APPENDIX

More Details about PATN

In order to protect user privacy on mobile devices, we generate adversarial perturbations for future accelerometer and gyroscope signals by leveraging historical sensor data. The validity of this approach can be justified from three theoretical perspectives.

1) Temporal Correlation and Local Stationarity. Inertial sensor readings can be modeled as

$$x_{t+1} = g(x_t) + \epsilon_t \tag{11}$$

where $g(\cdot)$ denotes a smooth dynamical process and ϵ_t is a noise term. Over short horizons, ϵ_t is approximately weakly stationary, implying

$$p(\epsilon_{t+k}) \approx p(\epsilon_t), \quad k \ll T. \tag{12}$$

Thus, the noise statistics extracted from historical data can approximate the distribution of future perturbations.

2) Direct Perturbation Generation without Explicit Prediction. Modern generative models (e.g., VAEs, conditional GANs, RNN-based predictors) are capable of learning the conditional distribution $p(x_{t+1}|x_{1:t})$ to synthesize future sensor data. Since our goal is not to reconstruct exact future trajectories but to utilize the underlying statistical properties, we can bypass explicit prediction and directly sample adversarial perturbations from the estimated historical noise distribution:

$$\delta_t \sim \hat{p}(\epsilon). \tag{13}$$

This avoids compounding prediction errors and yields perturbations more representative of the true signal dynamics.

3) Dynamic Adjustment via Motion Pattern Resampling. Historical sensor data can be decomposed into characteristic motion patterns $\phi_i(t)$. To adapt perturbations to current states, we dynamically combine these patterns:

$$\delta_t = \sum_i w_i(t)\phi_i(t), \tag{14}$$

where weights $w_i(t)$ are updated based on current inputs and optimization goals (e.g., maximizing inference error while controlling perturbation magnitude). This ensures the generated perturbations remain physically plausible while effectively degrading privacy-invasive inference.

Overall, by exploiting the local stationarity of sensor noise, leveraging the statistical modeling capability of generative frameworks, and applying dynamic resampling, we can efficiently construct adversarial perturbations for future sensor readings without explicit signal prediction.

More Details about the Datasets

MotionSense (Malekzadeh et al. 2018) is an open-source dataset that contains data captured from an accelerometer and a gyroscope at a constant frequency of 50 Hz, collected using an iPhone 6s placed in the participants' front pocket. In total, 24 participants performed six distinct activities—going downstairs, going upstairs, walking, jogging, sitting, and standing—over 15 trials, all conducted in a controlled environment. This dataset is suitable for studying privacy inference of gender based on motion patterns. However, as all participants are adults, it cannot be used for privacy inference related to distinguishing between children and adults.

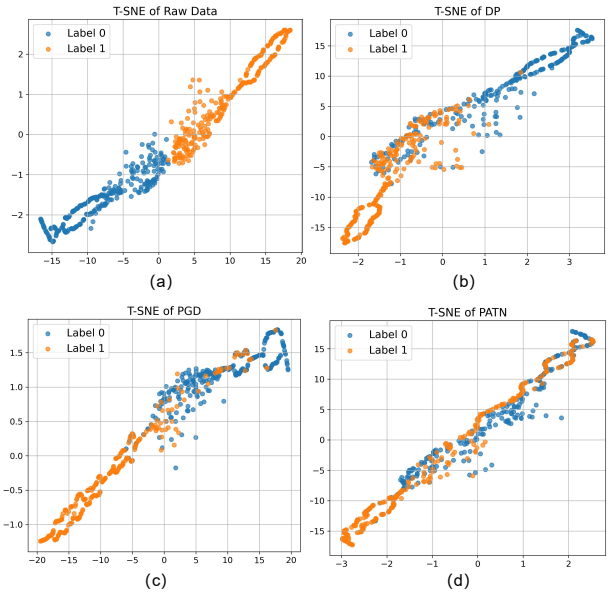

Figure 6: T-SNE visualization of feature representations under different privacy perturbation methods: (a) Raw Data, (b) DP, (c) PGD, and (d) PATN.

ChildShield (Lin et al. 2023) is a proprietary dataset, for which we obtained permission from the authors for research use. It contains motion sensor data collected from accelerometers and gyroscopes at a constant frequency of 60 Hz during gameplay across multiple mobile games. A total of 1,875 participants completed trials on five different games. This dataset enables studies on privacy inference of age attributes (child and adult users). However, as no gender labels were collected, it cannot be used to train or evaluate gender-based privacy inference models.

Interpretable Feature Perturbation via T-SNE Visualization

To better interpret the effect of different privacy defense mechanisms on the internal representations of the inference model, we visualize the latent feature space using T-SNE. Specifically, we compare the feature distributions resulting from the raw data, our proposed PATN method, and baseline methods: DP and PGD. These visualizations are based on the hidden representation extracted from the inference model and reflect how input perturbations alter the separability of private attributes in the feature space.

As shown in Figure 6, the raw data exhibits clear clustering patterns aligned with the private attribute labels, indicating that the inference model can easily exploit these structures. In contrast, the PATN method significantly disrupts these patterns, making the latent representations more entan-

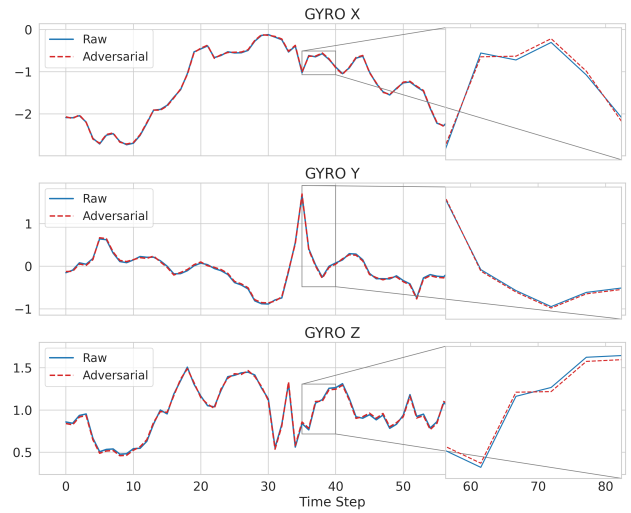

Figure 7: Comparison of Raw and Adversarial sensor data (gyroscope) on IMU Axes.

Method	RNN		LSTM	
	ASR(%)↑	EER(%)↑	ASR(%)↑	EER(%)↑
Raw data	\	14.76	\	13.54
DP	5.05	19.52	1.50	14.56
UAP	2.45	15.92	1.04	14.24
FGSM	8.52	20.90	6.74	18.44
PGD	8.52	20.90	6.11	17.73
PATN(our)	**31.59**	**29.26**	**25.35**	**25.01**

Table 9: Performance comparison with baseline methods.

gled and less linearly separable. Compared to DP and PGD, PATN achieves a more substantial deformation of the feature space while maintaining naturalness in the perturbed signals. This illustrates the interpretability advantage of PATN: rather than merely injecting noise, it learns to adaptively shift samples in the feature space to confuse the inference model effectively. The visual evidence thus supports the claim that PATN provides stronger and more targeted privacy protection through latent feature perturbation.

IMU Visualization of Gyroscope Data

To better interpret our method's effect in the sensor space, we also visualize gyroscope IMU signals, as shown in Figure 7. The gyroscope visualization shows that our perturbations preserve strong semantic fidelity: key motion patterns and temporal dynamics remain intact despite the added noise. This indicates that the method maintains signal realism essential for utility tasks, while effectively masking features used for privacy inference.

Effectiveness of PATN on Temporal Models

To further evaluate the generalizability of PATN, we extend our experiments to sequential models, including RNN and LSTM, using the MotionSense dataset. As shown in Table 9, PATN significantly outperforms all baseline methods in both ASR (Attack Success Rate) and EER (Equal Error Rate), across both model types. Notably, PATN achieves 31.59% ASR and 29.26% EER against the RNN-based inference model, while the best-performing baseline (FGSM/PGD) reaches only 8.52% ASR and 20.90% EER. Similarly, on the LSTM model, PATN attains 25.35% ASR and 25.01% EER, considerably higher than all baselines (e.g., PGD: 6.11% ASR, 17.73% EER).

We observe that attacking sequential models such as RNNs and LSTMs is generally more challenging than attacking CNNs, possibly due to their non-convex optimization landscape and temporally entangled feature representations (Ding et al. 2023). While the overall adversarial effectiveness is reduced in this setting, our method still provides a noticeable level of privacy protection, outperforming all baselines in both ASR and EER. These results suggest that PATN maintains a certain degree of robustness across different model architectures, even under complex temporal modeling conditions. In future work, we plan to further investigate specialized perturbation strategies tailored for recurrent models, in order to enhance privacy defense in sequential settings.