Supplementary Material for *SkeletonContext: Skeleton-side Context Prompt Learning for Zero-Shot Skeleton-based Action Recognition*

Ning Wang[1,2], Tieyue Wu[2], Naeha Sharif[3], Farid Boussaid[3], Guangming Zhu[2], Lin Mei[4], Mohammed Bennamoun[3], Liang Zhang[2*]
[1]Chang'an University,[2]Xidian University, [3]University of Western Australia,[4]Donghai Lab
ning@chd.edu.cn, gxyzd648@gmail.com, {gmzhu, liangzhang}@xidian.edu.cn,
{naeha.sharif,farid.boussaid,mohammed.bennamoun}@uwa.edu.au, meilin@donghailab.com

In this appendix, we present additional experimental results, representative failure cases, LLM-generated contextual descriptions, and more qualitative visualizations.

1. More Experimental Results

Methods	NTU-60		NTU-120	
	40/20	30/30	80/40	60/60
PURLS	31.1	23.5	28.4	19.6
TDSM	36.1	25.9	37.0	27.2
Ours	**39.3**	**30.1**	**39.1**	**28.4**

Table 1. Zero-shot action recognition accuracy (%) on NTU-60 and NTU-120 under more challenging unseen-class splits.

Validity on more challenging seen/unseen splits. Following the PURLS and TDSM protocols, we evaluate SkeletonContext on more challenging seen/unseen splits. As shown in Table 1, we achieves competitive performance under these settings.

Effect of Semantic Meaning in Context Prompts. Table 2 evaluates whether the model truly leverages the semantic content of the LLM-generated contextual descriptions. When we replace contextual descriptions with those from other categories by shuffling the class–context pairs (Random Context) or fill each contextual slot with randomly sampled words while keeping the slot structure intact (Random Context Slots), the performance drops substantially on both ZSL and GZSL metrics. This demonstrates that our approach does not merely exploit the prompt format, but depends on coherent semantic information to enhance cross-modal alignment. Using the original LLM-generated context achieves the best results.

Fine-Grained Action Recognition. To further compare the discriminative capability of our SkeletonContext with FS-VAE [3], we compute the per-class accuracy difference on unseen categories. As illustrated in Fig. 1, our method delivers notable improvements across several actions. In particular, actions such as "pushing", "touch pocket", and "clapping" exhibit clear performance gains owing to the proposed cross-modal context prompt module. By reconstructing contextual semantics (e.g., interaction intent, involved objects, and surrounding environment) from skeleton motion, our model enhances semantic completeness on actions where subtle pose variations alone are insufficient for accurate discrimination. Meanwhile, "walking towards" and "falling down" benefit significantly from our key-part decoupling module, which explicitly emphasizes the most informative body regions and mitigates interference from irrelevant joints. This allows the model to better capture the core motion dynamics of actions dominated by specific body-part movements.

However, several categories remain challenging. Actions such as "nausea or vomiting" remain difficult because their discriminative cues primarily lie in subtle, high-frequency torso and head motions. FS-VAE [3]'s frequency-domain refinement is well suited to capture these micro-dynamic signals, whereas our context-driven reconstruction provides limited benefit due to the inherently weak and ambiguous contextual cues associated with such actions.

Table 2. **Effect of contextual semantic validity on ZSL/GZSL performance.** Replacing LLM-generated contextual descriptions with randomized context or shuffled slots leads to clear performance degradation, indicating that SkeletonContext benefits from meaningful semantic cues.

Context	NTU-60 (55/5 split)	
	ZSL	GZSL
Random Context	82.8	73.7
Random Context Slots	82.5	74.3
Context	**89.6**	**77.1**

*Corresponding author.

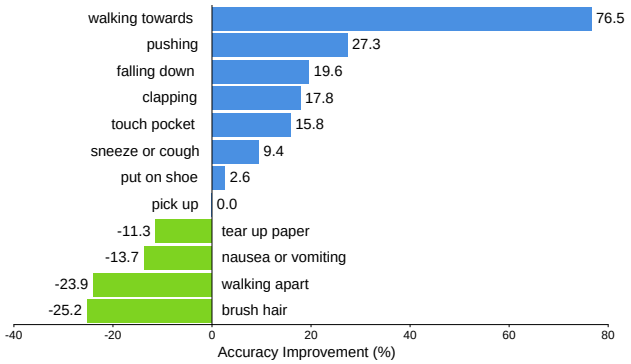


Figure 1. **Per-class performance comparison on unseen actions.** The results show that SkeletonContext performs well on many categories involving clear contextual cues.

Computation Efficiency Comparison As shown in Table 3, SkeletonContext achieves the best overall performance among the compared methods. Although the full model incurs a higher computational cost (2.26 GFLOPs) due to the context prompt reconstruction, it delivers the highest ZSL/GZSL accuracy. Notably, the lightweight variant Ours†, which uses pre-extracted BERT [1] template features, reduces computation to 0.73 GFLOPs, comparable to Neuron [2] and FS-VAE [3], while still maintaining competitive performance. This demonstrates that the framework can flexibly trade computation for accuracy depending on practical requirements.

Table 3. **Comparison of computational cost and performance.** Our model achieves the best accuracy with a still acceptable computational cost, while the lightweight variant (Ours†), which uses pre-extracted BERT [1] features, significantly reduces computation while still achieving competitive performance compared to recent methods.

Method	GFLOPs	NTU-60 (55/5 split)	
		ZSL	GZSL
Neuron [2]	0.65	86.9	71.4
FS-VAE [3]	0.52	86.9	75.7
Ours	2.26	**89.6**	**77.1**
Ours†	0.73	87.6	76.2

2. Failure Case Analysis

We observed that some actions with highly similar motion patterns remain difficult to distinguish, even when accurate contextual information is provided. As shown in Fig. 2, actions involving close hand contact and subtle movement differences can still be confused, indicating that contextual cues alone offer limited discrimination for such fine-grained cases. In the future, we will further focus on more fine-grained spatiotemporal skeleton modeling to better capture subtle motion distinctions.

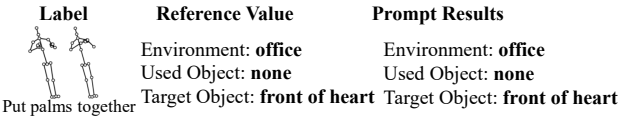


Figure 2. **An example of a failure case.** Despite providing the correct contextual prompt, the model still misidentifies the action as *Rub two hands*, owing to the subtle motion differences between *Rub two hands* and *Put palms together*.

3. Details of Context Description

To obtain context-rich semantic descriptions, we use the ChatGPT to generate structured action descriptions following the template: *"Please describe the following human skeleton actions in the form of "In [environment], [part of the body] uses [object] to do [action] on [object]". Please note that some actions may not require a used object or target object, in which case the description of [object] can be write [none]. Try to make the descriptions of the actions as different as possible, but [environment] and [object] can co-occur in different categories. The body parts of a person include: head, hand, arm, hip, leg, foot. For example, the relevant body parts for a person to perform the action "drink water" are [hand, head], and the used object is [cup]. The relevant body parts for a person to perform the action "reading" are [hand, head], and the used object is [book]. For each action name, please generate 10 different descriptions following the template. Don't output extra texts other than ones in the template.Please describe the following action:"*

As shown in Table 4, we generate multiple contextual descriptions for the same action to enrich its semantic diversity. Taking writing as an example, the LLM produces ten variants that differ in environment (e.g., classroom, office, library, studio) and interacting objects (e.g., pen–paper, stylus–tablet, chalk–board, marker–whiteboard), while preserving the core action semantics. This multi-description strategy provides the model with a broader range of plausible contexts for a single motion pattern, enabling it to learn more flexible and robust cross-modal associations. Consequently, the model becomes better equipped to reconstruct context from skeletal cues and generalize to unseen categories.

4. More Visualization Results

Figure 3 presents additional visualization results of our context reconstruction module. For each input skeleton sequence, the model predicts the corresponding contex-

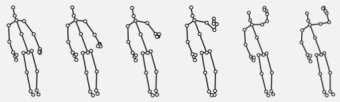
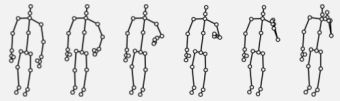
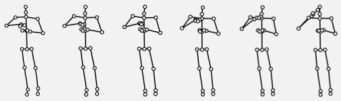
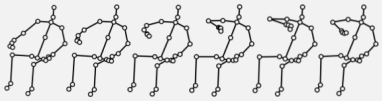
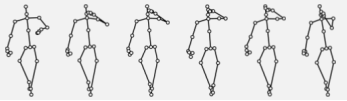
Label	Visualized Sequence	Reference Value	Prompt Results
Taking a selfie		tourist | camera | face	none | camera | none
Phone call		home | phone | ear	home | phone | none
Eat meal		dining | fork | plate	dining | fork | plate
Check time		street | watch | wrist	room | watch | none
Brush teeth		bathroom | toothbrush | sink	room | brush | tooth

Figure 3. **Additional visualization of reconstructed contextual semantics.** For various actions, the model predicts the environment, used object, and target object slots from skeleton-only inputs, demonstrating its ability to infer plausible action-related context.

Table 4. **Examples of LLM-generated contextual descriptions for the action writing.** Each description follows the structured template and provides diverse environments, interacting objects, and targets for the same action.

Action: writing
1. In **a classroom**, **hand and arm** use **pen** to do **writing** on **paper**.
2. In **an office**, **hand, arm and head** use **pencil** to do **writing** on **notebook**.
3. In **a library**, **hand and arm** use **marker** to do **writing** on **whiteboard**.
4. In **a studio**, **hand and arm** use **brush** to do **writing** on **canvas**.
5. In **a meeting room**, **hand and arm** use **stylus** to do **writing** on **tablet**.
6. In **a hallway**, **hand and arm** use **chalk** to do **writing** on **board**.
7. In **a bedroom**, **hand, arm and head** use **keyboard** to do **writing** on **computer**.
8. In **a café**, **hand and arm** use **pen** to do **writing** on **journal**.
9. In **a workshop**, **hand and arm** use **engraving tool** to do **writing** on **metal plate**.
10. In **a garden**, **hand and arm** use **stick** to do **writing** on **ground**.

tual slots and aligns them with the action label. As shown, actions such as eat meal, phone call, and check time are paired with coherent environment, used-object, and target-object predictions (e.g., "dining–plate–plate" for eat meal, "home–phone–ear" for phone call, and "street–watch–wrist" for check time). These examples illustrate that SkeletonContext can reliably infer plausible and action-relevant contextual semantics solely from motion cues, even when the interacting objects are not explicitly present in the skeleton modality. This demonstrates the model's ability to form consistent cross-modal associations and further validates the effectiveness of language-driven context reconstruction.

References

[1] Shivaji Alaparthi and Manit Mishra. Bidirectional encoder representations from transformers (bert): A sentiment analysis odyssey. *arXiv preprint arXiv:2007.01127*, 2020. 2

[2] Yang Chen, Jingcai Guo, Song Guo, and Dacheng Tao. Neuron: Learning context-aware evolving representations for zero-shot skeleton action recognition. *arXiv preprint arXiv:2411.11288*, 2024. 2

[3] Wenhan Wu, Zhishuai Guo, Chen Chen, Hongfei Xue, and Aidong Lu. Frequency-semantic enhanced variational autoencoder for zero-shot skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 1, 2