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, Japan
LREC 2026 · 12th Workshop on Sign Languages · Palma de Mallorca, Spain

1. Background & Motivation

- **Why is this important?:**
  Sign language is the primary communication medium for ~70 million deaf individuals worldwide.
  As sign language AI (recognition, translation) advances rapidly, large-scale motion data is increasingly collected and shared.
- **The Privacy Problem:**
  Sign language motions contain identity-specific kinematic features — velocity patterns, trajectory curvatures, joint coordination, and temporal rhythms — forming unique "**motion signatures**".
  Signers can be identified from body movements alone, even with complete facial occlusion (Bigand et al., 2020).
  Traditional anonymization (face swapping/blurring) only addresses facial features, ignoring these motion-level identity cues and providing a false sense of security.
- **Our Goal:**
  Remove WHO is signing while preserving WHAT is being signed
  → anonymize identity features in 3D motion space without degrading linguistic content.

2. Proposed Framework

Three-component pipeline operating directly in 3D pose space:

(1) Multi-Modal Encoder

- **Joint Encoder**:
  4-layer Transformer capturing spatiotemporal dependencies from 65-joint × 3D skeletal sequences.
- **Text Encoder:**
  Sentence-BERT providing semantic context.
- **Fusion:**
  Cross-modal attention (4-layer, 8-head) fuses these into unified latent representations.

(2) Semantic-Guided Diffusion Transformer (SG-DiT)

8 Transformer blocks with FiLM-conditioned semantic prompts and hierarchical gating that dynamically decide which features to preserve vs. modify.

(3) Hierarchical Decoder: Region-specific reconstruction with MANO hand priors.

Hands → greater transformation freedom **Trunk** → largely preserved

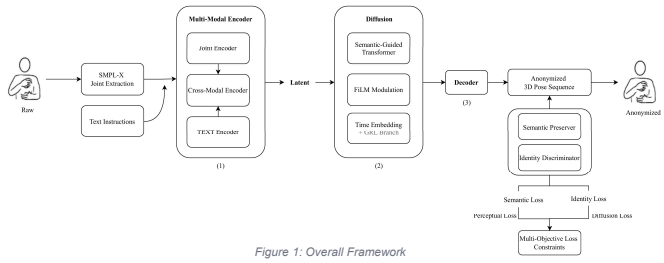

Figure 1: Overall Framework

3. Semantic-Guided Diffusion Transformer

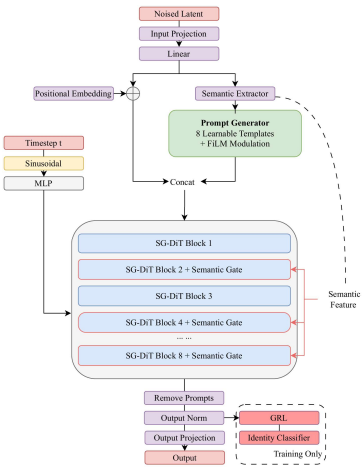

Figure 2: SG-DiT architecture

- **Semantic Prompt Generation**
  - 8 learnable prompt templates, each conditioned via FiLM on the input's semantic features.
  - Softmax attention weights dynamically select which prompts are relevant, enabling input-adaptive conditioning.

$$P_i = \alpha_i \cdot FiLM(p_i, s),\ \ \alpha_i = softmax(MLP(s))_i$$

- **Hierarchical Semantic Gating**
  - Semantic gates at even-numbered layers (2,4,6,8) control feature updates via sigmoid gating:
    - g → 0: preserve feature (semantic content)
    - g → 1: allow modification (identity features)

$$g_l = \sigma(W_g \cdot [FFN(h_l), s]),\ \ h_{l+1} = h_l + g_l \odot FFN(h_l)$$

- **Adversarial Identity Confusion**
  - GRL reverses gradient during backpropagation, forcing the diffusion module to produce identity-uninformative representations.
  - Combined with entropy maximization to encourage uniform predictions across all signer classes.
- **Multi-Objective Loss**
  - Multi-objective loss combining diffusion denoising, reconstruction, semantic preservation, and adversarial identity confusion losses.
  - Trained via 4-stage curriculum learning.

4. Dataset & Experimental Setup

- **Data Source & 3D Reconstruction**

(1) WLASL (Word-Level ASL): largest word-level ASL dataset with 2,000 glosses and 100+ signers.
(2) SignAvatars extends WLASL via SMPL-X fitting, reconstructing 3D body models with 65 joints per frame:

  - 25 body joints (pelvis, spine, limbs, head)
  - 30 hand joints (15 MANO-based per hand)
  - 10 face/foot joints

- **Experimental Subset**
  - ASL100 subset (top 100 glosses).
  - 22 signers selected (≥10 samples each), yielding 684 isolated sign samples. Split: 15 train / 4 validation / 3 test with no signer overlap.
- **Evaluation Models**
  - Identity classifier: Spatial encoder + BiLSTM (81.7% accuracy, vs. 4.5% random chance)
  - Semantic evaluator: Pre-trained Pose-TGCN (49.6% Top-1, 89.1% Top-10)
  - Quality metric: RMSE between original and anonymized joint coordinates

5. Results

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

Key Findings

- **Privacy:**
  - Identity accuracy **81.7%** → **38.0%** (−43.7pp)
    demonstrating effective disruption of identity-specific motion features.
- **Semantics:**
  - Cosine similarity of 0.994 indicates that 99.4% of linguistic information is retained, confirming minimal semantic degradation.
- **Quality:**
  - RMSE = 0.064 / Overall motion structure and naturalness preserved.
- **t-SNE:**
  - Clear clusters → overlapping distributions

Limitation: 38.0% identity accuracy remains above random chance (4.5%), indicating that some identity information persists and further improvement is needed.

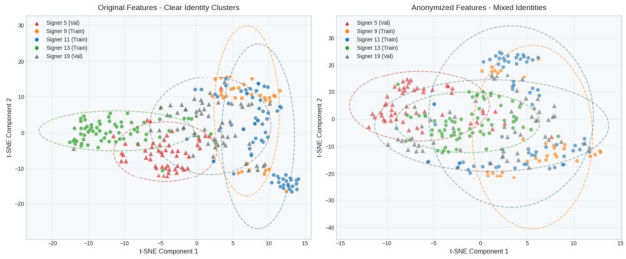

Figure 3: t-SNE visualization of identity features before and after anonymization

6. Qualitative Comparison

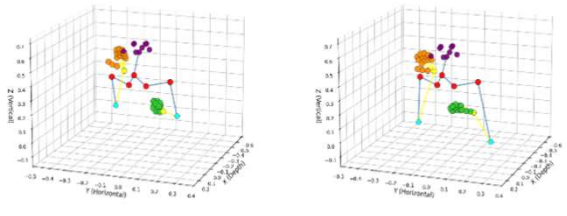

Figure 4: Qualitative comparison of original and anonymized motion

Semantics preserved, identity modified.

The overall signing posture, hand configuration, and trajectory are preserved after anonymization. Visible differences in hand elevation, arm extension angles, and shoulder orientation indicate that identity-revealing kinematic features have been selectively modified while leaving semantically relevant aspects intact.

7. Conclusions & Future Work

Conclusions

- First application of diffusion models to 3D sign language motion anonymization in pose space.
- Semantic-guided conditioning (SG-DiT with FiLM prompts + hierarchical gating) combined with adversarial identity confusion (GRL + entropy maximization) achieves meaningful privacy-utility balance.
- Identity accuracy −43.7pp while retaining 99.4% semantic fidelity..
- Current limitation: 38% identity accuracy remains above random chance (4.5%), and evaluation is constrained by small data scale (22 signers).

Future Work

- Larger and more diverse signer datasets..
- Extend from isolated signs to continuous signing and conversational dialogue, preserving paralinguistic features.
- Cross-linguistic evaluation on Japanese Sign Language (JSL) and Chinese Sign Language (CSL).
- Per-sample semantic conditioning.
- User-controllable anonymization levels.

This research was supported by JSPS KAKENHI Grant Numbers 23K11197, 23K17511, 25H00473, and 26K14865. | Contact: cnq14068@ict.nitech.ac.jp