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

Luca Cazzola[1][0009−0000−6285−8342] and Ahed Alboody[2][0000−0002−9555−7471]

[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 acquisition cost for large, annotated motion datasets is a critical bottleneck for skeletal-based Human Activity Recognition (HAR). Although powerful Text-to-Motion (T2M) generative models offer a compelling, scalable source of data, their training goals, which focus on general, artistic motion, and their dataset structure fundamentally differ from the requirements of HAR. This disparity results in a significant domain gap, making these models ill-equipped for generating kinematically precise actions. To address this challenge, we propose KineMIC (Kinetic Mining In Context), a transfer learning framework for few-shot action synthesis. KineMIC adapts a general T2M diffusion model for the HAR domain by hypothesizing that semantic correspondence in the pre-trained text encoding space can provide the soft supervision needed for kinematic distillation. We operationalize this via a kinetic mining strategy that leverages CLIP text embeddings to establish semantic pairings between sparse HAR labels and the T2M source data. This specialized process guides the fine-tuning of the diffusion model, transforming the generalist T2M backbone into a few-shot Action-to-Motion generator. KineMIC generates synthetic data from as little as 10 real samples per class, validating the framework by enhancing the accuracy of a downstream HAR classifier. 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 [4]. Skeletal-based HAR remains a fundamental modality, 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

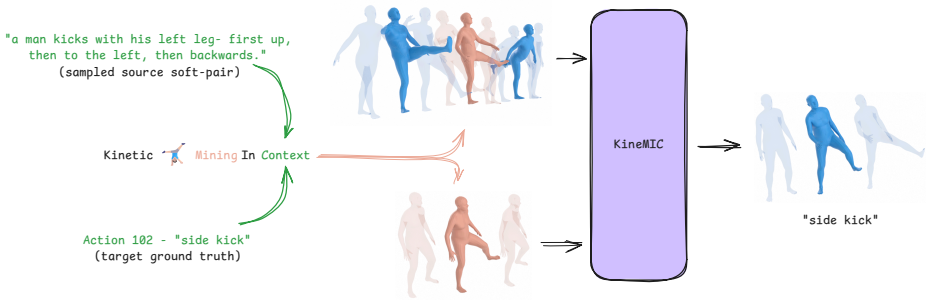

Fig. 1. Kinetic Mining in Context. The target action sample (bottom) is used to contextualize the search within the large source data sample (top), establishing soft pairs. The mining operation identifies a kinematically relevant segment (in orange) from the source data.

the availability of large-scale, accurately annotated datasets. The acquisition and precise labeling of such high-quality, task-specific data is notoriously expensive and labor-intensive, creating a significant bottleneck that hampers progress, particularly in few-shot settings [4,15]. 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]. An alternative approach 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 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 deep 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, is significantly easier to achieve for general diversity than collecting high-volume, kinematically specific data for HAR.

While the community moves towards developing foundation models for 3D humans [31,16], the challenge of effectively exploiting such general T2M priors for specialized, downstream tasks like HAR remains underexplored. Our work concentrates on this challenge: adapting a general T2M prior to function as an 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 motions of the source domain and the short, atomic motions required for HAR.

The pretrained T2M model, being a generalist, is therefore ill-equipped to meet requirements 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 teacher-student architecture, wherein a frozen teacher, pretrained on a source T2M dataset, guides the fine-tuning of a student model on the limited target HAR domain. Our method starts by establishing a semantic correspondence between the sparse target action labels and the rich source textual descriptions through CLIP [23] text encoder. Secondly, relevant motion sub-sequences, extracted from the vast source dataset, are used to turn the general-purpose student into a specialized generator. The main contributions of this work are as follows:

- To the best of the authors' knowledge, we are the first to tackle the specific challenge of adapting a T2M model into an A2M generator for HAR applications, moreover, addressing this within a few-shot setting.
- We introduce KineMIC, a teacher-student framework, which ultimately adapts a general T2M diffusion prior [30] to a specific HAR domain with minimal data, proving its effectiveness at improving HAR accuracy.

2 Related Works

2.1 Skeletal-based HAR

Recognizing human activity from skeletal data is a pivotal research area in computer vision. Early approaches modeled motion as a time series using RNNs and LSTMs [6,27]. A paradigm shift occurred with the introduction of Graph Convolutional Networks (GCNs) [33], which re-conceptualized the skeleton as a graph to model spatio-temporal dependencies. This led to rapid advancements using adaptive graph structures [28], refined GCNs [18,7,1,2], and more recently, specialized architectures like 3D convolutional networks [8] and Transformers [22,5]. Despite these advancements, the field still struggles with data scarcity, as the high capacity of modern models leads to severe overfitting in limited-data regimes. Our work addresses this bottleneck by proposing a novel synthesis framework leveraging deep T2M models, to create the necessary training data, supporting HAR applications in few-shot settings.

2.2 Generative Models for 3D Skeleton-based Motion

Realistic human motion synthesis evolved from early VAEs and GANs conditioned on discrete actions [13,20] to highly expressive models. This progress was fueled by large motion capture (MoCap) datasets [19] and richer conditioning signals, advancing from text [21] to even music [34] as a modality. The current state-of-the-art is dominated by Denoising Diffusion Probabilistic Models (DDPMs) [30,35] and masked generative models [11], which achieve impressive fidelity. However, these T2M models, trained for general character animation,

inherently lack the kinematic specificity required for downstream HAR tasks. Our proposed KineMIC framework addresses this domain gap by using a kinetic mining strategy to distill HAR-relevant knowledge from the T2M backbone, enabling its application as a specialized data generator.

2.3 Few-Shot HAR with Generative Models

The high cost of annotated motion data bottlenecks Few-Shot Human Activity Recognition (FSHAR). While classifier-side methods (e.g., metric/meta-learning) [32] are the most popular approach, they are inherently constrained by the limited kinematic diversity of the few-shot support set. A less-explored alternative is deep generative data augmentation. Most notably, recently Fukushi et al. [10] demonstrated this using a GAN with cross-domain regularization. To bypass GAN notorious training instability and explore a new avenue, we propose leveraging modern T2M diffusion models for few-shot motion synthesis. In direct contrast to [10], our approach employs a "semantics-first" matching strategy. By utilizing the rich pre-trained text encoding space, we effectively disentangle the synthesis process from the kinematic limitations of the few-shot 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 goal 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 $\boldsymbol{x} = \{x(j) \in \mathbb{R}^d\}_{j=1}^n$, 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 a large-scale dataset $\mathcal{D}^P$, which is a collection of pairs $(\boldsymbol{x}^P, c)$. Here, $\boldsymbol{x}^P$ represents a motion sequence, and c is its associated rich, free-form, descriptive text caption.
2. A target domain T, defined by a dataset $\mathcal{D}^T$, which is a collection of pairs $(\boldsymbol{x}^T, y)$. Here, $\boldsymbol{x}^T$ 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 subset of the target domain $\bar{T} \subset T$ is available at training time. Our goal is to use the 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; capable of synthesizing novel, class-conditional motion samples from the action set Y. The quality of synthetic samples is measured by their ability to improve classification accuracy on $\mathcal{D}^T$ test split, when used for data augmentation in a HAR classifier training. The core challenge lies in bridging the significant domain gap between $\mathcal{D}^P$ and $\mathcal{D}^T$, which manifests in two ways:

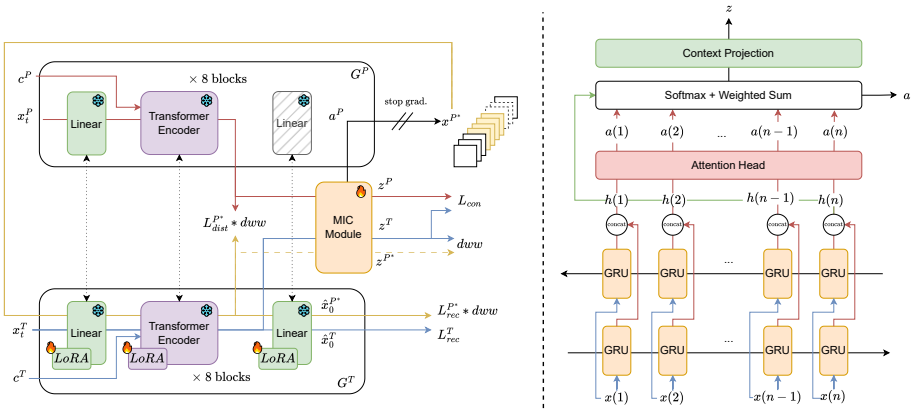

Fig. 2. The KineMIC model. **(Left)** Core information flow (Dashed lines indicate detached gradients. dotted lines denote shared weights): A frozen teacher G^P and a trainable student G^T are linked by the MIC module. The module performs contrastive alignment of latent motion features $(z^P, z^{\bar{T}})$ to identify and extract the most relevant kinematic sub-sequence, the prior window $\boldsymbol{x}^{P^*}$, from the long source motion $\boldsymbol{x}_t^P$. This extracted window is used to fine-tune G^T with the target conditioning c^T as a pseudo-label. **(Right)** The MIC Module: an Attention-enhanced bidirectional GRU encoder that converts the sequence of frame-wise motion tokens into a single context-aware latent vector.

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 may contribute to further disparity.

4 Methodology

We present our framework, named Kinetic Mining In Context (KineMIC) (Figure 2, Figure 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 the foundational knowledge of diverse human movements. In parallel, the student (i.e. target) stream G^T is a trainable copy of G^P. Both models share initialization weights from $\mathcal{D}^P$ pre-training. For the conditioning schema we stick to practices define by [30] and apply a small modification: while the student's text embedding layer remains frozen to leverage the deep semantic space learned by the teacher, we introduce an additional learnable

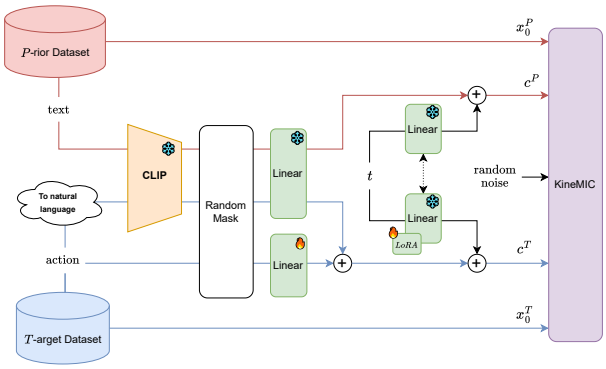

Fig. 3. Input handling and conditioning workflow. Source stream (red) conditioning c^P is derived from CLIP text embeddings and a timestep embedding. Target stream (blue) conditioning c^T uses a text embedding from the action label, a separate learnable action embedding, and its own timestep embedding. Note: Dotted lines denote weight sharing.

action embedding layer tailored to the discrete nature of $\mathcal{D}^T$ labels. Moreover, we employ Low-Rank Adaptation (LoRA) [14] on all layers of the G^T, which keeps frozen the vast majority of pre-training weights while efficiently tuning low rank matrices.

4.1 Kinetic Mining In Context

Soft Positive Search To bridge the semantic domain gap, we introduce a soft positive search strategy. This approach is founded on the inherent one-to-many relationship: a sparse target action label (y) can correspond to numerous specific motions described by rich source captions (c). Our strategy leverages the semantic space of the CLIP text encoder to find "soft" matches between target labels and source captions. We first unify the conditioning modalities by converting each $y \in Y$ into a natural language prompt (e.g., 'side kick' becomes 'a person does a side kick'). Next, we use the CLIP text encoder to embed both target prompts and the descriptive source captions into a shared vector space. Finally, we compute the cross-dataset pairwise cosine similarity and retain the top-k most similar source captions for each target action. This pairing is termed "soft" because it utilizes the depth of CLIP's semantic space, correlating concepts beyond exact keyword matches. This approach is based on the assumption that semantic alignment in the text space indicates a high probability that the source sample contains a sub-sequence correlated with the target sample from a high-level kinematic perspective, thereby guiding the subsequent mining process.

Contrastive Alignment During training, a batch is considered to be a set $\{(\boldsymbol{x}^{\bar{T}}, \boldsymbol{x}^P, y)^i\}_{i=0}^B$, such that y represents the class of $\boldsymbol{x}^{\bar{T}}$ and $\boldsymbol{x}^P$ is sampled

randomly from the pre-computed list of top-k soft positive matches w.r.t. y class. Each motion sample is processed by its respective model, G^T and G^P and its own conditioning signals (Figure 3, Figure 2). Before last linear projections, tokens, which encode frame-wise information, are then processed by the Mining In Context (MIC) module (Figure 2), to obtain latent vectors z^T and z^P. We proceed aligning latent representations through a Soft Nearest Neighbors loss [24,9]. We define the set of local soft positive matches $B^+(i)$ to be the set of all batch indexes j containing class y. The contrastive objective is defined as:

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

This loss is an intuitive choice, as it inherently manages the one-to-many relationship established by the soft positive search.

Kinematic Mining As a consequence of being trained with Equation 1, the MIC module, which is implemented to contain an attention head (Figure 2), will associate higher activations to $\boldsymbol{x}^P$ frames that are the most responsible to produce an alignment with $\boldsymbol{x}^{\bar{T}}$. We leverage this learned attention to perform the mining operation. For a defined soft pair $(\boldsymbol{x}^P, \boldsymbol{x}^{\bar{T}})$ where $\boldsymbol{x}^P$ has length n and $\boldsymbol{x}^{\bar{T}}$ has length m, we let $\boldsymbol{a}^P = \{a^P(k)\}_{k=1}^n$ be the attention scores over $\boldsymbol{x}^P$. The prior window $\boldsymbol{x}^{P^*}$ is the contiguous sub-sequence of length m that maximizes cumulative attention.

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

The extracted $\boldsymbol{x}^{P^*}$ is then treated as new training data for G^T.

4.2 Optimization goal

Denoising Reconstruction Objective The fundamental training objective for G^T is rooted in the denoising diffusion paradigm of MDM [30]. Sampled some random noise, the model is trained to predict the original, "clean" motion signal $\boldsymbol{x}_0$ given a noisy input $\boldsymbol{x}_t$, noised at timestep t. This objective is applied to both the ground-truth target samples $\boldsymbol{x}_0^{\bar{T}}$ and the pseudo-labeled prior windows $\boldsymbol{x}_0^{P^*}$. The reconstruction losses for their noised counterparts are formulated as:

$$L_{rec}^T = ||\boldsymbol{x}_0^{\bar{T}} - G^T(\boldsymbol{x}_t^{\bar{T}}, c^{\bar{T}})||_2^2 \tag{3}$$

$$L_{rec}^{P^*} = ||\boldsymbol{x}_0^{P^*} - G^T(x_t^{P^*}, c^{\bar{T}})||_2^2 \tag{4}$$

where $c^{\bar{T}}$ is the conditioning signal, obtained through a combination of the action ID, a simple natural language conversion of the action label, and the diffusion timestep t (Figure 3). this objective ensures G^T generates motions faithful to both the original target domain and the rich kinematic structures extracted from the source domain.

Window Distillation To ensure the student model internalizes the kinematic specificity from the pre-trained teacher, we introduce a feature-level distillation loss. This objective encourages the student model's internal representations to mimic those of the teacher model for the mined motion windows. Let $\boldsymbol{u} = \{u(1), \ldots, u(n)\}$ denote the pre-projection feature sequence. The features $\boldsymbol{u}^{P^*}$, from student model's output for the prior window $\boldsymbol{x}_t^{P^*}$, are compared against the corresponding features $\boldsymbol{u}^P$ from the teacher model, specifically for the frames associated with the window within the original prior motion $\boldsymbol{x}_t^P$:

$$L_{dist}^{P^*} = \frac{1}{m} \sum_{j=0}^{m-1} ||u^{P^*}(j) - u^P(i+j))||_2^2 \tag{5}$$

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

Dynamic Window Weighting. While soft positive search identifies semantically relevant motions, the extracted window ($\boldsymbol{x}^{P^*}$) isn't guaranteed to be kinematically coherent as the soft pairing relies purely on semantic match in the text encoding space. To address this, we introduce Dynamic Window Weighting (dww). This simple mechanism assesses the quality of each mined window to dynamically modulate its impact on training. After extraction, $\boldsymbol{x}^{P^*}$ is passed through the teacher generator G^T and the MIC module to compute a new latent representation, z^{P^*}. The match quality is then quantified by the cosine similarity between z^{P^*} and the target sample's latent representation, $z^{\bar{T}}$.

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

Crucially, we disable gradient contributions during the z^{P^*} computation. This prevents the MIC module from updating via the weighting process, ensuring it remains an unbiased, semantics-driven kinematic aligner trained solely by Equation 1. The dww scales the window-related losses (Equation 3, Equation 5).

Complete objective The complete optimization goal consists in a weighted sum of the all components.

$$L = \lambda_{rec}^T L_{rec}^T + \lambda_{con} L_{con} + dww * (\lambda_{rec}^{P^*} L_{rec}^{P^*} + \lambda_{dist} L_{dist}) \tag{7}$$

5 Experiments

5.1 Experimental Setup

Given the narrow scope of few-shot skeleton-based action synthesis for HAR, we utilize the experimental setup proposed by [10] for comparison. We define our source pre-train dataset ($\mathcal{D}^P$) as HumanML3D [12] and the target ($\mathcal{D}^T$) as a subset of NTU RGB+D 120 [27], analyzing 'running on spot' (A099), 'side

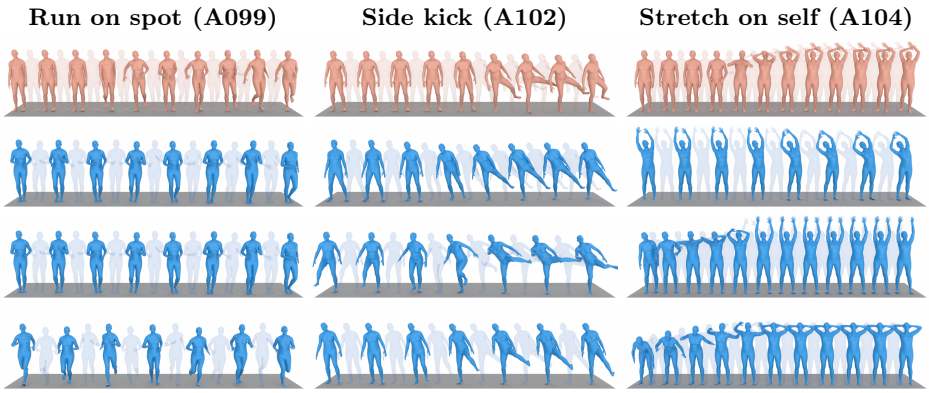

Fig. 4. Qualitative Comparison of Generated Motions. The three columns depict the three target action classes. The first row shows corresponding real ground truth samples from NTU RGB+D 120 extracted through VIBE [17] (in orange), while following rows show diverse samples generated by our KineMIC model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

kick' (A102), and 'stretch on self' (A104). The simplicity of these actions makes them a perfect testbed for evaluating a model's ability to move beyond broad synthesis and capture the kinematic details essential for recognition. Our setup is thus designed to measure how KineMIC overcomes this specificity gap. Following the few-shot protocol, we randomly select only 10 samples per action class (30 total) as our support set ($\mathcal{D}^T$) for framework training. Further details about the experimental setup are available in the supplementary material.

Dataset and Preprocessing. Following [10], we re-estimate 3D skeletons from RGB videos using VIBE [17]. To align the target data with our teacher model's pre-training on HumanML3D [12], we introduce two preprocessing steps: (1) Hand joints are excluded from the SMPL skeletons to match the joint topology; (2) Motion sequences are downsampled from 30 fps to 20 fps to match acquisitions frame rate. All skeletons undergo the HumanML3D normalization pipeline, which centers the root joint, sets feet height to zero, orients the initial pose to the positive Z-axis, and normalizes joint lengths across all frames. For motion representation, we adopt a 263-dimensional representation [30], concatenating: root-relative joint positions, 6D joint rotations, joint velocities, and estimated binary foot-contact labels.

Implementation Details. The pre-trained model G^P is an MDM 8-block transformer [30] trained on HumanML3D [12]. Our target model G^T shares the identical architecture and weights of G^P, with the sole addition of an action embedding layer to its conditioning workflow (Figure 3). We fine-tune G^T using LoRA [14] (rank=16, α=32, dropout=0.1) applied to all transformer layers, training only the low-rank matrices and the new action embedding layer. All

other optimization objective λ parameters are set to 1.0. For the soft positive search, we set the number of nearest neighbors $k = 250$ and a temperature of $\tau = 0.07$. This k value was chosen to effectively balance diversity and semantic relevance in the drawn samples. Search is pre-computed once, resulting in an effective training dataset size of (Number of Shots + k) $\times$ (Number of Classes). We employ classifier-free guidance with guidance scale of 2.5 [30], dropping the Action and Text conditioning modalities each with independent probability of 0.1 during training. Models are trained for 5000 steps using AdamW optimizer (LR $2 \cdot 10^{-5}$) with 100 diffusion steps and cosine noise scheduling. We apply gradient clipping (norm 1.0) for early training stability. Each training step processes all $\mathcal{D}^T$ samples (30 total), pairing each $(\boldsymbol{x}^T, y)$ with a randomly chosen soft-pair $\boldsymbol{x}^P$ drawn from the top-k matches in $\mathcal{D}^P$ relative to class y.

Evaluation Protocol Our evaluation focuses on maximizing synthetic data utility for HAR, centered on recognition accuracy, Multimodality, and Diversity. We employ a ST-GCN classifier [33] for all downstream HAR and generative evaluations, following PYSKL practices and implementation [7]. Each training set is composed of 30 real samples augmented with 1149 synthetically generated motions, both uniformly distributed per class. All experiments are repeated five times with different seeds. For comparison with prior work [10], we report the median top-1 accuracy; internal analyses use mean and standard deviation. Further details are provided in the supplementary material.

5.2 Baselines and Prior Work Comparison

We compare our proposed KineMIC framework against key prior work and critical baselines to contextualize our performance. Results are detailed in Table 1. First, we establish a foundational zero-shot baseline (83.9%) by using the pre-trained MDM model to generate motions from general text prompts describing target classes (e.g., class 'side kick' becomes 'a person does a side kick'), without any fine-tuning. This score confirms that the generalist MDM already posses strong motion knowledge regarding the chosen target actions. When the MDM model is trained from random initialization (73.9%) or by finetuning from a $\mathcal{D}^P$ checkpoint with LoRA (75.5%) we achieve poor results. We attribute this to overfitting of the diffusion model on the very limited data. In contrast, our KineMIC framework makes better use of pre-training knowledge while preventing overfitting, achieving a final accuracy of 86.2%. our results are similar to prior work by [10], but obtained through a substantially different approach, by exploiting T2M datasets. Crucially in Figure 5, we test how synthetic data produced by our model scale when more target data is made available. When the ST-GCN classifier is trained with only real data, the model exhibits poor performances and high variance, especially when little data is provided. On the other hand, when synthetic data is concatenated to available real data, performances are consistently better, and ultimately approach asymptotically the maximum.

Table 1. Comparative Performance on Few-Shot HAR. Median top-1 accuracy on the NTU RGB+D 120 dataset, following the evaluation protocol defined by the prior work of Fukushi et al. [10]. Results marked with † are reported directly from [10]. The up-arrow (↑) indicates that higher is better.

Source	Method	Top-1 Acc (%) ↑
Prior Work†	Real data only (30 samples)†	58.4
	ACTOR† [20]	73.6
	Kinetic-GAN† [3]	81.7
	Fukushi et al.† [10]	86.4
Our Analysis	Real data only (30 samples)	63.1
	MDM (pretrained)	83.9
	MDM (from scratch)	73.9
	MDM (LoRA finetune)	75.5
	KineMIC	86.2
	Ground Truth (all real data)	97.1

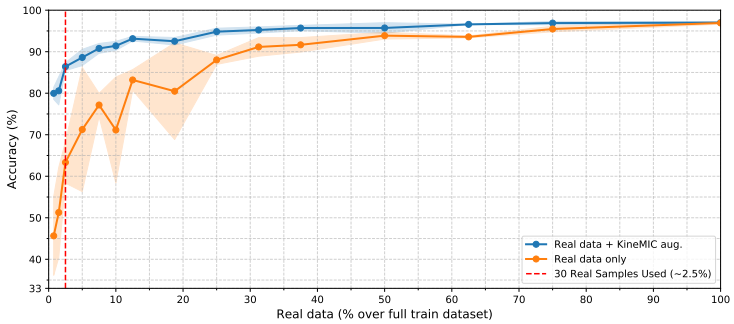

Fig. 5. Synthetic data scalability. When the ST-GCN classifier is trained using only real data training exhibits high variance and poor performances. When concatenating synthetic data from KineMIC to real data, accuracy is consistently much higher and converge asymptotically to the maximum.

5.3 Ablation Study

We conduct a systematic ablation study to evaluate the contribution of each component within our KineMIC framework. Starting from baselines, most importantly, inspecting the pretrained baseline we can notice that while it's associated to large diversity the accuracy score is also associated to a very large error bound. Visual inspection confirms that: while such samples possess great variety, they struggle adhering to the expected kinematics, producing noisy training signals for the HAR classifier. Please refer to the supplementary material for visuals. For our LoRA-finetune baseline, we notice instead that the model struggles with overfitting on the very limited data, producing limited variety. Introducing ablations for KineMIC, our base model performs similarly to the pretrained baseline, while reducing the error bound. While the distillation objective (L_{dist}) doesn't provide expected performance gains. Results enabling the *dww* component alone

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 Equation 1 and Equation 3 respectively. We then incrementally add L_{dist} and *dww*. Note: (↑) denotes that higher values are better.

Method	**Accuracy (%) ↑**	**Diversity ↑**	**Multimodality ↑**
MDM (pretrained)	$85.23^{\pm 3.32}$	$\underline{26.82}^{\pm 1.20}$	$6.10^{\pm 2.42}$
MDM (LoRA finetune)	$76.27^{\pm 1.51}$	$9.45^{\pm 0.39}$	$4.98^{\pm 0.36}$
KineMIC (Base)	$85.21^{\pm 2.24}$	$13.69^{\pm 1.10}$	$10.21^{\pm 0.60}$
+ L_{dist} only	$83.78^{\pm 2.38}$	$22.45^{\pm 1.56}$	$\underline{17.12}^{\pm 2.25}$
+ *dww* only	$\underline{86.41}^{\pm 0.95}$	$14.24^{\pm 1.09}$	$10.58^{\pm 0.77}$
+ L_{dist} + *dww*	$84.94^{\pm 2.37}$	$19.49^{\pm 1.30}$	$13.64^{\pm 0.91}$

are the best ones. This supports the hypothesis that, while similarity in semantics can be a guide towards similarity in kinematics, it's not always the case, and implementing a form of filtering yields to more stable HAR performances.

5.4 Qualitative Evaluation

As illustrated in Figure 4, which displays three generated animations per class, the motions synthesized by KineMIC are coherent with target examples, generalizing to meaningful, semantically related variations. This capacity for plausible diversity is supported by significant gains in both downstream accuracy and diversity metrics (detailed in Table 2). Diversity is particularly evident in the generated samples. For instance, for the 'side kick' class, KineMIC produces a range of motions, including some resembling more "combat-like" kicks, demonstrating an understanding of the broader action concept beyond the sparse few-shot examples. Similarly, for the 'stretch on self' action, our model synthesizes semantically richer variations, such as a stretch with 'hands behind the head' or 'lateral tilts of the torso'.

Motion Composition During our analysis, we discovered a compelling emergent property: when guided by a text prompt distinct from its action-conditioning class, KineMIC synthesizes a motion that effectively fuses the semantics of both. For instance, as shown in Figure 6, 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 overhead'. We observed that this blending is most successful when motions are not kinematically conflicting. Combining an upper-body dominant action ('stretch on self') with a lower-body one ('jumping') yields a coherent result, whereas combining two lower-body dominant actions ('running on spot' and 'jumping') often fails and produces unstable motion. We hypothesize this compositional ability emerges from the synergy between our dual-conditioning scheme (Figure 3) and LoRA, which help preserving both the fine-tuned action and the general capabilities of the pre-trained model,

similar to findings in [25]. While this emergent behavior is not fully robust and is beyond the scope of this work, we report it as a promising future direction.

Fig. 6. 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 Conclusions, limitations and future directions

In this work, we have formally introduced and addressed the challenge of few-shot action synthesis for HAR using T2M priors. We systematically investigated how T2M generative models can serve as powerful baselines but can lack the kinematic specificity required for specialized, atomic action classes. To tackle such limitations, we introduced KineMIC, a teacher-student framework that adapts a pre-trained diffusion model using as few as 10 samples per class. Our major technical contributions are a soft positive search strategy that leverages a shared semantic space to retrieve relevant motion knowledge from the source domain, paired with a kinematic mining strategy. Finally, KineMIC proved to increase downstream classifier performance on the NTU RGB+D 120 few-shot benchmark defined by [10].

The key finding of this work is that while fine-tuning generative models, more specifically diffusion models, on extremely limited data is highly susceptible to overfitting and often yields poor results, T2M datasets can be effectively leveraged. Specifically, when T2M priors are combined with a careful, context-aware filtering strategy, they represent a viable and scalable source of knowledge. This approach successfully mitigates model collapse and expands the kinematic diversity available for few-shot tasks, moving beyond the inherent constraints of the sparse target set.

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 actions, implying a practical limit on mining highly complex or novel atomic movements. 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.

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)
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. Degardin, B., Neves, J., Lopes, V., Brito, J., Yaghoubi, E., Proenca, H.: Generative adversarial graph convolutional networks for human action synthesis (2021)
4. Dentamaro, V., Gattulli, V., Impedovo, D., Manca, F.: Human activity recognition with smartphone-integrated sensors: A survey. Expert Syst. Appl. **246**, 123143 (2024)
5. Do, J., Kim, M.: Skateformer: Skeletal-temporal transformer for human action recognition (2024)
6. 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
7. Duan, H., Wang, J., Chen, K., Lin, D.: Pyskl: Towards good practices for skeleton action recognition (2022)
8. Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B.: Revisiting skeleton-based action recognition (2022)
9. Frosst, N., Papernot, N., Hinton, G.: Analyzing and improving representations with the soft nearest neighbor loss (2019)
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)
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. pp. 2021–2029. ACM (2020). https://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)
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)
16. Khirodkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for human vision models (2024)
17. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5252–5262 (2020). https://doi.org/10.1109/CVPR42600.2020.00530
18. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition (2020)
19. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes (2019)
20. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae (2021)
21. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions (2022)
22. Plizzari, C., Cannici, M., Matteucci, M.: Spatial Temporal Transformer Network for Skeleton-Based Action Recognition, pp. 694–701. Springer International Publishing (2021). https://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)
24. Salakhutdinov, R., Hinton, G.: Learning a nonlinear embedding by preserving class neighbourhood structure. In: Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics. pp. 412–419 (2007)
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)
26. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior (2023)
27. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+d: A large scale dataset for 3d human activity analysis (2016)
28. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition (2019)
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)
30. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model (2022)
31. Wang, Y., Sun, Y., Patel, P., Daniilidis, K., Black, M.J., Kocabas, M.: Prompthmr: Promptable human mesh recovery (2025)
32. Wanyan, Y., Yang, X., Dong, W., Xu, C.: A comprehensive review of few-shot action recognition (2025)
33. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition (2018)

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)
35. Zhao, M., Liu, M., Ren, B., Dai, S., Sebe, N.: Modiff: Action-conditioned 3d motion generation with denoising diffusion probabilistic models (2023)