

Energy-Efficient High-Performance Representation Learning for Sensor Data via Flow Matching

Kosuke Ukita and Tsuyoshi Okita

Kyushu Institute of Technology, Iizuka, Fukuoka, Japan

Keywords: Flow Matching, Self-Supervised Learning, Human Activity Recognition, Generative AI

Abstract

Diffusion models can inherently learn rich representations during their process of high-quality data generation, suggesting their potential for SSL. However, existing methods face a trade-off between generative quality and discriminative performance, alongside high computational costs. To address this, we propose Flow Matching-based Foundation Model (**FlowFM**), which jointly trains a representation encoder and a flow matching generator. This decoupled design achieves both high-fidelity generation and effective recognition. Experiments on wearable sensor data show **FlowFM reduces training time by 50.4%** compared to a diffusion-based approach and **surpasses the state-of-the-art SSL method** on five downstream datasets, **achieving up to a 51.0x inference speedup**.

Objectives

Problem 1: Performance Trade-off

- **Generation** ↔ **Discrimination**

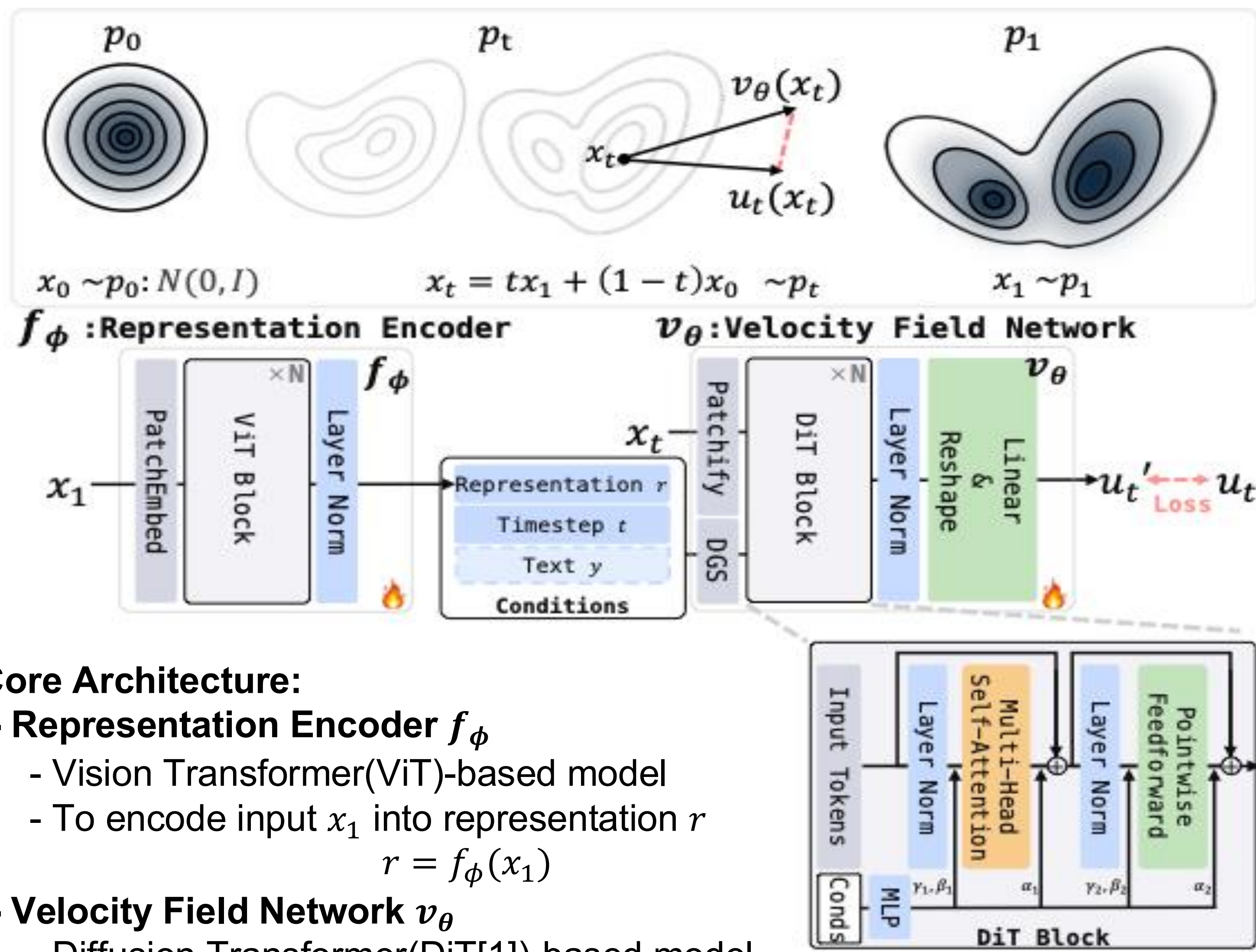
Problem 2: High Computational Cost

- The iterative sampling of diffusion models is computationally expensive

Our Goals:

- To develop a novel **foundation model** (→ **FlowFM**)
- To achieve both high **discriminative** and **generative** performance
- With high **computational efficiency**

Proposed Methods



Core Architecture:

Representation Encoder f_ϕ

- Vision Transformer(ViT)-based model
- To encode input x_1 into representation r
 $r = f_\phi(x_1)$

Velocity Field Network v_θ

- Diffusion Transformer(DiT[1])-based model
- To learn the flow from the p_0 to the p_1 (Flow Matching[2])

$$u_t' = v_\theta(x_t, t)$$

Joint Training:

- Encoder f_ϕ and generator v_θ are trained jointly
- Encoder learns the optimal representation to simplify generation task.

Loss Function:

$$\begin{aligned} \mathcal{L}_{\text{FlowFM}}(\theta, \phi) &= \mathbb{E}_{t, x_0, x_1} [\|v_\theta(x_t, t, r) - u_t(x_t | x_1, r)\|^2] \\ &= \mathbb{E}_{t, x_0, x_1} [\|v_\theta(x_t, t, f_\phi(x_1)) - u_t(x_t | x_1, f_\phi(x_1))\|^2] \\ &= \mathbb{E}_{t, x_0, x_1} [\|v_\theta^\phi(x_t, t) - u_t^\phi(x_t | x_1)\|^2] \end{aligned}$$

Key Technique: Dynamic Guidance Switching (DGS)

- Representation condition r is randomly masked during training (50% prob.)
- This acts as regularization
- To learn both conditional/unconditional generation
- To acquire more robust features.

Table1: Accuracy of classification results by fine-tuning with DGS.

Mask prob.	ADL	Opportunity	PAMAP2	REALWORLD
0%	.9370	.7775	.8868	.9004
50%	.9449	.7974	.8920	.9112

Experimental Results

Recognition Performance (HAR Task)

- FlowFM outperforms all baselines, including SOTA (SSL-Wearabls), BERT, DINO, and a diffusion-based counterpart (DiffFM), across all five datasets.

Table2: Classification results from human activity recognition tasks.

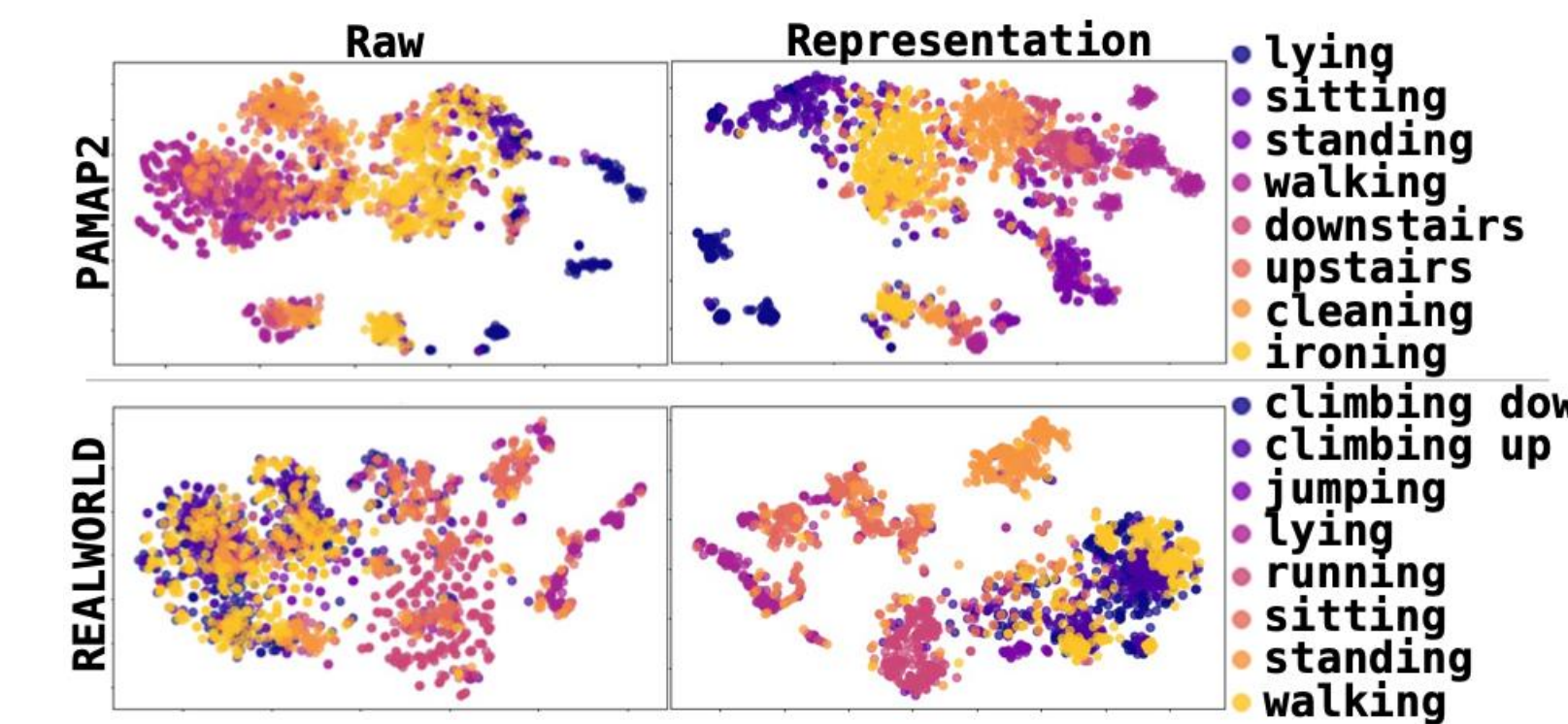
Method	ADL		Opportunity		PAMAP2		REALWORLD		WISDM	
	acc.	f1	acc.	f1	acc.	f1	acc.	f1	acc.	f1
Transfer learning										
SSL-Wearables [3]	-	.7540	-	.5470	-	.7250	-	.7710	-	.7230
DINO [4]	.8835	.8691	.7557	.7682	.8111	.7981	.8124	.8243	.7747	.7716
SEnvT-u4 [5]	.8394	.7936	.7103	.6826	.7157	.6912	.7470	.7512	.6667	.6618
SEnvT-u7 [5]	.8472	.7849	.7150	.6909	.7098	.6869	.7628	.7692	.6819	.6766
DiffFM (ours)	.9055	.8813	.7787	.7760	.8171	.8090	.8614	.8741	.8261	.8224
FlowFM (ours)	.9370	.9168	.7881	.7792	.8467	.8373	.8683	.8832	.8337	.8322
Fine-tuning										
Random Forest ¹	.8346	.6830	.6814	.4260	.6602	.5467	.6618	.5712	.4630	.4224
Transformer scratch ²	.8503	.8230	.6276	.6209	.7682	.7594	.8554	.8663	.8219	.8194
BERT [6]	.8583	.8553	.7272	.7245	.7561	.7527	.7807	.7819	.7923	.7924
SSL-Wearables [3]	-	.8290	-	.5950	-	.7890	-	.7920	-	.8100
DINO [4]	.9024	.8780	.7656	.7741	.8341	.8223	.8878	.8978	.8693	.8684
SEnvT-u4 [5]	.9087	.8780	.7489	.7411	.8488	.8454	.9015	.9124	.8693	.8679
SEnvT-u7 [5]	.8992	.8684	.7429	.7448	.8477	.8445	.9058	.9163	.8875	.8867
Diffusion Classifier [7]	.8818	.8515	.7693	.7526	.8763	.8768	.8959	.8574	.8076	.8075
DiffFM (ours)	.9291	.9118	.7857	.7782	.8902	.8877	.9084	.9197	.8839	.8832
FlowFM (ours)	.9449	.9286	.7974	.7925	.8920	.8877	.9112	.9211	.8918	.8910

- t-SNE visualization

- the learned representations from

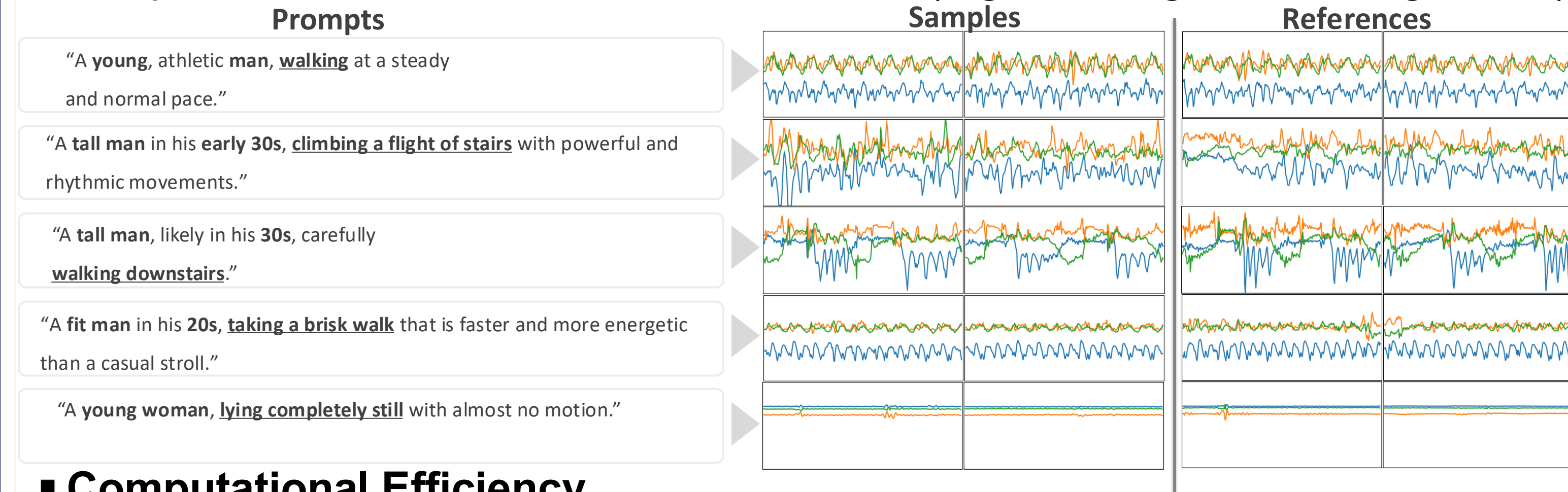
- distinct, well-separated clusters

- even on out-of-domain data



Generative Quality (Text-to-Signal Task)

- FlowFM successfully generates high-fidelity signals
- from complex text prompts
- capture nuances between similar activities (e.g., "walking" vs "walking down")



Computational Efficiency

Training:

- Time is reduced by **50.4%** compared to DiffFM (3.2h vs 6.5h)

Inference:

- Achieves up to **51.0x speedup** in generation compared to DiffFM
- maintaining superior or equivalent generative quality (FID: 5.042)

Table3: Generation speed and quality during inference.

Model	steps	time[s] / 1 sample↓	FID↓	IS↑
DiffFM	1000	4.62×10^{-2}	5.047	2.751
DiffFM	100	4.58×10^{-3}	8.837	2.800
DiffFM	50	2.29×10^{-3}	10.08	<u>2.819</u>
FlowFM	-	9.05×10^{-4}	<u>5.014</u>	2.626

Conclusion

We proposed **FlowFM**, a novel foundation model that successfully applies flow matching to self-supervised representation learning.

Key Achievements:

- Resolved the trade-off between generative and discriminative performance
- Achieved **SOTA** results on downstream HAR tasks
- Demonstrated high computational efficiency, **halving training costs** and dramatically accelerating inference speed by up to **51x**

Future Work:

- FlowFM is a promising framework for building efficient and powerful foundation models for various modalities.

References

- [1] Peebles, et al. "Scalable diffusion models with transformers.", *ICCV*, 2023, pp.4195-4205.
- [2] Lipman, et al. "Flow Matching for Generative Modeling", *ICLR*, 2023.
- [3] Yuan, et al. "Self-supervised learning for human activity recognition using 700,000 person-days of wearable data.", *NPJ digital medicine*, 2024, 1, (7), pp.91-100.
- [4] Caron, et al. "Emerging properties in Self-Supervised Vision Transformers.", *ICCV*, 2021, pp.9650-9660.
- [5] Okita, et al. "Towards LLMs for Sensor Data: Multi-Task Self-Supervised Learning.", *UbiComp/ISWC*, 2023, pp.499-504.
- [6] Devlin, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding.", *NAACL*, 2019, pp.4171-4186.
- [7] Li, et al. "Your Diffusion Model is Secretly a Zero-shot Classifier.", *ICCV*, 2023, pp.2206-2217.