Diffusion-Based 3D Sign Language Motion Anonymization: A Feasibility Study on Balancing Identity Confusion and Semantic Preservation

Zixuan Dai, Shinji Sako
Graduate School of Engineering, Nagoya Institute of Technology
Nagoya, Aichi, Japan
cnq14068@ict.nitech.ac.jp, s.sako@nitech.ac.jp

Abstract

Sign language motions contain individual-specific kinematic features. As the engineering applications of sign language become more widespread, privacy protection of sign language data has emerged as a new challenge. This paper proposes a diffusion model-based approach for sign language motion anonymization. The proposed framework combines conditional diffusion processes with adversarial training to transform identity features while preserving semantic information. For the design and preliminary validation of the proposed model, we conduct a proof-of-concept experiment using a subset of 22 signers from the ASL100 dataset of WLASL, which demonstrates the feasibility of the proposed approach for sign language anonymization.

Keywords: Sign Language Anonymization, Diffusion Model, Privacy Protection, Identity Confusion, Transformer

1. Introduction

Sign language serves as a vital means of communication for deaf and hard-of-hearing communities worldwide, constituting an independent visual-spatial language system with its own grammatical structures and expressive modalities. In building a truly equitable and inclusive digital society, the needs and rights of the deaf community cannot be overlooked. While recent advances in deep learning have substantially improved the performance of sign language recognition systems, they have simultaneously given rise to new privacy risks. Sign language, as a complex visual-spatial language, not only conveys semantic information through handshapes, positions, and movements, but also unconsciously transmits rich personal behavioral characteristics. These dynamic features, including hand movement velocity patterns, trajectory curvature, joint coordination, and temporal rhythm, form distinctive "motion signatures" unique to each individual. At the same time, these individual motion characteristics constitute an important aspect of each signer's personal identity and self-expression.

Figure 1: Motion style variation across signers. Two different signers performing the sign "AIM" at the same relative frame position.

Respecting such individuality is essential. However, in certain contexts, such as public dataset sharing or anonymous communication, the ability to conceal these identifying features becomes equally important. Balancing respect for personal signing identity with the need for privacy protection presents a unique challenge, one that has become increasingly urgent in the current technological landscape.

In today's highly digitized and intelligent era, sign language users face unprecedented privacy risks. Modern computer vision models can extract and analyze subtle motion features from seemingly innocuous videos, raising significant privacy concerns for sign language communication. As the social demand for sign language grows, engineering applications related to sign language are expected to become increasingly active. Accordingly, sign language datasets are likely to become larger in scale and closer to real-life conditions, which in turn amplifies the urgency of privacy protection.

Several application scenarios illustrate the severity of this concern. In educational settings, student practice videos collected by online sign language teaching platforms may inadvertently expose learners' identities, learning progress, and common error patterns. Similarly, sign language communication in telemedicine consultations involves highly sensitive information, including health conditions and psychological states, with serious consequences if personal identities are revealed. Furthermore, training sign language AI systems requires large-scale datasets, yet using raw video data without adequate consent mechanisms may conflict with privacy protection regulations across countries. Grow-

Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages, pages 93–99
16 May 2026.

ing awareness of privacy rights and biometric data protection regulations necessitates that sign language AI research balance technological advancement with privacy protection. These challenges highlight the urgent need for effective sign language motion anonymization techniques.

Several studies have demonstrated that signers can be identified from their motion patterns alone. (Bigand et al., 2021) showed that kinematic features serve as reliable cues for signer identification in French Sign Language. (Dai and Sako, 2025) analyzed motion-based individualization patterns in Japanese Sign Language, confirming the cross-linguistic nature of kinematic identity features. Furthermore, human perception experiments have demonstrated that observers can identify signers at significantly above-chance accuracy even when facial information is completely concealed (Bigand et al., 2020). Figure 1 illustrates this phenomenon using samples from our dataset: two different signers performing the same sign exhibit visibly distinct motion characteristics in hand height, arm extension, and overall posture, despite conveying identical semantic content.

Conventional video anonymization techniques have primarily focused on facial processing (Perea-Trigo et al., 2025). While approaches using face-swapping techniques and virtual avatars have been proposed, these methods suffer from two fundamental limitations. First, as the findings of (Bigand et al., 2021, 2020) and (Dai and Sako, 2025) indicate, simply concealing the face cannot remove the motion signatures embedded in body movements. These identity-revealing features manifest through individual-specific motion patterns such as arm swing amplitude, wrist rotation patterns, shoulder tilt angles, and subtle differences in finger curvature. Second, facial expressions themselves constitute a crucial component of sign language grammar, conveying essential linguistic information through non-manual signals such as eyebrow movements, mouth patterns, and head tilts. Removing or replacing facial information therefore risks degrading the semantic content of the signed message. More recent work has further confirmed that signer identity can be reliably inferred from pose-level skeletal data alone (Battisti et al., 2024), and that pose-based appearance transfer methods can modify visual identity while attempting to preserve signing content (Moryossef et al., 2025). Taken together, these findings demonstrate that facial concealment alone cannot achieve true anonymization in sign language, and may even provide users with a false sense of security while simultaneously compromising linguistic integrity. This situation calls for a new technical approach capable of effectively removing identity features at the motion level while preserving the semantic content of the movements.

Recent advances in human motion anonymization have leveraged foundation motion models such as VPoser (Pavlakos et al., 2019) and HuMoR (Rempe et al., 2021), which learn latent representations of everyday movements like walking and sitting. By operating in these learned latent spaces, such approaches can generate biomechanically plausible anonymized motions. However, sign language presents fundamentally different challenges. Unlike general body movements, sign language functions as a linguistic system where fine-grained finger configurations and hand shapes directly determine semantic meaning. Moreover, no foundation motion model currently exists for sign language, making it infeasible to directly apply these latent-space anonymization techniques. This motivates our approach of designing constraint mechanisms tailored to sign language characteristics, focusing on transforming existing motions rather than generating entirely new ones from a general prior.

Diffusion models offer a promising framework for such controlled transformation. In recent years, they have achieved remarkable success in image and video generation tasks (Ho et al., 2020; Tevet et al., 2023), and their application to sign language video anonymization has also been reported (Xia et al., 2024). Their conditional generation capability suggests the potential to selectively remove identity-specific features while preserving semantic information. Motivated by these observations, this paper presents a feasibility study on diffusion-based 3D sign language motion anonymization, proposing a framework built upon a Transformer-based diffusion architecture. The proposed method consists of three main components: a multi-modal encoder for joint representation learning that integrates joint coordinates with semantic information, a Transformer-based diffusion module for feature transformation in latent space, and a decoder for reconstructing anonymized 3D motion sequences. In particular, we introduce a semantic conditioning mechanism for preserving linguistic content and an adversarial identity confusion mechanism for actively disrupting identity information, thereby seeking a balance between privacy protection and semantic preservation. Through experiments using a subset of the WLASL dataset involving 22 signers, we demonstrate that the proposed approach can achieve a meaningful degree of identity confusion while maintaining high semantic fidelity, confirming the feasibility of diffusion-based sign language motion anonymization.

The remainder of this paper is organized as follows. Section 2 describes the dataset and the proposed framework. Section 3 presents the experimental setup and results. Section 4 discusses limitations and future directions.

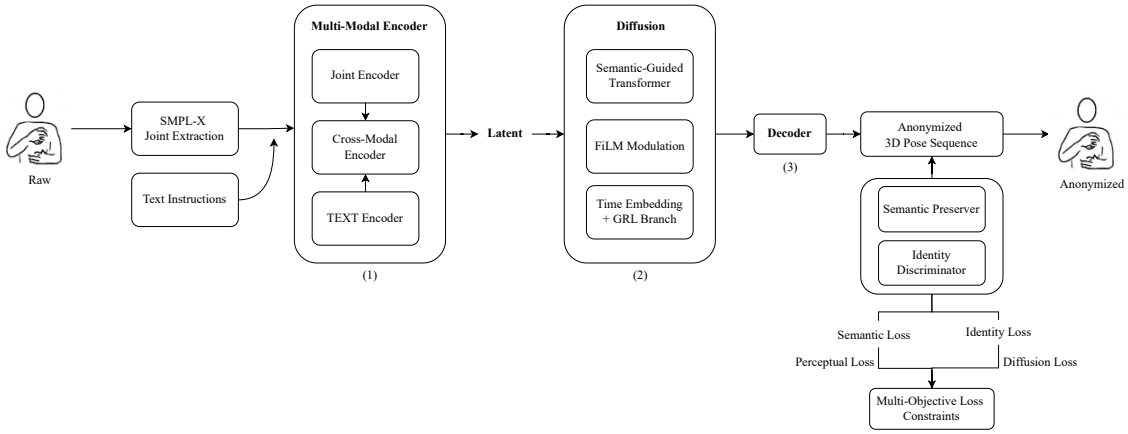

Figure 2: Overall Framework.

2. Methodology

2.1. Dataset

This study utilizes 3D joint coordinate sequences derived from the SignAvatars dataset (Yu et al., 2024), which applies 3D human body reconstruction techniques to video data from WLASL (Word-Level American Sign Language) (Li et al., 2020). WLASL is currently the largest word-level ASL recognition dataset, encompassing 2,000 sign language glosses recorded by over 100 signers in diverse environments. SignAvatars extends WLASL by reconstructing 3D human body models from the original video data through an automated annotation pipeline that performs hierarchical initialization followed by optimization with temporal smoothing and biomechanical constraints, yielding SMPL-X parameters from which sequences of 65 3D joint coordinates per frame (25 body, 20 left-hand, and 20 right-hand joints) are extracted. Since WLASL is a word-level dataset, each sample corresponds to an isolated sign rather than a continuous utterance. This 3D representation offers richer spatial information than 2D pose estimation, making it particularly suitable for analyzing the fine-grained motion characteristics relevant to both sign language understanding and anonymization.

While large-scale sign language datasets have been developed primarily for recognition and translation tasks, they present a structural challenge for anonymization research. In WLASL, although a large number of signers are represented, the distribution is highly imbalanced: a few signers contribute a large proportion of samples, while many signers appear only a handful of times with minimal overlap across glosses. This distribution is well-suited for vocabulary-level recognition tasks but makes it difficult to model individual-specific motion characteristics, which is essential for both identity modeling and anonymization evaluation. Moreover, no existing 3D sign language motion dataset has been specifically designed for anonymization research, and available resources remain far scarcer than their 2D video counterparts. Given these constraints, we select 22 signers from the ASL100 subset, each contributing a sufficient number of samples to enable meaningful identity modeling.

2.2. Framework Overview

Figure 2 illustrates the overall architecture of the proposed framework, which consists of three components: a multimodal encoder, a Transformer-based diffusion module, and a decoder.

The multimodal encoder takes as input 3D joint coordinate sequences (65 joints $\times$ 3 dimensions per frame) together with textual semantic information, and produces fused latent representations. Skeletal motion is encoded through a Transformer encoder that attends over all frames in the sequence to capture spatiotemporal dependencies, while semantic information is embedded via a pre-trained Sentence-BERT encoder (Reimers and Gurevych, 2019). In the current proof-of-concept implementation, the text input provides a fixed semantic category descriptor; extending this to per-sample gloss-level conditioning is planned for future work. A cross-modal attention mechanism integrates these heterogeneous modalities, producing latent representations that are both motion-aware and semantically informed. This joint encoding is essential because the downstream diffusion process must be able to distinguish between what a signer is saying and how they are saying it.

The core of the framework is the Transformer-based diffusion module, which operates on the encoded latent representations. A key observation motivating our design is that conventional anonymization techniques (such as geometric transformations or noise injection applied directly to joint coordinates) treat all motion features uniformly.

They lack the ability to selectively target identity-specific characteristics while leaving semantic content intact. Diffusion models offer a more principled alternative: by learning to iteratively denoise from a corrupted signal, they can be trained to generate motion sequences that satisfy specific constraints.

However, a standard unconditional diffusion process would modify all latent features indiscriminately, potentially degrading the linguistic content of the sign. To address this, we introduce a semantic conditioning mechanism that dynamically guides the denoising process based on the semantic information extracted by the encoder. This allows the model to be aware of which aspects of the motion carry linguistic meaning and should therefore be preserved throughout the transformation.

Semantic preservation alone, however, is not sufficient for effective anonymization. The model must also actively disrupt identity-specific features rather than merely hoping they are incidentally removed during generation. To this end, we incorporate an adversarial identity confusion mechanism using a Gradient Reversal Layer (GRL) (Ganin and Lempitsky, 2015). An identity classifier is attached to the diffusion module's output features; during backpropagation, the GRL reverses the gradient signal, encouraging the feature representations to become maximally uninformative with respect to signer identity. This creates an explicit adversarial objective: the classifier attempts to recover the original signer's identity, while the diffusion module learns to produce representations that confuse this classifier. Together, the semantic conditioning and adversarial confusion mechanisms establish a dual optimization objective: preserving what is said while obscuring who is saying it. The decoder reconstructs anonymized 3D joint coordinate sequences from the transformed latent representations, producing motion outputs that maintain the spatiotemporal structure of natural sign language.

The model is trained with a multi-objective loss function that combines diffusion denoising loss, reconstruction loss, semantic preservation loss, and adversarial identity confusion loss, with loss weights adjusted through a curriculum learning strategy to ensure stable convergence.

3. Experiments

To assess the feasibility of the proposed approach, we conduct preliminary experiments on the 22-signer subset described in Section 2.1. The signers are split into 15 for training, 4 for validation, and 3 for testing, yielding approximately 700 samples across 100 glosses. We evaluate the anonymization performance along three dimensions: identity protection (measured by the drop in identity classification accuracy after anonymization), semantic preservation (measured by cosine similarity between original and anonymized latent representations), and reconstruction quality (measured by root mean square error between joint coordinates).

While prior work has developed identity recognition models based on 2D pose or kinematic statistics (Bigand et al., 2021), no established model exists specifically for 3D sign language skeletal motion data. We therefore train a lightweight classifier based purely on skeletal dynamics features to capture each signer's unique motion patterns. This model achieves 81.7% accuracy on the validation set, substantially exceeding random chance, confirming that identity information is clearly encoded in the raw motion data. The same model serves dual purposes in our framework: providing quantitative evaluation of anonymization effectiveness and supplying adversarial training signals through the gradient reversal mechanism.

Category	Metric	Before	After
Privacy	Identity Acc.	81.7%	38.0%
Utility	Semantic Sim.	1.000	0.994
Quality	RMSE	–	0.064

Table 1: Evaluation results before and after anonymization.

Table 1 summarizes the evaluation results. After anonymization, identity classification accuracy drops from 81.7% to 38.0%, a reduction of 43.7 percentage points, demonstrating that the proposed method effectively disrupts identity-specific features. To evaluate semantic preservation, we employ a pre-trained graph convolutional network (GCN) for sign language recognition to extract motion features from both original and anonymized sequences, then compute their cosine similarity. The observed value of 0.994 suggests that 99.4% of the semantic information, as captured by the recognition model, is retained after anonymization. This high preservation rate indicates that the linguistic content of the signs remains largely intact despite the identity-related modifications. The RMSE of 0.064 measures the geometric deviation between original and anonymized joint coordinates, confirming that the overall motion structure is maintained.

Figure 3 visualizes the identity feature distributions using t-distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction. In the original data (left), samples show a general tendency to form clusters by signer identity, with each color representing a distinct signer, though some inter-cluster overlap is present. After anonymization (right), the clusters significantly overlap, confirming that identity-specific features have been effectively obfuscated while the overall structure of

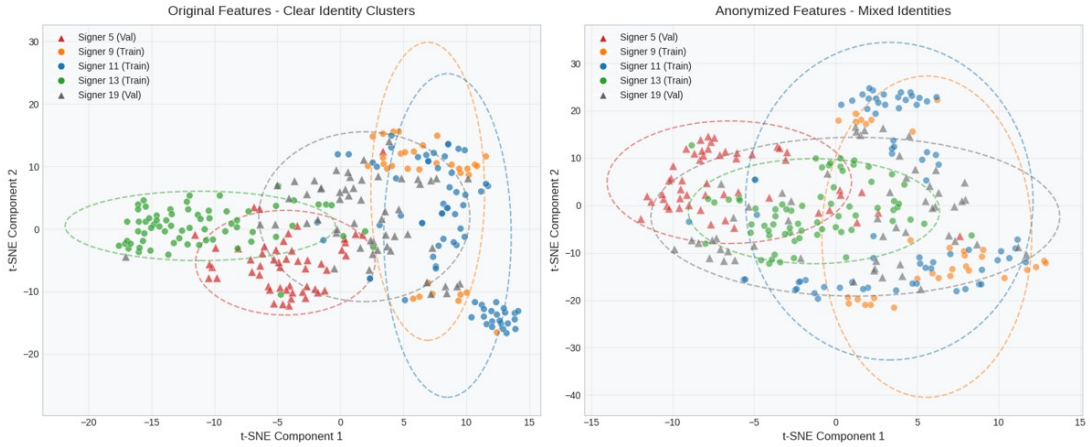

Figure 3: t-SNE visualization of identity features.

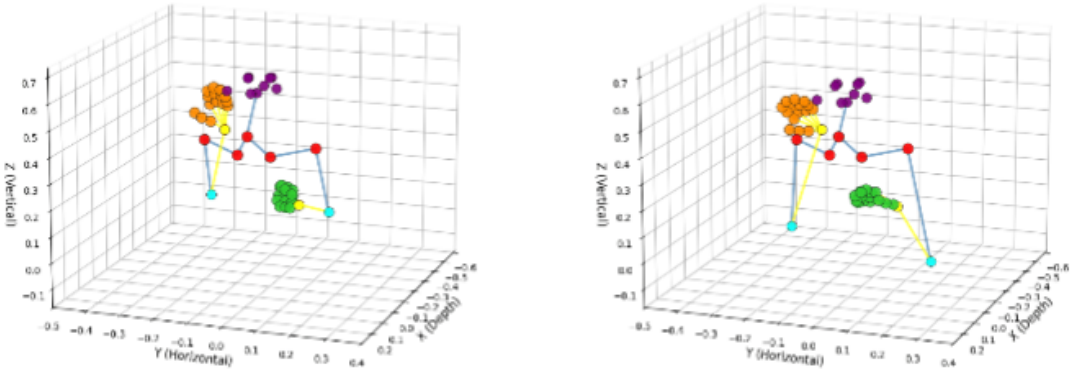

Figure 4: Visual comparison before and after anonymization.

the motion data is retained.

Figure 4 presents a qualitative comparison between original and anonymized motion for the same sign. The overall signing posture, hand configuration, and trajectory are preserved, maintaining semantic interpretability. At the same time, noticeable differences can be observed in hand elevation, arm extension angles, shoulder orientation, and subtle postural characteristics. These are precisely the regions where individual motion styles typically manifest. These modifications suggest that the model has learned to target identity-revealing features while leaving semantically relevant aspects largely intact. However, we note that formal validation by sign language experts remains necessary to confirm full semantic preservation from a linguistic perspective, which we leave for future work.

4. Conclusion and Future Work

This study presents a preliminary exploration of diffusion-based 3D sign language motion anonymization. While the results demonstrate the feasibility of balancing identity confusion with semantic preservation, several limitations should be acknowledged. The current evaluation is constrained by data scale. Our experiments involve only 22 signers. This limited scope may not fully capture the diversity of signing styles across broader populations. Furthermore, 3D sign language motion datasets remain scarce compared to 2D video resources, restricting opportunities for large-scale validation. Additionally, while quantitative metrics suggest high semantic preservation, formal validation by sign language experts has not yet been conducted, leaving open the question of whether linguistic meaning is fully retained from a native signer's perspective.

Several directions warrant further investigation.

Extending the framework to larger, more diverse datasets would enable better modeling of cross-demographic variations in signing styles. Adapting the approach to continuous sign language dialogue presents additional challenges, requiring preservation of paralinguistic features such as emotional expression and conversational dynamics. Extending evaluation to sign languages beyond ASL, such as Japanese Sign Language or Chinese Sign Language, could validate cross-linguistic generalizability and reveal universal patterns in gestural identity characteristics.

From a practical perspective, developing user-controllable anonymization levels would enable adaptive privacy protection based on specific use cases and risk profiles. Investigating the perceptual boundaries of identity recognition in sign language through human subject studies could further inform targeted anonymization strategies that minimize semantic disruption while maximizing privacy protection.

5. Acknowledgements

This research was supported by JSPS KAKENHI Grant Numbers: 23K11197, 23K17511, 25H00473, and 26K14865.

6. Bibliographical References

Alessia Battisti, Eveline van den Bold, Anja Göhring, Franziska Holzknecht, and Sarah Ebling. 2024. Person identification from pose estimates in sign language. In *11th Workshop on the Representation and Processing of Sign Languages: Evaluation of Sign Language Resources*, pages 17–23. ELRA and ICCL.

Félix Bigand, Elise Prigent, and Annelies Braffort. 2020. Person identification based on sign language motion: Insights from human perception and computational modeling. In *Proceedings of the 7th International Conference on Movement and Computing (MoCo)*, pages 1–7. ACM.

Félix Bigand, Elise Prigent, and Annelies Braffort. 2021. Machine learning of motion statistics reveals the kinematic signature of the identity of a person in sign language. *Frontiers in Bioengineering and Biotechnology*, 9.

Zixuan Dai and Shinji Sako. 2025. Motion-based analysis of personalization and kinematic features in Japanese Sign Language video data. In *Adjunct Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents*, New York, NY, USA. Association for Computing Machinery.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1180–1189. PMLR.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469.

Amit Moryossef, Gal Sant, and Zifan Jiang. 2025. Pose-based sign language appearance transfer. In *Proceedings of the Third International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 1–6. European Association for Machine Translation.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985.

Marina Perea-Trigo, Manuel Vázquez-Enríquez, Jose C. Benjumea-Bellot, Jose L. Alba-Castro, and Juan A. Álvarez-García. 2025. Sign language anonymization: Face swapping versus avatars. *Electronics*, 14(12).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. 2021. HuMoR: 3D human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11488–11499.

Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. 2023.

Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*.

Zhaoyang Xia, Yang Zhou, Ligong Han, Carol Neidle, and Dimitris N. Metaxas. 2024. DiffSLVA: Harnessing diffusion models for sign language video anonymization. In *Proceedings of the LREC-COLING 2024 11th Workshop on the Representation and Processing of Sign Languages*, pages 395–407. ELRA and ICCL.

Zhengdi Yu, Shaoli Huang, Yongkang Cheng, and Tolga Birdal. 2024. SignAvatars: A large-scale 3D sign language holistic motion dataset and benchmark. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–19.