Frequency-Semantic Enhanced Variational Autoencoder for Zero-Shot Skeleton-based Action Recognition

Wenhan Wu[1], Zhishuai Guo[2], Chen Chen[3], Hongfei Xue[1], Aidong Lu[1]
[1]University of North Carolina at Charlotte, Department of Computer Science
[2]Northern Illinois University, Department of Computer Science
[3]Center for Research in Computer Vision, University of Central Florida
{wwu25, hongfei.xue, aidong.lu}@charlotte.edu, zguo@niu.edu, chen.chen@crcv.ucf.edu

Abstract

*Zero-shot skeleton-based action recognition aims to develop models capable of identifying actions beyond the categories encountered during training. Previous approaches have primarily focused on aligning visual and semantic representations but often overlooked the importance of fine-grained action patterns in the semantic space (e.g., the hand movements in drinking water and brushing teeth). To address these limitations, we propose a **F**requency-**S**emantic Enhanced **V**ariational **A**uto**e**ncoder (**FS-VAE**) to explore the skeleton semantic representation learning with frequency decomposition. FS-VAE consists of three key components: 1) a frequency-based enhancement module with high- and low-frequency adjustments to enrich the skeletal semantics learning and improve the robustness of zero-shot action recognition; 2) a semantic-based action description with multilevel alignment to capture both local details and global correspondence, effectively bridging the semantic gap and compensating for the inherent loss of information in skeleton sequences; 3) a calibrated cross-alignment loss that enables valid skeleton-text pairs to counterbalance ambiguous ones, mitigating discrepancies and ambiguities in skeleton and text features, thereby ensuring robust alignment. Evaluations on the benchmarks demonstrate the effectiveness of our approach, validating that frequency-enhanced semantic features enable robust differentiation of visually and semantically similar action clusters, thereby improving zero-shot action recognition. Our project is publicly available at: https://github.com/wenhanwu95/FS-VAE.*

1. Introduction

Human action recognition has gained significant attention in computer vision due to its wide-ranging applications, including surveillance [39, 40], human-computer interaction [20, 33], and automated driving [20, 33]. Among the various modalities, skeleton-based action recognition stands out for its robustness to environmental variations, as it focuses on 3D structural poses derived from human joints. Although conventional supervised approaches [6, 9, 18, 21, 24, 30, 44, 45, 47] have achieved remarkable success in this domain, their reliance on extensive labeled data limits their performance to diverse and unseen action categories. Zero-shot skeleton-based action recognition (ZSSAR) addresses this challenge by enabling models to recognize unseen actions using knowledge from seen action categories and semantic descriptions.

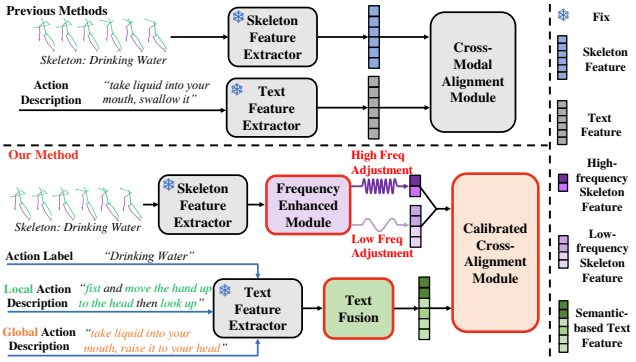


Figure 1. The overall design of our frequency-semantic enhanced variational autoencoder for zero-shot skeleton action recognition.

Existing ZSSAR methods [7, 15, 19, 27, 28, 50, 51] primarily align skeleton features with text embeddings within a shared latent space to enhance generalization. Specially, [15, 27, 28] employ Variational Auto Encoder (VAE) [22] as the training framework to learn structured and generalizable latent representations. However, these VAE-based approaches often yield less semantic information due to their coarse feature representations. Skeleton sequences, unlike raw videos, lack detailed appearance cues, making it difficult to encode the fine-grained semantics. As a result, critical motions (e.g., nuanced limb or hand movements) are often underrepresented, limiting the models' ability to

capture semantic distinctions between similar actions such as "drinking" and "eating". Additionally, skeleton data are inherently ambiguous due to occlusions and variations in camera viewpoints, further complicating motion interpretation. Traditional cross-modal alignments often treat all skeleton-text pairs as equally reliable, ignoring the uncertainty in skeleton representations. Since skeleton features can be noisy and ambiguous, rigid alignment [15, 27, 28] with text embeddings may lead to misalignment and degraded generalization in zero-shot scenarios.

Driven by these concerns, we raise two fundamental questions: ***Q1****: How can skeletal semantics be enriched to enhance the generalization of learned features?* ***Q2****: How can cross-modal alignment be improved by effectively leveraging enriched semantics?* To address these challenges, *we design a framework that enhances action semantics with frequency-based modeling and semantic action descriptions. Additionally, a calibrated cross-modal alignment module is proposed to bridge modality gaps, enabling robust zero-shot recognition of both global and fine-grained patterns.*

Firstly, our approach introduces a Frequency Enhanced Module, which employs the Discrete Cosine Transform (DCT) [2] to transform and enhance skeleton motions in the frequency domain. This decomposition enables a structured enhancement strategy, where low-frequency components (overall semantic structure of actions) and high-frequency components (fine-grained motion details) are selectively refined. Specifically, low-frequency coefficients undergo a progressively diminishing amplification effect to strengthen the global motion representation without distorting structural integrity. Meanwhile, high-frequency coefficients are subjected to an adaptive attenuation mechanism that gradually reduces their magnitude without excessive suppression. This adjustment allows us to preserve subtle actions, such as limb movements and micro-gestures, and simultaneously mitigates the influence of high-frequency noise caused by skeletal jitter. The enhancement not only enriches the semantic representations but also improves the model's robustness against noise and irrelevant variations.

Secondly, our framework incorporates a Semantic-based action Description (SD) mechanism to generate embeddings that capture both localized and global action semantics, enhancing cross-modal alignment in ZSSAR. This allows the model to leverage semantic consistency for recognizing unseen actions without direct supervision. The SD consists of Local action Description (LD), which encodes fine-grained motion details, and Global action Description (GD), which represents overall body posture and movement patterns. For example, LD highlights hand movement in "drinking water", while GD captures body coordination and the sequential flow of motion. This structured description ensures that both detailed actions and contextual motion are well-represented, enabling a precise alignment with frequency-enhanced skeleton features.

Thirdly, a calibrated cross-alignment loss is proposed for semantic embeddings with frequency-enhanced skeleton features, addressing modality gaps and skeleton ambiguities in ZSSAR tasks. It minimizes the disparity between true skeleton-semantic pairs while mitigating mismatched pairs. Unlike conventional uniform alignment losses, the calibrated loss employs a sigmoid-based distance measure to dynamically balance contributions from positive and negative pairs, ensuring robust learning even in the presence of noisy or ambiguous skeleton data. By integrating the frequency-enhanced structure of skeleton and text features, the loss further reinforces cross-modal correspondence. Actions with overlapping semantics, such as "drinking" and "eating", can be effectively distinguished by leveraging fine-grained details (e.g., hand trajectories) and global patterns (e.g., body movements). This approach mitigates overfitting to noisy alignments and reduces modality-specific biases, improving generalization to unseen actions in zero-shot scenarios. The overall method of our FS-VAE is illustrated in Fig. 1, and the contributions are as follows:

- We propose a **Frequency Enhanced Module** that employs Discrete Cosine Transform (DCT) to decompose skeleton motions into high- and low-frequency components, allowing adaptive feature enhancement to improve semantic representation learning in ZSSAR.
- We introduce a novel **Semantic-based action Description (SD)**, comprising Local action Description (LD) and Global action Description (GD), to enrich the semantic information for improving the model performance.
- A **Calibrated Cross-Alignment Loss** is proposed to address modality gaps and skeleton ambiguities by dynamically balancing positive and negative pair contributions. This loss ensures robust alignment between semantic embeddings and skeleton features, improving the model's generalization to unseen actions in ZSSAR.
- Extensive experiments on benchmark datasets demonstrate that our framework significantly outperforms state-of-the-art methods, validating its effectiveness and robustness under various seen-unseen split settings.

2. Related Works

2.1. Zero-Shot Skeleton Action Recognition

Traditional ZSSAR methods focus mainly on mapping skeleton features and semantic embeddings into a shared latent space for alignment [15, 19, 43, 50]. These approaches utilize techniques such as visual-textual correlation learning and adversarial training to reduce the modality gap. Recent works have explored enhanced feature representations and multi-modal alignment strategies to improve performance. For example, [51] leverages part-based feature

modeling to address prompting and partitioning issues in alignment, while [28] introduces semantic attention mechanisms to highlight irrelevant and related semantic features. Despite these advances, existing methods often overlook the semantics of frequency-domain features in capturing both fine-grained motions and global action patterns. Moreover, ambiguities in skeleton representations and noisy or mismatched skeleton-text pairs remain significant challenges.

Unlike prior works that focus solely on the spatial and temporal domain, our work differs fundamentally by leveraging frequency decomposition to model and enhance skeleton motions in the frequency domain, capturing both fine-grained motion variations and overarching action structure to provide richer skeletal semantics. Furthermore, our calibrated cross-alignment loss explicitly addresses skeleton-text ambiguities and modality gaps, ensuring robust alignment in zero-shot scenarios.

2.2. Skeleton-based Frequency Representation Learning

Pose-based approaches aim to directly extract motion patterns from human poses for applications such as motion prediction [25], pose estimation [49], and action recognition [9]. These methods rely on representations from the pose space, which naturally encode spatial structural relationships and temporal motion dependencies. However, effectively integrating these spatio-temporal aspects into a unified framework remains a significant challenge.

Recent research has taken advantage of frequency domain transformations to encode temporal information [3] compactly and smoothly. Studies such as [12, 26, 32, 42] utilize DCT to convert temporal motion signals into the frequency domain, facilitating frequency-specific representation learning. The decomposition of motion signals into high- and low-frequency components enables a fine-grained action analysis, preserving both subtle motion details and global movement patterns. Despite the success of frequency-based modeling, its application in skeleton-based action recognition remains limited and has not yet been explored in the context of zero-shot learning (ZSL). For example, [4] employed a wavelet transform-based approach to disentangle salient and subtle motion features, targeting fine-grained action recognition. Similarly, [44] proposed a frequency-aware transformer that enhances discriminative feature learning for fully supervised action recognition.

In contrast, our approach pioneers the use of DCT in ZSSAR, leveraging its ability to effectively enhance and redistribute motion signals across frequency coefficients. This ensures a robust representation of global motion patterns while preserving fine-grained movement details without amplifying noise. Additionally, a semantic-based action description further enriches action semantics by capturing both localized and holistic action patterns, bridging the semantic gap between textual and skeletal features.

3. FS-VAE: Frequency-Semantic Enhanced Variational Autoencoder

Our goal is to recognize actions from unseen categories using only seen class knowledge and their semantic representations. We adopt a generative VAE framework [15, 27, 28] to learn the skeleton and semantic cross-model features, which are used to generate unseen class representations in latent space. To further enhance generalization in ZSL, we propose a frequency-semantic enhanced framework that refines skeletal inputs through frequency decomposition, preserving essential motion patterns while improving alignment with semantic features. Below, we introduce the model details in FS-VAE.

3.1. Problem Formulation

Zero-shot skeleton-based action recognition (ZSSAR) aims to classify actions from unseen categories using knowledge from seen categories. A skeleton dataset is represented as $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{C}_i, \mathbf{A}_i)\}_{i=1}^{N}$, where $\mathbf{X}_i \in \mathbb{R}^{J\times 3\times F\times M}$ denotes the skeleton sequence of the i-th sample, composed of 3D joint coordinates over F frames for J joints and M subjects. The corresponding action category is $\mathbf{C}_i$, and $\mathbf{A}_i$ represents the GPT-generated semantic description. The dataset is partitioned into a training set $\mathcal{D}_{\text{tr}}^{s}$ with samples of seen categories $\mathcal{C}_s$, and two disjoint test sets: $\mathcal{D}_{\text{te}}^{u}$ for unseen categories $\mathcal{C}_u$ and $\mathcal{D}_{\text{te}}^{s}$ for seen categories. The category sets satisfy $\mathcal{C} = \mathcal{C}_s \cup \mathcal{C}_u$ and $\mathcal{C}_s \cap \mathcal{C}_u = \emptyset$, each action category is associated with $\mathbf{A}_i$.

The objective of ZSSAR is to learn a mapping function $f : \mathbb{R}^{J\times 3\times F\times M} \rightarrow \mathcal{C}_u$. To achieve this, we align the skeleton feature f_s extracted from $\mathbf{X}_i$ with the text feature f_t obtained from $\mathbf{A}_i$ in a shared latent space. During training, a feature alignment mechanism enforces the relationship between f_s and f_t for seen categories $\mathcal{C}_s$. Specifically, the skeleton feature f_s is encoded into a latent distribution z_s, while the text feature f_t is mapped to another latent distribution z_t. These latent variables z_s and z_t serve as a bridge for cross-modal knowledge transfer to ensure the learned representations capture both motion and semantic patterns within the cross-alignment. In Generalized Zero-Shot Skeleton-Based Action Recognition (GZSSAR), the model classifies actions from both $\mathcal{C}_s$ and $\mathcal{C}_u$ during testing.

3.2. Frequency Enhanced Module

Motivation for Frequency Enhancement in ZSL. In fully supervised learning, frequency-aware models [4, 44] capture both high- and low-frequency components from labeled data, where high-frequency details are particularly useful for recognizing subtle movements, and low-frequency motions are utilized to capture global movement patterns.

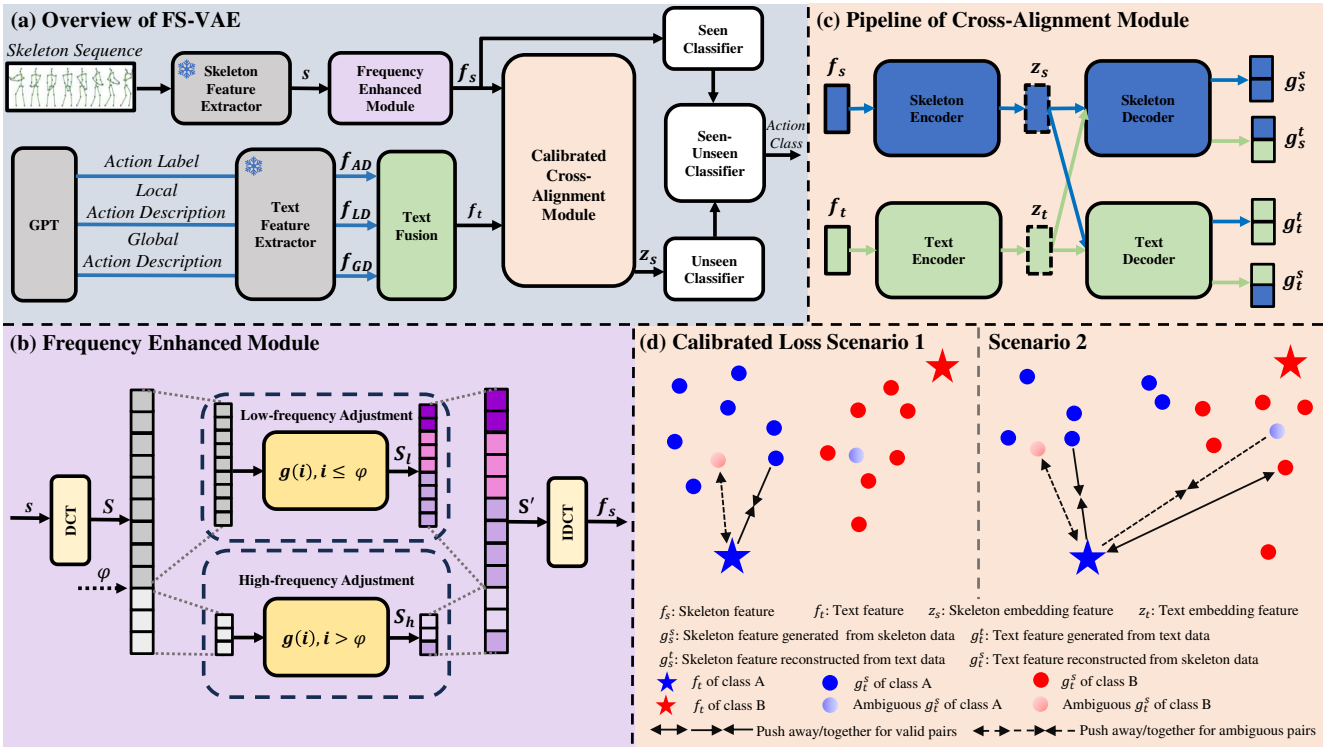


Figure 2. Overview of the proposed FS-VAE. The frequency-enhanced module integrates the global and fine-grained skeleton utilizing the low-frequency and high-frequency adjustments introduced in Section 3.2. The semantic-based action descriptions, including action labels, local action descriptions, and global action descriptions, are introduced in Section 3.3 to generate comprehensive semantic embeddings for cross-alignment. Moreover, the novel calibrated loss in the cross-alignment module is proposed in Section 3.4 for minimizing the disparity between semantic and skeletal features.

However, the lack of unseen class data prevents the direct learning of class-specific high-frequency distributions in ZSL, making these details more sparse and noisy. To address this, our approach enhances low-frequency components to extract richer semantic information, improving generalization to unseen categories. Meanwhile, we adaptively suppress high-frequency variations to preserve essential fine-grained details while mitigating noise (e.g., skeletal jitter or limb fluctuations). This adjustment reinforces the ZSL training to learn richer skeletal information and more structured semantics compared to conventional purely spatial-temporal modeling [15, 27, 28], which leads to more effective ZSSAR performance. The pipeline of the enhanced module is illustrated in Fig. 2 (b).

Frequency Division Formula. Let $\mathbf{s} \in \mathbb{R}^{J\times C\times F}$ denote the input joint sequence, where J represents the number of joints, C the coordinate dimension (e.g., x, y, z), and F the number of frames. The trajectory of the j-th joint across T frames is denoted as $T_j = (t_{j,1}, t_{j,2}, \ldots, t_{j,F})$. We apply the DCT [32, 44, 48] to obtain the frequency-domain representation for skeleton sequence: $\mathbf{S} = \mathrm{DCT}(\mathbf{s})$, the DCT decomposes the input skeleton sequence $\mathbf{s}$ into frequency components, producing the transformed representation $\mathbf{S}$ of the same length as the input. Each component in $\mathbf{S}$ corresponds to a specific frequency coefficient, where lower-indexed coefficients represent low-frequency (global) motion patterns, and higher-indexed coefficients capture high-frequency (fine-grained) details. For the trajectory T_j, the i-th DCT coefficient to each individual trajectory is calculated as:

$$C_{j,i} = \sqrt{\frac{2}{F}} \sum_{f=1}^{F} t_{j,f} \frac{1}{\sqrt{1+\delta_{i1}}} \cos\left[\frac{\pi(2f-1)(i-1)}{2F}\right] \quad (1)$$

where the *Kronecker delta* $\delta_{ij} = 1$ if $i = j$, and $\delta_{ij} = 0$ otherwise. In particular, $i \in \{1, 2, \ldots, F\}$, and the larger i corresponds to higher frequency coefficients. These coefficients enable us to represent skeleton motion effectively within the frequency domain by capturing both subtle dynamic details and global motion patterns [44].

Low-Frequency Adjustment. For the low-frequency range, the adjustment is applied using the piecewise scaling function $g(i)$:

$$\mathbf{S}_l \leftarrow \mathbf{S} \cdot g(i), i \le \varphi \quad (2)$$

where $g(i) = 1 + w_i\left(1 - \frac{i}{b}\right)$ for low-frequency components, φ is the low frequency threshold. $\mathbf{S}_l$ represents the low-frequency components, which capture global motion patterns such as large-scale movements of limbs and torso. The term b is an adjusting parameter intended to make $g(i)$ decrease gradually within the low-frequency range. The

fraction $\frac{i}{b}$ ensures a progressive reduction in enhancement strength as frequency increases, thereby maintaining the integrity of large-scale global motion. The learned weight w_i adaptively controls enhancement for different frequencies, amplifying the most distinguishing ones.

High-Frequency Adjustment. For the high-frequency range, the scaling function $g(i)$ is given by:

$$\mathbf{S}_h \leftarrow \mathbf{S} \cdot g(i), i > \varphi \tag{3}$$

where $g(i) = 1 - w_i \left(1 - \frac{i-b}{b}\right)$. $\mathbf{S}_h$ represents the high-frequency components of the skeleton features, which captures fine-grained details such as finger, wrist, and rapid limb movements. In high-frequency adjustment, b serves as a normalization factor that controls the suppression of high-frequency variables. Specifically, it ensures that attenuation decreases (i.e., $g(i)$ increases) smoothly as frequency increases, preventing excessive suppression of fine-grained motion details. By scaling the suppression term proportionally to $i - b$, this formulation mitigates skeletal noise while preserving essential micro-movements. Meanwhile, w_i adaptively modulates the suppression strength for each high-frequency component, up-weighting the most distinguishing frequencies.

Inverse Transform. The adjusted frequency-domain signal is reconstructed into the time domain using the Inverse Discrete Cosine Transform (IDCT), represented as: $f_s = \text{IDCT}(\mathbf{S}')$, where $\mathbf{S}'$ is the frequency-enhanced skeleton component in the frequency domain. The specific restoration for each joint trajectory is given by:

$$t_{j,f} = \sqrt{\frac{2}{F}} \sum_{i=1}^{F} C_{j,i} \frac{1}{\sqrt{1+\delta_{i1}}} \cos\left[\frac{\pi(2f-1)(i-1)}{2F}\right] \tag{4}$$

where $j \in \{1, 2, \ldots, J\}$ and $f \in \{1, 2, \ldots, F\}$. Here, $t_{j,f}$ represents the restored joint trajectory in the time domain, reconstructed from its frequency-domain coefficients $C_{j,i}$.

This process integrates enhanced global patterns and preserved fine-grained details, creating a comprehensive representation of the action. Using the enhanced time domain skeleton feature f_s, the model aligns these features with semantic embeddings, enabling robust recognition in zero-shot scenarios. More frequency analysis and method illustration can be found in Appendix F.

3.3. Semantic-based Action Description

Unlike the previous ZSL methods that focus on motion descriptions in a temporal way [27] and focus on the action prompting with GPTs [51] that ignores the semantic characteristics, we propose a novel strategy that leverages semantic decomposition and feature alignment to fully capture both the localized details and global semantic structures inherent in human actions. This method stems from the observation that actions can be naturally divided into components reflecting dynamic movements and overarching patterns, enabling a comprehensive and robust representation of action semantics.

The semantic text description consists of the Action Label (AL) and a Semantic-Based Description (SD). SD is further divided into two complementary components: Local action Description (LD) and Global action Description (GD). We adopt the pre-trained text-encoder of CLIP [34] to extract the corresponding semantic features. f_{AL} indicates the feature of action label. The local component f_{LD} captures fine-grained motion details, which are crucial for understanding localized dynamics and specific body-part interactions. For example, in the action of "drinking water," f_{LD} describes detailed movements such as "fist and move the hand up to the head, then look up," to emphasize precise body-part motions. In contrast, f_{GD} represents the overall movements that provide a high-level overview, such as "take liquid into your mouth, raise it to your head." By integrating f_{AL}, f_{LD} and f_{GD}, our representation enriches the semantic space, capturing both localized motion details and holistic action patterns. Extra examples of action descriptions and promptings are listed in Appendix D.

To unify these components into a cohesive semantic embedding, we concatenate the descriptions and normalize them as follows:

$$f_{\text{t}} = \frac{\text{Concat}(f_{\text{AL}}, f_{\text{LD}}, f_{\text{GD}})}{\|\text{Concat}(f_{\text{AL}}, f_{\text{LD}}, f_{\text{GD}})\|} \tag{5}$$

3.4. Calibrated Cross-Alignment Loss

Motivation for Calibrated Loss. Text data, especially the semantically rich descriptions we introduce in Section 3.3, are inherently clean and precise, offering a strong foundation for capturing action nuances. In zero-shot learning, the text features serve as the bridge between seen and unseen classes, enabling the model to generalize effectively. However, when the text encoder is affected by noisy skeleton data, it may struggle to retain these semantic details, leading to suboptimal performance.

Skeleton-based features (e.g., g_t^s), on the other hand, are noisy and unreliable for the following reasons. First, skeleton features omit crucial contextual information from the raw video data, including environmental context and fine-grained motion details. Second, variations in camera angles and viewpoints further exacerbate ambiguities in skeletons. Finally, skeletons for actions with similar motion patterns, such as "drinking water" and "eating a meal," are inherently difficult to distinguish.

As a result, rigidly aligning g_t^s with the text features f_t may result in poor updates to the text encoder, causing the model to overlook important semantic details in the text data. This misalignment is particularly problematic in zero-shot learning, where the model must generalize to unseen classes based on robust feature representations. To this end,

we propose the following calibrated alignment loss that enhances the resilience of the text encoder to noise, preserving the quality of learned representations. The illustration is presented in Fig. 2 (c) and (d).

Definition of Calibrated Loss. The calibrated loss adjusts the alignment by encouraging positive text-skeleton pairs to align while penalizing negative pairs, unlike [27, 28], which only encourages the alignment of positive pairs. The calibrated loss is defined as:

$$\mathcal{L}_{\text{Align}} = \frac{\lambda}{B}\sum_{i\in B}\frac{1}{1+\exp((\|f_t(i)-g_t^s(i^-)\|^2-\|f_t(i)-g_t^s(i)\|^2)/\lambda)} + \frac{\lambda}{B}\sum_{i\in B}\frac{1}{1+\exp((\|f_s(i)-g_s^t(i^-)\|^2-\|f_s(i)-g_s^t(i)\|^2)/\lambda)}, \tag{6}$$

where i^- denotes a negative sample to i in the batch, and λ is a temperature parameter that controls the sensitivity of the alignment loss.

Key Scenarios Addressed. The calibrated loss is robust to the following scenarios:

1. Mismatched Positive Pair with a Reliable Negative Pair: A mismatched positive pair arises when a skeleton feature (e.g., $g_t^s(i)$) is inherently ambiguous and resembles skeletons from other classes. For instance, the skeletal sequence for "reading" may be similar to "writing." Aligning such an ambiguous skeleton with the text ($f_t(i)$) of its own class ("reading") can degrade the text encoder. This problem can be balanced by introducing a reliable negative pair, i.e., a reliable negative pair ($f_t(i)$ and $g_t^s(i^-)$ ((e.g., the text feature for "reading" and a skeleton from "writing" that is clearly distinguished from with "reading")).

2. Mismatched Positive Pair with Mismatched Negative Pair: The calibrated loss remains robust even when both positive and negative pairs are mismatched. For instance, $f_t(i)$ and $g_t^s(i)$ form a mismatched positive pair of text feature for "reading" but the skeleton feature though labeled "reading", resembles "writing". Similarly, $f_t(i)$ and $g_t^s(i^-)$ form a mismatched negative pair where the text feature represents "reading" and a skeleton is labeled "writing" but inherently similar to "reading". The calibrated loss leverages correctly aligned pairs in the dataset to counteract these inconsistencies. This robustness stems from the symmetric property of the sigmoid function, i.e., $\ell(a)+\ell(-a)=1$, which has been utilized to handle noisy labels [5, 14]. In our case, let $a=(\|f_t(i)-g_t^s(i^-)\|^2-\|f_t(i)-g_t^s(i)\|^2)/\lambda$ represent a term where both pairs are mismatched. Correctly aligned pairs in the data set may contribute a corresponding term $-a$, leading to a natural counterbalance due to the symmetry of $\ell(\cdot)$. For non-symmetric losses, i.e., where $\ell(a)+\ell(-a)$ is not a constant, even a correctly aligned term cannot fully offset an incorrect one. A detailed analysis of the calibrated loss is provided in Appendix E. Notably, while our calibrated loss shares similarity with triplet losses [10, 11, 13, 16, 17, 23, 37], most of these do not satisfy the symmetric property and are, therefore, less robust to noisy scenarios. In Appendix E, we construct various alignment losses based on triplet loss and present experiments that demonstrate the advantages of our calibrated loss.

Overall Loss. Following the literature on variational autoencoder (VAE)-based architecture for skeleton recognition [15, 27], we use the evidence lower bound loss (ELBO) for reconstruction:

$$\mathcal{L}_{\text{VAE}}^s = \mathbb{E}_{q_\phi(z_s|f_s)}[\log p_\theta(f_s|z_s)] - \beta D_{KL}(q_\phi(z_s|f_s)\|p_\theta(z_s|f_s)), \tag{7}$$

where $p_\theta(\cdot)$ and $q_\phi(\cdot)$ represent the likelihood and the prior, respectively. β is a hyperparameter that controls the balance between reconstruction and regularization. $q_\phi(z_s|f_s)$ follows the multivariate Gaussian distribution $\mathcal{N}_s(\mu_s,\Sigma_s)$. L_{VAE}^t is symmetric to L_{VAE}^s. And thus, the VAE loss is $L_{VAE}=L_{VAE}^s+L_{VAE}^t$. The overall loss is

$$\mathcal{L}_{VAE}^{cali}=\mathcal{L}_{VAE}+\alpha\mathcal{L}_{Align}, \tag{8}$$

where α adjusts the trade-off between the VAE loss and the alignment loss.

4. Experiments

4.1. Implementation Details

Our work mainly follows [15] for data preprocessing. The data split strategy follows [15, 27]. In ZSL, a 55/5 split means training on 55 seen classes and testing on 5 unseen classes. In Generalized Zero-Shot Learning (GZSL), training remains the same, but testing includes both the 55 seen and 5 unseen classes. We adopt Shift-GCN [8] as the skeleton extractor. Meanwhile, GPT-4 [1] is utilized to generate action descriptions. More settings can be found in Appendix B, and please refer to Appendix C for additional experiments, including results on the PKU-MMD dataset [29].

4.2. Comparisons with State-of-the-Art Methods

To evaluate the effectiveness of our approach, we compare it with state-of-the-art methods under both ZSL and GZSL

Table 1. Zero-Shot Learning Results. The highest values are highlighted in red, while the second-highest values (from other works) are marked in blue. ↑ indicates the improvement over the second-highest value. * indicates the reproduced results of the released codes. † denotes the use of only w_i for frequency coefficients.

Methods	Venue	NTU-60 (ACC,%)		NTU-120 (ACC,%)	
		55/5 split	48/12 split	110/10 split	96/24 split
ReViSE[19]	ICCV2017	53.9	17.5	55.0	32.4
JPoSE[43]	ICCV2019	64.8	28.8	51.9	32.4
CADA-VAE[36]	CVPR2019	76.8	29.0	59.5	35.8
SynSE[15]	ICIP2021	75.8	33.3	62.7	38.7
SMIE[50]	ACMM2023	78.0	40.2	61.3	42.3
STAR[7]	ACMM2024	81.4	45.1	63.3	44.3
GZSSAR*[27]	ICIG2023	83.3	49.8	72.0	60.7
PURLS[51]	CVPR2024	79.2	41.0	72.0	52.0
SA-DVAE[28]	ECCV2024	82.4	41.4	68.8	46.1
Ours†	\	84.2	52.6	71.2	61.9
Ours	\	**86.9**$_{\uparrow 3.6}$	**57.2**$_{\uparrow 7.4}$	**74.4**$_{\uparrow 2.4}$	**62.5**$_{\uparrow 1.8}$

Table 2. Generalized Zero-Shot Learning Results. The highest values are highlighted in red, and the second-highest values (from other works) are marked in blue. H represents the harmonic mean. The result analysis is presented in Section 4.2.

Methods	Venue	NTU-60 (55/5 split)			NTU-60 (48/12 split)			NTU-120 (110/10 split)			NTU-120 (96/24 split)		
		Seen	Unseen	H	Seen	Unseen	H	Seen	Unseen	H	Seen	Unseen	H
ReViSE[19]	ICCV 2017	74.2	34.7	29.2	62.4	20.8	31.2	48.7	44.8	46.7	49.7	25.1	33.3
JPoSE[43]	ICCV 2019	64.4	50.3	56.5	60.5	20.6	30.8	47.7	46.4	47.0	38.6	22.8	28.7
CADA-VAE[36]	CVPR 2019	69.4	61.8	65.4	51.3	27.0	35.4	47.2	19.8	48.4	41.1	34.1	37.3
SynSE[15]	ICIP2021	61.3	56.9	59.0	52.2	27.9	36.3	52.5	57.6	54.9	56.4	32.2	41.0
STAR[7]	ACMM2024	69.0	69.9	69.4	62.7	37.0	46.6	59.9	52.7	56.1	51.2	36.9	42.9
GZSSAR*[27]	ICIG2023	66.8	70.7	68.7	54.8	41.4	47.1	58.1	57.8	58.0	59.2	45.9	51.7
SA-DVAE[28]	ECCV2024	62.3	70.8	66.3	50.2	36.9	42.6	61.1	59.8	60.4	58.8	35.8	44.5
Ours†	\	76.4	61.9	68.4	57.4	43.5	49.5	55.7	66.8	60.7	58.7	48.3	53.0
Ours	\	77.0	**74.5**$_{\uparrow 3.7}$	**75.7**$_{\uparrow 6.3}$	56.2	**48.6**$_{\uparrow 7.2}$	**52.1**$_{\uparrow 5.0}$	59.2	**67.9**$_{\uparrow 8.1}$	**63.3**$_{\uparrow 2.9}$	57.8	**51.9**$_{\uparrow 6.0}$	**54.7**$_{\uparrow 3.0}$

Table 3. Influence of different modules. Semantic-based action Descriptions (SD), Frequency-enhanced Module (FM), Calibrated Loss (CL).

Modules			NTU-60 (ACC,%)		NTU-120 (ACC,%)	
SD	FM	CL	55/5 split	48/12 split	110/10 split	96/24 split
✗	✗	✗	83.3	49.8	72.0	60.7
✓	✗	✗	85.4	52.7	73.0	61.3
✗	✓	✗	85.8	53.1	74.0	61.8
✗	✗	✓	84.4	54.1	72.8	60.0
✓	✓	✓	**86.9**	**57.2**	**74.4**	**62.5**

Table 4. Influence of different text descriptions. Action Label (AL), Local action Description (LD), and Global action Description (GD).

Descriptions			NTU-60 (ACC,%)		NTU-120 (ACC,%)	
AL	LD	GD	55/5 split	48/12 split	110/10 split	96/24 split
✓	✗	✗	81.9	38.7	70.3	47.3
✗	✓	✗	79.3	43.8	54.5	45.7
✗	✗	✓	82.0	48.6	64.7	59.2
✗	✓	✓	83.7	54.1	57.5	46.7
✓	✓	✓	**86.9**	**57.2**	**74.4**	**62.5**

settings. The results on NTU-60 [38] and NTU-120 [31] datasets, following established split protocols [15, 27], are summarized in Tables 1 and 2. Our model consistently achieves the highest accuracy in ZSL, demonstrating strong generalization to unseen actions. In GZSL, it outperforms existing methods, leading to the highest harmonic mean score[15] (H-score, $H = \frac{2\times \text{Seen}\times \text{Unseen}}{\text{Seen}+\text{Unseen}}$), highlighting the effectiveness of FS-VAE in learning a semantic-enhanced and well-calibrated representation. Notably, we focus on unseen accuracy and the H-score as they best reflect generalization. Unseen accuracy measures the model's ability to classify novel actions, while H-score balances seen and unseen performance to prevent biased predictions.

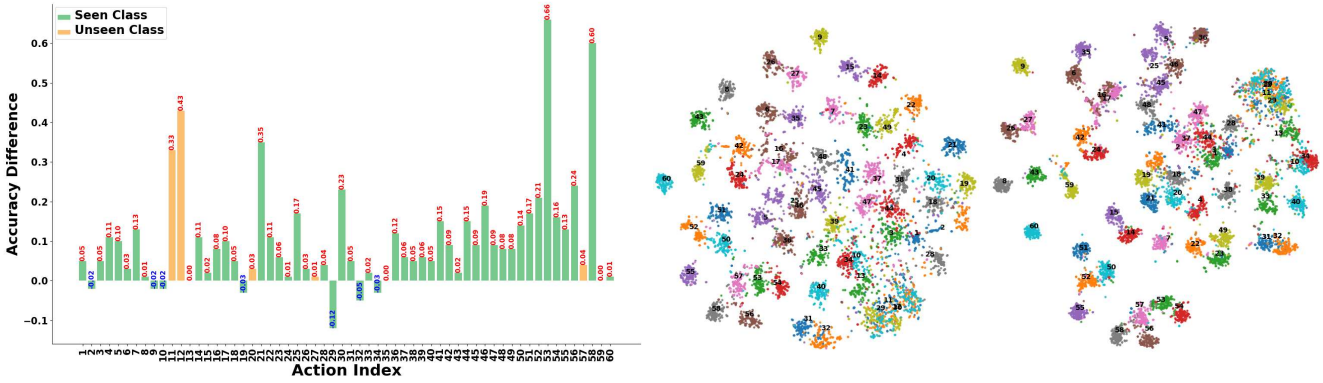
4.3. Ablation Study

Influence of Different Modules. Table 3 highlights the impact of each key component in our framework, including the Semantic-based Descriptions (SD), Frequency-enhanced Module (FM), and Calibrated Loss (CL). Adding SD improves the accuracy of the baseline to 85.4% in the NTU-60 dataset (e.g., 55/5 split), emphasizing the importance of semantic enrichment. Similarly, integrating FM or CL individually achieves 85.8% and 84.4%, respectively, demonstrating their individual contributions to frequency-specific feature learning and robust alignment. Combining all three components leads to the highest performance of 86.9%, which confirms their complementary effects in addressing the ZSSAR challenges.

Influence of Different Text Descriptions. Table 4 presents the influence of different text descriptions, including Action Label (AL), Local action Description (LD), and Global action Description (GD). Using AL alone achieves an accuracy of 81.9% on the NTU-60 dataset (55/5 split), showing its fundamental role in providing basic semantic information. Incorporating LD or GD results in 79.3% and 82.0%, respectively, suggesting that GD contributes more to performance as it provides more comprehensive semantics of the overall action. Combining LD and GD boosts the performance to 83.7%, highlighting the synergy between these two semantic-aware features. Integrating all three components achieves the highest accuracy of 86.9%, demonstrating their complementary contributions to enriching semantic embeddings in ZSSAR.

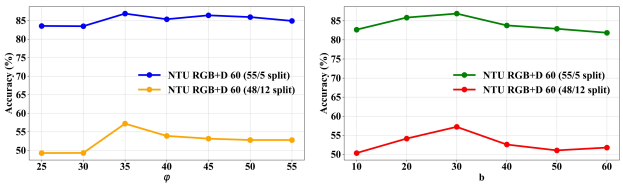
Influence of φ and b in Frequency Enhanced Module. In Fig. 4(a)-(b), we evaluate the impact of various hyperparameter settings of our frequency-enhanced module. The analysis reveals that φ and b play a crucial role in determining the overall accuracy of our method. The optimal configuration is achieved with $\varphi = 35$ and $b = 30$. These values effectively balance feature enhancement of both global action structures and fine-grained motion details. φ controls the separation between low- and high-frequency components, ensuring that structural semantics are preserved while capturing subtle motion variations. Meanwhile, b determines the intensity of enhancement, preventing the amplification of noise while enhancing the model's ability of discriminative representation learning.

Influence of α and λ for Calibrated Loss. Recall that α balances the reconstruction loss and alignment loss, playing a crucial role in ensuring that both losses contribute effectively to the overall objective. Meanwhile, λ controls the sensitivity of the alignment loss, with smaller values making it more responsive to misalignments. Specifically, a small λ places greater emphasis on larger misalignments compared to a large λ, which reduces this sensitivity, making the alignment loss less responsive to smaller misalignments. Notably, the alignment loss retains its symmetric property regardless of the choice of λ. As shown in Table 5, the analysis reveals that the optimal combination of parameters is $\alpha = 0.1$ and $\lambda = 100$ to yield the best performance.



(a) Accuracy difference (b) t-SNE for baseline [27] (c) t-SNE for our method

Figure 3. (a) shows the accuracy difference for seen-unseen actions compared to baseline [27] under the NTU-60 55/5 split, where the outperforming accuracies are marked in red, and others are in blue. (b) and (c) depict the t-SNE visualizations, the corresponding action indices (listed in Appendix G) are labeled in the clusters. Best viewed by zooming in.



(a) Influence of φ on NTU-60 split (b) Influence of b on NTU-60 split

Figure 4. Influence of φ and b in frequency enhanced module.

Impact of Removing Frequency Adjustment. In Tables 1 and 2 (denoted by 'Ours†'), we also evaluate the impact of removing explicit frequency adjustment and replacing it with purely learnable frequency weight: $\mathbf{S} \leftarrow \mathbf{S} \cdot g(i)$, where $g(i) = w_i$. w_i is the learnable weight applied directly to all frequency components. The performance drop highlights two key issues: (1) Without explicit frequency adjustment, the model fails to balance global structural patterns and fine-grained details, leading to weaker semantic alignment and increased sensitivity to noise. (2) Explicit frequency scaling provides a prior-informed enhancement, whereas purely learnable weights rely solely on data-driven optimization. This often results in inconsistent frequency adjustments across different training samples, leading to overfitting to seen categories and ineffective generalization to unseen actions.

4.4. Qualitative Analysis

We present the accuracy difference results compared to the baseline method for the NTU-60 55/5 split in Fig. 3a. Our approach consistently outperforms the baseline across both seen and unseen actions. The results highlight our method's effectiveness in not only enhancing overall accuracy but also improving recognition across most discriminative actions. Notably, FS-VAE surpasses the baseline in actions that require fine-grained motion understanding, such as "reading" and "writing" in unseen classes (orange) and most of the actions in seen classes (green). Additionally,

Table 5. Influence of α and λ in calibrated loss.

α	NTU-60 (ACC,%)		NTU-120 (ACC,%)	
	55/5 split	48/12 split	110/10 split	96/24 split
0.01	81.9	51.8	74.1	61.8
0.05	84.0	49.4	72.8	60.0
0.1	**86.9**	**57.2**	**74.4**	**62.5**
0.5	85.9	53.7	74.0	61.8
1	83.2	51.8	72.8	61.9
2	81.0	50.3	71.3	60.4

λ	NTU-60 (ACC,%)		NTU-120 (ACC,%)	
	55/5 split	48/12 split	110/10 split	96/24 split
70	84.1	51.9	71.7	60.7
80	84.9	53.1	73.0	61.9
90	86.0	53.8	73.0	60.3
100	**86.9**	**57.2**	**74.4**	**62.5**
110	86.2	53.8	74.0	61.9
120	85.7	52.4	74.2	60.2

Fig. 3b and 3c present t-SNE [41] visualization examples of NTU-60 dataset under 55/5 split. The results illustrate that our method improves the visual and semantic alignment (e.g., better inter-class separation between a pair of skeletal-similar actions, such as "reading" and "type on a keyboard"). Furthermore, it produces a tighter and semantically structured embedding space (e.g., stronger intra-class cohesion of "reading" and "pushing").

5. Conclusion

We introduce a novel framework for zero-shot skeleton-based action recognition (ZSSAR) that combines frequency-enhanced modeling with a calibrated alignment mechanism. The frequency-enhanced module leverages DCT to capture fine-grained details and preserves global patterns. The semantic-based action description enriches feature embeddings, while the calibrated cross-alignment loss dynamically addresses modality gaps and ambiguities. Extensive evaluations on the benchmarks demonstrate the state-of-the-art performance of our approach in recognizing unseen actions. This work establishes a robust ZSSAR framework, paving the way for future advances in frequency-aware action recognition.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6

[2] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1): 90–93, 1974. 2, 18

[3] Ijaz Akhter, Yaser Sheikh, Sohaib Khan, and Takeo Kanade. Nonrigid structure from motion in trajectory space. *Advances in neural information processing systems*, 21, 2008. 3

[4] Haochen Chang, Jing Chen, Yilin Li, Jixiang Chen, and Xiaofeng Zhang. Wavelet-decoupling contrastive enhancement network for fine-grained skeleton-based action recognition. *arXiv preprint arXiv:2402.02210*, 2024. 3, 16

[5] Nontawat Charoenphakdee, Jongyeong Lee, and Masashi Sugiyama. On symmetric losses for learning from corrupted labels. In *International Conference on Machine Learning*, pages 961–970. PMLR, 2019. 6

[6] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13359–13368, 2021. 1

[7] Yang Chen, Jingcai Guo, Tian He, Xiaocheng Lu, and Ling Wang. Fine-grained side information guided dual-prompts for zero-shot skeleton action recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 778–786, 2024. 1, 6, 7

[8] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 183–192, 2020. 6, 18

[9] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20186–20196, 2022. 1, 3

[10] Thanh-Toan Do, Toan Tran, Ian Reid, Vijay Kumar, Tuan Hoang, and Gustavo Carneiro. A theoretically sound upper bound on the triplet loss for improving the efficiency of deep distance metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10404–10413, 2019. 6

[11] Xingping Dong and Jianbing Shen. Triplet loss in siamese network for object tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 459–474, 2018. 6, 14

[12] Xuehao Gao, Shaoyi Du, Yang Wu, and Yang Yang. Decompose more and aggregate better: Two closer looks at frequency representation learning for human motion prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6451–6460, 2023. 3

[13] Weifeng Ge. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European conference on computer vision (ECCV)*, pages 269–285, 2018. 6

[14] Zhishuai Guo, Rong Jin, Jiebo Luo, and Tianbao Yang. Fedxl: Provable federated learning for deep x-risk optimization. In *International Conference on Machine Learning*, pages 11934–11966. PMLR, 2023. 6

[15] Pranay Gupta, Divyanshu Sharma, and Ravi Kiran Sarvadevabhatla. Syntactically guided generative embeddings for zero-shot skeleton action recognition. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 439–443. IEEE, 2021. 1, 2, 3, 4, 6, 7, 11

[16] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 6, 14

[17] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *Similarity-based pattern recognition: third international workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*, pages 84–92. Springer, 2015. 6, 14

[18] Yonghong Hou, Zhaoyang Li, Pichao Wang, and Wanqing Li. Skeleton optical spectra-based action recognition using convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(3):807–811, 2016. 1

[19] Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. Learning robust visual-semantic embeddings. In *Proceedings of the IEEE International conference on Computer Vision*, pages 3571–3580, 2017. 1, 2, 6, 7, 11

[20] Mohamad Kashef, Anna Visvizi, and Orlando Troisi. Smart city as a smart service system: Human-computer interaction and smart city surveillance systems. *Computers in Human Behavior*, 124:106923, 2021. 1

[21] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3288–3297, 2017. 1

[22] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013. 1

[23] Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5385–5394, 2016. 6, 14

[24] Inwoong Lee, Doyoung Kim, Seoungyoon Kang, and Sanghoon Lee. Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1012–1020, 2017. 1

[25] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 214–223, 2020. 3

[26] Maosen Li, Siheng Chen, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. Skeleton-parted graph scattering networks

for 3d human motion prediction. In *European conference on computer vision*, pages 18–36. Springer, 2022. 3

[27] Ming-Zhe Li, Zhen Jia, Zhang Zhang, Zhanyu Ma, and Liang Wang. Multi-semantic fusion model for generalized zero-shot skeleton-based action recognition. In *International Conference on Image and Graphics*, pages 68–80. Springer, 2023. 1, 2, 3, 4, 5, 6, 7, 8, 11, 12

[28] Sheng-Wei Li, Zi-Xiang Wei, Wei-Jie Chen, Yi-Hsin Yu, Chih-Yuan Yang, and Jane Yung-jen Hsu. Sa-dvae: Improving zero-shot skeleton-based action recognition by disentangled variational autoencoders. In *European Conference on Computer Vision*, pages 447–462. Springer, 2025. 1, 2, 3, 4, 6, 7, 11

[29] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017. 6, 11

[30] Haowei Liu, Yongcheng Liu, Yuxin Chen, Chunfeng Yuan, Bing Li, and Weiming Hu. Transkeleton: Hierarchical spatial-temporal transformer for skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1

[31] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019. 7, 11

[32] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9489–9497, 2019. 3, 4

[33] Satyajit Nayak, Bingi Nagesh, Aurobinda Routray, and Monalisa Sarma. A human–computer interaction framework for emotion recognition through time-series thermal video sequences. *Computers & Electrical Engineering*, 93:107280, 2021. 1

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5, 18

[35] K Ramamohan Rao and Ping Yip. *Discrete cosine transform: algorithms, advantages, applications*. Academic press, 2014. 15

[36] Edgar Schonfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 54–57, 2019. 6, 7, 11

[37] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 6, 14

[38] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 7, 11

[39] Mohammad Shorfuzzaman, M Shamim Hossain, and Mohammed F Alhamid. Towards the sustainable development of smart cities through mass video surveillance: A response to the covid-19 pandemic. *Sustainable cities and society*, 64: 102582, 2021. 1

[40] Roshan Singh, Alok Kumar Singh Kushwaha, and Rajeev Srivastava. Multi-view recognition system for human activity based on multiple features for video surveillance system. *Multimedia Tools and Applications*, 78:17165–17196, 2019. 1

[41] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 8

[42] Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers. *Advances in Neural Information Processing Systems*, 34:6036–6049, 2021. 3

[43] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 450–459, 2019. 2, 6, 7, 11

[44] Wenhan Wu, Ce Zheng, Zihao Yang, Chen Chen, Srijan Das, and Aidong Lu. Frequency guidance matters: Skeletal action recognition by frequency-aware mixed transformer. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4660–4669, 2024. 1, 3, 4, 16

[45] Wentian Xin, Qiguang Miao, Yi Liu, Ruyi Liu, Chi-Man Pun, and Cheng Shi. Skeleton mixformer: Multivariate topology representation for skeleton-based action recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2211–2220, 2023. 1

[46] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 11

[47] Songyang Zhang, Xiaoming Liu, and Jun Xiao. On geometric features for skeleton-based action recognition using multilayer lstm networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 148–157. IEEE, 2017. 1

[48] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8877–8886, 2023. 4

[49] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11656–11665, 2021. 3

[50] Yujie Zhou, Wenwen Qiang, Anyi Rao, Ning Lin, Bing Su, and Jiaqi Wang. Zero-shot skeleton-based action recognition via mutual information estimation and maximization. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5302–5310, 2023. 1, 2, 6, 11

[51] Anqi Zhu, Qiuhong Ke, Mingming Gong, and James Bailey. Part-aware unified representation of language and skeleton for zero-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18761–18770, 2024. 1, 2, 5, 6

A. Appendix

The supplementary material is organized into the following sections:

- Section B: **More experimental settings.** (i) Datasets introduction (NTU-60, NTU-120, PKU-MMD); (ii) training strategy; (iii) parameter settings.
- Section C: **Additional experiments.** (i) Results on PKU-MMD; (ii) results on different text feature extractors.
- Section D: **Semantic-based action descriptions.** (i) Prompting examples; (ii) description examples.
- Section E: **Calibrated alignment loss analysis.** (i) Calibrated alignment loss explanation; (ii) extra ablation study for calibrated alignment loss.
- Section F: **Frequency-based skeleton representation analysis.** (i) Frequency domain representation and energy preservation proof; (ii) semantic integrity with frequency adjustment; (iii) frequency-based enhancement mechanism; (iv) energy redistribution derivation; (v) illustration example of frequency enhanced method; (vi) codes.
- Section G: **Justification for choosing DCT.**
- Section H: **NTU-60 dataset action index.**

B. More Experiments Settings

B.1. Datasets

NTU RGB+D 60 [38]. The NTU-60 dataset is one of the most popular large-scale datasets designed for the analysis of 3D human actions. It comprises 56,880 human action sequences captured by three Kinect-V2 cameras, covering 60 distinct action classes. In this work, we use only the skeleton data. Each skeleton sequence consists of up to two skeletons per frame, with each skeleton containing 25 joints. In this paper, two seen/unseen splits are employed, following prior work [15]: 55 seen classes and 5 unseen classes, and 48 seen classes and 12 unseen classes. The unseen classes are randomly selected, maintaining consistency with previous studies.

NTU RGB+D 120 [31]. The NTU-120 dataset is an extended version of NTU-60. It includes 114,480 action sequences performed by 106 subjects from 155 distinct viewpoints, spanning 120 action classes. These 120 classes build upon the original 60 classes in NTU-60, offering a broader range of human actions. For zero-shot learning, the dataset adopts seen/unseen splits of 110 seen classes and 10 unseen classes, and 96 seen classes and 24 unseen classes, consistent with the splits defined in [15].

Table 6. Zero-Shot Learning (ZSL) and Generalized Zero-Shot Learning (GZSL) results on PKU-MMD (46/5 split).

Methods	Venue	ZSL (ACC,%)	GZSL (ACC,%)		
			Seen	Unseen	H
ReViSE[19]	ICCV2017	59.3	60.9	42.2	49.8
JPoSE[43]	ICCV2019	57.2	60.3	45.2	51.6
CADA-VAE[36]	CVPR2019	60.7	63.2	35.9	45.8
SynSE[15]	ICIP2021	53.9	63.1	40.7	49.5
SMIE[50]	ACMM2023	60.8	-	-	-
SA-DVAE[28]	ECCV2024	66.5	58.5	51.4	54.7
Ours	\	**71.2**$_{\uparrow 4.7}$	64.3	**54.5**$_{\uparrow 3.1}$	**59.0**$_{\uparrow 4.3}$

Table 7. Comparisons of different text feature extractors in ZSL.

Model	NTU-60 (ACC,%)		NTU-120 (ACC,%)	
	55/5 split	48/12 split	110/10 split	96/24 split
ViT-B/16	84.2	49.4	72.7	60.2
ViT-B/32	**86.9**	**57.2**	**74.4**	**62.5**

PKU-MMD [29]. The PKU-MMD dataset is a large-scale benchmark for multimodal action recognition, providing both 3D skeleton sequences and RGB+D recordings. It consists of 66 subjects and 51 classes. We conduct the experiments on Phase I following the protocols from [27, 28] and the skeleton features provided by [28] for a fair comparison (skeleton features are generated by ST-GCN[46], 46/5 split settings, 46 seen classes and 5 unseen classes).

B.2. Training Strategy

The training phase follows the same processing procedure as [27], which is systematically organized into four stages: training the skeleton feature extractor to capture spatio-temporal dependencies, optimizing the generative cross-modal alignment module to bridge the skeleton and semantic features, training the unseen class classifier for generalization, and the seen-unseen classification gate for accurate category differentiation.

B.3. Parameter Settings

Table 12 shows the parameter settings of our method, including the parameters applied during all the training stages mentioned in the main paper and [27].

C. More Experiments

Results on PKU-MMD. Table 6 presents the ZSL and GZSL performance on the PKU-MMD dataset under the 46/5 split settings [28]. Our approach consistently outperforms prior methods in both ZSL and GZSL settings, demonstrating its effectiveness in recognizing unseen actions while maintaining strong generalization.

Comparisons of Different Text Feature Extractors. We evaluate two CLIP-based text encoders, ViT-B/16 and ViT-B/32, for ZSSAR and GZSSAR tasks on NTU-60 and NTU-120 datasets. As shown in Table 7, ViT-B/32 achieves higher ZSL accuracies in all splits, e.g., 86.9% vs. 84.2% on the NTU-60 55/5 split. For GZSSAR in Table 8, ViT-B/32 also outperforms ViT-B/16 in harmonic mean (H-score),

Table 8. Comparisons of different text feature extractors in GZSL.

Model	NTU-60 (55/5 split)			NTU-60 (48/12 split)			NTU-120 (110/10 split)			NTU-120 (96/24 split)		
	Seen	Unseen	H	Seen	Unseen	H	Seen	Unseen	H	Seen	Unseen	H
ViT-B/16	65.1	71.0	67.9	61.0	39.4	47.9	55.5	**68.9**	61.4	56.6	47.7	52.6
ViT-B/32	77.0	**74.5**	**75.7**	56.2	**48.6**	**52.1**	59.2	67.9	**63.3**	57.8	**51.9**	**54.7**

e.g., 75.7% vs. 67.9% on the NTU-60 55/5 split. Based on these results, we use ViT-B/32 as the text feature extractor in subsequent experiments.

D. Semantic-based Action Descriptions

Global Action Description Prompting Examples. *"Describe the action of [ACTION NAME] by summarizing its overall motion pattern and intent. Focus on the key movements that define the action as a whole. Avoid excessive details about specific joints but ensure the description captures how the action is performed in a natural way. For example, describe how objects are manipulated, how body posture changes, or the general sequence of motion from start to finish."*

Local Action Description Prompting Examples. *"Describe the action of [ACTION NAME] by detailing the precise movements of the hands, arms, or other involved body parts. Provide a step-by-step breakdown of how the action is executed at a fine-grained level, emphasizing joint motion, hand positioning, and transitions. Ensure the description remains human-readable and avoids overly technical terminology."*

Description Examples. Table 13 illustrates how our method refines action descriptions by incorporating both global and local semantic components. Compared to the baseline[27], which provides a vague summary, our approach explicitly decomposes actions into structured representations.

For example, in the action "drinking water ", the baseline only mentions the ingestion process, whereas our Global action Description (GD) highlights the sequential motion of "grasping an object, raising it to the head, and simulating a drinking motion", capturing the structural essence of the action. Meanwhile, Local action Description (LD) provides finer details, such as "moving the fist up to the head and looking slightly downward", which are critical for distinguishing similar actions like "eating".

Similarly, for "Brushing Teeth", the baseline merely describes the purpose of the action ("to clean teeth with a brush"), but GD focuses on the characteristic motion of "moving a toothbrush back and forth", while the LD refines it further by specifying "hand movement towards the head followed by wrist tremble". This level of granularity ensures better alignment between textual descriptions and skeleton-based representations.

These examples demonstrate that our description method not only improves semantic precision, which is crucial for robust skeleton-based action recognition. By explicitly decomposing actions into structured representations that encompass both global motion patterns and localized details, the model gains a more comprehensive understanding of action semantics. This enriched textual description provides a stronger supervision signal for aligning skeleton features with semantic embeddings, thereby reducing ambiguities in action recognition.

E. Analysis of Calibrated Alignment Loss

E.1. Calibrated Loss Explanation

In this section, we break down the loss function to analyze how the calibrated alignment loss operates. Without loss of generality, consider a multi-class classification problem with three classes: Class 1, Class 2, and Class 3. Each class is associated with a ground truth distribution, denoted as P_1, P_2, and P_3. Assume we collect a dataset as follows: 1) S_1 with $n_1+\tilde{n}$ data points in total, where n_1 points are sampled from the distribution P_1, and we let $\tilde{S}$ denote $\tilde{n}$ points from P_2. 2) S_2, containing n_2 points sampled from P_2. 3) S_3, containing n_3 points sampled from P_3.

We identify two types of potential errors: (1) misaligning points in $\tilde{S}$ with the text features of Class 1, and (2) incorrectly enforcing $\tilde{S}$ to be far from the text features of Class 2.

For simplicity, we focus on the first term in $\mathcal{L}_{Align}$, as the second term follows a similar structure. Let f_t^k denote the text feature of Class k, where $k \in 1, 2, 3$. Denote

$$\mathcal{L}_{Align}^1 := \sum_{q=1}^{3} \lambda \sum_{m \neq q} \sum_{i \in S_q} \sum_{j \in S_m} \left[\frac{1}{1 + \exp((\|f_t^q - g_t^s(j)\|^2 - \|f_t^q - g_t^s(i)\|^2)/\lambda)} \right]. \tag{9}$$

Let

$$\mathcal{L}_{q,m} := \lambda \sum_{i \in S_q} \sum_{j \in S_m} \ell^q(i,j), \tag{10}$$

where

$$\ell^q(i,j) = \frac{1}{1 + \exp((\|f_t^q - g_t^s(j)\|^2 - \|f_t^q - g_t^s(i)\|^2)/\lambda)}. \tag{11}$$

Rearranging the terms, we can rewrite the loss function as

$$\mathcal{L}_{Align}^1 = \mathcal{L}_{1,2} + \mathcal{L}_{1,3} + \mathcal{L}_{2,1} + \mathcal{L}_{2,3} + \mathcal{L}_{3,1} + \mathcal{L}_{3,2}, \tag{12}$$

where

$$\mathcal{L}_{1,2} = \lambda \sum_{i \in S_1/\tilde{S}} \sum_{j \in S_2} \ell^1(i,j) + \lambda \underbrace{\sum_{i \in \tilde{S}} \sum_{j \in S_2} \ell^1(i,j)}_{(A)}. \quad (13)$$

$$\mathcal{L}_{1,3} = \lambda \sum_{i \in S_1/\tilde{S}} \sum_{j \in S_3} \ell^1(i,j) + \lambda \underbrace{\sum_{i \in \tilde{S}} \sum_{j \in S_3} \ell^1(i,j)}_{(B)}. \quad (14)$$

$$\mathcal{L}_{2,1} = \lambda \sum_{i \in S_2} \sum_{j \in S_1/\tilde{S}} \ell^2(i,j) + \lambda \underbrace{\sum_{i \in S_2} \sum_{j \in \tilde{S}} \ell^2(i,j)}_{(C)}. \quad (15)$$

$$\mathcal{L}_{2,3} = \lambda \sum_{i \in S_2} \sum_{j \in S_3} \ell^2(i,j) \quad (16)$$

$$\mathcal{L}_{3,1} = \lambda \sum_{i \in S_3} \sum_{j \in S_1/\tilde{S}} \ell^3(i,j) + \lambda \underbrace{\sum_{i \in S_3} \sum_{j \in \tilde{S}} \ell^3(i,j)}_{(D)} \quad (17)$$

$$\mathcal{L}_{3,2} = \lambda \sum_{i \in S_3} \sum_{j \in S_2} \ell^3(i,j) \quad (18)$$

We observe that the noisy subset $\tilde{S}$ is only involved in terms A, B, C, and D. Although term D involves $\tilde{S}$, it does not lead to misalignment, as it merely encourages the text of Class 3 to be similar to other text from Class 3 and dissimilar to $\tilde{S}$. Since $\tilde{S}$ is generated from P_2, this is a valid operation. Terms A and C can be addressed in the following theorem.

Theorem 1. *For the data sets generated as described above and the loss function defined accordingly, the terms A and C are equal to constants in expectation, i.e.,*

$$\mathbb{E}_{S_1,S_2,S_3}[A] = \mathbb{E}_{S_1,S_2,S_3}[C] = 1. \quad (19)$$

Proof. For term A, we have

$$\mathbb{E}_{\tilde{S},S_2}\left[\lambda \sum_{i \in \tilde{S}} \sum_{j \in S_2} \ell^1(i,j)\right] = \lambda \tilde{n} n_2 \mathbb{E}_{i \in P_2} \mathbb{E}_{j \in P_2} \ell^1(i,j)$$
$$= \lambda \tilde{n} n_2 \mathbb{E}_{i \in P_2} \mathbb{E}_{j \in P_2} \frac{\ell^1(i,j) + \ell^1(j,i)}{2}, \quad (20)$$

where

$$\begin{aligned}
&\frac{\ell^1(i,j) + \ell^1(j,i)}{2} \\
&= \frac{1}{1 + \exp((\|f_t^1 - g_t^s(j)\|^2 - \|f_t^1 - g_t^s(i)\|^2)/\lambda)} \\
&\quad + \frac{1}{1 + \exp((\|f_t^1 - g_t^s(i)\|^2 - \|f_t^1 - g_t^s(j)\|^2)/\lambda)} \\
&= \frac{\exp((\|f_t^1 - g_t^s(i)\|^2 - \|f_t^1 - g_t^s(j)\|^2)/\lambda)}{1 + \exp((\|f_t^1 - g_t^s(i)\|^2 - \|f_t^1 - g_t^s(j)\|^2)/\lambda)} \\
&\quad + \frac{1}{1 + \exp((\|f_t^1 - g_t^s(i)\|^2 - \|f_t^1 - g_t^s(j)\|^2)/\lambda)} \\
&= 1.
\end{aligned} \quad (21)$$

Similarly, for term C we obtain that

$$\mathbb{E}_{S_2,\tilde{S}}\left[\lambda \sum_{i \in S_2} \sum_{j \in \tilde{S}} \ell^2(i,j)\right] = \lambda n_2 \tilde{n} \mathbb{E}_{i \in P_2} \mathbb{E}_{j \in P_2} \ell^2(i,j)$$
$$= \lambda n_2 \tilde{n} \mathbb{E}_{i \in P_2} \mathbb{E}_{j \in P_2} \frac{\ell^2(i,j) + \ell^2(j,i)}{2}, \quad (22)$$

where

$$\begin{aligned}
&\frac{\ell^2(i,j) + \ell^2(j,i)}{2} \\
&= \frac{1}{1 + \exp((\|f_t^2 - g_t^s(j)\|^2 - \|f_t^2 - g_t^s(i)\|^2)/\lambda)} \\
&\quad + \frac{1}{1 + \exp((\|f_t^2 - g_t^s(i)\|^2 - \|f_t^2 - g_t^s(j)\|^2)/\lambda)} \\
&= \frac{\exp((\|f_t^2 - g_t^s(i)\|^2 - \|f_t^2 - g_t^s(j)\|^2)/\lambda)}{1 + \exp((\|f_t^2 - g_t^s(i)\|^2 - \|f_t^2 - g_t^s(j)\|^2)/\lambda)} \\
&\quad + \frac{1}{1 + \exp((\|f_t^2 - g_t^s(i)\|^2 - \|f_t^2 - g_t^s(j)\|^2)/\lambda)} \\
&= 1.
\end{aligned} \quad (23)$$

□

For term B, which is given by

$$\sum_{i \in \tilde{S}} \sum_{j \in S_3} \ell^1(i,j) = \frac{1}{1 + \exp((\|f_t^1 - g_t^s(j)\|^2 - \|f_t^1 - g_t^s(i)\|^2)/\lambda)}, \quad (24)$$

note that $\|f_t^1 - g_t^s(i)\|^2$ represents a misalignment term, but it can be partially balanced by $\|f_t^1 - g_t^s(j)\|^2$. Additionally, the term B does not exist in the case of a binary classification problem.

E.2. Extra Ablation Study for Calibrated Alignment Loss

In this subsection, we compare our results with those obtained using triplet losses as alignment losses. Although triplet losses also consider both positive and negative pairs, most of them do not satisfy the symmetric property, making them less robust to noisy features. The results are summarized in Table 9.

Specifically, the triplet alignment losses are developed based on popular triplet loss formulations, as follows. First, following the work of [37], we define:

$$\begin{aligned}\mathcal{L}_{\text{T},1} =& \frac{1}{B}\sum_{i\in B}\max(\|f_t(i)-g_t^s(i)\|^2-\|f_t(i)-g_t^s(i^-)\|^2+m,0)\\ &+\frac{1}{B}\sum_{i\in B}\max(\|f_s(i)-g_s^t(i)\|^2-\|f_s(i)-g_s^t(i^-)\|^2+m,0),\end{aligned} \tag{25}$$

which m is a margin term. It is not globally symmetric due to $\max(\cdot,0)$ function.

Second, following [11, 16], we define

$$\begin{aligned}\mathcal{L}_{\text{T},2} =& \frac{1}{B}\sum_{i\in B}\log\frac{1}{1+\exp((\|f_t(i)-g_t^s(i^-)\|^2-\|f_t(i)-g_t^s(i)\|^2)/\lambda)}\\ &+\frac{1}{B}\sum_{i\in B}\log\frac{1}{1+\exp((\|f_s(i)-g_s^t(i^-)\|^2-\|f_s(i)-g_s^t(i)\|^2)/\lambda)},\end{aligned} \tag{26}$$

which is non-symmetric due to the $\log$ function.

Third, following [17], we define

$$\begin{aligned}\mathcal{L}_{\text{T},3} =& \frac{\lambda}{B}\sum_{i\in B}\left(\frac{\exp(\|f_t(i)-g_t^s(i)\|_2}{\exp(\|f_t(i)-g_t^s(i)\|_2+\exp((\|f_t(i)-g_t^s(i^-)\|_2)}\right)^2\\ &+\frac{\lambda}{B}\sum_{i\in B}\left(\frac{\exp(\|f_s(i)-g_s^t(i)\|_2)}{\exp(\|f_s(i)-g_s^t(i)\|_2)+\exp(\|f_s(i)-g_s^t(i^-)\|_2)}\right)^2,\end{aligned} \tag{27}$$

which is non-symmetric due to the squared function.

Fourth, following [23], we define

$$\begin{aligned}\mathcal{L}_{\text{T},4} =& \frac{1}{B}\sum_{i\in B}\max\left(1-\frac{\|f_t(i)-g_t^s(i^-)\|^2}{\|f_t(i)-g_t^s(i)\|^2+m},0\right)\\ &+\frac{1}{B}\sum_{i\in B}\max\left(1-\frac{\|f_s(i)-g_s^t(i^-)\|^2}{\|f_s(i)-g_s^t(i)\|^2+m},0\right),\end{aligned} \tag{28}$$

which is also non-symmetric.

In the experiments of this subsection, the only distinction between our method and the others lies in the formulation of the alignment loss. As shown in Table 9, although most of these methods outperform the baselines in the literature of ZSSAR, they perform significantly worse than ours with the calibrated alignment loss due to their absence of symmetry. This emphasizes the effectiveness of our alignment loss design.

Table 9. ZSL accuracy with different alignment loss.

Alignment Loss	NTU-60 (ACC,%)		NTU-120 (ACC,%)	
	55/5 split	48/12 split	110/10 split	96/24 split
$\mathcal{L}_{T,1}$	84.4	45.3	**72.7**	58.6
$\mathcal{L}_{T,2}$	79.9	32.0	59.1	38.7
$\mathcal{L}_{T,3}$	83.8	**49.5**	71.8	**60.7**
$\mathcal{L}_{T,4}$	**85.3**	42.2	69.0	49.7
Ours	**86.9**	**57.2**	**74.4**	**62.5**

F. Frequency-based Representation Analysis for Skeleton Sequences

F.1. Motivation

The Discrete Cosine Transform (DCT) enables lossless feature enhancement through energy-preserving manipulation. The key sight is that the strict energy preservation of DCT and Inverse-DCT (IDCT) between the frequency and time domains: **enhanced components in the frequency domain can be transferred to the time-domain features through IDCT without information loss**. This allows dual semantic enhancements: 1) amplifying low-frequency coefficients enhances global motion patterns (e.g., overarching torso coordination), 2) refining high-frequency components preserves fine-grained kinematics (e.g., hand trajectories) while mitigating the noise. Moreover, this energy-invariant enhancement provides richer information representations for further alignment, where cross-modal correspondences can be learned from both global and local action semantics.

F.2. Frequency Domain Representation and Energy Preservation Proof

Let $\mathbf{s}\in\mathbb{R}^{J\times C\times F}$ denote a skeleton sequence in the time domain, where J is the number of body joints (e.g., 25 joints in NTU-RGB+D dataset), C is the number of coordinate dimensions ($C=3$ for x,y,z coordinates), and F is the temporal length (number of frames). The frequency-domain representation $\mathbf{C}\in\mathbb{R}^{J\times C\times F}$ is obtained through the orthogonal DCT. For each joint $j\in\{1,\dots,J\}$, coordinate $c\in\{1,\dots,C\}$, and frequency index $i\in\{0,\dots,F-1\}$, the transformation is defined as:

$$C_{j,c,i}=\sum_{f=0}^{F-1}s_{j,c,f}\cdot\phi_i(f) \tag{29}$$

where the normalized DCT basis functions $\phi_i(f)$ are given by:

$$\phi_i(f)=\sqrt{\frac{2-\delta_{i0}}{F}}\cdot\cos\left[\frac{\pi}{F}\left(f+\frac{1}{2}\right)i\right], \tag{30}$$

with δ_{i0} denoting the Kronecker delta function (i.e., $\delta_{i0}=1$ when $i=0$ and $\delta_{i0}=0$ otherwise), and $f\in\{0,\dots,F-1\}$.

For any joint j and coordinate c, the energy equivalence between the time and frequency domains is proved as follows:

$$\begin{aligned} E_{\text{freq},j,c} &= \sum_{i=0}^{F-1} C_{j,c,i}^2 \\ &= \sum_{i=0}^{F-1} \left(\sum_{f=0}^{F-1} s_{j,c,f}\, \phi_i(f) \right)^2 \\ &= \sum_{i=0}^{F-1} \sum_{f=0}^{F-1} \sum_{f'=0}^{F-1} s_{j,c,f}\, s_{j,c,f'}\, \phi_i(f)\phi_i(f') \\ &= \sum_{f=0}^{F-1} \sum_{f'=0}^{F-1} s_{j,c,f}\, s_{j,c,f'} \sum_{i=0}^{F-1} \phi_i(f)\phi_i(f') \\ &= \sum_{f=0}^{F-1} s_{j,c,f}^2 = E_{\text{time},j,c}. \end{aligned} \tag{31}$$

The orthogonality relationship [35]

$$\sum_{i=0}^{F-1} \phi_i(f)\phi_i(f') = \begin{cases} 1, & \text{if } f = f' \\ 0, & \text{if } f \neq f' \end{cases}$$

eliminates cross-terms between different frames ($f \neq f'$). Consequently, the energy preservation holds globally:

$$\sum_{j=1}^{J} \sum_{c=1}^{C} \sum_{f=0}^{F-1} s_{j,c,f}^2 = \sum_{j=1}^{J} \sum_{c=1}^{C} \sum_{i=0}^{F-1} C_{j,c,i}^2. \tag{32}$$

F.3. Semantic Integrity with Frequency Adjustment

Given modified coefficients $C'_{j,c,i} = C_{j,c,i} \cdot g(i)$ with scaling function $g(i)$, the reconstructed signal becomes:

$$s'_{j,c,f} = \sum_{i=0}^{F-1} C'_{j,c,i}\phi_i(f) = \sum_{i=0}^{F-1} g(i) C_{j,c,i}\phi_i(f) \tag{33}$$

The modified energy preserves the relationship:

$$\begin{aligned} E'_{\text{time},j,c} &= \sum_{f=0}^{F-1} (s'_{j,c,f})^2 \\ &= \sum_{f=0}^{F-1} \left(\sum_{i=0}^{F-1} g(i) C_{j,c,i}\phi_i(f) \right)^2 \\ &= \sum_{i=0}^{F-1} \sum_{k=0}^{F-1} g(i)g(k) C_{j,c,i} C_{j,c,k} \underbrace{\sum_{f=0}^{F-1} \phi_i(f)\phi_k(f)}_{\delta_{ik}} \\ &= \sum_{i=0}^{F-1} g(i)^2 C_{j,c,i}^2 = E'_{\text{freq},j,c} \end{aligned} \tag{34}$$

This derivation demonstrates three key properties: First, the orthogonal basis eliminates cross-frequency interference during adjustment (δ_{ik} removes terms where $i \neq k$), ensuring distortion-free modifications. Second, energy redistribution follows $E'_{\text{time}} = \sum_i g(i)^2 C_i^2$, allowing controlled enhancement ($g(i) > 1$) or suppression ($g(i) < 1$) of specific frequency. Third, semantic integrity is maintained through the physical meaning of frequency components - low frequencies ($i \leq \varphi$) encode global motion trajectories, while high frequencies ($i > \varphi$) capture local kinematic details (φ is the low-frequency threshold), enabling targeted manipulation without corrupting overall motion semantics.

F.4. Frequency-based Enhancement Mechanism

Since semantic information in skeleton motion is inherently tied to frequency components, higher energy indicates richer information, while energy distribution across frequencies highlights different motion scales. Thus, enhancing skeleton-based frequency components in the frequency domain enriches semantic representation in the time domain (proved above, semantic integrity is preserved during DCT-IDCT), leading to improved generalization in ZSL. This mechanism consists of two adjustments:

Low-Frequency Enhancement. The amplification term $w_i \left(1 - \frac{i}{b}\right)$ is designed to emphasize fundamental movement patterns in skeletal dynamics. By progressively reducing the enhancement effect as frequency increases, this mechanism ensures that low-frequency components, which encode the overall motion structure, are strengthened without distorting the natural motion flow. For whole-body actions such as "walking" or "clapping," it enhances limb coordination and preserves joint continuity.

High-Frequency Suppression. The attenuation term $-w_i \left(1 - \frac{i-b}{b}\right)$ is designed to progressively reduce the suppression effect as frequency increases. This ensures that while high-frequency components are attenuated to mitigate noise and skeletal jitter, fine-grained and rapid motion details are not excessively diminished. The parameter b controls the rate of suppression decay, allowing higher frequency components to retain essential micro-movements, such as finger and wrist gestures.

F.5. Illustration

We also provide the illustration example of our frequency-enhanced mechanism in Fig. 5. Assume the number of the DCT coefficients is 20, the low-frequency threshold φ is 15. As shown in the figure, in the low-frequency range ($i \leq \varphi$), the enhancement applied to the low-frequency coefficients gradually decreases, allowing a smooth transition while preserving global motion integrity. Meanwhile, in the high-frequency range ($i > \varphi$), the suppression of high-frequency coefficients diminishes progressively, allowing essential fine-grained motion details to be retained while

Property	DCT	Wavelet
Energy Compaction	Strong global compaction	Localized
Coefficient Control	Easy frequency separation	Requires multi-scale design
Integration	Simple matrix operations	Needs wavelet basis selection
Usage	Semantic enrichment	Fine-grained separation

Table 10. Comparison between DCT and Wavelet in terms of structural properties and usage for representation learning.

mitigating noise.

F.6. Code

The key part of the implementation of the frequency-enhanced module in our method is presented in Fig. 6. The code snippet provided illustrates the core mechanism of our frequency-aware enhancement strategy within the skeleton decoder. The codes for frequency adjustment with purely learnable weight are also provided in Fig. 7. Extra ablation study and discussion are provided in the main paper.

G. Justification for Choosing DCT

We adopt the Discrete Cosine Transform (DCT) as our frequency encoding method due to its strong energy compaction property and its ability to flexibly separate low- and high-frequency components. These characteristics make it particularly effective for semantic representation learning in zero-shot settings, where training data is limited and fine-grained generalization is critical. Specifically, DCT helps preserve global motion information while enabling localized modulation. This frequency-aware modulation enriches latent representations without requiring strict temporal alignment, aligning well with the post-encoded features.

As shown in Table 10, while wavelet transforms are also viable for signal analysis, they are primarily designed for multi-scale, localized analysis and often require more complex basis selection and hierarchical decomposition. In contrast, DCT is lightweight, easily integrable through matrix operations, and offers more straightforward control over frequency bands for modulation. Our use of DCT is not intended as a traditional frequency separation mechanism, as in prior fully-supervised methods[4, 44], but as a semantic enhancement strategy to improve generalization under zero-shot learning.

H. NTU-60 Dataset Action Index

We also provide the list of action indices from the NTU-60 dataset in Table 11.

Table 11. NTU-60 action classes and their corresponding indices.

Index	Action
1	Drink water
2	Eat meal
3	Brush teeth
4	Brush hair
5	Drop
6	Pick up
7	Throw
8	Sit down
9	Stand up
10	Clapping
11	Reading
12	Writing
13	Tear up paper
14	Put on jacket
15	Take off jacket
16	Put on a shoe
17	Take off a shoe
18	Put on glasses
19	Take off glasses
20	Put on a hat/cap
21	Take off a hat/cap
22	Cheer up
23	Hand waving
24	Kicking something
25	Reach into pocket
26	Hopping
27	Jump up
28	Phone call
29	Play with phone/tablet
30	Type on a keyboard
31	Point to something
32	Taking a selfie
33	Check time (from watch)
34	Rub two hands together
35	Nod head/bow
36	Shake head
37	Wipe face
38	Salute
39	Put palms together
40	Cross hands in front
41	Sneeze/cough
42	Staggering
43	Falling down
44	Headache
45	Chest pain
46	Back pain
47	Neck pain
48	Nausea/vomiting
49	Fan self
50	Punch/slap
51	Kicking
52	Pushing
53	Pat on back
54	Point finger
55	Hugging
56	Giving object
57	Touch pocket
58	Shaking hands
59	Walking towards
60	Walking apart

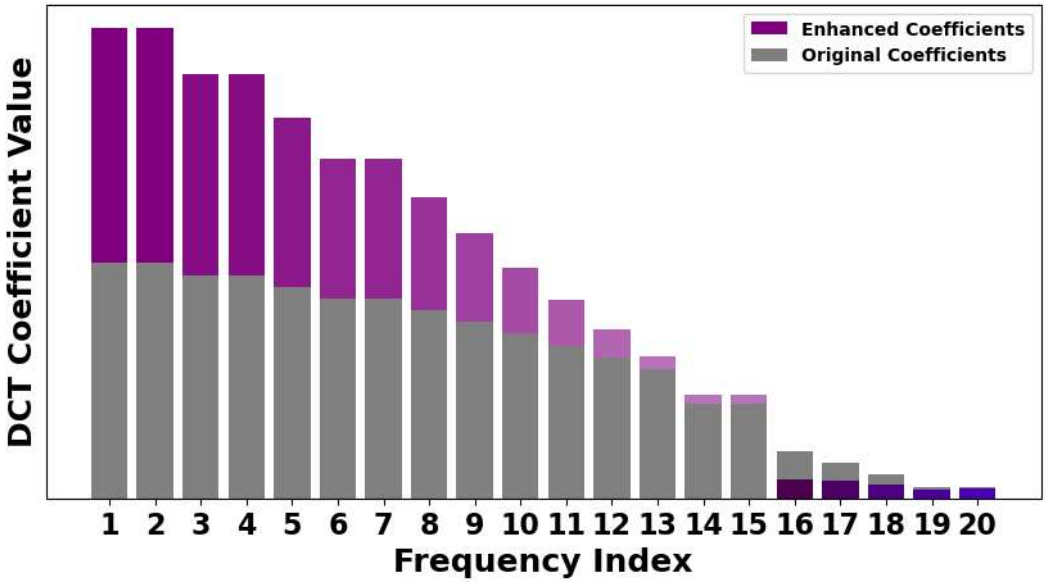


Figure 5. The illustration example of the frequency-enhanced method.

Table 12. Implementation details and parameter settings.

Datasets	**NTU-60**	**NTU-120**
Skeleton Feature Extractor	Shift-GCN [8]	
Text Feature Extractor	CLIP-ViT-B32/16 [34]	
Latent Embedding Dim (Stage 1)	256	512
Latent Embedding Dim (Stage 2)	100	200
Optimizer	Adam	
Learning Rate (Stage 2)	1.0×10^{-4}	
Batch Size (Stage 2)	64	
Training Epochs (Stage 2)	1900	
Unseen Class Features Dim (Stage 3)	500	
Unseen Classifier Epochs (Stage 3)	300	
Unseen Classifier Learning Rate	1.0×10^{-3}	
Classification Gate	Logistic Regression (LBFGS, $C = 1$)	
Frequency Module	DCT-IDCT [2]	
Frequency Parameters	$\varphi = 35$, $b = 30$	
Semantic Descriptions	GPT-4 Generated (LD+GD)	
Calibrated Loss α	0.1	
Calibrated Loss λ	100	
Hardware	NVIDIA A100 $\times$ 1	

Table 13. Examples of action descriptions between baseline and our method.

Action	**Baseline Description**	**Global Description (Ours)**	**Local Description (Ours)**
Eating Meal/Snack	to put food in your mouth, bite it, and swallow it	to pick up food with your hand or utensil, move it to the mouth, and chew	pinch and move the hand up to the head
Brushing Teeth	to clean, polish, or make teeth smooth with a brush	to move a toothbrush back and forth inside your mouth	move the hand up to the head, then tremble the wrist
Brushing Hair	to clean, polish, or make hair smooth with a brush	to run a brush or comb through your hair to smooth it	move the hand up to the head, then move the hand downward
Dropping an Object	to allow something to fall by accident from your hands	to release an object, letting it fall freely to the ground	release the hand in front of the middle of the body

```
# x = input data
# dct = Discrete Cosine Transform function
# b = adjusting parameter
# freq_weight = learnable weight for frequency
# split_freq = threshold for low- and high-frequency adjustment
    def dct_enhance(self, x):
        # Apply DCT to transform input to the frequency domain
        x_dct = dct.dct(x, norm='ortho')
        # Frequency enhancement
        for i in range(self.length_input):
            start = self.split_points[i]
            end = self.split_points[i + 1]
            freq_weight = self.freq_weight[i]
            # Low-frequency adjustment
            if end <= self.split_freq:
                # Scaling function for low frequency
                decay_factor = 1 - i / self.b
                x_dct[:, start:end] *= (1 + freq_weight * decay_factor)
            # High-frequency adjustment
            else:
                # Scaling function for high frequency
                decay_factor = 1 - (i - self.b) / self.b
                x_dct[:, start:end] *= (1 - freq_weight  * decay_factor)
        # Inverse DCT to transform back to the time domain
        return dct.idct(x_dct, norm='ortho')
```

Figure 6. PyTorch codes for frequency enhancement in the encoder.

```
    def dct_enhance(self, x):
        # Apply DCT to transform input to frequency domain
        x_dct = dct.dct(x, norm='ortho')
        for i in range(self.length_input):
            start = self.split_points[i]
            end = self.split_points[i + 1]
            freq_weight = self.freq_weight[i]
            # Apply learnable weight directly
            x_dct[:, start:end] *= freq_weight
        # Inverse DCT to transform back to time domain
        return dct.idct(x_dct, norm='ortho')
```

Figure 7. PyTorch codes for frequency enhancement with pure learnable weights.