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

Wenhan Wu[1], Zhishuai Guo[2], Chen Chen[3], Hongfei Xue[1], Aidong Lu[1]
[1]Department of Computer Science, University of North Carolina at Charlotte
[2]Department of Computer Science, Northern Illinois University
[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

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 [18]. 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 [5]: 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.

Table 1. 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[8]	ICCV2017	59.3	60.9	42.2	49.8
JPoSE[19]	ICCV2019	57.2	60.3	45.2	51.6
CADA-VAE[16]	CVPR2019	60.7	63.2	35.9	45.8
SynSE[5]	ICIP2021	53.9	63.1	40.7	49.5
SMIE[22]	ACMM2023	60.8	-	-	-
SA-DVAE[11]	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 2. 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**

NTU RGB+D 120 [13]. 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 [5].

PKU-MMD [12]. 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 [10, 11] and the skeleton features provided by [11] for a fair comparison (skeleton features are generated by ST-GCN[21], 46/5 split settings, 46 seen classes and 5 unseen classes).

B.2. Training Strategy

The training phase follows the same processing procedure as [10], which is systematically organized into four stages: training the skeleton feature extractor to capture spatio-temporal dependencies, optimizing the generative cross-

Table 3. 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**

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 7 shows the parameter settings of our method, including the parameters applied during all the training stages mentioned in the main paper and [10].

C. More Experiments

Results on PKU-MMD. Table 1 presents the ZSL and GZSL performance on the PKU-MMD dataset under the 46/5 split settings [11]. 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 2, 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 3, ViT-B/32 also outperforms ViT-B/16 in harmonic mean (H-score), 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 8 illustrates how our method refines action descriptions by incorporating both global and local semantic components. Compared to the baseline[10], 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

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

Let

$$\mathcal{L}_{q,m} := \lambda \sum_{i \in S_q} \sum_{j \in S_m} \ell^q(i,j), \tag{2}$$

where

$$\ell^q(i,j) = \frac{1}{1 + \exp((\|f_t^q - g_t^s(j)\|^2 - \|f_t^q - g_t^s(i)\|^2)/\lambda)}. \tag{3}$$

Rearranging the terms, we can rewrite the loss function as

$$\mathcal{L}^1_{Align} = \mathcal{L}_{1,2} + \mathcal{L}_{1,3} + \mathcal{L}_{2,1} + \mathcal{L}_{2,3} + \mathcal{L}_{3,1} + \mathcal{L}_{3,2}, \tag{4}$$

where

$$\mathcal{L}_{1,2} = \lambda \sum_{i \in S_1/\tilde{S}} \sum_{j \in S_2} \ell^1(i,j) + \lambda \underbrace{\sum_{i \in \tilde{S}} \sum_{j \in S_2} \ell^1(i,j)}_{(A)}. \tag{5}$$

$$\mathcal{L}_{1,3} = \lambda \sum_{i \in S_1/\tilde{S}} \sum_{j \in S_3} \ell^1(i,j) + \lambda \underbrace{\sum_{i \in \tilde{S}} \sum_{j \in S_3} \ell^1(i,j)}_{(B)}. \tag{6}$$

$$\mathcal{L}_{2,1} = \lambda \sum_{i \in S_2} \sum_{j \in S_1/\tilde{S}} \ell^2(i,j) + \lambda \underbrace{\sum_{i \in S_2} \sum_{j \in \tilde{S}} \ell^2(i,j)}_{(C)}. \tag{7}$$

$$\mathcal{L}_{2,3} = \lambda \sum_{i \in S_2} \sum_{j \in S_3} \ell^2(i,j) \tag{8}$$

$$\mathcal{L}_{3,1} = \lambda \sum_{i \in S_3} \sum_{j \in S_1/\tilde{S}} \ell^3(i,j) + \lambda \underbrace{\sum_{i \in S_3} \sum_{j \in \tilde{S}} \ell^3(i,j)}_{(D)} \tag{9}$$

$$\mathcal{L}_{3,2} = \lambda \sum_{i \in S_3} \sum_{j \in S_2} \ell^3(i,j) \tag{10}$$

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.,*

$$\mathbb{E}_{S_1,S_2,S_3}[A] = \mathbb{E}_{S_1,S_2,S_3}[C] = 1. \tag{11}$$

Proof. For term A, we have

$$\begin{aligned} &\mathbb{E}_{\tilde{S},S_2}\left[\lambda \sum_{i \in \tilde{S}} \sum_{j \in S_2} \ell^1(i,j)\right] = \lambda \tilde{n} n_2 \mathbb{E}_{i \in P_2} \mathbb{E}_{j \in P_2} \ell^1(i,j) \\ &= \lambda \tilde{n} n_2 \mathbb{E}_{i \in P_2} \mathbb{E}_{j \in P_2} \frac{\ell^1(i,j) + \ell^1(j,i)}{2}, \end{aligned} \tag{12}$$

where

$$\begin{aligned} &\frac{\ell^1(i,j) + \ell^1(j,i)}{2} \\ &= \frac{1}{1 + \exp((\|f_t^1 - g_t^s(j)\|^2 - \|f_t^1 - g_t^s(i)\|^2)/\lambda)} \\ &\quad + \frac{1}{1 + \exp((\|f_t^1 - g_t^s(i)\|^2 - \|f_t^1 - g_t^s(j)\|^2)/\lambda)} \\ &= \frac{\exp((\|f_t^1 - g_t^s(i)\|^2 - \|f_t^1 - g_t^s(j)\|^2)/\lambda)}{1 + \exp((\|f_t^1 - g_t^s(i)\|^2 - \|f_t^1 - g_t^s(j)\|^2)/\lambda)} \\ &\quad + \frac{1}{1 + \exp((\|f_t^1 - g_t^s(i)\|^2 - \|f_t^1 - g_t^s(j)\|^2)/\lambda)} \\ &= 1. \end{aligned} \tag{13}$$

Similarly, for term C we obtain that

$$\begin{aligned} &\mathbb{E}_{S_2,\tilde{S}}\left[\lambda \sum_{i \in S_2} \sum_{j \in \tilde{S}} \ell^2(i,j)\right] = \lambda n_2 \tilde{n} \mathbb{E}_{i \in P_2} \mathbb{E}_{j \in P_2} \ell^2(i,j) \\ &= \lambda n_2 \tilde{n} \mathbb{E}_{i \in P_2} \mathbb{E}_{j \in P_2} \frac{\ell^2(i,j) + \ell^2(j,i)}{2}, \end{aligned} \tag{14}$$

where

$$\begin{aligned}
&\frac{\ell^2(i,j)+\ell^2(j,i)}{2} \\
&= \frac{1}{1+\exp((\|f_t^2-g_t^s(j)\|^2-\|f_t^2-g_t^s(i)\|^2)/\lambda)} \\
&\quad + \frac{1}{1+\exp((\|f_t^2-g_t^s(i)\|^2-\|f_t^2-g_t^s(j)\|^2)/\lambda)} \\
&= \frac{\exp((\|f_t^2-g_t^s(i)\|^2-\|f_t^2-g_t^s(j)\|^2)/\lambda)}{1+\exp((\|f_t^2-g_t^s(i)\|^2-\|f_t^2-g_t^s(j)\|^2)/\lambda)} \\
&\quad + \frac{1}{1+\exp((\|f_t^2-g_t^s(i)\|^2-\|f_t^2-g_t^s(j)\|^2)/\lambda)} \\
&= 1.
\end{aligned} \tag{15}$$

□

For term B, which is given by

$$\sum_{i\in\tilde{S}}\sum_{j\in S_3}\ell^1(i,j) = \frac{1}{1+\exp((\|f_t^1-g_t^s(j)\|^2-\|f_t^1-g_t^s(i)\|^2)/\lambda)}, \tag{16}$$

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 4.

Specifically, the triplet alignment losses are developed based on popular triplet loss formulations, as follows. First, following the work of [17], we define:

$$\begin{aligned}
\mathcal{L}_{\text{T},1} =& \frac{1}{B}\sum_{i\in B}\max(\|f_t(i)-g_t^s(i)\|^2-\|f_t(i)-g_t^s(i^-)\|^2+m,0) \\
&+\frac{1}{B}\sum_{i\in B}\max(\|f_s(i)-g_s^t(i)\|^2-\|f_s(i)-g_s^t(i^-)\|^2+m,0),
\end{aligned} \tag{17}$$

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

Second, following [4, 6], we define

$$\begin{aligned}
\mathcal{L}_{\text{T},2} =& \frac{1}{B}\sum_{i\in B}\log\frac{1}{1+\exp((\|f_t(i)-g_t^s(i^-)\|^2-\|f_t(i)-g_t^s(i)\|^2)/\lambda)} \\
&+\frac{1}{B}\sum_{i\in B}\log\frac{1}{1+\exp((\|f_s(i)-g_s^t(i^-)\|^2-\|f_s(i)-g_s^t(i)\|^2)/\lambda)},
\end{aligned} \tag{18}$$

which is non-symmetric due to the log function.

Third, following [7], we define

$$\begin{aligned}
\mathcal{L}_{\text{T},3} =& \frac{\lambda}{B}\sum_{i\in B}\left(\frac{\exp(\|f_t(i)-g_t^s(i)\|_2}{\exp(\|f_t(i)-g_t^s(i)\|_2+\exp((\|f_t(i)-g_t^s(i^-)\|_2)}\right)^2 \\
&+\frac{\lambda}{B}\sum_{i\in B}\left(\frac{\exp(\|f_s(i)-g_s^t(i)\|_2)}{\exp(\|f_s(i)-g_s^t(i)\|_2)+\exp(\|f_s(i)-g_s^t(i^-)\|_2)}\right)^2,
\end{aligned} \tag{19}$$

which is non-symmetric due to the squared function.

Fourth, following [9], we define

$$\begin{aligned}
\mathcal{L}_{\text{T},4} =& \frac{1}{B}\sum_{i\in B}\max\left(1-\frac{\|f_t(i)-g_t^s(i^-)\|^2}{\|f_t(i)-g_t^s(i)\|^2+m},0\right) \\
&+\frac{1}{B}\sum_{i\in B}\max\left(1-\frac{\|f_s(i)-g_s^t(i^-)\|^2}{\|f_s(i)-g_s^t(i)\|^2+m},0\right),
\end{aligned} \tag{20}$$

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 4, 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 4. 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:

$$C_{j,c,i} = \sum_{f=0}^{F-1} s_{j,c,f} \cdot \phi_i(f) \tag{21}$$

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

$$\phi_i(f) = \sqrt{\frac{2-\delta_{i0}}{F}} \cdot \cos\left[\frac{\pi}{F}\left(f+\frac{1}{2}\right)i\right], \tag{22}$$

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:

$$\begin{aligned}
E_{\text{freq},j,c} &= \sum_{i=0}^{F-1} C_{j,c,i}^2 \\
&= \sum_{i=0}^{F-1} \left(\sum_{f=0}^{F-1} s_{j,c,f}\, \phi_i(f) \right)^2 \\
&= \sum_{i=0}^{F-1} \sum_{f=0}^{F-1} \sum_{f'=0}^{F-1} s_{j,c,f}\, s_{j,c,f'}\, \phi_i(f)\phi_i(f') \\
&= \sum_{f=0}^{F-1} \sum_{f'=0}^{F-1} s_{j,c,f}\, s_{j,c,f'} \sum_{i=0}^{F-1} \phi_i(f)\phi_i(f') \\
&= \sum_{f=0}^{F-1} s_{j,c,f}^2 = E_{\text{time},j,c}.
\end{aligned} \tag{23}$$

The orthogonality relationship [15]

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

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

$$\sum_{j=1}^{J}\sum_{c=1}^{C}\sum_{f=0}^{F-1} s_{j,c,f}^2 = \sum_{j=1}^{J}\sum_{c=1}^{C}\sum_{i=0}^{F-1} C_{j,c,i}^2. \tag{24}$$

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:

$$s'_{j,c,f} = \sum_{i=0}^{F-1} C'_{j,c,i}\phi_i(f) = \sum_{i=0}^{F-1} g(i) C_{j,c,i}\phi_i(f) \tag{25}$$

The modified energy preserves the relationship:

$$\begin{aligned}
E'_{\text{time},j,c} &= \sum_{f=0}^{F-1} (s'_{j,c,f})^2 \\
&= \sum_{f=0}^{F-1} \left(\sum_{i=0}^{F-1} g(i) C_{j,c,i}\phi_i(f) \right)^2 \\
&= \sum_{i=0}^{F-1}\sum_{k=0}^{F-1} g(i)g(k) C_{j,c,i}C_{j,c,k} \underbrace{\sum_{f=0}^{F-1} \phi_i(f)\phi_k(f)}_{\delta_{ik}} \\
&= \sum_{i=0}^{F-1} g(i)^2 C_{j,c,i}^2 = E'_{\text{freq},j,c}
\end{aligned} \tag{26}$$

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

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

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. 1. 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 mitigating noise.

F.6. Code

The key part of the implementation of the frequency-enhanced module in our method is presented in Fig. 2. 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. 3. 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 5, 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[2, 20], 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 6.

Table 6. 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

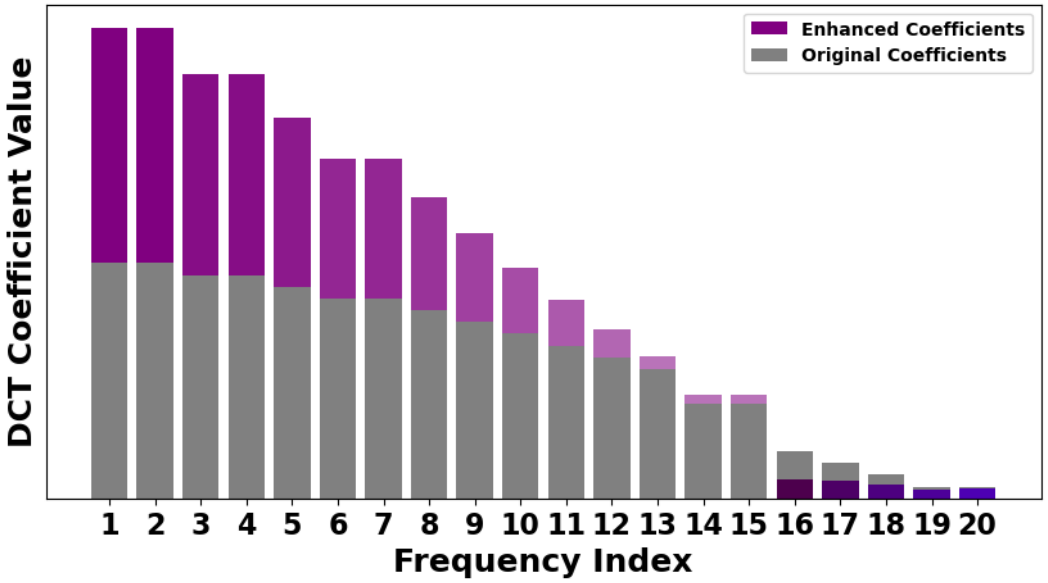

Figure 1. The illustration example of the frequency-enhanced method.

Table 7. Implementation details and parameter settings.

Datasets	**NTU-60**	**NTU-120**
Skeleton Feature Extractor	Shift-GCN [3]	
Text Feature Extractor	CLIP-ViT-B32/16 [14]	
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 [1]	
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 8. 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 2. 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 3. PyTorch codes for frequency enhancement with pure learnable weights.

References

[1] Nasir Ahmed, T_ Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1): 90–93, 1974. 8

[2] 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. 6

[3] 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. 8

[4] 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. 4

[5] 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

[6] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 4

[7] 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. 4

[8] 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

[9] 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. 4

[10] 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

[11] 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

[12] 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. 1

[13] 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. 1

[14] 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. 8

[15] K Ramamohan Rao and Ping Yip. *Discrete cosine transform: algorithms, advantages, applications*. Academic press, 2014. 5

[16] 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. 1

[17] 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. 4

[18] 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. 1

[19] 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. 1

[20] 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. 6

[21] 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. 1

[22] 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