LLM-Guided Test-Time Adaptation for Human Activity Recognition under Domain Shifts

Anonymous ACL submission

Abstract

Human activity recognition (HAR) remains challenging due to evolving activity patterns and distribution shifts in real-world environments. Test-time adaptation (TTA) provides a promising way to adapt models to unlabeled target data without accessing source data, but existing methods are vulnerable to erroneous updates caused by unreliable pseudo-labels, leading to negative adaptation. Large language models (LLMs) offer useful semantic knowledge for reasoning about unseen activities. However, their predictions may be unreliable and inefficient. We propose an LLM-guided test-time adaptation framework for IMU-based HAR that incorporates semantic activity descriptions as external priors. Our method learns task-adaptive text representations and a selector offline, and performs batch-based selective adaptation at test time based on cross-modal fusion of sensor signals and semantic descriptions. Experiments on multiple HAR benchmarks under both single-domain and multi-domain settings show that our method improves robustness by reducing negative adaptation, especially in multi-domain scenarios where erroneous updates can accumulate over time.

1 Introduction

Human activity recognition (HAR) from inertial measurement unit (IMU) signals is widely used in wearable and mobile sensing systems, but it remains highly vulnerable to domain shifts caused by subject variation, sensor placement, device heterogeneity, and environmental changes. Consequently, deep learning models trained on source users or devices often suffer substantial performance degradation when deployed in unseen target environments, even with existing adaptation methods (Wang et al., 2020; Jang et al., 2023; Iwasawa and Matsuo, 2021). A key limitation is that these methods typically adapt the model based only on unlabeled target IMU data, whose distribution can be noisy, ambiguous, and severely shifted. Without external knowledge about what each activity semantically means, the model may reinforce incorrect predictions during adaptation.

This motivates the use of semantic knowledge to complement data-driven adaptation. Human activities are naturally described by language: for example, walking, running, sitting, and standing differ not only in sensor patterns but also in their underlying motion semantics. Large language models (LLMs), which encode rich knowledge about human activities through natural language, provide a promising source of such semantic priors (Cao et al., 2024). Recent studies have explored fine-tuning LLMs for sensor-based data interpretation, demonstrating their potential in sensor understanding (Li et al., 2025b; Hong et al., 2025; Wei et al., 2025). However, directly deploying LLMs for wearable HAR remains impractical due to the substantial cost of continuous inference. This raises a natural question: *Can we leverage the semantic knowledge of LLMs to guide the adaptation of deep learning-based HAR models without deploying LLMs during inference?*

A straightforward way to answer this question is to use LLMs offline to generate pseudo-labels or semantic supervision for unlabeled target data, and then use such supervision to adapt a lightweight HAR model. This idea is related to recent LLM-generated pseudo-labeling methods, where LLMs are prompted with task descriptions or few-shot examples and their generated labels are directly used to train smaller models, support semi-supervised learning, or provide weak supervision (Pangakis and Wolken, 2024; Rahman and Caragea, 2025). These methods show that LLMs can serve as powerful annotators when human labels are scarce.

However, directly trusting all LLM-generated pseudo-labels is risky for wearable HAR under domain shifts. Although LLMs encode rich semantic knowledge about activities, they are not trained on

IMU sensor signals and thus cannot be regarded as HAR experts. As a result, the semantic knowledge may not be consistent with the IMU samples. This issue is further amplified in test-time adaptation, where erroneous pseudo-labels can be repeatedly reinforced and lead to error accumulation. Therefore, the key challenge is not merely how to obtain pseudo-labels from LLMs, but *how to determine what kind of pseudo-labels or activity descriptions should be trusted to guide adaptation.*

To address this challenge, we propose a reliability-aware LLM-guided test-time adaptation framework (LGT for short) for IMU-based HAR. Rather than directly asking an LLM to classify a given IMU sample, we use the LLM offline to construct a memory of diverse activity descriptions, enabling efficient adaptation without online LLM inference. During test-time adaptation, the HAR model retrieves the most relevant descriptions according to the IMU representation and uses them to augment the IMU features. The resulting text-enhanced features produce multiple candidate predictions, offering complementary semantic views of the same sensor signal.

Since the retrieved descriptions may not always align with target-domain IMU patterns, blindly trusting all text-enhanced predictions can introduce noisy supervision. We therefore introduce a reinforcement learning based selector that adaptively determines which candidate predictions should be trusted. This selector allows the model to exploit LLM-derived semantic knowledge while reducing error accumulation under domain shifts.

Our contributions are summarized as follows:

- We propose an LLM-guided test-time adaptation framework for IMU-based HAR that exploits LLM-generated activity descriptions as an offline semantic memory, without online LLM inference.
- We introduce a reinforcement learning based selector that selectively trusts different text-enhanced predictions and adaptively fuses them with the original IMU prediction, reducing noisy pseudo-label supervision and error accumulation under domain shifts.
- Extensive experiments under single and multiple target domains settings on multiple HAR benchmarks demonstrate that our method improves adaptation robustness and mitigates negative adaptation over existing baselines.

2 Method

We study unsupervised test-time adaptation for human activity recognition (HAR) under target-domain shift. Let f_θ denote a HAR model trained on labeled source-domain IMU data. At test time, the model is adapted using only unlabeled target samples. To improve adaptation reliability, we augment IMU predictions with auxiliary textual semantics. Our framework consists of four components: an attribute-aware text description memory, a policy-based text selection and fusion module, offline learning of text representations and the selection policy, and a text-guided selective test-time adaptation objective.

2.1 Attribute-Aware Text Description Memory

For each activity class $c \in \{1, \dots, C\}$, we prompt a large language model (LLM) to generate multiple attribute-aware natural-language descriptions,

$$\mathcal{D}_c = \{d_{c,1}, d_{c,2}, \dots, d_{c,N_c}\}, \tag{1}$$

where each description captures one plausible realization of activity c under attributes such as gender, age, or motion status. These descriptions provide finer-grained semantic variations than class labels alone.

Each description $d_{c,j}$ is encoded into a text embedding $e_{c,j}$, and all description embeddings are stored in a global text description memory:

$$\mathcal{E} = \{e_{c,j} \mid c = 1, \dots, C,\ j = 1, \dots, N_c\}. \tag{2}$$

2.2 Policy-Based Text Selection

Given a target IMU sample x, the HAR model first produces the IMU logits z^{imu} and feature representation h. We retrieve the top-K relevant text descriptions from the text memory $\mathcal{E}$ based on their similarity to h:

$$\{T_1, \dots, T_K\} = \mathrm{TopK}_{e \in \mathcal{E}}\, s(e, h), \tag{3}$$

where $s(e, h)$ denotes cosine similarity.

For each retrieved text embedding T_k, we generate a text-conditioned feature view and obtain its corresponding logits:

$$\tilde{h}_k = \mathrm{Aug}(h, T_k), \qquad z_k = W_{cls}(\tilde{h}_k). \tag{4}$$

A policy network predicts the text-augmented prediction fusion weights w and the gate g used to fuse original and text-augmented prediction:

$$(w, g) = \pi_\theta(h). \tag{5}$$

The text-conditioned logits are then aggregated as $z^{fuse} = \sum_{k=1}^{K} w_k z_k$.

To adaptively balance sensor prediction and text guidance, we compute a confidence-aware gate

$$\gamma = g \odot \mathrm{FusionGate}(z^{imu}), \tag{6}$$

and obtain the final logits by

$$z^{final} = (1-\gamma) \odot z^{imu} + \gamma \odot z^{fuse}. \tag{7}$$

The final prediction p^{final} is produced by applying softmax to z^{final}.

Adaptation Gain Reward. The selector is trained to make long-term beneficial choices, rather than focusing only on immediate entropy reduction. Thus, compared to directly minimizing the entropy on the current whole batch, each batch is split into two halves. The updating half is used to perform a one-step TTA update on the classifier under the policy action (a_{upd}). The evaluation half is reserved to assess whether the selected action yields a more effective fused prediction on unseen samples.

On the evaluation split, we measure the change in fusion quality before and after the update using the source classifier and the temporarily updated classifier, respectively. We obtain

$$\Delta Q = \mathbb{E}[H(z^{\text{before}})] - \mathbb{E}[H(z^{\text{after}})].$$

This formulation rewards actions that reduce the entropy of the fused prediction, thereby encouraging more confident and less uncertain updates. The reinforcement learning objective is

$$\mathcal{L}_{\mathrm{RL}} = -\mathbb{E}_{a_{\mathrm{upd}} \sim \pi_\theta(\cdot|h)}[\Delta Q]\,,$$

2.3 Offline Learning and Selective Test-Time Adaptation

We optimize the text representations and the selector offline using labeled source-domain data. For each source sample, we retrieve top-K text embeddings from $\mathcal{E}$, construct text-conditioned feature views, and compute the final fused prediction p^{final} using the policy-controlled selection and fusion defined above. The offline objective is

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{fuse} + \lambda_2 \mathcal{L}_{align} + \lambda_3 \mathcal{L}_{RL}, \tag{8}$$

where $\mathcal{L}_{cls}$ is the supervised IMU classification loss, $\mathcal{L}_{fuse}$ supervises the fused prediction, $\mathcal{L}_{align}$ encourages consistency between IMU and text-conditioned representations, and $\mathcal{L}_{RL}$ optimizes the selector policy using the look-ahead reward defined above.

At test time, for each uncertain sample (i.e, samples with low confidence), we retrieve top-K text embeddings, construct text-conditioned feature views, and compute the final prediction p^{final} using the policy-controlled fusion defined above. The selected samples are then used to compute the test-time adaptation loss L_{tta}, and we update the HAR classifier with one gradient step. During this stage, the policy π_θ remains frozen. This design avoids unnecessary updates on confident samples while using text guidance to refine uncertain predictions, thereby improving adaptation stability under target-domain shift.

3 Experiments

3.1 Setup

Datasets and metrics. We evaluate on three IMU-based HAR benchmarks: **UCI** (Anguita et al., 2012), **OPPO** (Roggen et al., 2010), and **UniMiB** (Micucci et al., 2017). We consider both single-domain TTA and multi-domain TTA. In single-domain TTA, each target stream comes from one unseen domain (subject), while in multi-domain TTA, each target stream contains samples from multiple mixed target domains. We use classification accuracy as the evaluation metric.

We compare with strong TTA baselines including **TENT**(Wang et al., 2020), **T3A**(Iwasawa and Matsuo, 2021), and **TAST**(Jang et al., 2023). Additionally, we compare with LLM-based methods, including direct LLM inference with **DeepSeek-v4**(DeepSeek-AI, 2026) and the text-retrieval method **mmCLIP**(Cao et al., 2024).

Implementation. Considering the deployment on the mobile devices, we use lightweight HAR models. For the UCI and UniMiB datasets, we employ a six-layer convolutional neural network, whereas a three-layer convolutional neural network is used for the OPPO dataset. The text description is generated by DeepSeek-v4 (DeepSeek-AI, 2026) and retrieve Top-10 descriptions for each IMU sample, the prompt template is shown in APPENDIX B.2.

3.2 Overall Performance

As shown in Tables 1 and 2, LGT achieves the best performance on several single and multiple target domains while remaining competitive on the others, demonstrating strong and stable adaptation behav-

Dataset	Target Domain	Original	TENT	T3A	TAST	DeepSeek	mmCLIP	LGT
UCI	0	93.08	94.23	**94.81**	94.23	33.43	93.41	94.52
	1	85.09	82.45	85.76	**88.07**	32.45	82.53	82.78
	2	97.36	98.53	97.65	98.55	28.45	98.50	**98.82**
	3	78.86	83.59	79.49	77.91	31.86	83.62	**84.22**
	4	76.82	87.74	84.10	87.08	30.13	87.46	**90.06**
	AVG.	86.24	89.31	88.36	89.17	31.50	89.10	**90.08**
OPPO	S1	83.54	**83.81**	82.02	82.61	9.33	83.69	83.74
	S2	73.22	74.15	74.03	**75.05**	5.83	73.72	74.60
	S3	70.18	70.56	72.26	**72.70**	3.33	70.60	70.76
	S4	80.33	81.24	79.28	78.78	8.33	81.13	**81.35**
	AVG.	76.82	77.44	76.89	77.29	6.71	77.29	**77.61**
UniMiB	1	60.41	66.14	61.71	63.02	9.90	66.14	**67.18**
	2	66.38	68.26	66.55	**72.55**	8.75	67.80	68.27
	3	75.65	76.64	80.59	**80.92**	11.84	76.95	77.00
	5	**41.61**	40.93	40.93	41.27	9.15	40.34	41.27
	AVG.	61.02	63.00	62.45	**64.44**	9.91	62.83	63.43

Table 1: Accuracy (%) comparison on single domains.

Dataset	Target Domain	Original	TENT	T3A	TAST	mmCLIP	LGT
UCI	0→1→2→3	50.95	65.79	66.94	68.24	63.21	**71.30**
	1→2→3→4	61.01	67.74	64.02	67.11	68.35	**68.46**
	2→3→4→0	74.90	78.04	77.58	**79.04**	76.69	78.65
	0→1→3→4	72.00	75.78	66.40	67.90	76.33	**76.90**
	0→1→2→4	49.61	63.69	62.23	61.13	61.13	**66.25**
	AVG.	61.69	70.21	67.43	68.68	69.14	**72.31**
OPPO	$S1 \to S2 \to S3$	67.55	67.91	66.87	66.14	67.82	**68.83**
	$S1 \to S3 \to S4$	65.55	**66.90**	64.52	65.40	66.61	66.72
	$S2 \to S3 \to S4$	67.02	**67.96**	65.38	65.75	67.49	67.88
	AVG.	66.71	67.59	65.59	65.76	67.31	**67.81**
UniMiB	1→2→3	24.07	25.57	17.07	16.75	25.83	**28.73**
	1→2→5	46.00	**46.71**	36.20	38.73	46.50	46.61
	1→3→5	55.78	58.01	56.69	**58.92**	58.11	58.51
	2→3→5	35.94	38.05	39.57	35.78	38.82	**39.92**
	AVG.	40.45	42.09	37.38	37.55	42.33	**43.44**

Table 2: Accuracy(%) comparison on multiple domains.

ior. In the single-domain setting, LGT obtains the highest average accuracy on UCI and OPPO, with 90.08% and 77.61%, respectively. On UniMiB, although TAST achieves the best average accuracy, LGT remains competitive with 63.43%, outperforming Original, TENT, T3A, and mmCLIP. The advantage of LGT is more pronounced in the multi-domain setting. These results suggest that LGT better resists negative updates during sequential adaptation and maintains more robust performance under evolving domain shifts.

Additionally, directly applying DeepSeek leads to poor performance, likely because it has not been trained on HAR datasets and therefore has limited ability to interpret IMU signals. Moreover, DeepSeek requires nearly 10 s per sample for inference, whereas LGT takes only 0.46 ms, making LGT substantially more efficient than LLM-based activity recognition. Although mmCLIP directly leverages textual descriptions for prediction, it still underperforms LGT. These results further demonstrate that activity descriptions benefit sensor-based activity recognition, and that selecting trustworthy samples for HAR model adaptation yields greater improvements across multiple target domains.

3.3 Effect of Policy-based Selection.

To evaluate policy-based selection for TTA, we compare our method with two variants: directly voting with the retrieved Top-K descriptions and directly fusing retrieved aggregated logits with IMU logits without selection. Results on multiple target domains are reported in Table 3, with single-domain results provided in Appendix B.3.

From the results, retrieved text descriptions are useful in test time, as *only retrieve* improves over the source model on average across all datasets. However, direct fusion can underperform retrieval alone, suggesting that trusting all fused samples without selection may be misled by over-confident incorrect predictions. LGT achieves the best average results, demonstrating that the policy-based selector effectively identifies samples that are beneficial for long-term model updates.

Dataset	Target domain	Original	Only retrieve	Directly retrieve & fusion	LGT
UCI	$0 \to 1 \to 2 \to 3$	50.96	67.18	66.02	**71.31**
	$1 \to 2 \to 3 \to 4$	61.01	68.07	66.95	**68.46**
	$2 \to 3 \to 4 \to 0$	74.90	77.00	76.66	**78.65**
	$3 \to 4 \to 0 \to 1$	72.00	75.63	76.81	**76.90**
	$4 \to 0 \to 1 \to 2$	49.61	63.93	61.84	**66.25**
	AVG.	61.70	70.36	69.66	**72.31**
OPPO	$S1 \to S2 \to S3$	67.55	67.68	66.16	**68.83**
	$S1 \to S3 \to S4$	65.55	65.73	64.69	**66.72**
	$S2 \to S3 \to S4$	67.02	66.92	66.36	**67.88**
	AVG.	66.71	66.78	65.74	**67.81**
UniMiB	$1 \to 2 \to 3$	24.08	26.99	25.43	**28.73**
	$1 \to 2 \to 5$	46.01	46.25	44.42	**46.61**
	$1 \to 3 \to 5$	55.78	58.11	58.01	**58.51**
	$2 \to 3 \to 5$	35.95	38.31	38.22	**39.92**
	AVG.	40.45	42.42	41.52	**43.44**

Table 3: Comparison of different selection strategies

4 Conclusion

We presented a LLM-guided test-time adaptation framework for IMU-based human activity recognition under both single-domain and multi-domain distribution shifts. The framework combines offline learning of task-adaptive text representations with online batch-based selective adaptation guided by cross-modal fusion between sensor signals and semantic activity descriptions. Experiments across multiple HAR benchmarks show that textual guidance improves adaptation reliability, maintains stable performance under continuous multi-domain adaptation, and mitigates negative adaptation.

Limitations

One limitation of this work is the dataset-specific training of text encoders. In our current framework, independent text encoders need to be trained for different datasets, even when the same activity labels share identical or highly similar textual descriptions across datasets. This design may limit the cross-dataset transferability of the learned textual representations and introduce additional training costs when adapting the framework to new datasets.

Another limitation lies in the construction of textual descriptions. Ideally, the designed descriptions should sufficiently reflect the key characteristics of each activity, including object interactions, motion patterns, and contextual cues. However, when descriptions are manually designed or generated by external language models, their granularity, specificity, and amount of information may vary across activity categories. This variation may lead to different levels of semantic guidance for different activities. As shown in Figure 3 in APPENDIX, textual descriptions generally improve recognition performance, and the improvement is more evident for some fine-grained activities, such as close fridge and close dishwasher. This suggests that descriptions containing informative and discriminative cues can help distinguish activities with subtle differences. Therefore, constructing standardized and sufficiently informative textual descriptions remains an important issue for future investigation.

References

Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L Reyes-Ortiz. 2012. Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine. In *International workshop on ambient assisted living*, pages 216–223. Springer.

Qiming Cao, Hongfei Xue, Tianci Liu, Xingchen Wang, Haoyu Wang, Xincheng Zhang, and Lu Su. 2024. mmclip: Boosting mmwave-based zero-shot har via signal-text alignment. In *Proceedings of the 22nd ACM conference on embedded networked sensor systems*, pages 184–197.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence. Technical Report.

Zhiqing Hong, Yiwei Song, Zelong Li, Anlan Yu, Shuxin Zhong, Yi Ding, Tian He, and Desheng Zhang. 2025. Llm4har: Generalizable on-device human activity recognition with pretrained llms. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '25, page 4511–4521, New York, NY, USA. Association for Computing Machinery.

Yusuke Iwasawa and Yutaka Matsuo. 2021. Test-time classifier adjustment module for model-agnostic domain generalization. *Advances in Neural Information Processing Systems*, 34:2427–2440.

Minguk Jang, Sae-Young Chung, and Hye Won Chung. 2023. Test-time adaptation via self-training with nearest neighbor information. *ICLR*.

Zengyuan Lai, Jiarui Yang, Songpengcheng Xia, Lizhou Lin, Lan Sun, Renwen Wang, Jianran Liu, Qi Wu, and Ling Pei. 2026. Radarllm: Empowering large language models to understand human motion from millimeter-wave point cloud sequence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 5791–5799.

Juanhui Li, Sreyashi Nag, Hui Liu, Xianfeng Tang, Sheikh Muhammad Sarwar, Limeng Cui, Hansu Gu, Suhang Wang, Qi He, and Jiliang Tang. 2025a. Learning with less: knowledge distillation from large language models via unlabeled data. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2627–2641.

Miaomiao Li, Jiaqi Zhu, Yang Wang, Yi Yang, Yilin Li, and Hongan Wang. 2024. Ruleprompt: Weakly supervised text classification with prompting plms and self-iterative logical rules. In *Proceedings of the ACM Web Conference 2024*, pages 4272–4282.

Zechen Li, Shohreh Deldari, Linyao Chen, Hao Xue, and Flora D Salim. 2025b. Sensorllm: Aligning large language models with motion sensors for human activity recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 354–379.

Daniela Micucci, Marco Mobilio, and Paolo Napoletano. 2017. Unimib shar: A dataset for human activity recognition using acceleration data from smartphones. *Applied Sciences*, 7(10):1101.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. 2022. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pages 16888–16905. PMLR.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. 2023. Towards stable test-time adaptation in dynamic wild world. *ICLR*.

Nicholas Pangakis and Sam Wolken. 2024. Knowledge distillation in automated annotation: Supervised text classification with LLM-generated training labels. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, pages 113–131, Mexico City, Mexico. Association for Computational Linguistics.

Md Mezbaur Rahman and Cornelia Caragea. 2025. Llm-guided co-training for text classification. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31092–31109.

Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, and 1 others. 2010. Collecting complex activity datasets in highly rich networked sensor environments. In *2010 Seventh international conference on networked sensing systems (INSS)*, pages 233–240. IEEE.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2020. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*.

Shuoyuan Wang, Jindong Wang, Huajun Xi, Bob Zhang, Lei Zhang, and Hongxin Wei. 2024. Optimization-free test-time adaptation for cross-person activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(4):1–27.

Qingxin Wei, Jiaming Huang, Yi Gao, and Wei Dong. 2025. One model to fit them all: Universal imu-based human activity recognition with llm-assisted cross-dataset representation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):1–22.

A Appendix

A.1 Detailed Related Work

Test-Time Adaptation. Test-time adaptation (TTA) provides a practical paradigm for improving source-trained models under unlabeled target domains without accessing source samples, which can also be used in the HAR task. Since labels are unavailable at test time, the model has to adapt using information generated by itself (Niu et al., 2023, 2022; Wang et al., 2024). For example, entropy-minimization methods such as TENT (Wang et al., 2020) adapt models by encouraging confident predictions on unlabeled target samples and updating normalization statistics. Beyond entropy minimization, classifier-adjustment methods (Iwasawa and Matsuo, 2021) update the classifier using target-domain feature representations, while nearest-neighbor self-training methods (Jang et al., 2023) improve pseudo-label reliability by exploiting neighborhood information.

However, most general TTA methods were originally designed for image classification and typically assume independent test samples. Such an assumption may not fully exploit the temporal dependency and sequential structure inherent in time-series data.

LLMs for Cross-domain HAR. Recent studies have explored LLMs for sensor understanding and cross-domain HAR. They often transform the sensor data into semantic tokens as the input of LLMs. For example, SensorLLM(Li et al., 2025b) aligns sensor data with language semantics, while LLM4HAR(Hong et al., 2025) introduces an LLM-based framework for cross-domain activity recognition. RadarLLM (Lai et al., 2026) presents an LLM for radar signal understanding, which divides the radar signal into semantic tokens. Although these methods show the value of language priors, adapting a full LLM is often too costly for lightweight sensor-based HAR scenarios.

LLM-Generated Pseudo-Labels. Recent studies leverage LLMs as automatic annotators to generate pseudo-labels for unlabeled data, particularly for text classification. By prompting LLMs with task descriptions or few-shot examples, these methods obtain labels to train smaller classifiers or support semi-supervised learning (Pangakis and Wolken, 2024; Li et al., 2025a). For example, LLM-guided co-training uses LLM predictions as an additional supervision source to iteratively improve text classifiers (Rahman and Caragea, 2025). Wei et al.(Wei et al., 2025) leverage LLM to generate synthetic activity data to help the neural network learn cross-dataset representation. Other works treat LLM-generated labels as weak supervision or knowledge distillation signals, reducing reliance on costly human annotation (Li et al., 2024).

Different from these methods, our work does not use LLMs to directly generate pseudo-labels for model training. Instead, we assess whether a pseudo-label is trustworthy during test-time adaptation. This policy-based selection mitigates error accumulation caused by noisy pseudo-labels and enables more robust adaptation for IMU-based HAR under domain shifts.

B More Experiments

B.1 Detailed Datasets Settings

UCI contains data from 30 subjects performing 6 activities. OPPO includes data collected from multiple subjects under a richer set of 17 daily activities. UniMiB contains acceleration-based signals from 30 subjects covering 17 activities, including both daily activities and falls. Each subject is considered as one domain in our experiments. We selected a subset of subjects for TTA following the practice of previous studies (Wang et al., 2024). This experimental setting is designed to simulate subject-wise distribution shifts. Specifically, we use 70% of the IMU samples from the selected subjects for model pre-training and the remaining samples for evaluating TTA.

B.2 Prompt Template for Description Generation

For each activity class, we use the following prompt template to generate attribute-aware descriptions as shown in Figure 1. The examples of generated results are shown in Figure 2

B.3 Comparison of Different Selection Strategies

Table 4 provides a finer-grained evaluation of different selection strategies under single-domain TTA. From the results, textual retrieval brings large gains on UCI target domains 3 and 4, while it is less effective or even harmful on UCI target domain 1 and UniMiB target domain 5. This suggests that not all text-enhanced predictions are equally suitable for model updating.

Prompt Template for Attribute-Aware Text Description Generation

Given the human activity label `[ACTIVITY]`, generate N short and diverse natural-language descriptions. Each description should correspond to a plausible execution of this activity under different human attributes, including:

1. gender (e.g., man, woman),
2. age group (e.g., young, middle-aged, elderly),
3. motion status (e.g., slow, normal, brisk),
4. body style or posture when appropriate.

The descriptions should focus on how the activity is performed, rather than external scene details. They should remain semantically faithful to the activity label and be useful for distinguishing this activity from other daily activities in IMU-based human activity recognition. Keep each description concise and natural. Output one description per line.

Figure 1: Prompt template used to construct attribute-aware text descriptions for each activity.

Examples of Generated Activity Description

Fall:	A student falls right while turning quickly. A worker falls to the right after a sudden misstep. ...
Sit:	A person sits down from standing onto a chair. A woman sits down at a desk after walking. ...
Jump:	A person jumps repeatedly in a workout session. A user performs vertical jumps with both feet. ...
Walk:	A person walks forward at a normal pace on a sidewalk. A subject performs normal walking in daily life.
...	...

Figure 2: Examples of LLM-generated activity description.

Compared with fixed strategies, LGT achieves more stable improvements across datasets by adaptively selecting samples that are useful for long-term adaptation. In particular, LGT obtains the best average accuracy on UCI, OPPO, and UniMiB, and avoids severe degradation in challenging domains where retrieved predictions may be unreliable. Together with the multi-domain results, this indicates that the selector is not only beneficial in complex mixed-domain streams, but also important in single-domain TTA, where it mitigates subject-specific unreliable pseudo labels and supports more stable model updates.

Dataset	Target domain	Original	Only retrieve	Directly retrieve & fusion	LGT
UCI	0	93.08	93.66	93.65	**94.52**
	1	85.10	82.12	82.45	**82.78**
	2	97.36	**98.83**	98.28	**98.83**
	3	78.86	83.60	83.59	**84.23**
	4	76.82	86.76	86.75	**90.06**
	AVG.	86.24	88.99	88.94	**90.08**
OPPO	$S1$	83.54	**83.95**	83.28	83.74
	$S2$	73.23	74.08	73.63	**74.60**
	$S3$	70.18	70.47	70.95	**70.76**
	$S4$	80.33	81.27	80.45	**81.35**
	AVG.	76.82	77.44	77.08	**77.61**
UniMiB	1	60.42	66.15	66.92	**67.19**
	2	66.38	**68.27**	67.74	**68.27**
	3	75.66	76.97	76.31	**77.00**
	5	**41.61**	40.60	40.60	41.27
	AVG.	61.02	63.00	62.90	**63.43**

Table 4: Comparison of different selection strategies

B.4 Comparison of Different Fusion Strategies

Algorithm 1 summarizes the full online adaptation procedure of our method. Given a target batch, the model first retrieves top-K text embeddings from the description memory using the IMU feature, constructs text-conditioned feature views, and computes view-specific logits. The frozen policy then produces the view weights and fusion gate, which are used to obtain the final fused prediction. Based on this prediction and the reliability branch, the HAR classifier is updated with one test-time gradient step, while the policy remains fixed.

Algorithm 2 presents the logit fusion module used in our framework. It aggregates the text-conditioned logits with policy-predicted weights and combines them with the original IMU logits through a confidence-aware gate. The resulting fused logits are then converted into the final prediction.

Table 5 shows the results of different fusion strategies under both single-target and multi-target settings on 3 datasets. Overall, the proposed adaptation strategy consistently improves over the source-only baseline, and the dual fusion strategy achieves the best average performance across almost all settings.

Under the single-target setting, *Dual fusion* yields the strongest average performance on all

Dataset	Setting	Target Domain(s)	Original	logit fusion	Dual fusion
UCI	Single-target ($n_{\text{target}} = 1$)	0	93.08	93.66	**94.52**
		1	**85.10**	82.78	82.78
		2	97.36	**98.83**	98.83
		3	78.86	83.60	**84.23**
		4	76.82	86.76	**90.06**
		Avg.	86.24	89.13	**90.08**
	Multi-target ($n_{\text{target}} = 4$)	0→1→2→3	50.96	64.73	**71.31**
		1→2→3→4	61.01	66.16	**68.46**
		2→3→4→0	74.90	76.74	**78.65**
		3→4→0→1	72.00	76.8	**76.90**
		4→0→1→2	49.61	64.78	**66.25**
		Avg.	61.96	69.84	**72.31**
OPPO	Single-target ($n_{\text{target}} = 1$)	$S1$	83.54	83.52	**83.74**
		$S2$	73.23	73.56	**74.60**
		$S3$	70.18	70.69	**70.76**
		$S4$	80.33	80.95	**81.35**
		Avg.	76.82	77.18	**77.61**
	Multi-target ($n_{\text{target}} = 3$)	$S1 \to S2 \to S3$	67.87	68.23	**68.83**
		$S1 \to S3 \to S4$	65.56	66.57	**66.72**
		$S2 \to S3 \to S4$	67.02	67.05	**67.88**
		Avg.	66.81	67.28	**67.81**
UniMiB	Single-target ($n_{\text{target}} = 1$)	1	60.42	67.19	**67.19**
		2	66.38	66.90	**68.27**
		3	75.66	76.32	**77.00**
		5	**41.61**	40.27	41.27
		Avg.	61.01	62.67	**63.43**
	Multi-target ($n_{\text{target}} = 3$)	1→2→3	24.08	25.49	**28.73**
		1→2→5	46.01	44.51	**46.61**
		1→3→5	55.78	**58.52**	58.51
		2→3→5	35.95	38.65	**39.92**
		Avg.	40.45	41.79	**43.44**

Table 5: Accuracy comparison of source-only and adaptation methods under single-target and multi-target settings on UCI, UniMiB, and OPPO. All numbers are reported in accuracy (%). The best result in each row is in bold.

three datasets, reaching 90.08% on UCI, 63.43% on UniMiB, and 77.61% on OPPO. The gains are particularly pronounced on UCI, where the average accuracy improves from 89.13% with *logit fusion* to 90.08% with *Dual fusion*, and on target domain 4 the accuracy increases substantially from 76.82% (source-only) to 90.06%. Similar improvements are also observed on OPPO, where *Dual fusion* consistently outperforms the source-only baseline on all target domains. On UniMiB, the improvements are smaller but still stable on average, suggesting that text-guided adaptation remains beneficial even in a more challenging and noisier setting.

Under the multi-target setting, the advantage of *Dual fusion* becomes more evident. On UCI, its average accuracy reaches 72.30%, outperforming both the source-only baseline (61.70% on average) and *logit fusion* (69.84%). On UniMiB, *Dual fusion* improves the average accuracy from 41.79% to 43.11% over *logit fusion*, while also outperforming the source-only baseline across all target combinations. On OPPO, the gains are more moderate but still consistent, with *Dual fusion* achieving the best average performance of 67.81%.

These results suggest two main findings. First, incorporating retrieved textual semantics at test time is consistently more effective than relying on the source model alone, especially under more severe cross-domain shifts. Second, compared with directly applying a residual correction at the logit level, the *Dual fusion* design provides a more robust way to aggregate text-guided evidence, leading to stronger and more stable adaptation performance across datasets and domain settings.

C Effect of Description

We evaluate the effect of descriptions for each activity, which compares the accuracy of LGT and TENT. The results are shown in Figure 3.

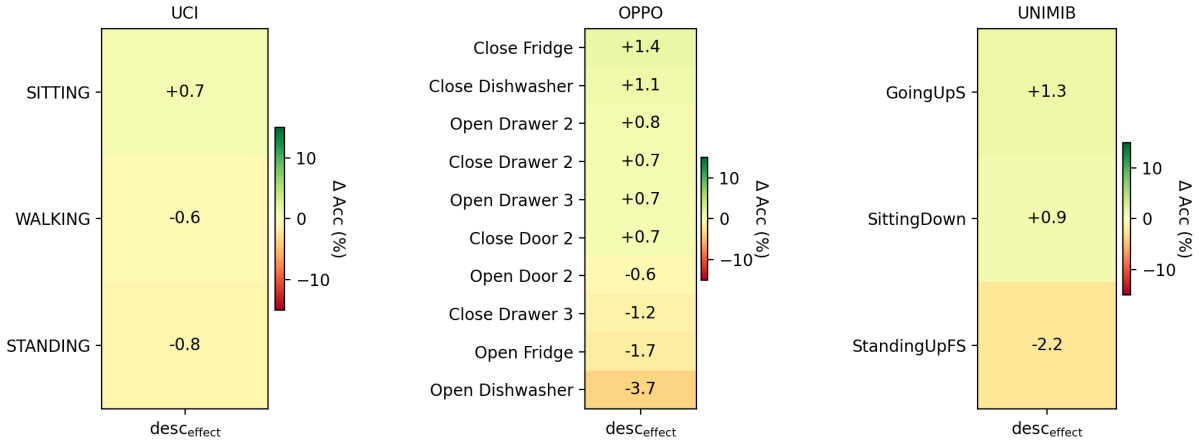

Figure 3: Comparison of description effect ($\Delta Acc = Acc_{LGT} - Acc_{TENT}$) on activity level across datasets.

Algorithm 1 Dual Fusion with Text Description

Require: Target batch x, classifier W_{cls}, text description memory $\mathcal{E}$, frozen policy π_θ, hyperparameters K, p, fusion and loss configurations

Ensure: Updated predictions

$(z^{imu}, h) \leftarrow f_\theta(x)$
$\{T_1, \ldots, T_K\} \leftarrow \textsc{RetrieveTopK}(h, \mathcal{E})$
for $k = 1, \ldots, K$ **do**
  $\tilde{h}_k \leftarrow \textsc{AugmentFeature}(h, T_k; p)$
  $z_k \leftarrow W_{cls}(\tilde{h}_k)$
end for
$(w, g) \leftarrow \pi_\theta(h)$
$z^{fuse} \leftarrow \sum_{k=1}^{K} w_k z_k$
$\gamma \leftarrow g \odot \textsc{FusionGate}(z^{imu})$
$z^{final} \leftarrow (1-\gamma) \odot z^{imu} + \gamma \odot z^{fuse}$
Compute L_{tta} from z^{final}, u, and the reliability branch
Update W_{cls} with one gradient step
Keep π_θ frozen
return updated predictions

Algorithm 2 Solely logit fusion

Require: IMU logits z^{imu}, text-conditioned logits $\{z_1, \ldots, z_K\}$, frozen policy π_θ, IMU feature h

Ensure: Final prediction p^{final}

$(w, g) \leftarrow \pi_\theta(h)$
$z^{fuse} \leftarrow \sum_{k=1}^{K} w_k z_k$
$\gamma \leftarrow g \odot \textsc{FusionGate}(z^{imu})$
$z^{final} \leftarrow (1-\gamma) \odot z^{imu} + \gamma \odot z^{fuse}$
$p^{final} \leftarrow \mathrm{softmax}(z^{final})$
return p^{final}