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

Luca Cazzola[1][0009−0000−6285−8342] and Ahed Alboody[2][?−?−?−?]

[1] University of Trento, Via Sommarive 5, 38123 Trento, Italy
luca.cazzola-1@studenti.unitn.it
[2] CESI Lineact, 13 avenue Simone Veil, 06200 Nice, France
aalboody@cesi.fr

Abstract. The high cost of acquiring large-scale annotated motion datasets creates a critical bottleneck for skeletal-based Human Activity Recognition (HAR). While large-scale Text-to-Motion generative models offer a compelling, scalable source of data, their primary development focus on artistic and general animation workflows means their massive training datasets lack the kinematic specificity and atomic action structure required for downstream HAR tasks, resulting in a significant domain gap. To effectively bridge this gap, we propose KineMIC (Kinetic Mining In Context), a novel transfer learning framework for few-shot action synthesis. KineMIC operates on the hypothesis that semantic correspondence in the rich, pre-trained text encoding space—comparing sparse HAR labels to motion captions—can provide essential "soft" supervision to distill the needed kinematic information. We operationalize this with a novel kinetic mining strategy that leverages CLIP text embeddings to establish these soft pairings between target actions and the source data. This process guides the fine-tuning of the diffusion model, identifying and extracting relevant motion sub-sequences to transform the generalist Text-to-Motion model into a specialized, few-shot Action-to-Motion generator. By synthesizing data from as few as 10 real samples per class, KineMIC provides a data augmentation strategy that we demonstrate to be effective at enhancing accuracy of downstream HAR classifiers. Animated illustrations and additional material available at https://lucazzola.github.io/publications/kinemic.

Keywords: Human Activity Recognition · Data Augmentation · Motion Synthesis

1 Introduction

Human Activity Recognition (HAR) has become a cornerstone in a multitude of fields, including sports performance analysis, human-robot collaboration, and intelligent surveillance [5]. Skeletal-based HAR remains a fundamental modality,

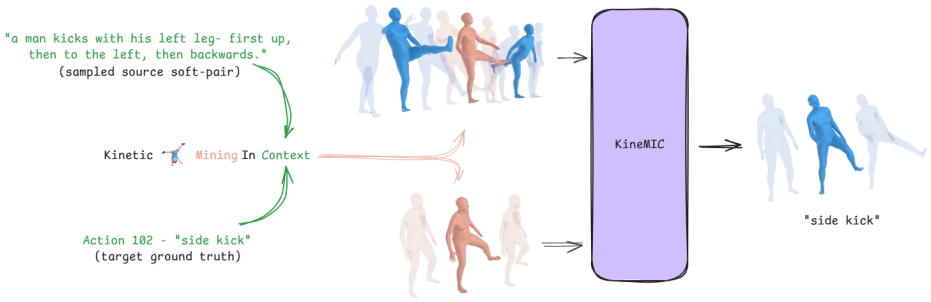

Fig. 1. Kinetic Mining in Context. The target action sample (bottom) is used to contextualize the search within the large source data sample (top), establishing soft pairs. The mining operation identifies a kinematically relevant segment (in orange) from the source data.

frequently employed due to its lightweight representation, robustness to environmental variations, and inherently privacy-preserving nature [33]. However, the performance of deep learning models for this task is fundamentally limited by the availability of large-scale, accurately annotated datasets. The acquisition and precise labeling of such high-quality, task-specific data are notoriously expensive and labor-intensive, creating a significant bottleneck that hampers progress, particularly in few-shot settings [5,15]. This core challenge of data scarcity in HAR is what we aim to address.

To mitigate this fundamental data bottleneck, much of modern few-shot recognition research focuses on applying strategies such as meta-learning and metric-learning frameworks directly to the action recognition model [32]. A significantly less popular, yet open alternative leverages generative models to create novel synthetic samples, thereby augmenting the small training set [10]. When it comes to motion synthesis, current research is dominated by Text-to-Motion (T2M) synthesis models, due to the growing interest in text as a powerful conditioning modality. This interest has driven the creation and collection of large-scale T2M datasets [12,19], which has, in turn, led to the development of powerful T2M models [30,11,12] that can be employed as general motion priors for other generative tasks [26,25,29]. T2M priors are strategically appealing as a generative solution because the scalability of their annotation, using free-form, descriptive text (captions), is significantly easier to achieve for general diversity than collecting the high-volume, kinematically specific data requiring precise, atomic action labels for HAR.

While the community moves towards developing foundation models for 3D humans [31,16], it remains a significantly underexplored avenue whether such general T2M priors can be effectively exploited for specialized, downstream tasks like HAR. Our work concentrates on this critical transfer challenge: adapting a general T2M prior to function as an effective Action-to-Motion (A2M) synthetic data generator for a specific target HAR domain. This transformation is non-trivial due to a significant domain gap characterized by two key factors. First, a

semantic discrepancy exists where source T2M data uses descriptive text, while target HAR requires generation based on discrete action labels. Second, a kinematic gap exists between the broad, fluid MoCap motions of the source domain and the short, atomic motions required for HAR. The pre-trained T2M model, being a generalist, is therefore ill-equipped to meet the rigorous kinematic specificity needed for reliable HAR classification.

To address these challenges and bridge this domain gap for few-shot action synthesis, we propose KineMIC (Kinetic Mining In Context, Figure 1). Our method employs a dual-stream, teacher-student architecture wherein a frozen, pre-trained T2M model acts as a teacher, guiding the fine-tuning of a student model for the target HAR domain. The core innovation is a novel soft positive mining strategy that leverages CLIP [23] text embeddings to establish a semantic correspondence between the sparse target action labels and the rich textual descriptions from the source domain. This process identifies and extracts the relevant, specific motion sub-sequences from the vast source dataset, effectively transforming the general-purpose T2M model into a specialized A2M generator 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 T2M model into an A2M generator for HAR applications, moreover, addressing this within a few-shot setting.
- We introduce KineMIC, a novel teacher-student transfer learning framework incorporating a Kinetic Mining In Context strategy that effectively adapts a general T2M diffusion prior [30] to a specific HAR domain with minimal data, proving its effectiveness in boosting downstream HAR model accuracy.

2 Related Works

2.1 Skeletal-based HAR

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 [7,27]. 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 [28] and more sophisticated GCN architectures [18,8,1,2]. More recently, 3D convolutional networks [9] and transformers [22] have set new performance benchmarks. Finally, hybrid architectures have emerged, attempting to capture both short-range skeletal dependencies and long-range temporal context [6]. Despite these advancements, the field still struggles when data is scarce, as the high capacity of modern deep learning architectures often leads to severe overfitting on the limited samples. Our work addresses this bottleneck by proposing a novel synthesis framework to create kinematically-specific, high-quality training data, thereby supporting HAR models in limited-data regimes.

2.2 Generative Models for 3D Skeleton-based Motion

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 [13,20,4]. A significant leap in quality was enabled by two key factors: the curation of larger, more diverse Motion Capture (MoCap) datasets [19,12] and the introduction of richer conditioning signals like free-form text [21]. 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 [30,35]. The current state-of-the-art is largely characterized by a competition between these diffusion models and masked generative models [11], with recent research pushing the boundaries of conditional synthesis even further by incorporating modalities like music and images [34]. While these models achieve impressive fidelity and diversity, they are primarily trained for general animation and text-to-motion applications, often overlooking the fine-grained specificity required by downstream HAR tasks. We bridge this domain gap by introducing a kinetic mining strategy to distill the necessary HAR-specific information from a pre-trained generative backbone.

2.3 Few-Shot HAR with Generative Models

Few-Shot Human Activity Recognition (FSHAR) is bottlenecked by the high cost of annotated motion datasets. While historical solutions focused on classifier-side methods (e.g., metric/meta-learning) [32], these approaches are inherently constrained by the limited kinematic diversity of the few-shot support set. A less explored alternative is deep generative data augmentation. Fukushi et al. [10] demonstrated this with a GAN, notably showing results with as few as 10 samples per class. Their core mechanism for few-shot learning involves cross-domain regularization of the discriminator using reference motion patches from a large source set, which effectively prevents overfitting to the limited few-shot data. Crucially, this augmentation strategy is fundamentally tied to kinematic feature matching in the motion domain, which ties the generated variations closely to the specific kinematics of the few-shot support set. Furthermore, their reliance on GANs introduces notorious challenges like training instability. To bypass these issues, we propose leveraging modern T2M diffusion models for few-shot motion synthesis, a novel direction in this domain. In direct contrast to the kinematic feature matching of [10], our method employs a "semantics-first" matching strategy. By establishing meaningful semantic correspondences via the rich pre-trained text encoding space, we can effectively disentangle the synthesis process from the tight kinematic limitations of the few-shot support set.

3 Problem Formulation

The core challenge we address is the adaptation of a pre-trained T2M generative model for data augmentation within the context of few-shot HAR. Our primary

objective is to leverage the extensive kinematic knowledge contained in a large source (i.e. prior) domain to synthesize a high volume of diverse, class-specific motion sequences for a target domain, thereby enhancing the performance of a downstream HAR classifier. Let a skeletal motion sequence be defined 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 consider two distinct domains:

1. A prior domain P, characterized by the large-scale dataset $\mathcal{D}^P$, which is a collection of pairs (x^P, c). Here, x^P represents a motion sequence, and c is its associated rich, free-form, descriptive text caption.
2. A target domain T, defined by the small, action-specific dataset $\mathcal{D}^T$, which is a collection of pairs (x^T, y). Here, x^T represents a motion sequence, and y is its associated discrete action label from a set of action classes Y.

Working in a few-shot setting implies that only a small, labeled subset of the target data, $\mathcal{D}^{\bar{T}} \subset \mathcal{D}^T$, is available for adaptation and training. Our central goal is to utilize this limited set $\mathcal{D}^{\bar{T}}$ to adapt the generative model G^P, pre-trained on $\mathcal{D}^P$, yielding a new model G^T. This model G^T must be capable of synthesizing novel, class-conditional motion samples from the action set Y. The generated synthetic data is then employed to train a HAR classification model, which is subsequently evaluated on the test/validation splits of $\mathcal{D}^T$. The quality and success of these synthetic samples are measured by their ability to improve the accuracy and robustness of a subsequent HAR classifier when used for data augmentation. The central difficulty lies in bridging the significant domain gap between $\mathcal{D}^P$ and $\mathcal{D}^T$, which manifests in two critical ways:

1. Conditioning modalities differ fundamentally (semantic gap). The source domain uses high-variance, free-form text, while the target domain employs concise, semantically unambiguous discrete action labels.
2. There is a discrepancy in the motion distributions. Target motions tend to be more atomic and short, contrasting with the generally longer motions in the source domain. Furthermore, differences in data acquisition methods contribute to this disparity.

Our work is focused on solving this multifaceted adaptation problem to effectively utilize large-scale generative models for few-shot HAR data augmentation.

4 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 (2, 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) [30] with transformer encoder backbone, which acts as a static repository of rich motion priors learned from the extensive T2M dataset. This model provides

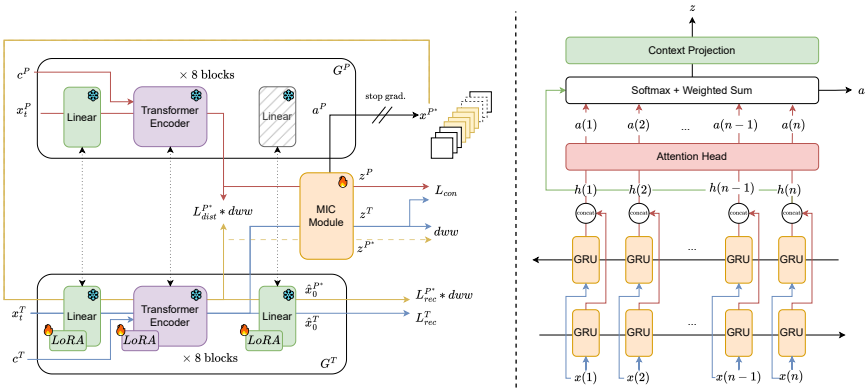

Fig. 2. The KineMIC teacher-student framework. **(Left)** The overall architecture. A frozen teacher processes a noisy, long source motion x_t^P, which acts as a "soft pair" to a given noisy target motion x_t^T. The MIC module, trained via a contrastive objective to align their latent representations (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.

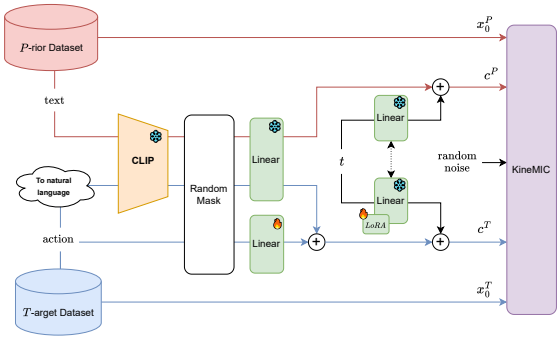

Fig. 3. The input handling and conditioning workflow. Source stream conditioning (c^P) is derived from CLIP text embeddings and a timestep embedding. Target stream conditioning (c^T) uses a text embedding from the action label, a separate learnable action embedding, and its own timestep embedding.

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 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) [14] on all layers of the target stream, which keeps frozen the vast majority of pre-training weights while efficiently tuning low rank matrices.

4.1 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 [23] 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.

4.2 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 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 [24,?] 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:

$$L_{con} = -\sum_{i \in B} \log \frac{\sum_{j \in B^+(i)} \exp\left(\text{sim}(z^{T_i}, z^{P_j})/\tau\right)}{\sum_{k \in B} \exp\left(\text{sim}(z^{T_i}, z^{P_k})/\tau\right)} \tag{1}$$

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 [3,?] 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 Equation 1, 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:

$$x^{P^*} = \underset{\{x^P(i),\ldots,x^P(i+m)\} \subset x^P}{\arg\max} \sum_{k=i}^{i+m} a^P(k) \tag{2}$$

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.

4.3 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 [30], 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:

$$L_{rec}^T = ||x_0^T - G^T(x_t^T, c^T, t)||_2^2 \tag{3}$$

$$L_{rec}^{P^*} = ||x_0^{P^*} - G^T(x_t^{P^*}, c^T, t)||_2^2 \tag{4}$$

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.

4.4 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:

$$L_{dist}^{P^*} = \frac{1}{m} \sum_{j=1}^{m} ||u^{P^*}(j) - u^P(i+j-1))||_2^2 \tag{5}$$

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

4.5 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

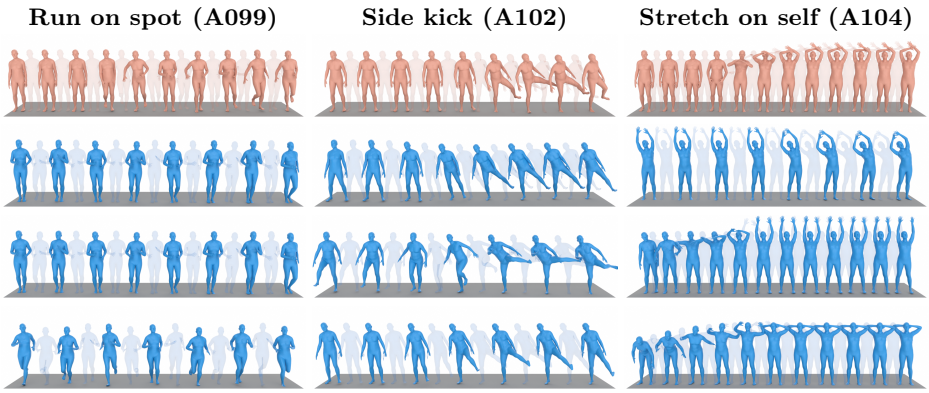

Fig. 4. Qualitative Comparison of Generated Motions. The three columns depict the three target action classes. The first row shows corresponding real ground truth samples from NTU RGB+D 120 extracted through VIBE [17] (in orange), while following rows show diverse samples generated by our KineMIC model (in blue).

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:

$$dww = \frac{\text{sim}(z^T, z^{P^*}) + 1}{2} \tag{6}$$

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 Equation4 and Equation5, 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.

5 Experiments

5.1 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 [10]. 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 [27]: "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.

Dataset and Preprocessing. As in [10], we re-estimate 3D skeletons from RGB videos through VIBE [17]. To align the topology and temporal characteristics of the target data with our teacher model's pre-training on the HumanML3D [12] 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 [30]. 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.

Implementation Details. Our KineMIC framework employs a teacher-student architecture, where the teacher is a frozen, pre-trained 8-block MDM transformer encoder [30] trained on HumanML3D. The student model utilizes an identical architecture and shares its initialization with the teacher. For finetuning, we use LoRA [14], 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 [30]. 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.

5.2 Evaluation Protocol

Our primary objective is to maximize the utility of synthesized data as an augmentation source for a HAR classifier. Therefore, our core evaluation metrics focus on recognition accuracy, Multimodality, and Diversity, aligning with the evaluation scope of other few-shot 3D motion synthesis works [10]. For complete analysis of generative fidelity, including the Fréchet Inception Distance (FID), we refer the reader to the supplementary material. 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 three times with different random seeds. Following the protocol of [10], 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.

5.3 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 1. First, we establish a **zero-shot baseline** by using the pre-trained MDM model directly to generate motions from general text prompts describing the target classes (e.g., class 'side kick' becomes '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.

5.4 Ablation Study

We conduct a systematic ablation study to evaluate the contribution of each component within our KineMIC framework, assessing both downstream HAR

Table 1. State-of-the-Art Comparison. Median top-1 accuracy on the NTU RGB+D 120 few-shot benchmark defined in [10]. The up-arrow (↑) indicates that higher is better. Results marked with † are reported from Fukushi et al. [10] as we adopt an identical evaluation protocol.

Source	Method	Top-1 Acc (%) ↑
Prior Work†	Real data only (30 samples)†	58.4
	ACTOR† [20]	73.6
	Kinetic-GAN† [4]	81.7
	Fukushi et al.† [10]	86.4
Our Analysis	Real data only (30 samples)	63.1
	MDM (Zero-shot)	83.1
	MDM (from scratch)	81.2
	MDM (LoRA finetune)	80.7
	KineMIC (Full)	<u>87.2</u>
	Ground Truth (all real data)	97.1

accuracy and generative quality. As presented in Table 2, the MDM baseline fine-tuned with LoRA achieves only 80.7% accuracy, confirming its high susceptibility to mode collapse. The introduction of the distillation loss ($\mathcal{L}_{dist}$) alone yields a substantial performance increase to 86.32%, establishing its role as a crucial latent space regularizer that effectively prevents overfitting. While distillation provides this strong regularization, the full KineMIC model, which also incorporates *dww*, achieves the highest Top-1 accuracy (87.28%). The full model achieves this superior accuracy despite exhibiting lower Diversity and Multimodality scores than the $\mathcal{L}_{dist}$-only variant. This aligns with the known generative trade-off where the *dww* component prioritizes "meaningful" diversity for discriminative tasks rather than mere volume, by actively discouraging kinematically poor soft positives. However, we must note a practical trade-off: the dynamic selective nature of the *dww* component, while relevant for improving kinematic specificity, introduces optimization variance during the motion synthesis phase. This variance is reflected by the comparatively higher error intervals observed in the downstream HAR accuracy results in Table 2, suggesting that the quality of the synthesized training data is less consistently reliable across different model initialization runs. Thus, the *dww* effectively accentuates the core contribution of the distillation objective and the MIC module by demonstrating the necessity of filtering for kinematic quality over sheer generation volume. Given the observed variance, future work is warranted to develop a more robust and lower-variance filtering strategies.

5.5 Qualitative Evaluation

As illustrated in Figure 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

Table 2. 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, Equation Equation 1 and Equation Equation 3 respectively. We then incrementally add the distillation loss (L_{dist}) and Dynamic Window Weighting (*dww*). **Note:** (↑) denotes that higher values are better.

Method	**Accuracy (%) ↑**	**Diversity ↑**	**Multimodality ↑**
Baseline (MDM + LoRA)	$80.89^{\pm 0.82}$	$9.45^{\pm 0.39}$	$4.98^{\pm 0.36}$
KineMIC (Base)	$81.05^{\pm 1.26}$	$13.69^{\pm 1.10}$	$10.21^{\pm 0.60}$
+ L_{dist} only	$86.32^{\pm 1.14}$	$\underline{22.45}^{\pm 1.56}$	$\underline{17.12}^{\pm 2.25}$
+ *dww* only	$80.87^{\pm 2.05}$	$14.24^{\pm 1.09}$	$10.58^{\pm 0.77}$
+ L_{dist} + *dww* (Full)	$\underline{87.28}^{\pm 1.48}$	$19.49^{\pm 1.30}$	$13.64^{\pm 0.91}$

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 Figure 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 [25]. While a full exploration is beyond the scope of this work, this suggests a promising avenue for creating stylized motions from few-shot examples.

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

6 Conclusion

In this work, we addressed the critical challenge of few-shot action synthesis for HAR. We systematically investigated how T2M generative models can serve as powerful priors but often lack the kinematic specificity required for specialized, atomic action classes, resulting in a significant domain gap. To effectively bridge this, we introduced KineMIC, a novel teacher-student framework that adapts a pre-trained diffusion model using 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 distill 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, significantly boosting downstream classifier performance on the NTU RGB+D 120 few-shot benchmark defined by [10]. Despite these advances, the KineMIC framework presents conceptual and practical limitations that inform future research. The primary constraint lies in our core assumption that semantic correspondence is an effective proxy for kinematic relevance during the mining process. This assumption does not always hold; for instance, text encoders may find high correlation between ’punch’ and ’kick’ due to shared concepts (e.g., ’fighting’), despite their substantial kinematic disparity, increasing the importance of an effective filtering strategy. Consequently, the method’s robustness is intrinsically linked to (a) the scale and diversity of the T2M source data and (b) the specificity of the target few-shot action, implying a practical limit on mining highly complex or novel atomic movements. Future efforts should therefore focus on developing a more proactive, learned kinematic quality score to enhance mining purity. Furthermore, we note that we did not explore prompt augmentation strategies. Integrating these techniques is expected to be highly beneficial for enhancing the

semantic richness of the target conditioning, a critical component for leveraging the full potential of the T2M prior.

Acknowledgements ????????

References

1. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition (2021), https://arxiv.org/abs/2107.12213
2. Chi, H.G., Ha, M.H., Chi, S., Lee, S.W., Huang, Q., Ramani, K.: Infogcn: Representation learning for human skeleton-based action recognition. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20154–20164 (2022). https://doi.org/10.1109/CVPR52688.2022.01955
3. Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR **abs/1412.3555** (2014), http://arxiv.org/abs/1412.3555
4. Degardin, B., Neves, J., Lopes, V., Brito, J., Yaghoubi, E., Proença, H.: Generative adversarial graph convolutional networks for human action synthesis (2021), https://arxiv.org/abs/2110.11191
5. Dentamaro, V., Gattulli, V., Impedovo, D., Manca, F.: Human activity recognition with smartphone-integrated sensors: A survey. Expert Syst. Appl. **246**, 123143 (2024), https://api.semanticscholar.org/CorpusID:266926424
6. Do, J., Kim, M.: Skateformer: Skeletal-temporal transformer for human action recognition (2024), https://arxiv.org/abs/2403.09508
7. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1110–1118 (2015). https://doi.org/10.1109/CVPR.2015.7298714
8. Duan, H., Wang, J., Chen, K., Lin, D.: Pyskl: Towards good practices for skeleton action recognition (2022), https://arxiv.org/abs/2205.09443
9. Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B.: Revisiting skeleton-based action recognition (2022), https://arxiv.org/abs/2104.13586
10. Fukushi, K., Nozaki, Y., Nishihara, K., Nakahara, K.: 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). pp. 3934–3943 (2024). https://doi.org/10.1109/WACV57701.2024.00390
11. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions (2023), https://arxiv.org/abs/2312.00063
12. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5142–5151 (2022). https://doi.org/10.1109/CVPR52688.2022.00509
13. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. p. 2021–2029. MM '20, ACM (Oct 2020). https://doi.org/10.1145/3394171.3413635, http://dx.doi.org/10.1145/3394171.3413635

14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021), https://arxiv.org/abs/2106.09685
15. Hung-Cuong, N., Nguyen, T.H., Scherer, R., Le, V.H.: Deep learning for human activity recognition on 3d human skeleton: Survey and comparative study. Sensors (Basel, Switzerland) **23** (2023), https://api.semanticscholar.org/CorpusID:258957145
16. Khirodkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for human vision models (2024), https://arxiv.org/abs/2408.12569
17. Kocabas, M., Athanasiou, N., Black, M.J.: VIBE: video inference for human body pose and shape estimation. CoRR **abs/1912.05656** (2019), http://arxiv.org/abs/1912.05656
18. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition (2020), https://arxiv.org/abs/2003.14111
19. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes (2019), https://arxiv.org/abs/1904.03278
20. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae (2021), https://arxiv.org/abs/2104.05670
21. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions (2022), https://arxiv.org/abs/2204.14109
22. Plizzari, C., Cannici, M., Matteucci, M.: Spatial Temporal Transformer Network for Skeleton-Based Action Recognition, p. 694–701. Springer International Publishing (2021). https://doi.org/10.1007/978-3-030-68796-0_50, http://dx.doi.org/10.1007/978-3-030-68796-0_50
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), https://arxiv.org/abs/2103.00020
24. Salakhutdinov, R., Hinton, G.: Learning a nonlinear embedding by preserving class neighbourhood structure. In: Meila, M., Shen, X. (eds.) Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 2, pp. 412–419. PMLR, San Juan, Puerto Rico (21–24 Mar 2007), https://proceedings.mlr.press/v2/salakhutdinov07a.html
25. Sawdayee, H., Guo, C., Tevet, G., Zhou, B., Wang, J., Bermano, A.H.: Dance like a chicken: Low-rank stylization for human motion diffusion (2025), https://arxiv.org/abs/2503.19557
26. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior (2023), https://arxiv.org/abs/2303.01418
27. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+d: A large scale dataset for 3d human activity analysis (2016), https://arxiv.org/abs/1604.02808
28. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition (2019), https://arxiv.org/abs/1805.07694
29. Tevet, G., Raab, S., Cohan, S., Reda, D., Luo, Z., Peng, X.B., Bermano, A.H., van de Panne, M.: Closd: Closing the loop between simulation and diffusion for multi-task character control (2024), https://arxiv.org/abs/2410.03441
30. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model (2022), https://arxiv.org/abs/2209.14916

31. Wang, Y., Sun, Y., Patel, P., Daniilidis, K., Black, M.J., Kocabas, M.: Prompthmr: Promptable human mesh recovery (2025), https://arxiv.org/abs/2504.06397
32. Wanyan, Y., Yang, X., Dong, W., Xu, C.: A comprehensive review of few-shot action recognition (2025), https://arxiv.org/abs/2407.14744
33. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition (2018), https://arxiv.org/abs/1801.07455
34. Zhang, Z., Wang, Y., Mao, W., Li, D., Zhao, R., Wu, B., Song, Z., Zhuang, B., Reid, I., Hartley, R.: Motion anything: Any to motion generation (2025), https://arxiv.org/abs/2503.06955
35. Zhao, M., Liu, M., Ren, B., Dai, S., Sebe, N.: Modiff: Action-conditioned 3d motion generation with denoising diffusion probabilistic models (2023), https://arxiv.org/abs/2301.03949