SG-DiT: Semantic-Guided Sign Language Anonymization Balancing Privacy and Linguistic Fidelity

Zixuan Dai[1][0009−0001−0521−816X] and Shinji Sako[1][0000−0001−8436−5768]

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 the primary communication medium for approximately 70 million deaf individuals worldwide. In the era of rapid AI advancement, protecting the privacy rights of sign language users while leveraging technological conveniences has emerged as a critical issue. Recent advances in Sign Language Recognition (SLR) and Translation (SLT) [23, 12, 9] have demonstrated the substantial potential of deep learning in sign language processing. However, privacy-preserving research [16] indicates that identity information embedded in visual data poses a non-negligible risk. DiffSLVA [21] applies diffusion models to sign language video anonymization in the 2D domain, but may not fully capture the 3D spatiotemporal structure of sign language movements.

Sign language not only conveys semantic information through handshapes, locations, and movements but also unconsciously carries individual behavioral characteristics, including movement velocity patterns, trajectory curvatures, joint coordination, and temporal rhythms, forming unique "motion signatures". This phenomenon has been validated in other biometric domains such as gait recognition [7] and speaker recognition [4]. In sign language, [1] demonstrated that

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_42

kinematic features serve as identity recognition cues, [2] confirmed that observers can identify signers above chance even with facial occlusion, and [5] validated the cross-linguistic nature of these identity features in Japanese Sign Language.

Traditional video anonymization techniques primarily focus on facial processing [20], but neglect identity features embedded in body movements. Merely occluding facial information cannot achieve genuine anonymization and may provide users with a false sense of security. Modern computer vision can extract identity-revealing motion features from seemingly innocuous videos, raising privacy concerns across educational, medical, judicial, and dataset collection contexts. Growing awareness of privacy rights and biometric data protection regulations [6, 13] necessitates effective anonymization techniques. Recent work on human motion anonymization using foundation models [10] inspires consideration of applying diffusion models to sign language anonymization. Based on these observations, this research proposes a diffusion model-based 3D sign language motion anonymization framework.

The main contributions include: (1) the first application of diffusion models to 3D sign language motion space, achieving balance between semantic preservation and identity removal through semantically guided conditional generation; (2) adversarial identity obfuscation through gradient reversal layers for actively learning identity-confounding representations; (3) comprehensive evaluation demonstrating superior privacy-utility trade-offs compared to statistical feature matching baselines.

2 Preliminaries

2.1 Dataset

This research builds upon WLASL (Word-Level American Sign Language) [14], currently the largest word-level ASL recognition dataset, encompassing 2,000 sign language glosses recorded by over 100 signers. SignAvatars [22] extends WLASL by reconstructing 3D human body models from the original video data using SMPL-X-based fitting techniques, yielding sequences of 65 3D joint coordinates per frame. The 65-joint skeleton comprises 25 body joints (pelvis, spine, limbs, and head), 30 hand joints (15 MANO [18]-based joints per hand), and 10 additional joints for facial landmarks and foot details. For this study, we utilize the ASL100 subset, which consists of the top 100 glosses. Preliminary analysis reveals highly imbalanced signer distributions. To ensure effective identity modeling, we select 22 signers with at least 10 samples each, totaling 684 samples. We adopt a 15:4:3 split ratio for training, validation, and test sets, ensuring no signer overlap across sets to evaluate cross-identity generalization.

2.2 Evaluation Models

The identity discriminator serves dual purposes: quantifying anonymization effectiveness and providing adversarial training signals through gradient reversal

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_42

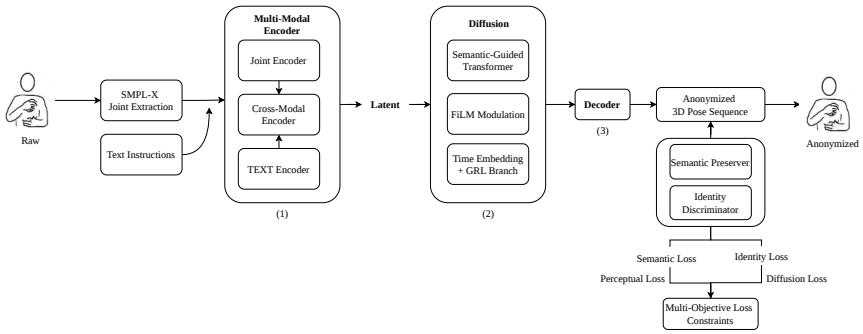

Fig. 1. Overall Framework.

layer (GRL) [8]. We design a lightweight classifier based on 3D skeletal dynamics features, employing a spatial encoder for pose configurations and a bidirectional LSTM for temporal dynamics. The model achieves 81.7% Top-1 accuracy on the 22-signer validation set, significantly exceeding random chance (4.5%), confirming that identity information is reliably encoded in skeletal motion patterns.

For semantic preservation evaluation, we adopt the pre-trained Pose-TGCN from the WLASL benchmark, achieving 49.6% Top-1 and 89.1% Top-10 accuracy on the ASL100 subset. The model parameters remain frozen during anonymization training. We evaluate semantic preservation through cosine similarity between intermediate feature representations of original and anonymized sequences, measuring structural semantic similarity in the learned embedding space rather than relying on classification outputs.

3 Methodology

3.1 Framework Overview

Fig. 1 illustrates the end-to-end diffusion-based architecture comprising three core components.

Multimodal Encoder. The encoder extracts semantically-aware latent representations through joint modeling of skeletal sequences and textual instructions. The joint encoder employs Transformer layers to capture spatiotemporal dependencies. The text encoder embeds a task-specific instruction; in the current implementation, the text input provides a fixed semantic category descriptor. Cross-modal attention mechanisms fuse these modalities into unified latent representations.

Semantic-Guided Diffusion Module. Built upon Transformer architecture, this module performs the core anonymization through a conditional diffusion process, selectively preserving semantic content while transforming identity-related features. It incorporates learnable semantic prompts and FiLM-based [17] feature modulation that dynamically adjust the transformation.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_42

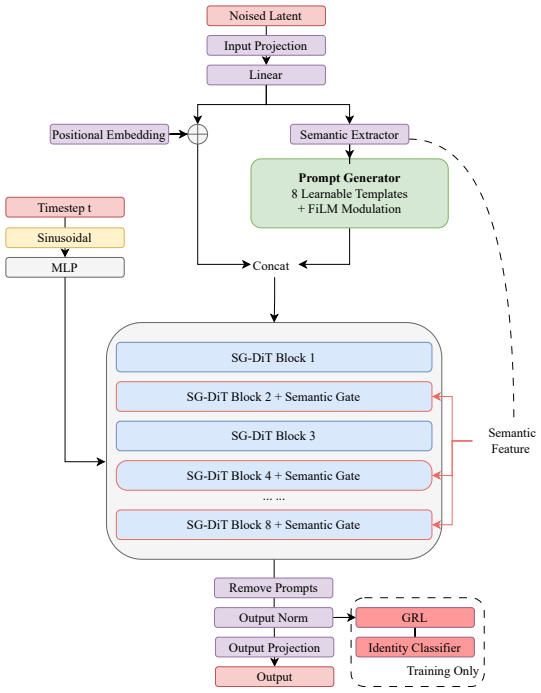

Fig. 2. Semantic-Guided Diffusion Transformer.

Hierarchical Decoder. The decoder reconstructs anonymized joint trajectories from the transformed latent representations. Hand configurations integrate MANO parametric priors with biomechanical constraints. The hierarchical design applies region-specific processing, with hands receiving greater transformation freedom while the trunk remains largely preserved. The framework employs cosine noise scheduling with standard forward-reverse diffusion for training and inference.

3.2 Semantic-Guided Diffusion Transformer

Building upon diffusion probabilistic models [11] and motion diffusion approaches [19], and inspired by learnable conditioning mechanisms such as Q-Former [15] and prompt tuning techniques, we propose a Semantic-Guided Diffusion Transformer (SG-DiT). As illustrated in Fig. 2, the architecture comprises two synergistic components: a semantic prompt generator and a hierarchical semantic gating mechanism.

Semantic Prompt Generation. Given semantic features $\mathbf{s}$, we generate $K = 8$ learnable prompts through semantic-adaptive weighting and FiLM conditioning:

$$\mathbf{P}_i = \alpha_i \cdot \text{FiLM}(\mathbf{p}_i, \mathbf{s}), \quad \alpha_i = \text{softmax}(\text{MLP}(\mathbf{s}))_i \tag{1}$$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_42

where $\mathbf{p}_i$ denotes learnable prompt templates. These prompts are prepended to the input sequence as context-aware prefix tokens.

Hierarchical Semantic Gating. The SG-DiT employs 8 Transformer blocks with semantic gates at even-numbered layers:

$$\mathbf{g}_l = \sigma(\mathbf{W}_g \cdot [\text{FFN}(\mathbf{h}_l), \mathbf{s}]), \quad \mathbf{h}_{l+1} = \mathbf{h}_l + \mathbf{g}_l \odot \text{FFN}(\mathbf{h}_l) \tag{2}$$

The gate values $\mathbf{g}_l \in [0, 1]$ control feature updates, ensuring semantic preservation while permitting modification of identity-specific features.

3.3 Adversarial Identity Confusion

The framework implements adversarial training by connecting the identity discriminator to SG-DiT output features through GRL. During backpropagation, GRL multiplies gradients by $-\lambda$, compelling the feature extractor to learn identity-confounding representations. Beyond cross-entropy loss, the framework maximizes prediction entropy, encouraging uniform identity predictions. Signers with different occurrence frequencies employ differentiated adversarial weights to prevent overfitting.

3.4 Loss Function

The total loss combines eight weighted components: denoising MSE loss, hierarchical reconstruction loss with higher weights on hand joints, semantic feature distance loss, perceptual similarity loss, adversarial identity confusion loss with entropy maximization, temporal coherence loss, velocity smoothness loss, and MANO biomechanical regularization. Training adopts a three-stage curriculum: reconstruction learning, progressive diffusion introduction, and full multi-objective optimization.

4 Experiment

Effective evaluation of sign language anonymization requires balancing privacy protection, semantic preservation, and motion quality. We establish a three-dimensional evaluation framework to comprehensively assess these objectives.

Privacy protection is quantified through identity recognition accuracy, prediction confidence, and entropy on anonymized sequences. Semantic preservation is evaluated using the pre-trained Pose-TGCN model through Top-K recognition accuracy and cosine similarity between feature representations of original and anonymized sequences. Motion quality is assessed through joint position accuracy (MPJPE - Mean Per Joint Position Error), temporal smoothness (acceleration consistency), and biomechanical plausibility.

These individual metrics are aggregated into composite scores. The Privacy Score S_{priv} is a weighted combination of identity accuracy drop (15%), per-sample confusion quality (45%), confidence drop (10%), normalized entropy increase (5%), and class change rate (25%). The Utility Score S_{util} averages cosine similarity, hand motion preservation, and trajectory similarity. The Quality

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_42

Table 1. Comparison with baseline on test set.

Method	Privacy	Utility	Quality	Tradeoff
Stat-Match	0.293	0.891	0.957	0.422
SG-DiT w/o Guidance	0.658	0.628	0.735	0.470
SG-DiT	**0.672**	0.752	0.761	**0.539**

Score S_{qual} is a weighted sum of reconstruction accuracy, temporal smoothness, bone preservation, joint validity, hand motion fidelity, and spatial consistency. The overall Tradeoff Score computes the harmonic mean of Privacy and Utility, scaled by Quality, penalizing imbalanced performance across dimensions.

Existing sign language anonymization methods have focused on 2D video processing, while motion-based anonymization in 3D skeletal space remains largely unexplored. Direct comparison with DiffSLVA [21], which operates in 2D video space, or Pantomime [10], which relies on foundation motion models not available for sign language, is not feasible due to fundamental differences in data representation. We therefore implement a statistical feature matching baseline (Stat-Match) inspired by [3], which anonymizes motion through kinematic feature manipulation.

Table 1 presents quantitative results on the test set of 3 unseen signers,where Privacy, Utility, and Quality correspond to the three evaluation dimensions described in Section 4: identity protection, semantic preservation, and motion quality, respectively. Our SG-DiT achieves a Privacy Score of 0.672, representing a 129% improvement over the statistical baseline, while maintaining competitive Utility (0.752) and Quality (0.761). The balanced performance across all three dimensions yields a Tradeoff Score of 0.539, demonstrating effective privacy-utility-quality balance.

The statistical baseline achieves high semantic preservation but minimal privacy protection due to conservative kinematic modifications. Removing semantic guidance (SG-DiT w/o Guidance) degrades both Privacy and Utility, validating the critical role of semantic conditioning. We note that the post-anonymization identity accuracy (38.0%) remains above random chance (4.5% for 22 signers), indicating room for further improvement. The semantic recognizer's moderate Top-1 accuracy (49.6%) places a ceiling on classification-based evaluation; however, Top-10 accuracy (89.1%) and our primary reliance on feature-space cosine similarity provide a more robust assessment of semantic fidelity.

Fig. 3 presents a qualitative comparison between original and anonymized motion. The overall signing posture, hand configuration, and trajectory are preserved, while noticeable differences in hand elevation, arm extension angles, and shoulder orientation indicate that identity-revealing features have been selectively modified.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_42

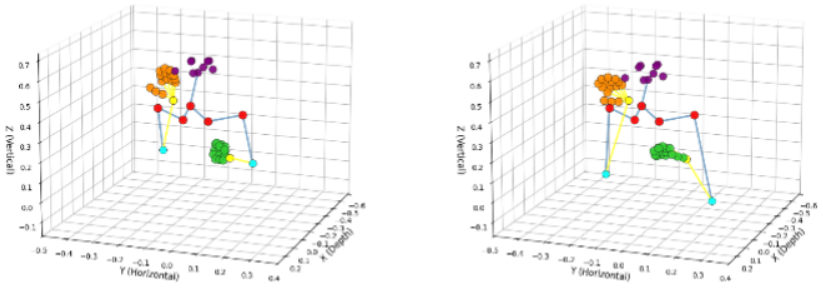

Fig. 3. Visual comparison between original and anonymized motion for the same sign.

5 Conclusions and Future Work

This work proposes a diffusion-based framework for sign language anonymization in 3D pose space, integrating semantic-guided transformers, adversarial training, and biomechanical constraints. The current evaluation is constrained by data scale: available large-scale 3D sign language datasets remain limited, and our experiments involve only 22 signers. Construction and evaluation of datasets with a larger number of signers is an important direction for future work. Other promising directions include adapting the approach to continuous signing, extending evaluation to other sign languages such as Japanese Sign Language (JSL) or Chinese Sign Language (CSL), and developing user-controllable anonymization levels for adaptive privacy protection.

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

References

1. Bigand, F., Prigent, E., Berret, B., Braffort, A.: Machine learning of motion statistics reveals the kinematic signature of the identity of a person in sign language. Frontiers in Bioengineering and Biotechnology **9** (2021)
2. Bigand, F., Prigent, E., Braffort, A.: 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). pp. 1–7. ACM (2020)
3. Bigand, F., Prigent, E., Braffort, A.: Synthesis for the kinematic control of identity in sign language. In: Proceedings of the 7th International Workshop on Sign Language Translation and Avatar Technology: The Junction of the Visual and the Textual: Challenges and Perspectives. pp. 1–6 (2022)
4. Campbell Jr., J.P.: Speaker recognition: A tutorial. Proceedings of the IEEE **85**(9), 1437–1462 (1997). https://doi.org/10.1109/5.628714
5. Dai, Z., Sako, S.: 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. Association for Computing Machinery, New York, NY, USA (2025)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_42

6. European Parliament and Council: General data protection regulation (gdpr) (2016), regulation (EU) 2016/679
7. Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., Huang, Y., Li, Q., He, Z.: Gaitpart: Temporal part-based model for gait recognition. Pattern Recognition **102**, 107208 (2020)
8. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: Proceedings of the 32nd International Conference on Machine Learning (ICML). pp. 1180–1189 (2015)
9. Gong, J., Foo, L.G., He, Y., Rahmani, H., Liu, J.: Llms are good sign language translators. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18362–18372 (2024)
10. Hanisch, S., Todt, J., Strufe, T.: Pantomime: Motion data anonymization using foundation motion models. arXiv preprint arXiv:2501.07149 (2025)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020)
12. Hu, L., Gao, L., Liu, Z., Feng, W.: Continuous sign language recognition with correlation network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2529–2539 (2023)
13. Illinois General Assembly: Biometric information privacy act (bipa) (2008), 740 ILCS 14
14. Li, D., Rodriguez, C., Yu, X., Li, H.: 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. pp. 1459–1469 (2020)
15. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the 40th International Conference on Machine Learning (2023)
16. Mi, Y., Huang, Y., Zhong, Z., Ji, J., Xu, J., Wang, J., Wang, S., Ding, S., Zhou, S.: Privacy-preserving face recognition using trainable feature subtraction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 297–307 (2024)
17. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.C.: Film: Visual reasoning with a general conditioning layer. In: AAAI (2018)
18. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: modeling and capturing hands and bodies together. ACM Trans. Graph. **36**(6) (Nov 2017)
19. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2023)
20. Wang, X., et al.: Sign language anonymization: Face swapping versus avatars. Electronics **14**(12), 2360 (2025)
21. Xia, Z., Zhou, Y., Han, L., Neidle, C., Metaxas, D.N.: 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. pp. 395–407. ELRA and ICCL (2024)
22. Yu, Z., Huang, S., Cheng, Y., Birdal, T.: Signavatars: A large-scale 3d sign language holistic motion dataset and benchmark. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 1–19 (2024)
23. Zhang, H., Guo, Z., Yang, Y., Liu, X., Hu, D.: C2st: Cross-modal contextualized sequence transduction for continuous sign language recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 21053–21062 (2023)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_42