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

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 remains a critical bottleneck for skeletal-based Human Activity Recognition (HAR). Although Text-to-Motion (T2M) generative models offer a compelling, scalable source of synthetic data, their training objectives, which emphasize general artistic motion, and dataset structures fundamentally differ from HAR's requirements for kinematically precise, class-discriminative actions. This disparity creates a significant domain gap, making generalist T2M models ill-equipped for generating motions suitable for HAR classifiers. To address this challenge, we propose KineMIC (Kinetic Mining In Context), a transfer learning framework for few-shot action synthesis. KineMIC adapts a T2M diffusion model to an HAR domain by hypothesizing that semantic correspondences in the text encoding space can provide soft supervision for kinematic distillation. We operationalize this via a kinetic mining strategy that leverages CLIP text embeddings to establish correspondences between sparse HAR labels and T2M source data. This process guides fine-tuning, transforming the generalist T2M backbone into a specialized few-shot Action-to-Motion generator. We validate KineMIC using HumanML3D as the source T2M dataset and a subset of NTU RGB+D 120 as the target HAR domain, randomly selecting just 10 samples per action class. Our approach generates significantly more coherent motions, providing a robust data augmentation source that delivers a +23.1% accuracy points improvement. Animated illustrations and supplementary materials are available at https://lucazzola.github.io/publications/kinemic.

Keywords: Human Motion Synthesis, Few-Shot Action-to-Motion Generation, Human Activity Recognition, Synthetic Data Generation,

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 [35]. However, the

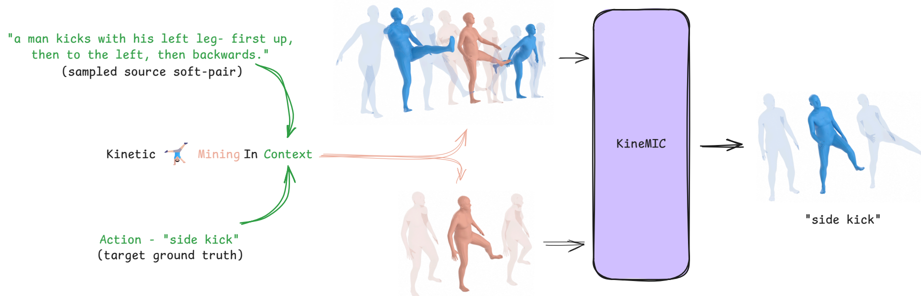

Fig. 1. Kinetic Mining in Context. The target action sample (bottom) is used to contextualize the search within the large source data sample (top), establishing soft pairs. The mining operation identifies a kinematically relevant segment (in orange) from the source data.

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 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 [34]. 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 [12,21], which has, in turn, led to the development of powerful deep models [11,12,32] that can be employed as general motion priors for other generative tasks [27,28,31]. 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 [16,33], 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 pre-trained 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, Fig. 1). Our method employs teacher-student architecture, wherein a frozen teacher, pre-trained 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 [25] 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 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 [32] 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 [6,29]. A paradigm shift occurred with the introduction of Graph Convolutional Networks (GCNs) [35], which re-conceptualized the skeleton as a graph to model spatio-temporal dependencies. This led to rapid advancements using adaptive graph structures [30], refined GCNs [1,2,7,19], and more recently, specialized architectures like 3D convolutional networks [8] and Transformers [5,24]. 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 [13,22] to highly expressive models. This progress was fueled by large motion capture datasets [12,21] and richer conditioning signals, advancing from text [23] to even music [36] as a modality. The current state-of-the-art is dominated by Denoising Diffusion Probabilistic Models (DDPMs) [32,37] and masked generative models [11], which achieve impressive generative 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) [34] 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 [18,20]. 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 CLIP [25] text encoding space, we 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 $\mathbf{x} = \{x(j) \in \mathbb{R}^d\}_{j=1}^{n}$, 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 $(\mathbf{x}^P, w)$. Here, $\mathbf{x}^P$ represents a motion sequence, and w is its associated rich, free-form, descriptive text caption, from a set W.
2. A target domain T, defined by a dataset $\mathcal{D}^T$, which is a collection of pairs $(\mathbf{x}^T, y)$. Here, $\mathbf{x}^T$ 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:

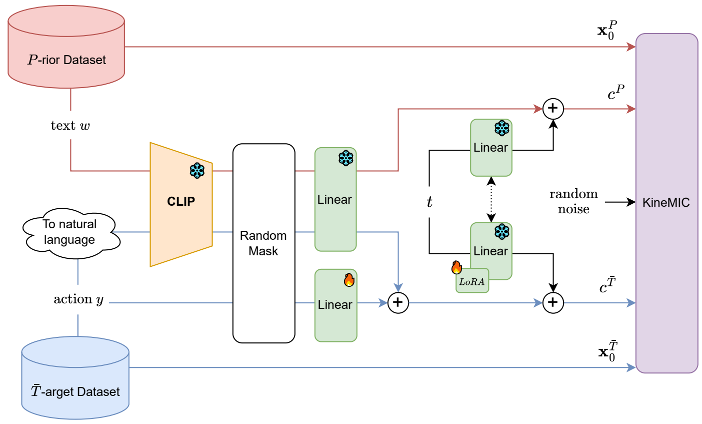

Fig. 2. Source stream (red): c^P from CLIP text embeddings + timestep. **Target stream (blue):** c^T from action-label text embedding + learnable action embedding + timestep. Dotted lines denote weight sharing.

1. Conditioning modalities differ fundamentally (semantic gap). The source domain uses high-variance, free-form text, while the target domain employs discrete action labels.
2. There is a discrepancy in the motion distributions (kinematic gap). 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 introduce KineMIC, our proposed architecture (Fig. 2 and 3). KineMIC adapts a general T2M diffusion prior into a few-shot A2M generator, by combining: semantic retrieval, kinematic alignment, kinematic mining, and student adaptation, all within a unified teacher-student framework. Below, we describe each component in detail, following the order in which information flows through the model.

4.1 Teacher-Student Architecture

KineMIC builds on two identical MDM [32] models: a frozen teacher G^P and a trainable student G^T. Both models retain MDM's core structure and diffusion paradigm: predicting clean motion $\mathbf{x}_0$ from noised input $\mathbf{x}_t$ under a conditioning signal c. We refer to [32] for full diffusion process details. To preserve prior knowledge, G^T is fine-tuned with Low-Rank Adaptation (LoRA) [14], which updates only a small fraction of weights through low-rank matrices while keeping the vast majority of pre-trained parameters frozen. Furthermore, while the student

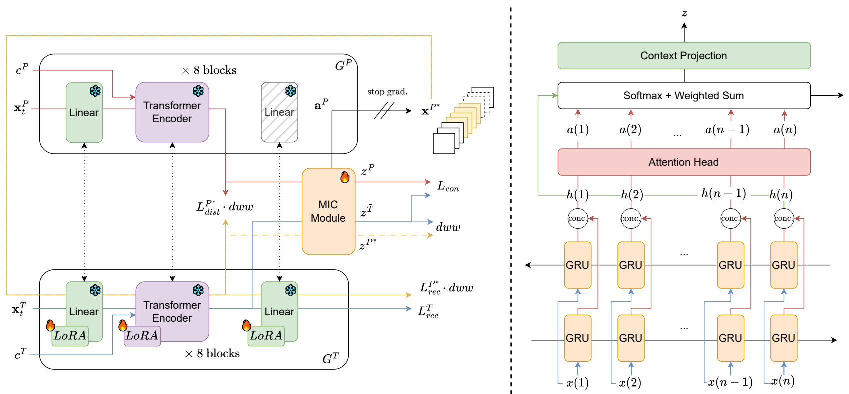

Fig. 3. (Left) Core pipeline: dashed lines show detached gradients; dotted lines show weight sharing. MIC receives frame-wise tokens from teacher (G^P) and student (G^T) streams, producing latents z^P and $z^{\bar{T}}$ for contrastive alignment and window mining. **(Right)** MIC module: attention-enhanced biGRU encoder aggregating frame-wise motion tokens into context-aware latents.

inherits the frozen text embedding layer used in T2M, we introduce a learnable action embedding (Fig. 2) that enables G^T specialization to discrete HAR labels.

4.2 Semantic Alignment via Soft Positive Search

Target labels in $\mathcal{D}^T$ lack the descriptive richness of T2M captions in $\mathcal{D}^P$, creating a one-to-many mapping (single label → multiple text prompts). We bridge this via CLIP-based [25] semantic retrieval. For each y action label:

1. Turn y into natural-language (e.g. *'side kick'* → *'a person does a side kick'*).
2. Encode both the target prompt and all captions in $\mathcal{D}^P$.
3. Compute pairwise cosine similarities and retain top-k closest samples.

The resulting set of retrieved captions defines the *soft positive matches* for each target label. This exploits CLIP's semantic space to correlate concepts even without explicit keyword overlap, modeling the one-to-many label-to-prompt mapping by identifying multiple semantically compatible descriptions that correspond to distinct yet potentially related motion patterns.

4.3 Kinematic Alignment via MIC and Contrastive Learning

Semantic similarity does not guarantee kinematic relevance. Moreover, $\mathcal{D}^P$ motions, designed for motion understanding through natural language, span longer time periods than atomic $\mathcal{D}^T$ actions designed for activity recognition, creating a structural disparity (*see supplementary material for sequence-length analysis*). We address both via the Mining In Context (MIC) module, which learns frame-level alignments for targeted window extraction.

MIC processing : For each soft pair of two motion sequences $(\mathbf{x}^P, \mathbf{x}^{\bar{T}})$:

1. A random timestep t is sampled, and the forward diffusion process produces noised sequences $\mathbf{x}_t^P$ and $\mathbf{x}_t^{\bar{T}}$, starting from the same random noise.
2. These are processed by G^P and G^T to obtain frame-wise feature tokens.
3. The tokens from both streams are fed into the MIC module, a bidirectional GRU with attention, producing context-aware latent vectors z^P and $z^{\bar{T}}$.

These latent representations summarize the motion dynamics of each sample in a manner sensitive to both temporal structure and global context.

Contrastive Alignment MIC is trained to align $z^{\bar{T}_i}$ with z^{P_j} from soft-positive pairs sharing the same target class y. Given a training batch of size B, we use the Soft Nearest Neighbors loss [26,9], which naturally accommodates the one-to-many soft positives:

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

where $B^+(i)$ contains batch indices j whose z^{P_j} was retrieved as a soft positive for the same class y as $z^{\bar{T}_i}$. This enforces kinematic alignment within semantically-matched pairs via contrastive learning, while MIC's attention enables subsequent window extraction.

4.4 Kinematic Mining

With MIC trained via contrastive loss, its attention mechanism learns to focus on kinematically relevant frames in $\mathbf{x}^P$ to minimize the enforced alignment loss with $\mathbf{x}^{\bar{T}}$. Let $\mathbf{a}^P = \{a_k^P\}_{k=1}^n$ denote attention weights over $\mathbf{x}^P$ of length n, and let m be the length of $\mathbf{x}^{\bar{T}}$. We extract the most relevant contiguous subsequence (the "prior window") as the m-frames segment maximizing cumulative attention:

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

The identified prior window $\mathbf{x}^{P^*}$ serves as a pseudo-labeled training sample for G^T under the same target conditioning c^T (Fig. 3).

4.5 Optimization of Multi-Objective Loss Function

Reconstruction Loss. The student G^T is trained to reconstruct both target motion and the extracted prior window:

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

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

These losses ensure that G^T learns both target-specific dynamics and source-domain motion structures consistent with the target action. Notice that both samples are processed enforcing the same $c^{\bar{T}}$ conditioning signal.

Window Distillation. We distill the teacher's internal representations relative to the mined window into the student. Let $\mathbf{u}^P$ and $\mathbf{u}^{P^*}$ denote pre-projection feature sequences for G^P and G^T, respectively. For a prior window starting at frame index i, the distillation loss is:

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

Dynamic Window Weighting. Mining is not perfect: a caption containing "kick" word may retrieve front-kick samples for a side-kick target, yielding sub-optimal kinematic windows despite semantic correlation. To modulate their influence, we compute a window quality score:

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

where z^{P^*} is computed with gradients detached, ensuring MIC remains trained exclusively by Eq. 1. The score dww scales both reconstruction (Eq. 4) and distillation (Eq. 5) losses, giving higher weight to more reliable windows.

Complete Multi-Objective Loss Function. The final optimization objective is a weighted combination of all components:

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

5 Experiments

5.1 Experimental Setup

Given the narrow scope of few-shot skeleton-based action synthesis for HAR, we position ourselves in the same experimental setup proposed by [10] to contextualize our results. 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 [29], analyzing actions: *'running on spot'* (A099), *'side kick'* (A102), and *'stretch on self'* (A104) from cross subject benchmark. Given the simplicity of these actions, generalist T2M models can reasonably generate them from broad text prompts. However, this setting precisely tests the challenge of capturing fine, class-discriminative kinematics critical for HAR. We choose these actions to evaluate KineMIC's ability to close this kinematic specificity gap. Following the few-shot protocol, we randomly select only 10 samples per action class (30 total) as our support set ($\mathcal{D}^{\bar{T}}$) for framework training. *Further details in supplementary material.*

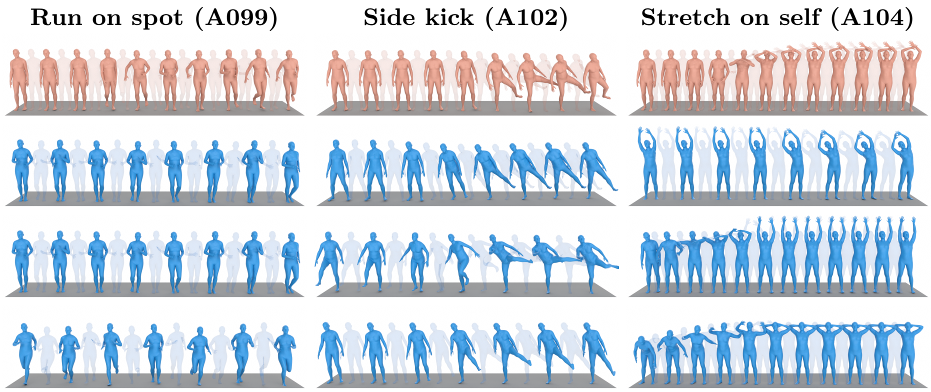

Fig. 4. Qualitative Comparison of Generated Motions. The three columns depict the three target action classes. The first row shows corresponding real ground truth samples from NTU RGB+D 120 extracted through VIBE [17] (in orange), while following rows show diverse samples generated by our KineMIC model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

Dataset and preprocessing Following [10], 3D skeletons are re-estimated from RGB with VIBE [17]. To align NTU-derived target data with the HumanML3D pretraining [12] we apply two preprocessing steps: (1) drop hand joints from SMPL skeletons to match joint topology; (2) downsample sequences from 30→20 fps. All skeletons then follow the HumanML3D normalization pipeline: root centering, feet height set to zero, initial pose oriented to +Z, and joint-length normalization across frames. Motions are represented with the 263-dim vector used in MDM [32], concatenating root-relative joint positions, 6D joint rotations, joint velocities, and binary foot-contact flags.

Implementation Details The pre-trained model G^P is an MDM 8-block transformer encoder [32] 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 (Fig. 2) and the MIC module. 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, the MIC module 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 [32], 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 $(\mathbf{x}^{\bar{T}}, y)$ with a randomly chosen soft-pair $(\mathbf{x}^P, w)$ 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, Diversity (motion distribution variability) and MultiModality (per-prompt variance). We employ a ST-GCN classifier [35] for all downstream HAR and generative evaluations, following PYSKL practices and implementation [7]. Each training set is composed of 30 real samples augmented with 1152 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 in supplementary material.*

5.2 Baselines and Prior Work Comparison

In Table 1 we compare our proposed KineMIC framework against key prior work and critical baselines to contextualize performance. First, we establish a foundational baseline using the pre-trained MDM model to generate motions directly from general text prompts describing our target classes (e.g., converting class *'side kick'* to *'a person does a side kick'*), without any fine-tuning. This achieves 83.9%, confirming that the generalist MDM already possesses substantial motion knowledge relevant to our chosen HAR actions and providing a strong starting point. However, training MDM from random initialization (73.9%) or fine-tuning from a $\mathcal{D}^P$ checkpoint with LoRA (75.5%) both result in significantly lower performance. We attribute this to severe overfitting on the extremely limited few-shot data, causing the diffusion model to collapse. In stark contrast, our KineMIC framework better leverages pre-training knowledge while preventing this overfitting, achieving 86.2%. This gets close to the performance reported by [10], but through a substantially different approach that leverages T2M diffusion priors rather than GAN-based regularization. Relative to real-data-only training on $\mathcal{D}^{\bar{T}}$ (63.1%), KineMIC delivers a substantial +23.1 percentage point improvement. Crucially, Figure 5 demonstrates scalability: real-data-only training shows poor performance and high variance in low-data regimes, while KineMIC augmentation consistently improves accuracy and asymptotically approaches full $\mathcal{D}^T$ training performance.

5.3 Ablation Study

In Table 2 we conduct a systematic ablation study to evaluate the contribution of each component within our KineMIC framework. Starting from baselines, the pre-trained MDM shows highest diversity (26.82) but largest error bounds (85.23±3.32%), as its generic samples deviate from desirable $\mathcal{D}^T$ kinematics despite visual variety (*75 samples shown in supplement material*), producing

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† [22]	73.6
	Kinetic-GAN† [3]	81.7
	Fukushi et al.† [10]	<u>86.4</u>
Our Analysis	Real data only (30 samples)	63.1
	MDM (pre-trained)	83.9
	MDM (from scratch)	73.9
	MDM (LoRA fine-tune)	75.5
	KineMIC	86.2

Table 2. Ablation Study. We evaluate the contribution of each component to our framework. Baselines are pre-trained MDM and LoRA fine-tuned MDM. Our core KineMIC (Base) model builds on this by adding Eq. 1 and 3 respectively. We then incrementally add L_{dist} and dww. (↑) denotes that higher values are better.

Method	**Accuracy (%) ↑**	**Diversity ↑**	**MultiModality ↑**
MDM (pre-trained)	$85.23^{\pm 3.32}$	$\underline{26.82^{\pm 1.20}}$	$6.10^{\pm 2.42}$
MDM (LoRA fine-tune)	$76.27^{\pm 1.51}$	$9.45^{\pm 0.39}$	$4.98^{\pm 0.36}$
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	$84.94^{\pm 2.37}$	$19.49^{\pm 1.30}$	$13.64^{\pm 0.91}$

noisy training signals for the downstream classifier. In contrast, our LoRA fine-tuning baseline overfits on the extremely limited few-shot data, collapsing diversity to 9.45. Introducing KineMIC components, our base model (contrastive alignment + reconstruction) matches pre-trained accuracy while reducing variance (85.21±2.24%). Interestingly, distilling teacher features (L_{dist}) actually reduces accuracy (83.78%) despite achieving peak multimodality (17.12). This suggests L_{dist} enforces the teacher's generalist T2M behavior rather than promoting the HAR-specific kinematic consistency we seek. Ultimately, enabling dynamic window weighting (dww) alone produces the best, most consistent results (86.41±0.95%). This strongly supports our core hypothesis: while semantic similarity provides a valuable guide toward kinematic alignment, it is not always reliable and requires intelligent quality filtering to deliver stable HAR performance.

5.4 Qualitative Evaluation

We conduct a comprehensive visual assessment of 75 generated samples from our best KineMIC model, alongside 75 outputs from the pre-trained MDM base-

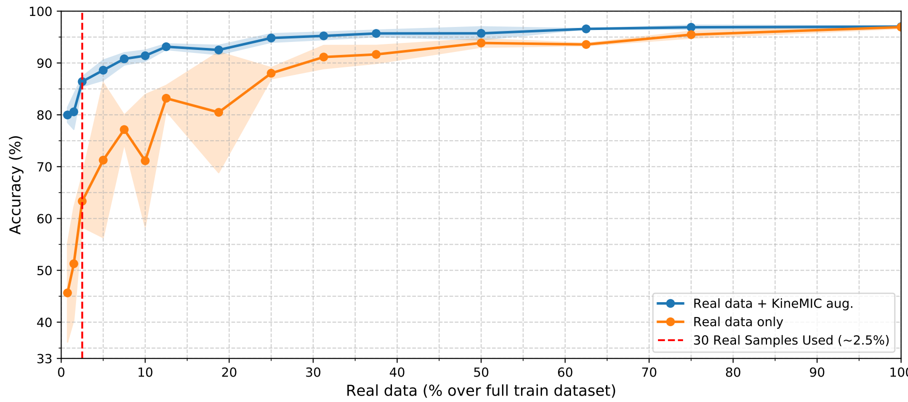

Fig. 5. Synthetic data scalability. When the ST-GCN classifier is trained using only real $\mathcal{D}^{T}$ data, training exhibits high variance and poor performances. When synthetic data from KineMIC is concatenated, accuracy is consistently much higher.

line and all 30 real $\mathcal{D}^{\bar{T}}$ samples (*see supplementary material*). Figure 4 presents select examples, demonstrating KineMIC's ability to produce motions coherent with target kinematics while exhibiting rich, semantically meaningful variations, as supported by ablation gains in accuracy and diversity (Tab. 2). This diversity appears across action classes. For the *'side kick'* class, KineMIC generates motions ranging from $\mathcal{D}^{\bar{T}}$ replicas to dynamic combat-style kicks, demonstrating nuanced understanding beyond the sparse $\mathcal{D}^{\bar{T}}$ data. The *'stretch on self'* action produces enriched variations including stretches with hands behind the head or lateral torso tilts, while *'run on spot'* yields natural pacing and arm swing differences. Unlike pre-trained MDM's generic variability, KineMIC outputs more consistently align with desirable $\mathcal{D}^{T}$ high level kinematics.

Motion Composition KineMIC also exhibits an emergent compositional behavior when guided by a text prompt that differs from its action-conditioning class. When conditioned on *'stretch on self'* action id but prompted with the text *'a person is jumping'*, for example, the model generates a plausible animation of a figure *'jumping while stretching its arms overhead'* (Fig. 6). We observe this property to be most successful when the two controls are not kinematically conflicting: combining an upper-body dominant action (*'stretch on self'*) with a lower-body one (*'jumping'*) produces coherent motions, whereas combining two lower-body dominant actions (*'running on spot'* and *'jumping'*) often leads to unstable results. This compositional ability likely emerges from the synergy between our dual-conditioning scheme (Fig. 2) and LoRA, which together help preserve both the fine-tuned action behavior and the general capabilities of the pre-trained model, in line with observations from [27]. Although this behavior is not fully robust and a detailed analysis is beyond the scope of this work, it points to a promising avenue for future exploration.

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 inputs into a coherent, novel animation.

6 Conclusions, limitations and future directions

In this work, the challenge of few-shot action synthesis for HAR using T2M priors is formally introduced and addressed. The study shows that T2M generative models can provide a strong baseline, yet often lack the kinematic specificity required for specialized, atomic action classes. To tackle these limitations, KineMIC, a teacher–student framework that adapts a pre-trained diffusion model using as few as 10 samples per class, is introduced. The main technical contributions are a soft positive search strategy that leverages a shared semantic space to retrieve relevant motion knowledge from the source domain, combined with a kinematic mining strategy. Empirically, KineMIC improves downstream classifier performance on the NTU RGB+D 120 few-shot benchmark [10]. The key finding is that T2M priors represent a viable approach for few-shot HAR: naive fine-tuning leads to collapse (73.9-75.5%), whereas KineMIC's context-aware mining delivers stable +23.1 percentage point gains over real-data-only baselines (63.1% $\rightarrow$ 86.2%).

Despite these advances, KineMIC has conceptual and practical limitations that point to future research directions. The primary constraint lies in the core assumption that semantic correspondence is a reliable proxy for kinematic relevance during the mining process. This assumption does not always hold; for instance, text encoders may assign high similarity to *'punch'* and *'kick'* due to shared concepts such as *'fighting'*, despite their substantial kinematic disparity, which increases the importance of effective filtering. As a result, 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 when mining highly complex or novel atomic movements. Furthermore, prompt augmentation strategies were not explored. Integrating such techniques is expected to enhance the semantic richness of the target conditioning, a critical factor for fully exploiting the potential of T2M priors.

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. Leng, Z., Iyer, A., Plotz, T.: Scaling human activity recognition: A comparative evaluation of synthetic data generation and augmentation techniques (2025)
19. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition (2020)
20. Lupion, M., Cruciani, F., Cleland, I., Nugent, C., Ortigosa, P.: Data augmentation for human activity recognition with generative adversarial networks. IEEE Journal of Biomedical and Health Informatics **28**(4), 2350–2361 (2024). https://doi.org/10.1109/JBHI.2024.3364910
21. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes (2019)
22. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae (2021)
23. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions (2022)
24. 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
25. 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)
26. 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)
27. 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)
28. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior (2023)
29. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+d: A large scale dataset for 3d human activity analysis (2016)
30. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition (2019)
31. 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)
32. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model (2022)
33. Wang, Y., Sun, Y., Patel, P., Daniilidis, K., Black, M.J., Kocabas, M.: Prompthmr: Promptable human mesh recovery (2025)
34. Wanyan, Y., Yang, X., Dong, W., Xu, C.: A comprehensive review of few-shot action recognition (2025)
35. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition (2018)
36. 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)
37. Zhao, M., Liu, M., Ren, B., Dai, S., Sebe, N.: Modiff: Action-conditioned 3d motion generation with denoising diffusion probabilistic models (2023)