(DRAFT)
Kinetic Mining in Context: Few-Shot Action Synthesis via Text-to-Motion Distillation

1st Luca Cazzola
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Alboody Ahed
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—The high cost of acquiring large-scale annotated motion datasets fundamentally limits skeletal-based Human Activity Recognition (HAR). While text-to-motion generative models offer a promising solution, their development has focused on artistic workflows like avatar animation. Consequently, they are trained on vast datasets that, despite immense variety, lack the kinematic specificity required for HAR tasks, creating a significant domain gap. To bridge this gap, we propose KineMIC (Kinetic Mining In Context), a transfer learning framework for few-shot action synthesis. Our approach is built on the hypothesis that semantic similarity in the text encoding space—comparing sparse action labels to rich captions—can provide "soft" guidance towards kinematic similarity in the motion domain. To operationalize this, we introduce a kinetic mining strategy that uses CLIP text embeddings to create "soft pairings" between target actions and rich source descriptions. These pairings guide a training process that identifies relevant sub-sequences from the large-scale dataset, transforming the generalist text-to-motion model into a specialized, few-shot action-to-motion generator. Ultimately, our framework generates synthetic motion from as few as 10 samples per class, providing a robust data augmentation strategy that significantly enhances the performance of downstream HAR models. Additional material available at https://lucazzola.github.io/publications/kinemic.**

***Index Terms*—Human Activity Recognition, Data Augmentation, Generative Models, Diffusion Models, Knowledge Transfer, Few-Shot Learning, Motion Synthesis.**

I. Introduction

Human Activity Recognition (HAR) has become a cornerstone technology in a multitude of fields, including sports performance analysis, human-robot collaboration, and intelligent surveillance. While methods using RGB video have seen significant progress, largely driven by advancements in image processing and their ability to capture rich contextual information that skeletal data inherently lacks, skeletal-based HAR remains a fundamental modality. It is frequently employed both in multimodal systems and as a standalone "motion format" due to its distinct advantages: its lightweight representation is not only more robust to environmental variations but also inherently privacy-preserving. However, the performance of deep learning models for this task is fundamentally tied to the availability of large-scale, accurately annotated datasets. The acquisition and annotation of such data are notoriously expensive and labor-intensive, creating a significant bottleneck that hampers progress.

Recent advancements in generative modeling, particularly text-to-motion synthesis, present a promising avenue for mitigating this data scarcity. In this work, we focus on adapting a pre-trained model from a large-scale text-to-motion "prior" dataset, HumanML3D [13], to a "target" HAR dataset, NTU RGB+D [29]. However, this adaptation is non-trivial due to a significant domain gap, which can be characterized along two primary axes: semantic and statistical. The first is a semantic discrepancy in the conditioning inputs. HumanML3D is conditioned on free-form, descriptive sentences (e.g., "a person does martial arts") that can encompass multiple distinct actions within a single long sequence. In contrast, NTU RGB+D utilizes discrete, one-hot encoded action labels (e.g., "kicking"), creating a disparity in both semantic richness and granularity. Second, a stark statistical discrepancy arises from different data acquisition methods. HumanML3D, derived from high-fidelity Motion Capture (MoCap), contains broad, fluid motions with smooth trajectories, often exceeding 150 frames. Conversely, NTU RGB+D, captured with Kinect depth sensors, consists of shorter, more atomic motions (averaging 40-50 frames) that are inherently noisier. This durational and qualitative mismatch means the pre-trained model is ill-equipped to generate the concise motions required for HAR.

To address these challenges, we propose KineMIC (Kinetic Mining In Context), a transfer learning framework engineered to bridge this domain gap. Our method employs a dual-stream, teacher-student architecture wherein a frozen, pre-trained text-to-motion model acts as a "teacher," guiding the fine-tuning of a "student" model for the target HAR domain. The core of our approach is a novel soft positive mining strategy that leverages CLIP text embeddings to establish a semantic correspondence between the target action labels and the rich textual descriptions from the source domain. This process identifies relevant motion sub-sequences within the vast source dataset, effectively transforming the general-purpose model into a specialized action-to-motion generator

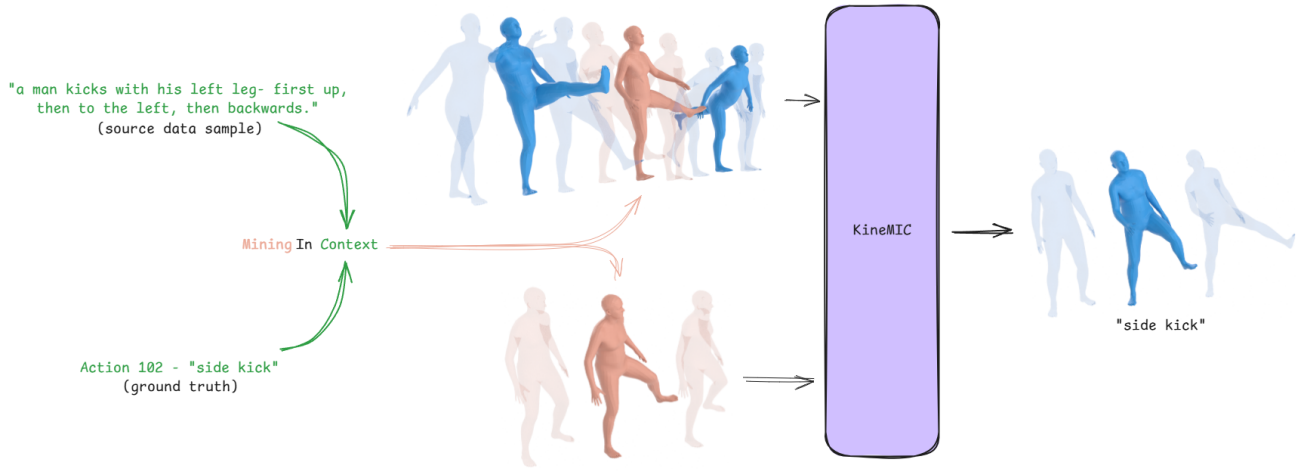

Fig. 1: **Kinetic Mining in Context**. The target action sample (bottom) is used to contextualize the search within the large source data sample (top). The mining operation identifies the semantically relevant segment (in orange) from the source, thereby providing semantically correlated "soft pairs" samples for the KineMIC few-shot synthesis pipeline.

capable of producing synthetic data tailored for HAR.

The main contributions of this work are as follows:

- To the best of the authors' knowledge, we are the first to tackle the challenge of adapting a large-scale Text-to-Motion model into an Action-to-Motion generator for HAR applications, moreover, addressing this within a few-shot setting.
- We introduce KineMIC, a novel teacher-student transfer learning framework that adapts a pre-trained model to a specific target domain with minimal data, proving the effectiveness of our approach in boosting downstream HAR accuracy.

II. Related Works

A. Skeletal-based Human Activity Recognition

Recognizing human activity from skeletal data is a pivotal research area in computer vision, evolving from early methods using RNNs and LSTMs to model motion as a time series [7], [29]. A paradigm shift occurred when the skeleton was re-conceptualized as a graph, enabling Graph Convolutional Networks (GCNs) to model spatio-temporal dependencies in a unified framework [33]. This approach was rapidly advanced by methods introducing adaptive, data-driven graph structures [30] and more sophisticated GCN architectures [2], [3], [8], [19]. More recently, 3D convolutional networks [9] and Transformers [24] have set new performance benchmarks, leading to promising hybrid architectures. These models merge the two paradigms to effectively capture both short-range skeletal dependencies and long-range temporal context [6].

B. Generative Models for Skeleton-based Motion Synthesis

The synthesis of realistic human motion has progressed in parallel with recognition. Early deep learning successes in this area were driven by Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), which proved effective at generating motions conditioned on discrete action classes [5], [14], [22]. A significant leap in quality was enabled by two key factors: the curation of larger, more diverse Motion Capture (MoCap) datasets [13], [21] and the introduction of richer conditioning signals like free-form text [23]. This shift paved the way for Denoising Diffusion Probabilistic Models (DDPMs), which learn to reverse a gradual noising process to generate high-fidelity and diverse human motions [31], [35]. The current state-of-the-art is largely characterized by a competition between these diffusion models and masked generative models [12], with recent research pushing the boundaries of conditional synthesis even further by incorporating modalities like music and images [34].

C. Motion Diffusion Model (MDM)

The Motion Diffusion Model (**MDM**) [31] is a generative architecture based on denoising diffusion probabilistic models (DDPMs), designed for high-quality skeleton-based motion synthesis. Its operation is governed by a Markov chain, which models a two-phase process: a forward noising process and a backward denoising process. Given a sample of skeletal motion $x = \{x^1, x^2, \ldots, x^n\}$ composed of n frames, the forward process, denoted as $q(x_t|x_{t-1})$, progressively adds Gaussian noise to the clean signal x_0, following a defined variance schedule α_t, to produce a noisy sequence x_t. This process is defined as:

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)I) \tag{1}$$

Unlike traditional DDPMs that predict the noise component [15], the core innovation lies in training the network to directly predict the original signal [26]. The model's main objective is to minimize a simple reconstruction loss L_{simple} between the ground truth motion x_0 and the network's prediction $\hat{x}_0 = G(x_t, t, c)$ where x_t denotes the noisy input at timestep t and c is a conditioning signal.

$$L_{\text{simple}} = \mathbb{E}_{x_0 \sim q(x_0|c), t \sim [1,T]}[||x_0 - \hat{x}_0||_2^2] \quad (2)$$

The most successful implementations of MDM are based on a vanilla transformer encoder architecture [32], in which each token in input represents a single frame in the motion sequence. The conditioning signal c and the noise timestep t are projected to the transformer's dimension using separate feed-forward networks and then summed to create a token z_{tk}, which is processed by the transformer encoder blocks along with the motion tokens from x_t. The model is trained with classifier-free guidance [16] where the conditioning c can be text, motion or nothing, enabling to perform both conditional and unconditional generation. To ensure the physical plausibility and quality of the generated motions, MDM incorporates three additional geometric losses: a position loss L_{pos}, a velocity loss L_{vel}, and a foot-ground contact loss L_{foot}. These losses enforce constraints on key kinematic properties, preventing artifacts such as unnatural joint movements and foot sliding. The ensembled training objective is a weighted sum of all loss components:

$$L = L_{\text{simple}} + \lambda_{\text{pos}} L_{\text{pos}} + \lambda_{\text{vel}} L_{\text{vel}} + \lambda_{\text{foot}} L_{\text{foot}} \quad (3)$$

This comprehensive loss function allows MDM to generate motions that are not only semantically consistent with the conditioning input but also physically realistic, making it a stable, robust backbone for our work.

D. Synthetic data generation for HAR

To mitigate the data scarcity inherent in HAR, the generation of synthetic data has emerged as a relevant research direction. Generative Adversarial Networks (GANs) have been a primary tool for this task, successfully applied to augment datasets across different modalities, from wearable sensor signals [20] to dynamic skeleton sequences [**?**]. However, the challenge intensifies in **few-shot learning** scenarios, where standard generative models tend to overfit. A state-of-the-art approach by [11] directly tackles this by proposing a cross-domain adversarial learning framework. Their method transfers motion diversity from a large-scale source dataset to a few-shot target by introducing novel cross-domain and entropy regularization losses, enabling the synthesis of varied and realistic actions from limited samples. Broadening the application of generative models beyond direct data synthesis, recent work has also explored their use for creating semantic guidance. For instance, [**?**] leverages Large Language Models (LLMs) to generate rich, descriptive text prompts for different actions, using this textual information to supervise and improve the representation learning of skeleton-based recognition models.

III. Problem Formulation

The central challenge we address is the adaptation of pre-trained text-to-motion generative models for data augmentation in the context of few-shot Human Action Recognition (HAR). Our goal is to leverage the rich kinematic knowledge embedded in a large source dataset to generate high-fidelity, class-specific motions for a target domain. Let a skeletal motion sequence be denoted as $x = \{x(j) \in \mathbb{R}^d\}_{j=1}^n$, where n is the number of frames and d is the dimensionality of the pose representation. We define two distinct domains: a source (or prior) domain P, and a target domain T. The source dataset, $\mathcal{D}^P = \{(x_i^P, c_i^P)\}_{i=1}^{N^P}$, is a large-scale collection of motion-text pairs, where each motion sequence x^P is paired with a rich, descriptive text caption c^P. For our experiments, we consider HumanML3D [13] as our $\mathcal{D}^P$. In contrast, the target dataset, $\mathcal{D}^T = \{(x_j^T, c_j^T)\}_{j=1}^{N^T}$, is a smaller, action-specific dataset where each motion x^T is associated with a discrete action label c^T from a set Y of action classes. We operate in a few-shot setting, meaning only a handful of examples per class from $\mathcal{D}^T$ are available. This corresponds to a small subset of the NTU RGB+D dataset [29]. Our primary objective is to adapt a generative model G^P, pre-trained on $\mathcal{D}^P$, to create a new model G^T capable of synthesizing novel, class-conditional motion samples within D^T. The success of these synthetic samples is measured by their effectiveness as a data augmentation resource that improves the accuracy and robustness of a downstream HAR classifier. The core difficulty of this task lies in a significant domain gap between $\mathcal{D}^P$ and $\mathcal{D}^T$. This gap is two-fold. First, a **semantic gap** arises from the different conditioning modalities. The source domain uses free-form text (c^P), which exhibits high variance in specificity, ranging from broad descriptions like "doing martial arts" to detailed kinematic instructions. In contrast, the target domain's action labels (c^T), while less descriptive, provide a concise and semantically unambiguous scope for each motion class. Second, a **statistical gap** exists in the motion distributions themselves. This disparity stems from different data acquisition methods, as source datasets ($\mathcal{D}^P$) often contain clean motion from optical capture systems, while HAR datasets ($\mathcal{D}^T$) frequently use consumer-grade sensors or pose estimation models that produce inherently noisier data [18]. More importantly, the motion characteristics diverge: source motions are typically long and fluid, whereas target motions are shorter, more structured, and class-specific. Bridging this multifaceted domain gap is the central problem our work aims to solve.

IV. Methodology

We present our framework, named **Kinetic Mining In Context (KineMIC)**, designed to tackle the specific challenges that occurs when bridging between large-scale text-to-motion datasets and HAR datasets (fig. 2, fig. 3). The architecture is built around a teacher-student paradigm for knowledge distillation. The teacher (i.e. prior) stream G^P is a frozen, pre-trained Motion Diffusion Model (MDM) [31] with transformer encoder backbone, which acts as a static repository of rich, high-fidelity motion priors learned from the extensive HumanML3D dataset [13]. This model provides the foundational knowledge of diverse human movements. In parallel, the student (i.e. target) stream G^T is a trainable copy of the prior stream. Both model streams share initialization weights from HumanML3D pre-training. The target stream is

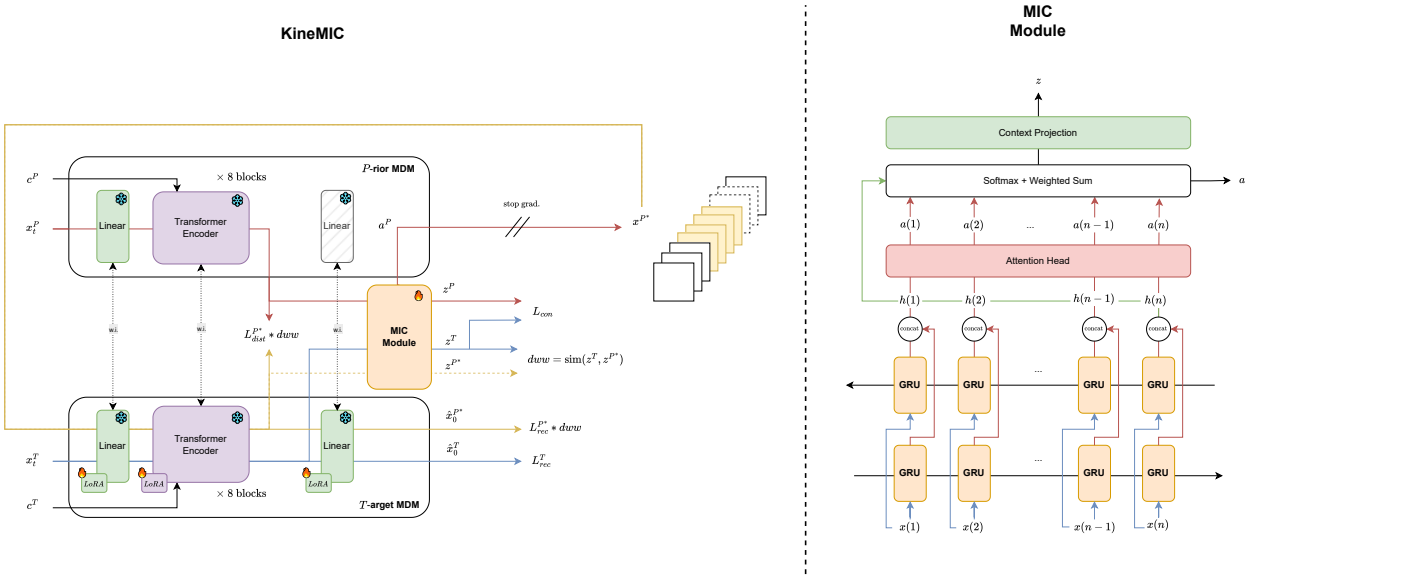

Fig. 2: The KineMIC teacher-student framework. **(Left)** The overall architecture. A frozen teacher processes a noisy, long source motion x_t^P, which acts as a "soft pair" to a given noisy target motion x_t^T. The MIC module, trained via a contrastive objective to align their latent representations (z^P, z^T), is thereby enabled to identify and extract the most semantically relevant window (x^{P^*}) from within the long source sequence. This mined window is then fed to the trainable student, using the target action class (c^T) as a pseudo-label. The student is trained with a multi-objective loss including reconstruction of both motions and a final distillation objective. Dashed lines indicate paths where gradients are not computed, and "w.i." denotes weight initialization. **(Right)** The MIC module architecture, which employs a BiGRU and attention head to produce a context-aware summary vector from a sequence of motion features.

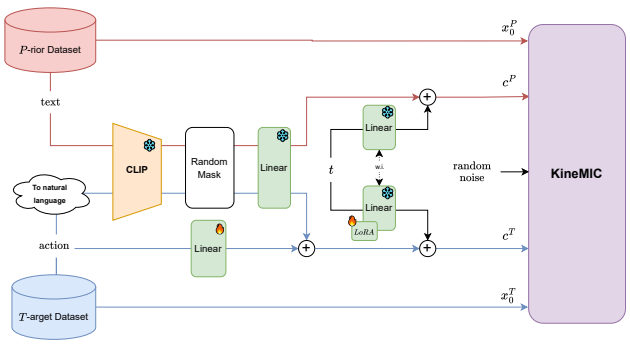

Fig. 3: The input handling and conditioning workflow. Source stream conditioning (c^P) is derived from CLIP text embeddings and a timestep embedding. Target stream conditioning (c^T) uses a text embedding from the action label, a separate learnable action embedding, and its own timestep embedding.

responsible for generating short, specific, and class-conditional actions. While the student's text embedding layer remains frozen to leverage the deep semantic space learned by the teacher, we introduce an additional learnable action embedding layer tailored to the discrete action classes of the target dataset. Moreover, we employ Low-Rank Adaptation (LoRA) [17] on all layers of the target stream, which keeps frozen the vast majority of pre-training weights while efficiently tuning low rank matrices.

A. *Soft Positive Mining*

To bridge the semantic gap between the two domains, we introduce a **soft positive mining** strategy. This approach is founded on the intuition that a single action label (e.g., "side kick") is a high-level concept that can encompass a multitude of specific physical executions, which might be described by a wide variety of text captions. This inherent one-to-many relationship means that while a source text provides a specific description, a target action label can correspond to a diverse set of motions. Our strategy leverages this by searching for semantically relevant "soft matches" between the target labels c^T and the source captions c^P for each target action. The process begins by unifying the conditioning modalities into a single semantic space. We first convert each discrete action label in c^T into a simple natural language prompt, for instance, transforming "side kick" into the sentence "a person performs a side kick." Subsequently, we use the CLIP [25] text encoder to embed both these target prompts and the descriptive source captions into a shared vector space. Within this space, we compute the cross-dataset pairwise cosine similarity for each target prompt and retain the top-k most semantically aligned source captions. We refer to this pairing as "soft" because a perfect one-to-one correspondence is neither expected nor required. The strength of this method lies in the richness of CLIP's semantic space, which captures

conceptual relationships beyond exact keywords. For example, while finding an exact textual match for "side kick" may be unlikely in a general motion dataset, the concept will be highly correlated with descriptive captions like "a person is doing martial arts." While not the entire "martial arts" sequence is relevant, it is highly probable that it contains a sub-sequence where the semantics and kinematics strongly correlate with the desired "kicking" dynamic. Our approach is designed to effectively mine these relevant contextual motions from the source dataset.

B. Mining In Context

To create a meaningful correlation between the kinematic representations of the two streams, we introduce the **Mining In Context (MIC)** module, depicted in fig. 2. This module processes the final sequence of hidden states from the transformer encoders of both the prior and target models, which we denote as $F^P(\cdot)$ and $F^T(\cdot)$ respectively (with $(\cdot)$ representing some input x_t given to the stream). Its primary training objective is to align these representations in a shared latent space based on the semantic similarity established by our soft-pairing strategy. To achieve this, we employ a contrastive objective. The choice of a **Soft Nearest Neighbors loss** [10], [27] is deliberate, as it is particularly well-suited for our one-to-many problem setting. It provides robustness against the inevitable "bad pairs" that can arise from a matching purely driven by semantics, where a source motion may be a semantic match, but still not sufficiently align with the target action from a kinematics perspective. Let $\text{cls}(x)$ be a function mapping a sample to its action class. For any target sample i in the batch B, we define its set of positive matches as $B^+(i) = \{j \mid \text{cls}(x^{T_i}) = \text{cls}(x^{P_j}), \forall j \in B\}$ such that in the case of source data, which is text conditioned and not action conditioned, we consider as pseudo-class the one of the target sample it was paired with i.e. $\text{cls}(x^{T_i}) = \text{cls}(x^{P_j})$ s.t. $(i = j)\ \forall i, j \in B$. The contrastive objective is then defined as:

$$L_{con} = -\sum_{i \in B} \log \frac{\sum_{j \in B^+(i)} \exp\left(\text{sim}(z^{T_i}, z^{P_j})/\tau\right)}{\sum_{k \in B} \exp\left(\text{sim}(z^{T_i}, z^{P_k})/\tau\right)} \tag{4}$$

where z^P and z^T are the latent vector representations produced by the MIC module and τ temperature parameter. This objective compels the module to map target samples and their corresponding soft positive prior samples to nearby points in the latent space. The MIC module consists of an attention-guided bidirectional GRU encoder [1], [4] that condenses the frame-wise transformer outputs into the single context vectors, z^P and z^T. As a direct consequence of being trained with eq. (4), the attention mechanism will implicitly weight more the frames within the long prior motion that are most responsible for the alignment with the target context. We leverage this learned attention to perform the "mining" operation. Let $a^P = \{a^P(k)\}_{k=1}^{|x^P|}$ be the sequence of attention scores computed by the module over a prior motion x^P. We use these scores to identify the **prior window** x^{P^*}. This window is a contiguous sub-sequence of x^P with a length m set to be equal to the length of the paired target sample x^T within the current batch. The window is extracted by finding the segment that maximizes the cumulative attention, thereby isolating the most relevant motion segment:

$$x^{P^*} = \underset{\{x^P(i),\ldots,x^P(i+m)\} \subset x^P}{\arg\max} \sum_{k=i}^{i+m} a^P(k) \tag{5}$$

The extracted prior motion windows are then treated as new, high-quality training data for the target stream. Crucially, the conditioning for this new sample x^{P^*} is the label c^T from the paired target sample in the current batch, creating a pseudo-labeled training example.

C. Denoising Reconstruction Objective

The fundamental training objective for our target model, G^T, is rooted in the denoising diffusion paradigm. Following the methodology of MDM [31], our model is trained to directly predict the original, "clean" motion signal, denoted as x_0, from a noisy input x_t, rather than predicting the noise vector ϵ as in traditional diffusion models training. This reconstruction objective is applied to both the ground-truth target samples from the original dataset and the pseudo-labeled prior windows mined by the MIC module. Let x_0^T and $x_0^{P^*}$ represent the ground-truth clean motions for a target sample and a prior window, respectively. The reconstruction losses for their noised counterparts, x_t^T and $x_t^{P^*}$, are formulated as a direct comparison with the model's output:

$$L_{rec}^T = ||x_0^T - G^T(x_t^T, c^T, t)||_2^2 \tag{6}$$

$$L_{rec}^{P^*} = ||x_0^{P^*} - G^T(x_t^{P^*}, c^T, t)||_2^2 \tag{7}$$

where c^T is the corresponding target action label. These two losses ensure that the target model G^T learns to generate motions that are faithful to both the target domain's original data and the rich kinematic structures extracted from the source domain.

D. Window Distillation

To ensure that the target model not only learns to reconstruct mined motion windows but also internalizes the kinematic representations from the pre-trained teacher, we introduce a feature-level distillation loss to encourage the target model's internal representations to mimic those of the prior model. We establish a distillation objective between the features produced by the final transformer block of the two streams. Let $u = \{u(1), \ldots, u(n)\}$ denote the sequence of features extracted by the final transformer encoder block outputs. The features $u^{P^*} = F^T(x_t^{P^*})$ relative to the prior window, after being processed by the target model $F^T(\cdot)$, are compared against the corresponding features $u^P = F^P(x_t^P)$ from the prior model $F^P(\cdot)$ that belong to the same frames associated to the window:

$$L_{dist}^{P^*} = \frac{1}{m} \sum_{j=1}^{m} ||u^{P^*}(j) - u^P(i + j - 1))||_2^2 \tag{8}$$

where i is the starting frame index of the window x^{P^*} within the original prior motion x^P and m the window size.

E. Dynamic Window Weighting

While the soft positive mining strategy is effective at identifying semantically relevant source motions, the window extraction process is driven purely by this semantic alignment and does not guarantee kinematic fidelity. To address this and prevent the model to learn from "bad matches," we introduce a **Dynamic Window Weighting** (dww) mechanism. This allows the model to dynamically assess the quality of each mined window and modulate its impact on the training process. After a prior window x^{P^*} is extracted, it is processed through the target stream. The resulting features are then passed through the MIC module to compute a new latent representation for the window, z^{P^*}. The quality of the match is then quantified by the cosine similarity between this new latent vector and the latent representation of the target sample z^T paired to x^P within the batch. This similarity score becomes our sample-level weight:

$$dww = \frac{\text{sim}(z^T, z^{P^*}) + 1}{2} \tag{9}$$

A small yet important detail is that we disable gradient contributions during the computation of z^{P^*}. This ensures that the MIC module is not updated through this weighting process, preventing it from trivially maximizing the similarity score. The only signal that trains the MIC module remains the contrastive objective L_{con}, preserving its role as an unbiased semantics driven kinematic aligner. This dynamically calculated weight is then used to scale eq. (7) and eq. (8), discouraging the model from learning from poor prior window choices. This mechanism ensures that only the most kinematically consistent and contextually relevant motion windows significantly contribute to G^T training.

V. Experimental Results

A. Experimental Setup

Considering how narrow is the state-of-the-art for few-shot skeleton-based action synthesis, we position our work within the experimental setup proposed by [11]. This choice of methodology is strategic for two primary reasons. First, it allows for a direct and explicit comparison with prior work by using the same three action classes from the NTU RGB+D 120 dataset [29]: "running on spot" (A099), "side kick" (A102), and "stretch on self" (A104). Second, this particular selection of actions provides a novel perspective on the limitations of large pre-trained models. Because these motions are general, a baseline model can generate them plausibly. However, this very generality makes them an excellent testbed for evaluating a model's ability to move beyond generic synthesis and capture the fine-grained, kinematic specificity required for high-accuracy recognition. Our setup, therefore, is designed not only for comparison but also to precisely measure how KineMIC overcomes this specificity gap. Following the few-shot protocol, we utilize as little as 10 randomly selected samples per action class for training our framework.

B. Dataset and Preprocessing

As in [11], we re-estimate 3D skeletons from RGB videos through VIBE [18]. To align the topology and temporal characteristics of the target data with our teacher model's pre-training on the HumanML3D [13] dataset, we introduce two minor preprocessing modifications. First, we exclude the hand joints from the SMPL skeletons to match the joint topology. Second, we downsample the motion sequences from their original 30 fps to 20 fps. Finally, to ensure full consistency, all skeletons undergo the same normalization pipeline as the HumanML3D dataset: the root joint (pelvis) is centered at the origin, feet are placed at a height of zero, the initial pose is oriented to face the positive Z-axis, and joint lengths are normalized across all frames to maintain consistent bone lengths throughout the animation. For the motion representation itself, we adopt the same input feature vector used by our teacher model, MDM [31]. This vector is a 263-dimensional feature set for each frame, concatenating root-relative joint positions, joint rotations (represented as 6D vectors), joint velocities, and estimated binary foot-contact labels.

C. Implementation Details

Our KineMIC framework employs a teacher-student architecture, where the teacher is a frozen, pre-trained 8-block MDM transformer encoder [31] trained on HumanML3D. The student model utilizes an identical architecture and shares its initialization with the teacher. For finetuning, we use LoRA [17], applying adapters (rank=16, alpha=32, dropout=0.1) to all student network layers. A critical exception is the action embedding layer; as this layer was not present in the original text-to-motion checkpoint, it is trained fully (i.e., its weights are updated directly) while all other pre-trained weights remain frozen. For our soft-positive mining, we set the number of nearest neighbors $k = 250$ and the temperature $\tau = 0.07$ for all experiments. This k hyperparameter controls the trade-off between the diversity of mined examples and their semantic relevance; a small k would limit diversity, while a large k risks noise from dissimilar motions. We selected $k = 250$ as a sufficiently large pool for drawing diverse, relevant examples and leave a detailed sensitivity analysis to future work. For conditioning, we employ classifier-free guidance with a parameter of 2.5, following [31]. As we utilize multiple conditioning signals (Action, Text), each modality is independently dropped with a disjoint probability, training the model on combinations of Action+Text, Action-only, Text-only, or no condition. We train our models for 5000 steps using the AdamW optimizer with a learning rate of $2 \cdot 10^{-5}$ and apply gradient clipping at 1.0, which we found significantly helps stabilize early training without a staged warmup. At each training step, a batch the model is trained on all 30 few-shot samples from the target dataset D^T (10 per class), combined with 30 randomly sampled soft-pairs from D^P (drawn from the top-k matches for the D^T samples), resulting in a total of 60 samples per training step.

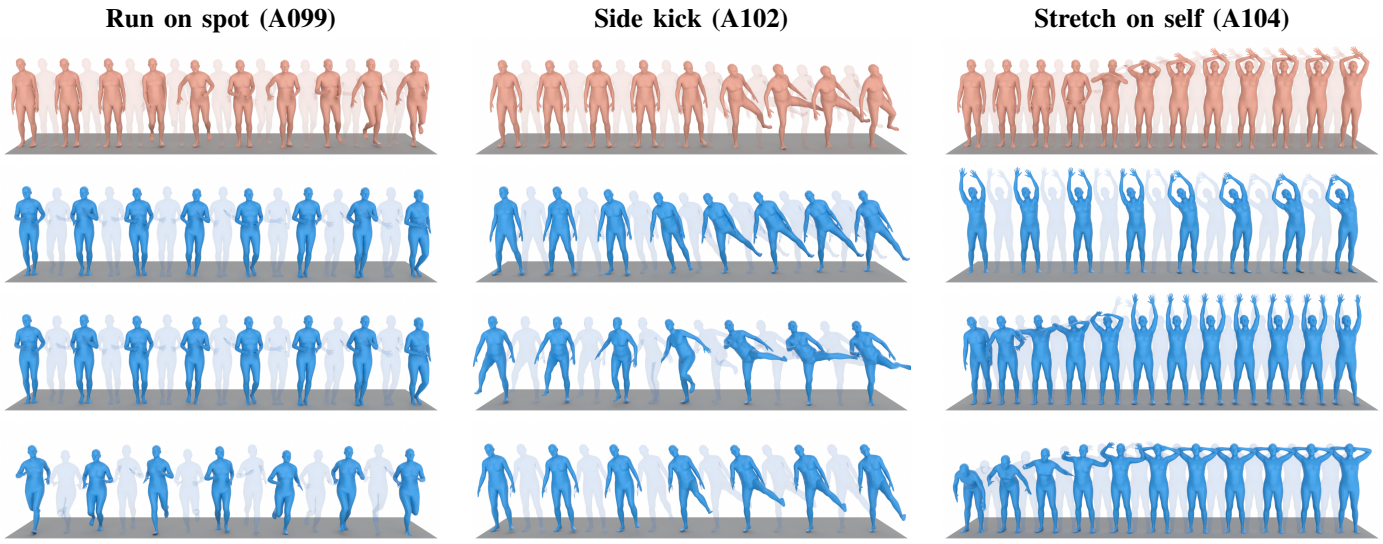

Fig. 4: **Qualitative Comparison of Generated Motions.** The three columns depict the three target action classes. The first row shows corresponding real ground truth samples from NTU RGB+D 120 extracted through [18] (in orange), while following rows show diverse samples generated by our KineMIC model (in blue).

TABLE I: **Ablation Study.** We evaluate the contribution of each component to our framework. The baseline is a LoRA fine-tuned MDM. Our core **KineMIC (Base)** model builds on this by adding the contrastive and prior window reconstruction objectives, eq. (4) and eq. (6) respectively. We then incrementally add the distillation loss (L_{dist}) and Dynamic Window Weighting (dww).

Method	Top-1 Acc (%) ↑	Diversity ↑	Multimodality ↑
Baseline (LoRA Finetune)	$80.89^{\pm 1.61}$	$9.4545^{\pm 0.388}$	$4.983^{\pm 0.358}$
KineMIC (Base)	$81.05^{\pm 1.56}$	$13.693^{\pm 1.099}$	$10.207^{\pm 0.599}$
+ L_{dist} only	$86.32^{\pm 1.05}$	$\underline{22.451}^{\pm 1.563}$	$\underline{17.115}^{\pm 2.252}$
+ dww only	$80.87^{\pm 2.05}$	$14.238^{\pm 1.092}$	$10.582^{\pm 0.772}$
+ L_{dist} + dww (Full)	$\underline{87.28}^{\pm 1.18}$	$19.489^{\pm 1.296}$	$13.636^{\pm 0.909}$

VI. Experimental Results

A. Evaluation Protocol

We evaluate the quality of our synthesized motions based on their utility in our primary objective: downstream, few-shot HAR. During our experiments, we observed a weak correlation between Fréchet Inception Distance (FID), a common metric for generative quality, and this downstream objective. We thus concluded that FID is not a reliable indicator of model utility in this specific few-shot context. Therefore, our primary evaluation metric is the top-1 classification accuracy of a downstream HAR classifier. We supplement this with metrics for Diversity and Multimodality to analyze the qualities of the generated motion. For the classifier evaluation, we employ a standard ST-GCN [33] from the PYSKL toolbox [8]. For each experiment, its training set is composed of the 30 original real samples (10 per class) combined with 1149 synthetically generated motions, distributed uniformly (383 per class). The network is trained for 80 epochs with a batch size of 64, using an SGD optimizer (LR=0.1, momentum=0.9, weight decay=0.0005) and a cosine annealing schedule. To ensure robustness, all experiments are repeated five times with different random seeds. Following the protocol of [11], we report the median top-1 accuracy across these runs for state of the art comparison, while we report mean and standard deviation for our internal ablation study.

B. State-of-the-Art Comparison

We compare our proposed KineMIC framework against prior state-of-the-art methods as well as a series of baselines. The results are detailed in table II. First, we establish a **zero-shot baseline** by using the pre-trained MDM model directly to generate motions from text prompts describing the target classes (e.g., "a person is doing a side kick"), without any fine-tuning. Baseline performances (83.1%) confirm that the chosen actions are well-represented within the HumanML3D pre-training data. However, it also reveals the core challenge: the pre-trained model is a strong generalist but lacks the specific kinematic style of the target actions. This lack of specialization sets a performance ceiling that naive adaptation methods fail to overcome. The observed quality degradation when training a randomly initialized model (81.2%) or fine-tuning with LoRA (80.7%) can be largely attributed to catastrophic forgetting and overfitting on the limited few-shot data. In contrast, our KineMIC framework better navigates this challenge, achieving a final accuracy of **87.2%**. These results are comparable and

slightly superior to existing methods, validating our hypothesis that a carefully guided adaptation is crucial for specializing a generalist model, even for seemingly "simple" actions.

TABLE II: **State-of-the-Art Comparison.** Median top-1 accuracy on the NTU RGB+D 120 few-shot benchmark. The arrow (↑) indicates that higher is better. Results marked with † are directly reported from Fukushi et al. [11] as we adopt an identical evaluation protocol and data setup.

Source	Method	Top-1 Acc (%) ↑
Prior Work†	Real data only (30 samples)†	58.4
	ACTOR† [22]	73.6
	Kinetic-GAN† [5]	81.7
	Fukushi et al.† [11]	86.4
Our Analysis	Real data only (30 samples)	63.1
	MDM (Zero-shot)	83.1
	MDM (from scratch)	81.2
	MDM (LoRA fine-tune)	80.7
	KineMIC (Full)	87.2
	Ground Truth (all real data)	97.1

C. Ablation Study

We conduct a systematic ablation study to evaluate the contribution of each component within our KineMIC framework, assessing both downstream HAR accuracy and generative quality. As presented in table I, the baseline model, consisting in MDM fine-tuned with LoRA, achieves only 80.7% accuracy and exhibits low Diversity and Multimodality scores. This result confirms that despite the use of parameter efficient fine-tuning, a simple adaptation is highly susceptible to mode collapse. The introduction of the distillation loss (L_{dist}) alone yields a substantial performance increase to 86.32%. This finding is significant, as it suggests the primary challenge in few-shot adaptation of the diffusion model is not merely sample scarcity but rather a collapse of the model's internal feature space. The distillation loss acts as latent space regularizer, anchoring the student's representations to the teacher's well-structured manifold and effectively preventing overfitting. While distillation provides this crucial regularization, the full KineMIC model, which also incorporates Dynamic Window Weighting (dww), achieves the highest Top-1 accuracy (87.28%). It is noteworthy that the full model achieves this superior accuracy despite exhibiting lower diversity and multimodality scores compared to the L_{dist}-only variant. This observation aligns with known trade-offs when employing generative models for data augmentation: for discriminative tasks, sheer generative volume is less critical than "meaningful" diversity. That is, generated samples should primarily serve to sharpen and expand decision boundaries, rather than reinforcing bias towards the original few-shot samples. Our dww component directly addresses this by discouraging the model from learning kinematically poor soft positives. This mechanism ensures that lower-quality samples have a diminished impact during training, thereby prioritizing the "meaningfulness" of the augmentation over its raw quantity.

D. Qualitative Evaluation

As illustrated in fig. 4, which displays three generated animations per class, the motions synthesized by KineMIC are both fluid and stylistically consistent with the target few-shot examples. A key strength of our framework is its ability to generate motions that not only adhere to the kinematic properties of the real data but also generalize to meaningful, semantically related variations. This is particularly evident in the diversity of the generated samples. For instance, within the *side kick* class, KineMIC produces a range of motions, including some that resemble more "combat-like" kicks, demonstrating an understanding of the broader action concept beyond the provided few-shot examples. This capacity for plausible variation is more apparent with the *stretch on self* action, arguably the most abstract of our three target classes. Our model synthesizes a richer set of semantically meaningful variations, including novel movements such as a stretch with "hands behind the head" or "lateral tilts of the torso." This generated diversity, however, highlights a potential bias within the NTU RGB+D dataset itself. The ground truth for *stretch on self*, for instance, is mostly skewed towards upper-body dominant motions. Consequently, valid interpretations like lower-body stretches, which are semantically coherent matches may not align with the narrow scope of the validation set. Despite this evaluation challenge, the overall improvement in our quantitative metrics is a healthy sign, suggesting the increased diversity provides a net benefit for training a more robust classifier.

During our analysis, we discovered a compelling emergent property. When guided by a text prompt distinct from its action-conditioning class, the model synthesizes a motion that fuses the semantics of both. For instance, as shown in fig. 5, a model conditioned on *stretch on self* but prompted with the text *a person is jumping* generates a plausible animation of a figure jumping while stretching its arms. We observed that this blending is most successful when the motions are not kinematically conflicting. For example, combining an upper-body dominant action (stretch on self) with a lower-body one (jumping) produces a coherent result, whereas combining two lower-body dominant actions (running on spot and jumping) often fails. We hypothesize this compositional ability emerges from the synergy between our dual-conditioning scheme and LoRA, which effectively preserves both the fine-tuned action and the general capabilities of the pre-trained model, similar to the findings in [28]. While a full exploration is beyond the scope of this work, this suggests a promising avenue for creating stylized motions from few-shot examples.

VII. Conclusion

In this work, we addressed the critical challenge of few-shot action synthesis for HAR. We systematically investigated how text-to-motion generative models can serve as powerful priors but often lack the kinematic specificity required for specialized action classes, resulting in a significant domain gap. To bridge this gap, we introduced KineMIC, a novel teacher-student framework that effectively adapts a pre-trained model using

Fig. 5: **Motion composition via dual conditioning.** The generated motion is conditioned on the action ‘stretch on self‘ and the text prompt ‘a person is jumping‘, demonstrating the model’s ability to blend kinematic and semantic inputs into a coherent, novel animation

as few as 10 samples per class. Our core contribution is a soft positive mining strategy that leverages a shared semantic space to intelligently retrieve and adapt relevant motion knowledge from the source domain, closing the gap to the target action classes. By combining this with a multi-objective loss featuring latent space distillation, KineMIC successfully generates diverse and high-quality synthetic data, establishing a new state-of-the-art on the NTU RGB+D 120 few-shot benchmark established in [11] and significantly boosting downstream classifier performance. For future work, we did not explore the possibility of employing prompt augmentation techniques, which could be a compelling direction for enhancing the semantic richness of the target conditioning. Furthermore, the emergent property of our model for controllable motion composition suggests a promising reframing of the architecture. Instead of a unified reconstruction space, the framework could be viewed as a style transfer mechanism, where the target samples provide the ”style” and source, text conditioned data provide the ”content.” This opens up exciting new avenues for fine-grained, controllable, and few-shot action synthesis.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

[2] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition, 2021.

[3] Hyung-Gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20154–20164, 2022.

[4] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014.

[5] Bruno Degardin, João Neves, Vasco Lopes, João Brito, Ehsan Yaghoubi, and Hugo Proença. Generative adversarial graph convolutional networks for human action synthesis, 2021.

[6] Jeonghyeok Do and Munchurl Kim. Skateformer: Skeletal-temporal transformer for human action recognition, 2024.

[7] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, 2015.

[8] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin. Pyskl: Towards good practices for skeleton action recognition, 2022.

[9] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition, 2022.

[10] Nicholas Frosst, Nicolas Papernot, and Geoffrey Hinton. Analyzing and improving representations with the soft nearest neighbor loss, 2019.

[11] Kenichiro Fukushi, Yoshitaka Nozaki, Kosuke Nishihara, and Kentaro Nakahara. Few-shot generative model for skeleton-based human action synthesis using cross-domain adversarial learning. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3934–3943, 2024.

[12] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions, 2023.

[13] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5142–5151, 2022.

[14] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM ’20, page 2021–2029. ACM, October 2020.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[17] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[18] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: video inference for human body pose and shape estimation. *CoRR*, abs/1912.05656, 2019.

[19] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition, 2020.

[20] Marcos Lupion, Federico Cruciani, I Cleland, CD Nugent, and Pilar Ortigosa. Data augmentation for human activity recognition with generative adversarial networks. *IEEE Journal of Biomedical and Health Informatics*, 28(4):2350–2361, February 2024.

[21] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes, 2019.

[22] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae, 2021.

[23] Mathis Petrovich, Michael J. Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions, 2022.

[24] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. *Spatial Temporal Transformer Network for Skeleton-Based Action Recognition*, page 694–701. Springer International Publishing, 2021.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[26] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[27] Ruslan Salakhutdinov and Geoff Hinton. Learning a nonlinear embedding by preserving class neighbourhood structure. In Marina Meila and Xiaotong Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 412–419, San Juan, Puerto Rico, 21–24 Mar 2007. PMLR.

[28] Haim Sawdayee, Chuan Guo, Guy Tevet, Bing Zhou, Jian Wang, and Amit H. Bermano. Dance like a chicken: Low-rank stylization for human motion diffusion, 2025.

[29] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis, 2016.

[30] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition, 2019.

[31] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model, 2022.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.
[33] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition, 2018.
[34] Zeyu Zhang, Yiran Wang, Wei Mao, Danning Li, Rui Zhao, Biao Wu, Zirui Song, Bohan Zhuang, Ian Reid, and Richard Hartley. Motion anything: Any to motion generation, 2025.
[35] Mengyi Zhao, Mengyuan Liu, Bin Ren, Shuling Dai, and Nicu Sebe. Modiff: Action-conditioned 3d motion generation with denoising diffusion probabilistic models, 2023.