion sequences for compositional and descriptive text prompts, demonstrating the model's ability to handle diverse action categories including transitional movements, interactive actions, and expressive gestures. Red circles highlight subtle interactive actions (e.g., door opening) for clarity.

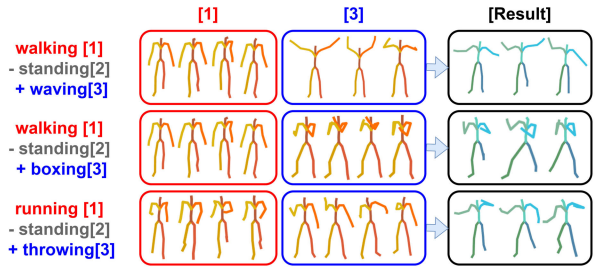

FIGURE 7. Motion editing via latent arithmetic. Combining text latents enables intuitive motion composition. For example, walking − standing + boxing yields a walking-boxing motion, demonstrating dynamic behavior emerging from simple latent operations.

5 warmup passes, in evaluation mode with gradients disabled and GPU timing synchronized at each measurement boundary. Both pipelines cover text encoding → motion decoder → 3D-joint conversion; TRIAD additionally passes through the dual-router MoE module. GIF rendering and disk I/O are excluded.

Our dual-router MoE module adds only ≈ 0.81 ms ($\approx 14.7\%$) of per-sample inference overhead over the MotionCLIP baseline (TRIAD: 6.30 ± 0.14 ms vs. MotionCLIP: 5.49 ± 0.16 ms), confirming that the disentanglement module imposes a negligible inference cost. The two models are directly comparable because they share the same VAE-based Transformer motion decoder; the overhead is attributable to the additional dual-router MoE module (≈ 3.16 M parameters, with the three experts of each branch evaluated in parallel).

This comparison is limited to methods with publicly released implementations; other trimodal frameworks (e.g., [12]) are therefore excluded.

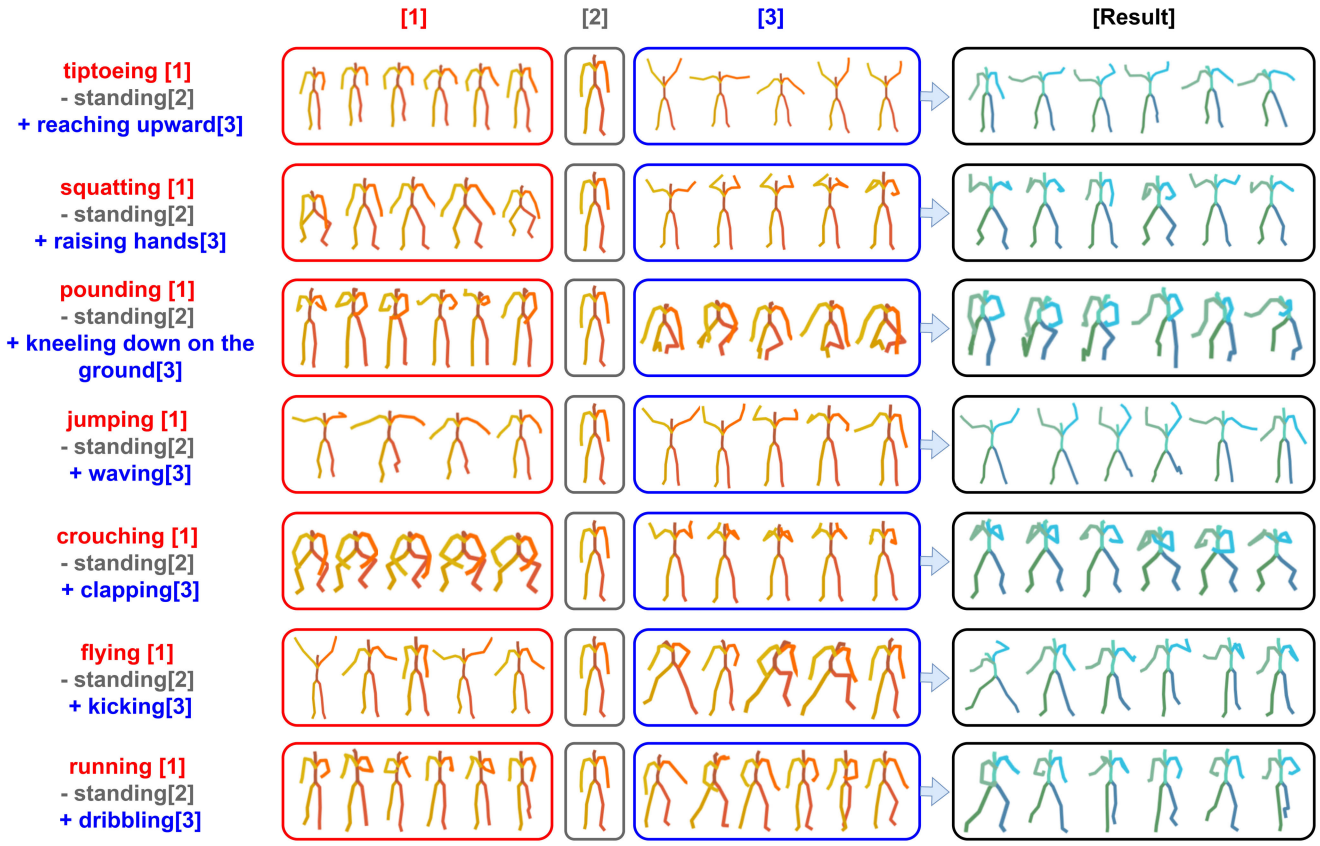

FIGURE 8. Additional motion editing examples. Various motion combinations achieved through latent arithmetic: $z_{\text{motion1}} - z_{\text{standing}} + z_{\text{motion2}}$, demonstrating that the learned representation preserves motion semantics under arithmetic operations.

TABLE 7. Ablation studies on HumanML3D test set. We evaluate the contribution of architectural components and loss functions.

Configuration	FID↓	Diversity ↑
Architecture Ablation		
w/o MoE (Baseline)	9.85	0.030
Loss Function Ablation		
MoE + GRAM only	7.34	0.991
MoE + GRAM + CE	5.96	1.150
MoE + CE only (**ours**)	**5.16**	0.936

TABLE 8. Cross-task representation validation. We verify task-specific specialization by cross-task evaluation with $\lambda_{te} = \lambda_{vi} = 0$ to isolate representation effects without fusion.

Representation	Generation (FID↓)	t2m Retrieval (R@5↑)
z^{gen}	**5.16**	3.10
z^{disc}	5.63	**8.74**

TABLE 9. Ablation on number of experts in DoubleRouterMoE (HumanML3D).

Total Experts	Per Branch	FID ↓	MoE Params
4	2 gen + 2 disc	5.72	2.11M
6 (ours)	**3 gen + 3 disc**	**5.16**	**3.16M**
8	4 gen + 4 disc	5.27	4.21M

V. CONCLUSION AND FUTURE WORKS

We introduced the first trimodal motion understanding system that simultaneously supports generative and discriminative capabilities through explicit representation disentanglement. Our core insight is that generation benefits from modality-independent representations, whereas retrieval requires tight cross-modal coupling. To reconcile these fundamentally different objectives, we decompose motion embeddings into task-specific components (z^{gen} and z^{disc}) and route them via a shallow dual-router Mixture-of-Experts architecture, effectively mitigating the trade-off that has constrained prior work.

Our framework outperforms MotionCLIP—the only existing trimodal text-video-motion baseline—in both generation and retrieval, and the learned embedding space further enables downstream applications such as motion editing through latent arithmetic. Current limitations include reliance on synthetic video data, which may limit generalization to real-world scenarios, and a compact 512-dimensional embedding that may constrain fine-grained representational capacity.

While closing the gap with task-dedicated methods remains an open challenge, we believe this dual-task paradigm holds significant potential for practical motion understanding systems. Future work will explore extending the disentanglement strategy to additional motion understanding tasks such as action recognition, where task-specific subspaces could separate temporal pattern features from fine-grained pose features. We also plan to investigate discrete or hybrid motion representations—such as VQ-VAE-based tokenization, which has shown strong results

in dedicated generation methods [8], [17]—to bridge the generation quality gap while preserving the unified framework's retrieval capability. More broadly, realizing the full potential of unified generative-discriminative frameworks will require exploring diverse architectural strategies beyond MoE-based decomposition—including hierarchical embedding spaces and integration with real RGB video.

ACKNOWLEDGMENT

The authors used Claude AI (Anthropic) to assist with LaTeX formatting and template conversion for this manuscript.

REFERENCES

[1] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa, "AMP: Adversarial motion priors for stylized physics-based character control," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–20, Aug. 2021.

[2] X. B. Peng, Y. Guo, L. Halper, S. Levine, and S. Fidler, "ASE: Large-scale reusable adversarial skill embeddings for physically simulated characters," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 1–17, Jul. 2022, doi: 10.1145/3528223.3530110.

[3] A. Serifi, R. Grandia, E. Knoop, M. Gross, and M. Bächer, "Robot motion diffusion model: Motion generation for robotic characters," in *Proc. SIGGRAPH Asia Conf. Papers*, Dec. 2024, pp. 1–9, doi: 10.1145/3680528.3687626.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[5] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent.*, 2014.

[6] J. Ho, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2024, pp. 6840–6851.

[7] M. Petrovich, M. J. Black, and G. Varol, "TEMOS: Generating diverse human motions from textual descriptions," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 13682, 2022, pp. 480–497.

[8] C. Guo, Y. Mu, M. G. Javed, S. Wang, and L. Cheng, "MoMask: Generative masked modeling of 3D human motions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 1900–1910.

[9] M. Kocabas, N. Athanasiou, and M. J. Black, "VIBE: Video inference for human body pose and shape estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5252–5262.

[10] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, "MotionBERT: A unified perspective on learning human motion representations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 15039–15053.

[11] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or, "Motionclip: Exposing human motion generation to clip space," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2022, pp. 358–374.

[12] K. Yin, S. Zou, Y. Ge, and Z. Tian, "Tri-modal motion retrieval by learning a joint embedding space," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 1596–1605.

[13] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," in *Proc. Int. Conf. Neural Netw.*, vol. 2, 1993, pp. 1339–1344.

[14] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. V. Le, G. E. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[15] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," 2022, *arxiv:2205.01917*.

[16] H. Luo, W. Xi, and D. Tang, "MLUG: Bootstrapping language-motion pre-training for unified motion-language understanding and generation," *Sensors*, vol. 24, no. 22, p. 7354, Nov. 2024.

[17] J. Zhang, Y. Zhang, X. Cun, S. Huang, Y. Zhang, H. Zhao, H. Lu, and X. Shen, "T2M-GPT: Generating human motion from textual descriptions with discrete representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023.

[18] C. Ahuja and L.-P. Morency, "Language2Pose: Natural language grounded pose forecasting," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2019, pp. 719–728.

[19] M. Petrovich, M. J. Black, and G. Varol, "Action-conditioned 3D human motion synthesis with transformer VAE," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10965–10975.

[20] G. Tevet, S. Raab, B. A. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano, "Human motion diffusion model," in *Proc. 11th Int. Conf. Learn. Represent.*, 2023. [Online]. Available: https://openreview.net/forum?id=SJ1kSyO2jwu

[21] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu, "MotionDiffuse: Text-driven human motion generation with diffusion model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 6, pp. 4115–4128, Jun. 2024.

[22] C. Guo, Z. Xinxin, S. Wang, and C. Li, "TM2T: Stochastic and tokenized modeling for the reciprocal generation of 3D human motions and texts," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 13695, 2022, pp. 580–597.

[23] L. Kovar, M. Gleicher, and F. Pighin, "Motion graphs," in *Proc. ACM SIGGRAPH*, 2008, pp. 1–10.

[24] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 8–13.

[25] M. Petrovich, M. J. Black, and G. Varol, "TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 9454–9463.

[26] T. Ren, W. Li, Z. Jiang, X. Li, Y. Huang, and J. Peng, "Video-based human motion capture data retrieval via MotionSet network," *IEEE Access*, vol. 8, pp. 186212–186221, 2020.

[27] L. Zhong, C. Guo, Y. Xie, J. Wang, and C. Li, "Sketch2Anim: Towards transferring sketch storyboards into 3D animation," *ACM Trans. Graph.*, vol. 44, no. 4, pp. 1–15, Aug. 2025.

[28] K. A. Kinfu and R. Vidal, "MotionBind: Multi-modal human motion alignment for retrieval, recognition, and generation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 38, 2025, pp. 122518–122548.

[29] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "Beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[30] H. Kim and A. Mnih, "Disentangling by factorising," in *Proc. Int. Conf. Mach. Learn.*, vol. 80, 2018, pp. 2649–2658.

[31] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and -specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1122–1131.

[32] Y. Zhang, M. Chen, J. Shen, and C. Wang, "Tailor versatile multi-modal learning for multi-label emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, vol. 36, no. 8, pp. 9100–9108.

[33] Y. Zhou, X. Liang, H. Chen, Y. Zhao, X. Chen, and L. Yu, "Triple disentangled representation learning for multimodal affective analysis," 2024, *arXiv:2401.16119*.

[34] P. Wang, Q. Zhou, Y. Wu, T. Chen, and J. Hu, "DLF: Disentangled-language-focused multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2025, vol. 39, no. 20, pp. 21180–21188.

[35] S. T. Wasim, M. Naseer, S. Khan, F. S. Khan, and M. Shah, "Vita-CLIP: Video and text adaptive CLIP via multimodal prompting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23034–23044.

[36] M. Kim, D. Han, T. Kim, and B. Han, "Leveraging temporal contextualization for video action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 15079, 2024, pp. 74–91.

[37] S. Ahmad, S. Chanda, and Y. S. Rawat, "EZ-CLIP: Efficient zeroshot video action recognition," 2023, *arXiv:2312.08010*.

[38] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[39] S. Ahmad, S. Chanda, and Y. S. Rawat, "T2l: Efficient zero-shot action recognition with temporal token learning," in *Proc. Trans. Mach. Learn. Res.*, 2025.

[40] K. Soomro, A. Roshan Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[41] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics*, vol. 34, no. 6, pp. 851–866, Oct. 2023.

[42] J. Romero, D. Tzionas, and M. J. Black, "SMPL+H: Modeling hands and bodies together," in *Proc. SIGGRAPH Asia Tech. Briefs*, vol. 36, no. 6, Nov. 2017, pp. 1–17.
[43] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5738–5746.
[44] V. Papyan, X. Y. Han, and D. L. Donoho, "Prevalence of neural collapse during the terminal phase of deep learning training," 2020, *arXiv:2008.08186*.
[45] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 41–48, doi: 10.1145/1553374.1553380.
[46] A. V. D. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018.
[47] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.
[48] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating diverse and natural 3D human motions from text," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5142–5151.
[49] M. Plappert, C. Mandery, and T. Asfour, "The KIT motion-language dataset," 2016, *arXiv:1607.03827*.
[50] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "AMASS: Archive of motion capture as surface shapes," in *Proc. Int. Conf. Comput. Vis.*, Oct. 2019, pp. 5441–5450.
[51] C. Guo, X. Zuo, S. Wang, S. Zou, Q. Sun, A. Deng, M. Gong, and L. Cheng, "Action2Motion: Conditioned generation of 3D human motions," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2021–2029.
[52] X. Zuo, S. Wang, J. Zheng, W. Yu, M. Gong, R. Yang, and L. Cheng, "SparseFusion: Dynamic human avatar modeling from sparse RGBD images," *IEEE Trans. Multimedia*, vol. 23, pp. 1617–1629, 2021.
[53] Meshcapade. (2023). *Smpl Blender Add-on*. [Online]. Available: https://github.com/Meshcapade/SMPLblenderaddon
[54] D. Casas and M. C. Trinidad, "SMPLitex: A generative model and dataset for 3D human texture estimation from single image," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2023, Paper 272.
[55] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2017.
[56] G. Jocher, A. Chaurasia, and J. Qiu. (2023). *Ultralytics YOLOv8*. [Online]. Available: https://github.com/ultralytics/ultralytics
[57] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. Thirty-second AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 7444–7452.
[58] S. Yu, Z.-A. Wang, K. Yin, Z. Tian, M. Zhang, W. Si, and S. Zou, "Multi-modal motion retrieval by learning a fine-grained joint embedding space," 2025, *arXiv:2507.23188*.
[59] G. Cicchetti, E. Grassucci, L. Sigillo, and D. Comminiello, "Gramian multimodal representation learning and alignment," in *Proc. 13th Int. Conf. Learn. Represent.*, 2025.

MYUNGIN KIM received the B.S. degree in mathematics from the University of California, Los Angeles, and the M.S. degree in statistics from The University of Chicago. He is currently pursuing the Ph.D. degree in artificial intelligence with Ajou University, South Korea. He has more than eight years of industry experience as a Machine Learning Engineer in computer vision, since 2017. His current research interests include computer graphics and character animation, with a particular focus on text-to-motion generation and motion synthesis.

RI YU is currently an Assistant Professor with the Department of Software and Computer Engineering, Ajou University. Her research interests include human motion reconstruction, physics-based simulation, humanoid modeling and control, human motion analysis, and deep learning.