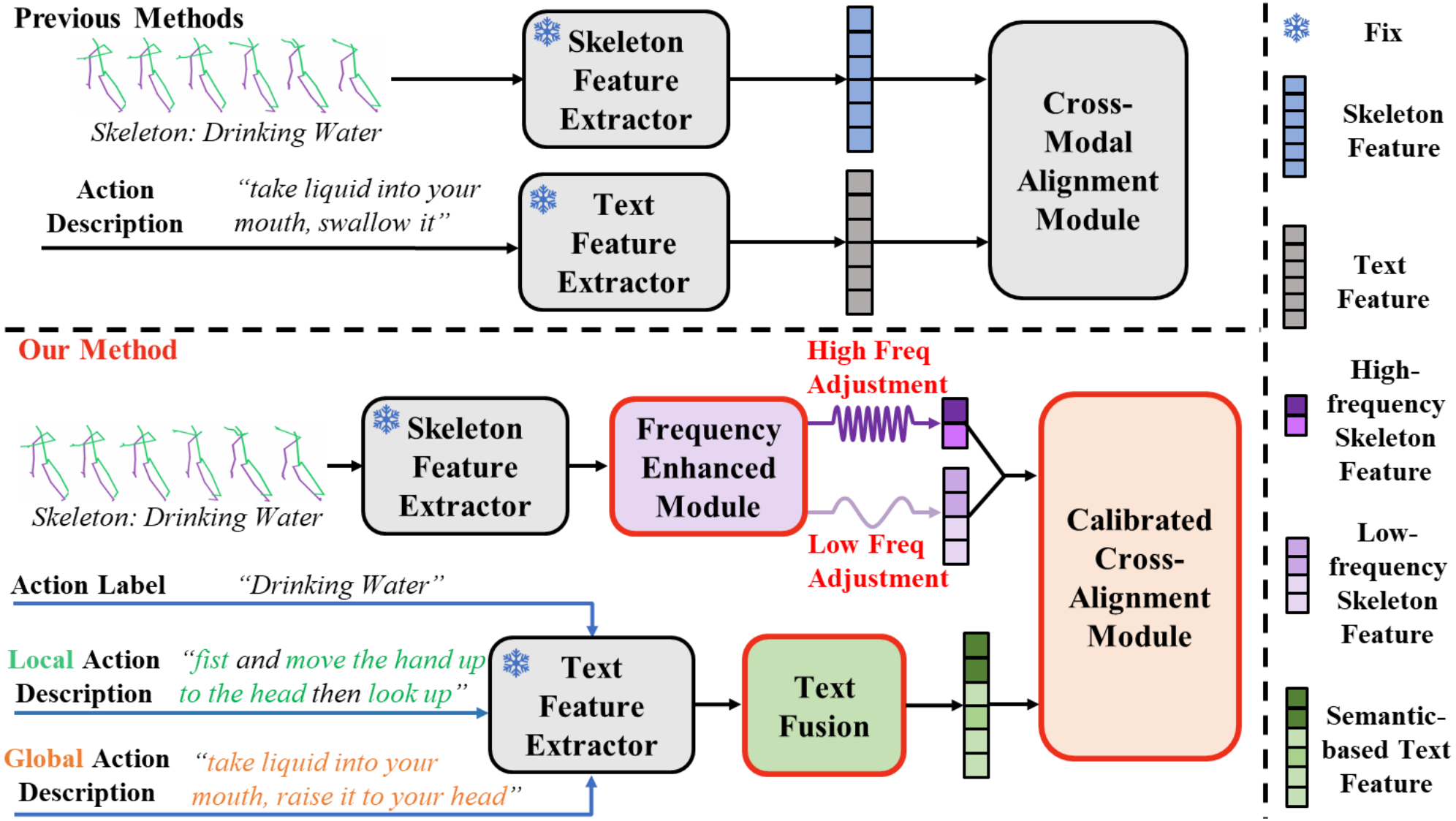
ICCV OCT 19-23, 2025 HONOLULU HAWAII

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]UNC Charlotte [2]Northern Illinois University [3]University of Central Florida

FS-VAE： Background



Contributions

- 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.

Problem Formulation

Task. Zero-Shot Skeleton-based Action Recognition (ZSSAR) aims to classify actions from **unseen** categories using knowledge learned from **seen** categories.

Dataset and notation.

$$\mathcal{D} = \{(\mathbf{X}_i, \mathbf{C}_i, \mathbf{A}_i)\}_{i=1}^{N}, \qquad \mathbf{X}_i \in \mathbb{R}^{J\times 3\times F\times M}.$$

$\mathbf{X}_i$: skeleton sequence (with J joints, 3D coords, F frames, M subjects); $\mathbf{C}_i$: action category; $\mathbf{A}_i$: GPT-generated semantic description.

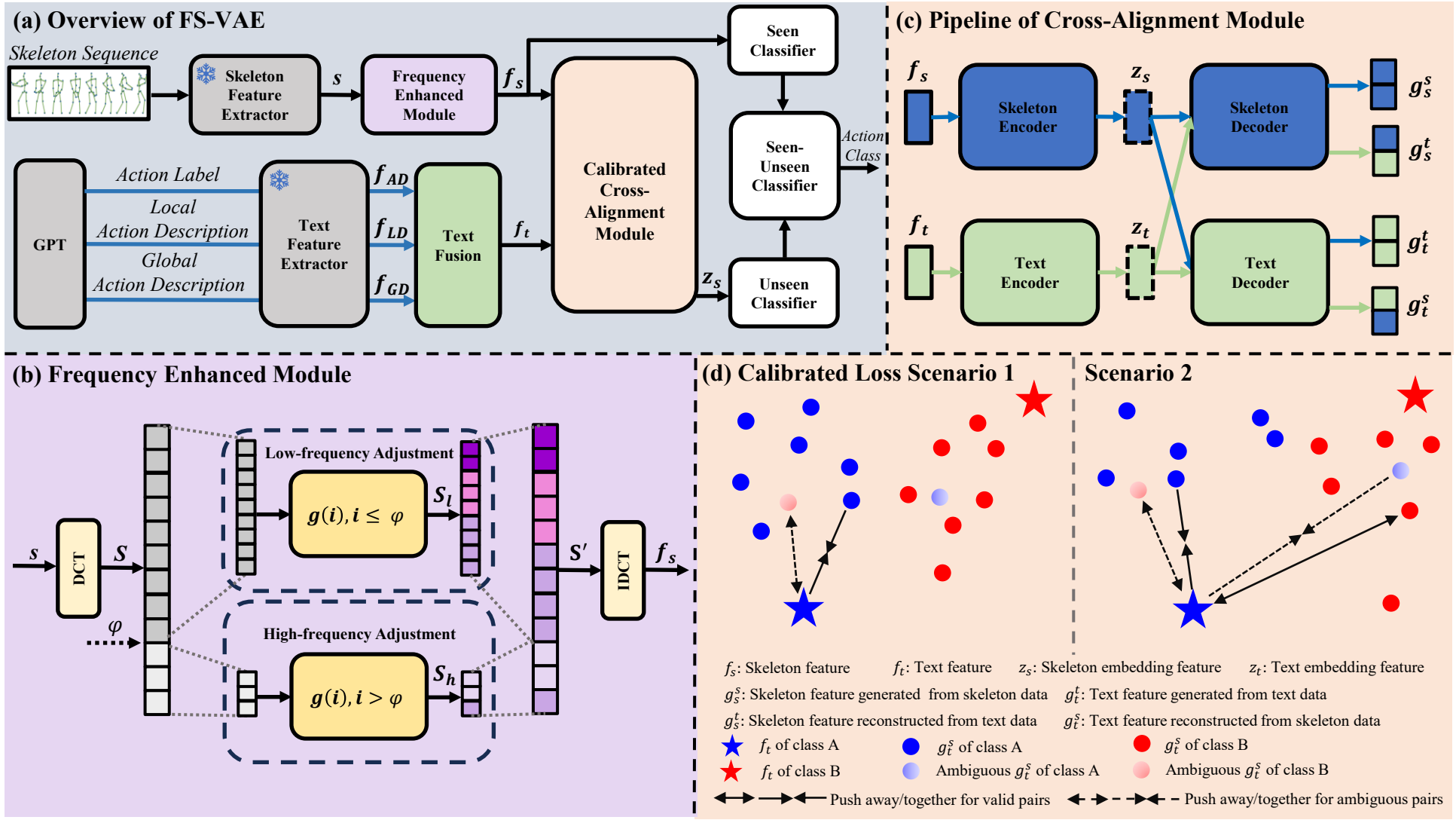
Seen/unseen splits.

$$\mathcal{D}^{s}_{\mathrm{tr}} \text{ (seen classes } \mathcal{C}_s), \quad \mathcal{D}^{u}_{\mathrm{te}} \text{ (unseen classes } \mathcal{C}_u), \quad \mathcal{D}^{s}_{\mathrm{te}} \text{ (seen classes)}.$$

$$\mathcal{C} = \mathcal{C}_s \cup \mathcal{C}_u, \qquad \mathcal{C}_s \cap \mathcal{C}_u = \varnothing.$$

Each action category is associated with a semantic description $\mathbf{A}_i$.

Overview


(a) Overview of FS-VAE
Skeleton Sequence
Skeleton Feature Extractor
s
Frequency Enhanced Module
f_s
Seen Classifier
Seen-Unseen Classifier
Action Class
Calibrated Cross-Alignment Module
GPT
Action Label
Local Action Description
Global Action Description
Text Feature Extractor
f_{AD}
f_{LD}
f_{GD}
Text Fusion
f_t
z_s
Unseen Classifier
(c) Pipeline of Cross-Alignment Module
Skeleton Encoder
Skeleton Decoder
Text Encoder
Text Decoder
z_s
z_t
g_s^s
g_s^t
g_t^t
g_t^s
(b) Frequency Enhanced Module
DCT
S
φ
Low-frequency Adjustment
$g(i), i \leq \varphi$
S_l
High-frequency Adjustment
$g(i), i > \varphi$
S_h
S'
IDCT
(d) Calibrated Loss Scenario 1
Scenario 2
f_s: Skeleton feature
f_t: Text feature
z_s: Skeleton embedding feature
z_t: Text embedding feature
g_s^s: Skeleton feature generated from skeleton data
g_t^t: Text feature generated from text data
g_s^t: Skeleton feature reconstructed from text data
g_t^s: Text feature reconstructed from skeleton data
f_t of class A
f_t of class B
g_t^s of class A
Ambiguous g_t^s of class A
g_t^s of class B
Ambiguous g_t^s of class B
Push away/together for valid pairs
Push away/together for ambiguous pairs

Frequency Adjustment

Scaling function for low-frequency components:

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

$$g(i) = 1 + w_i \left(1 - \frac{i}{b}\right)$$

- $\mathbf{S}_l$: low-frequency components capturing **global motion patterns** (e.g., torso and limb movement).
- $g(i)$: gradually decreasing scaling function, ensuring smooth enhancement within the low-frequency range.
- b: adjusting parameter that controls how enhancement strength reduces as frequency increases.
- w_i: learned weight adaptively amplifying the most discriminative frequencies.

Frequency Adjustment

Scaling function for high-frequency components:

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

$$g(i) = 1 - w_i \left(1 - \frac{i-b}{b}\right)$$

- $\mathbf{S}_h$: high-frequency components capturing **fine-grained details** (e.g., finger, wrist, rapid motions).
- $g(i)$: progressively reduces suppression as frequency increases, avoiding over-attenuation.
- b: normalization factor ensuring smooth suppression transition across frequencies.
- w_i: learned weight adaptively modulating suppression for different high-frequency components.

Semantic-based Action Description

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

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

Calibrated Cross-Alignment Loss

1. Calibrated Alignment Loss

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

2. VAE Reconstruction Loss

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

3. Overall Loss

$$\mathcal{L}_{VAE}^{cali} = \mathcal{L}_{VAE} + \alpha\mathcal{L}_{\text{Align}}$$

Notation:

- $f_t(i), f_s(i)$: text and skeleton encoders
- $g_t^s(i), g_s^t(i)$: skeleton $\rightarrow$ text and text $\rightarrow$ skeleton projections
- i^-: negative sample of i in the batch
- λ: temperature parameter controlling alignment sensitivity

Results

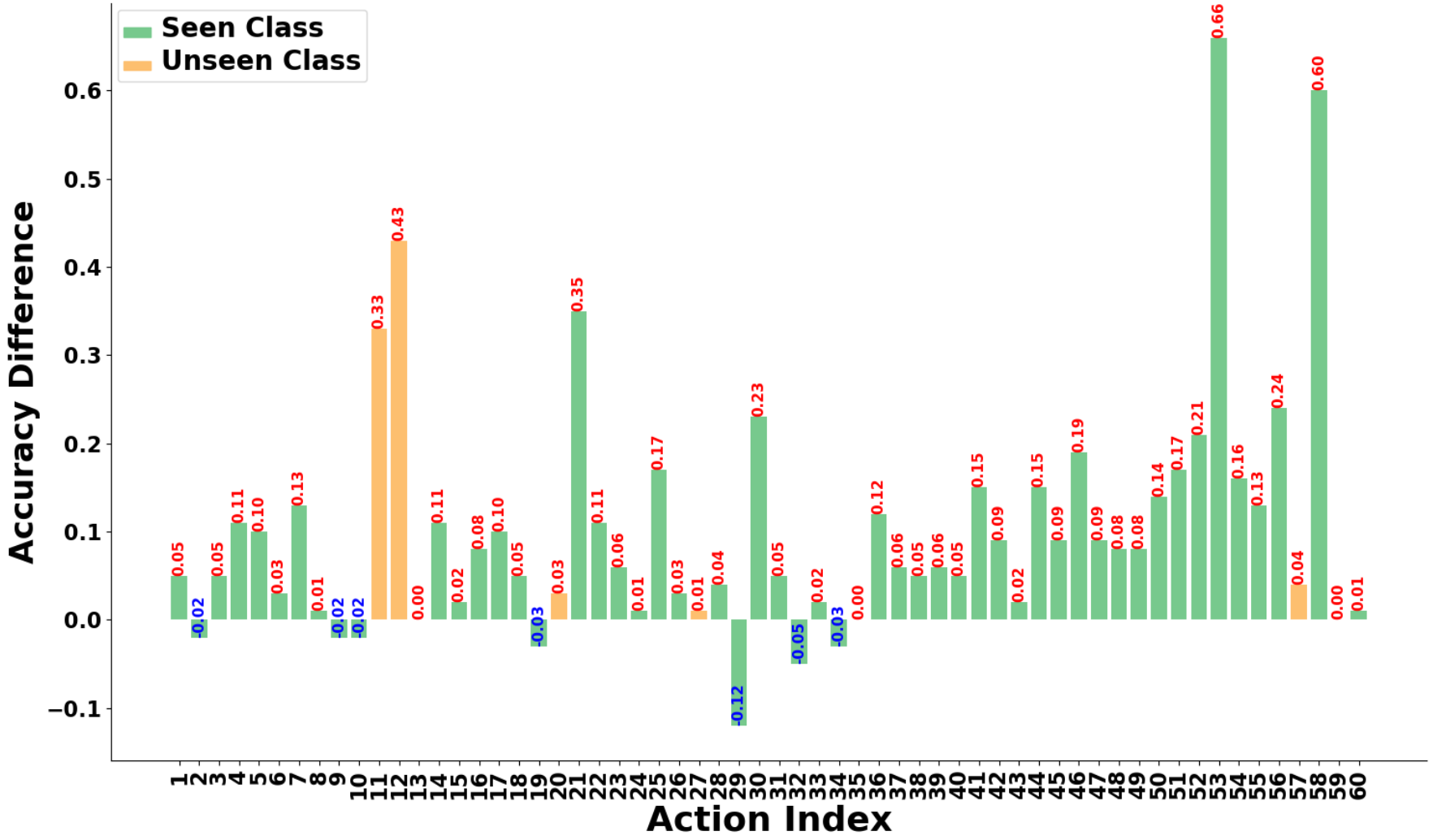
Methods	Venue	NTU-60 (ACC,%)		NTU-120 (ACC,%)	
		55/5 split	48/12 split	110/10 split	96/24 split
ReViSE[19]	ICCV2017	53.9	17.5	55.0	32.4
JPoSE[42]	ICCV2019	64.8	28.8	51.9	32.4
CADA-VAE[35]	CVPR2019	76.8	29.0	59.5	35.8
SynSE[15]	ICIP2021	75.8	33.3	62.7	38.7
SMIE[48]	ACMM2023	78.0	40.2	61.3	42.3
STAR[7]	ACMM2024	81.4	45.1	63.3	44.3
GZSSAR*[27]	ICIG2023	83.3	49.8	72.0	60.7
PURLS[49]	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**$_{\uparrow 1.8}$

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 reproduced results of the released codes. † denotes the use of only wi for frequency coefficients.

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[42]	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[35]	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**$_{\uparrow 3.0}$

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.

Results



Accuracy difference for seen-unseen actions compared to baseline, it show outperforming recognition results on the most of unseen actions.

Thank you !

Project GitHub