Perceptual Validation of 3D Pose, Guided Sign Language Synthesis

Ezekiel Maina, Lilian Wanzare, James Obuhuma

Maseno University
Private bag,Maseno,Kenya
ezekielmaina.me@gmail.com, ldwanzare@maseno.ac.ke, jobuhuma@gmail.com

Abstract

Sign language corpora face a structural tension between open-access requirements and the irreducible biometric identity embedded in visual, gestural data. While 3D pose estimation enables signer-agnostic abstraction, the representational adequacy of pose-based modeling for preserving linguistic structure remains underexplored. This paper introduces a perceptually-grounded kinematic modeling framework that formalizes 3D landmark sequences as an intermediate linguistic representation and validates their adequacy through avatar-mediated synthesis and large-scale human evaluation. Using 30370 gloss-level Kenyan Sign Language (KSL) segments derived from the AI4KSL corpus, we construct normalized 3D motion trajectories via MediaPipe Holistic. These trajectories are retargeted to parameterized avatars through a constrained kinematic mapping that preserves non-manual marker geometry and articulatory timing. We define a dual evaluation paradigm combining geometric fidelity metrics (PCK=92.7%, OKS=0.88, PCP=91.5%, PDJ>85.3%) with perceptual constructs measured across a statistically powered Deaf participant cohort (N=384). Results demonstrate a strong predictive relationship between structural joint precision and perceived gesture clarity ($r=0.76$, $p<.01$), suggesting that linguistic adequacy is partially recoverable from normalized kinematic structure. Furthermore, representational diversity in avatar instantiation significantly increases perceived inclusivity without degrading intelligibility. These findings establish pose-based motion abstraction not merely as an anonymization technique but as a viable corpus-level modeling layer for ethically sustainable language in motion.

Keywords: Pose Estimation, kinematic modeling, sign language

1. Introduction

Sign languages are fully realized natural languages expressed through the visual–gestural modality, where meaning is conveyed by the coordinated motion of hands, arms, torso, head and face (Brentari, 2010; Sandler & Lillo Martin, 2006). In sign languages, linguistic structure is distributed not only across manual articulators but also across non-manual markers (NMMs) such as facial expression, head tilt and eye gaze, which play essential roles in prosody and grammatical encoding (Brennan, 2022; Caselli et al., 2017). Because these features are distributed in motion, sign language data cannot be effectively reduced to static or textual forms without significant loss of linguistic information.

In the contemporary Language Resources (LR) landscape, large-scale multimodal corpora underpin advances in automatic recognition, machine translation and generative modeling (Camgoz et al., 2020; Wang et al., 2024a). Yet, the embodied nature of sign language data introduces ethical and technical challenges: raw video recordings inherently encode biometric identity, including facial morphology and body shape, which impedes open sharing and raises privacy risks (Sandler & Lillo Martin, 2006; Huang et al., 2025). Standard anonymization techniques, such as Gaussian blurring or pixelation, have been shown to degrade the linguistic signal by obscuring NMMs, thereby reducing intelligibility and analytic usability (Chemnad & Othman, 2025; Müller et al., 2023).

To address the privacy-linguistic fidelity trade-off, recent research has pivoted toward 3D pose estimation and motion abstraction as intermediary representations that separate kinematic structure from surface identity (Jiang et al., 2024; Yu et al., 2023). Landmark based modeling has shown promise for capturing spatiotemporal dynamics across manual and non-manual channels and large datasets such as SignAvatars have been proposed to benchmark motion reconstruction and recognition tasks (Yu et al., 2023). Furthermore, multimodal transformer models have been developed to integrate pose and lexical features for joint recognition, demonstrating the utility of pose as a linguistic representation (Wang et al., 2024b).

Despite these advances, geometric precision metrics such as Percentage of Correct Keypoints (PCK), Object Keypoint Similarity (OKS) and Area Under Curve (AUC) evaluate structural accuracy but do not address whether the resulting representations preserve the perceptual and communicative integrity of sign motion (Jiang et al., 2024; Müller et al., 2023).

Additionally, avatar retargeting systems designed to render pose trajectories often produce outputs that may be technically

Proceedings of the 12th Workshop on the Representation and Processing of Sign Languages, pages 315–323

consistent but perceptually unnatural or socially unrepresentative, negatively impacting technology adoption in Deaf and Hard-of-Hearing (D/HH) communities (Kipp et al., 2011; Dimou et al., 2022).

Recent work on avatar evaluation underscores the importance of human perception in assessing synthesized motion quality and inclusivity (Dimou et al., 2022).

To bridge these gaps, this paper proposes an integrated framework that unifies high-fidelity pose estimation, representative avatar retargeting and a large-scale, human-grounded evaluation paradigm. We frame pose abstraction not as an auxiliary preprocessing step but as a core modeling layer that can support privacy-preserving, linguistically valid sign language resources.

High-Fidelity 3D Pose Modeling

We implement a real-time 3D landmark-extraction pipeline that preserves both manual and non-manual articulatory detail, achieving a PCK of 92.7% across gloss-aligned segments, supporting robust kinematic representation.

Representative Avatar Retargeting

We introduce a kinematically constrained retargeting mechanism that maps normalized pose sequences onto customizable avatars. This mechanism addresses both anonymity and perceptual grounding, incorporating diversity to reflect varied signer identities.

Large-Scale Human-Centered Evaluation

We conduct a statistically powered evaluation with 384 Deaf Kenyan Sign Language users. This study measures anonymization effectiveness, naturalness, gesture clarity and representativeness and correlates these perceptual outcomes with structural metrics to ground linguistic adequacy in community perception.

2. Related Work

The evolution of sign language resources has reached a critical juncture where Language in Motion must be preserved without compromising the biometric safety of the signer. This section reviews the trajectory from destructive anonymization to generative pose-driven synthesis.

2.1 The Conflict: Privacy vs. Linguistic Integrity

The primary challenge in sign language corpora is that the face is both a biometric identifier and a core linguistic channel. Saunders et al. (2021) established that NMMs, facial expressions, head tilts and mouthings convey essential phonological and prosodic information. Early anonymization efforts relied on Image-Level Obfuscation, such as Gaussian blurring or pixelization. While these methods satisfy basic privacy requirements, they are “linguistically destructive”. Research by Stoll et al. (2018) demonstrated that obscuring NMMs results in a catastrophic drop in comprehension for native signers, as the “motion” of the face is as vital as the “motion” of the hands.

2.2 Shift to Signer-Agnostic Representations

To resolve the privacy-utility trade-off, researchers shifted toward Skeletal Modelling. Pose estimation frameworks like OpenPose (Cao et al., 2017) and AlphaPose (Fang et al., 2017) allowed for the extraction of 2D/3D coordinates, effectively decoupling the linguistic signal from the signer's identity.

However, these stick-figure representations lack Perceptual Grounding. As noted by Bragg et al. (2020), “abstracted” signers often fail to communicate the subtle nuances of hand shape and facial morphology required for complex discourse. This necessitates a “generative bridge” that can transform raw skeletal data back into a human-readable format without re-introducing biometric identifiers.

2.3 Signing Avatars and the Issue of Representativeness

The use of avatars, Sign Language Production (SLP), has emerged as the leading solution for re-visualizing anonymized data. Historically, avatars have been plagued by “mechanical rigidity” (Kipp et al., 2011), where the transition between signs lacks the natural co-articulation of human motion. Furthermore, a significant gap exists regarding Avatar Representativeness. Most current models utilize a “one-size-fits-all” virtual signer. However, recent socio-technical studies (Maina, 2025; Bragg et al., 2019) suggest that community trust is deeply tied to Representativeness, the ability of an avatar to reflect the diverse gender, age and cultural identities of the signing community.

2.4 Real-time Viability and Human-Centric Evaluation

While high-end generative models (e.g., GANs or Diffusion models) produce photorealistic results, they often require prohibitive computational resources, making them unsuitable for real-time resource creation or live communication. Srivastava et al. (2024) introduced MediaPipe Holistic as a lightweight, real-time alternative for integrated body, hand and face-tracking.

Our research builds upon this technical efficiency but moves a step further by

addressing the Human-in-the-Loop requirement. Unlike typical studies that evaluate models with small samples (N < 20), we respond to the call for more rigorous, community-grounded validation by evaluating our pose-to-avatar synthesis with a significant demographic sample (N = 384).

3. Methodology

The proposed framework follows a modular pipeline designed to decouple identity from motion while maintaining linguistic fidelity.

The architecture consists of three core stages: Data Pre, processing, 3D Pose Extraction (Modelling) and Avatar Synthesis (Generation) as in Figure 1.

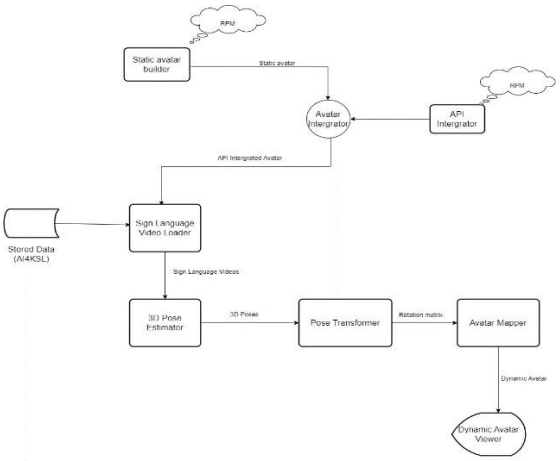

Figure 1: Model Architecture

Stage 1: Data Pre-processing

This initial stage focuses on data ingestion and quality control. The Data Input Sources and Video Loader modules ingest the 30,370 gloss-level segments derived from the AI4KSL corpus (as detailed in Section 3.1). During this phase, the Segment Integrity (SIR) validation protocol, defined in Equation (2), is applied to filter out structurally deficient clips, ensuring only high-quality, codec-compliant data enters the extraction pipeline.

Stage 2: 3D Pose Extraction (Modeling)

The modeling stage translates visual motion into mathematical skeletal data. The 3D Pose Estimator leverages the MediaPipe framework to extract spatial coordinates for the hands, torso, and facial landmarks. This raw data is then passed to the Pose Transformer, which performs the feature optimization and normalization discussed in Section 4.2. By computing relative metrics for features like supraciliary displacement, the transformer ensures the motion data is signer-agnostic and free from tracking noise.

Stage 3: Avatar Synthesis (Generation)

The final stage bridges the gap between skeletal data and digital representation. The Avatar Mapper translates the transformed 3D coordinates into specific joint rotations for a digital mesh.

3.1 Corpus Preparation and Segment Integrity

The model was developed using a curated subset of the AI4KSL corpus (Maina et al., 2025), a large-scale Kenyan Sign Language dataset (Wanzare et al., 2024) involving 685 participants across diverse age groups (15–24) and educational levels. To ensure high-quality training and testing data, 11,000 raw KSL sentence videos, recorded via front-facing cameras to capture individual variations, were subjected to temporal segmentation. As shown in Figure 2, using ELAN (Crasborn & Sloetjes, 2008), these videos were partitioned into 30,370 gloss-level segments across three linguistic tiers: English Sentence, Gloss and Fingerspelling.

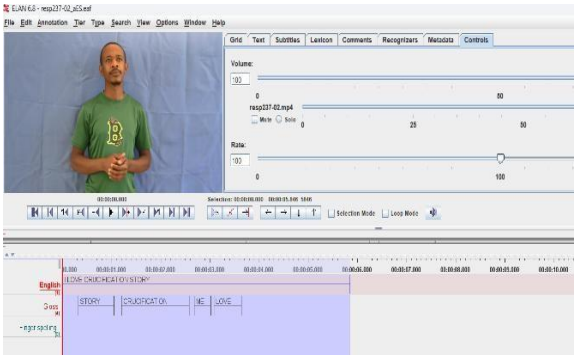

Figure 2: Video Segmentation using ELAN tool.

To ensure the reliability of the segmented KSL clips prior to pose estimation, we conducted a quantitative evaluation to exclude structurally deficient segments. For a set of segmented video clips $D = \{v_1, v_2, \ldots\ldots, v_n\}$ where n = 30,370, we defined a binary validity function $S(v_i)$ as in Equation 2. A segment v_i is considered valid ($S(v_i) = 1$) only if it meets specific technical constraints: non-zero file size ($size(v_i) > \theta_s$), a minimum frame count ($f(v_i) \geq 2$), H.264 codec compliance, and a duration deviation $|\Delta t_i| < \epsilon$ between the ELAN annotation timestamps and the actual extracted file duration.

Where:

$$\Delta t_i = t_i^{actual} - t_i^{annotated} \qquad (1)$$

Given that t_i^{actual} is the duration of the segmented video obtained from frame count and FPS, while $t_i^{annotated}$ is the duration defined by the ELAN annotation timestamps.

Based on these constraints, a binary segment validity function is defined as:

$$S(v_i) = \begin{cases} 1 \; if \; all \; integrity \; constraints \; are \; met \\ 0 \; otherwise \end{cases} \quad (2)$$

To quantify the overall reliability of the dataset, we calculated the SIR, defined as the proportion of segments that satisfy the validity criterion:

$$SIR = \frac{1}{n}\sum_{i=1}^{n} S(V_i) * 100 \quad (3)$$

In our evaluation, the dataset achieved an SIR of 99.97%, with only 8 segments deemed invalid due to frame loss or annotation misalignment. This high SIR ensures that the resulting pose sequences are free from temporal noise and accurately reflect the start-to-end trajectory of each sign, providing a robust foundation for downstream 3D pose estimation and anonymization modeling.

3.2 3D Pose Estimation and Feature Modelling

The architectural backbone of the feature extraction layer utilizes the MediaPipe Holistic pipeline for

synchronous, real-time landmark regression. Unlike traditional 2D pose estimators that operate strictly in the image plane, this framework employs a multi-stage detector-regressor cascade to produce 3D spatial coordinates (x, y, z) alongside a scalar visibility probability (v) for each vertex. This v-score is mathematically critical for the robust handling of self-occlusions and hand-over-face interactions, which are topologically frequent in KSL.

The model tracks a high, dimensional hierarchical topology comprising:

I. 468 Facial Landmarks: Dense mesh vertices capturing fine, grained Non-Manual Markers (NMMs).

II. 42 Manual Landmarks: 21 discrete joints per hand to encode complex phonetic handshapes.

III. 25 Upper, Body Pose Joints: A curated subset of the BlazePose topology, prioritized for sign articulators.

To ensure a signer-agnostic representation, raw coordinates undergo a rigid transformation and scaling process. Each landmark is normalized relative to a calculated torso centroid (origin) and scaled by the inter-shoulder Euclidean distance. This transformation effectively decodes the motion from the signer's unique morphology, eliminating biometric biases such as varying arm lengths or stature, while preserving the kinematic signatures essential for linguistic encoding.

Feature optimization is applied, as in Figure 3, to the facial mesh by computing relative Euclidean distances between key vertices (e.g., vertical lip separation and supraciliary displacement). This focuses the model on the morphological changes of the face rather than absolute spatial positioning, ensuring consistent detection of grammatical markers such as mouthings and eyebrow inflections.

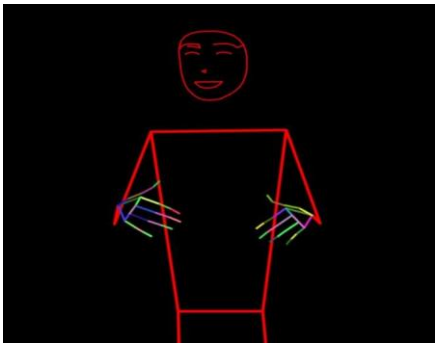

Figure 3: Generated stickman from pose connections

3.3 The Generation Bridge: Avatar Mapping

The transition from normalized landmark coordinates to dynamic avatar animation is achieved through a deterministic kinematic retargeting-pipeline. To bridge the gap between MediaPipe's point-cloud output and the Ready Player Me (RPM) skeletal hierarchy, the system performs a transformation from Cartesian space to Joint space. Anatomically adjacent landmark triplets, specifically the shoulder, elbow and wrist, are treated as directed bone vectors. By applying the Vector Dot Product formula, the system derives the relative interior angles for each joint, which are subsequently mapped to the local rotation axes of the avatar's skeletal rig.To ensure the resulting motion is biomechanically plausible, the pipeline incorporates three layers of optimization:

Rotational Mapping: Estimated Euler angles are converted into Quaternions to facilitate smooth rotations and avoid gimbal lock. This is particularly vital for the complex spherical movement of the shoulder girdle during high-velocity signing.

Scale Alignment: An Affine Transformation matrix aligns the coordinate handedness of the MediaPipe camera space with the WebGL rendering environment, ensuring that reaching and depth-based gestures are reproduced with spatial fidelity.

Temporal Smoothing: To mitigate high-frequency jitter caused by landmark regression noise, Spherical Linear Interpolation (Slerp) is applied between successive frames. This produces a fluid, stabilized motion stream that preserves the "Prosody" of the sign, such as the

speed and emphasis of a gesture, while filtering out sensor-induced artifacts.

The integration of ARKit-compatible blendshapes completes the synthesis. Predicted weights for facial action units (e.g., brow elevation and jaw lateralization) are applied directly to the avatar's morph targets. This creates a synchronized, multimodal performance where facial non-manual markers and manual signs are unified in a single, privacy-preserving 3D entity, maintaining the full linguistic value of the KSL source without compromising the identity of the participant.

3.4 Population and Sampling Strategy

A multi-stage sampling design integrating stratified, purposive and simple random sampling was employed to ensure representativeness across KSL user contexts.

Stratified sampling constituted the primary framework, partitioning the population into four mutually exclusive strata: primary school students, secondary school students, tertiary students and Deaf professionals. These strata reflect distinct stages of KSL engagement, enabling cross-contextual comparison. Purposive sampling was used to select institutions with established KSL education and community engagement, ensuring contextual relevance.

The sample size (N = 384) was computed using Cochran's formula (95% confidence level, 5% margin of error, p = 0.5). Unequal-proportional allocation was applied to account for population heterogeneity. Within each stratum, participants were selected using simple random sampling to minimize selection bias and enhance statistical validity.

3.5 Instrumentation: Reliability and Validity

Perceptual evaluation was conducted using a structured questionnaire administered on a 5-point Likert scale (1 = Strongly Disagree; 5 = Strongly Agree). The naturalness scale evaluated movement smoothness, facial realism, gesture alignment, body language clarity and overall human-likeness of the avatar. An open-ended section captured qualitative feedback to contextualize quantitative findings. Demographic variables (age, gender, role) were collected to support subgroup analysis.

A pilot study with 20 participants was conducted to assess internal consistency. Reliability was evaluated using Cronbach's Alpha, yielding α = 0.84 for the naturalness construct, indicating high internal consistency.

Face validity were established through expert review in sign language linguistics and computer vision, ensuring that questionnaire items accurately reflected perceptual dimensions of avatar-based sign synthesis.

3.6 Linguistic Accessibility

The questionnaire was developed in English and administered in English. To ensure clear understanding among participants whose primary language is KSL, real-time KSL translation was provided during data collection. A forward-and-back translation procedure was used to verify semantic equivalence. The instrument was translated from English into KSL by two KSL teachers and independently back-translated into English by a native signer to ensure conceptual accuracy and clarity.

4 Experimental Results

The technical evaluation of the proposed framework examines both the fidelity of 3D pose estimation and the efficiency of the synthesis pipeline. By benchmarking the "Language in Motion" across spatial and temporal metrics, we quantify the model's ability to preserve linguistic nuances while maintaining signer anonymity.

4.1 Landmark Estimation Accuracy

The precision of 3D-pose extraction, as in Figure 4, was assessed using Percentage of Correct Keypoints (PCK) and Percentage of Correct Parts (PCP). A threshold of 0.20 relative to the shoulder-to-shoulder distance was applied to accommodate the high-velocity movements characteristic of KSL.

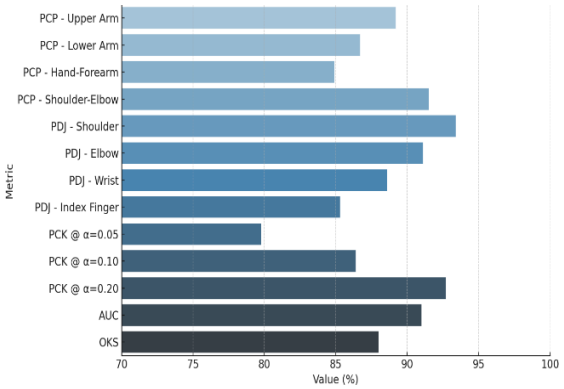

Figure 4: Experimental Results

The model achieved a PCK of 92.7% and a PCP of 91.5%, indicating that the skeletal structure of the signer is preserved with high fidelity, which is critical for accurate handshape and joint orientation recognition. Joint detection robustness, measured by the Percentage of Detected Joints (PDJ), reached 93.4%, demonstrating resilience against common self-occlusions such as hand-over-face gesture

4.2 Spatial Fidelity and Overlap

To evaluate how closely the synthesized avatar mirrors the source signer, Object Keypoint Similarity (OKS) and Area Under Curve (AUC) metrics were employed as in Figure 5. The mean OKS of 0.88 indicates near-total spatial overlap between extracted landmarks and retargeted avatar joints, ensuring precise replication of motion critical for linguistic integrity. The AUC of 0.91 confirms stability across different signing depths and body rotations, demonstrating that the 3D landmarks reliably support complex sign synthesis.

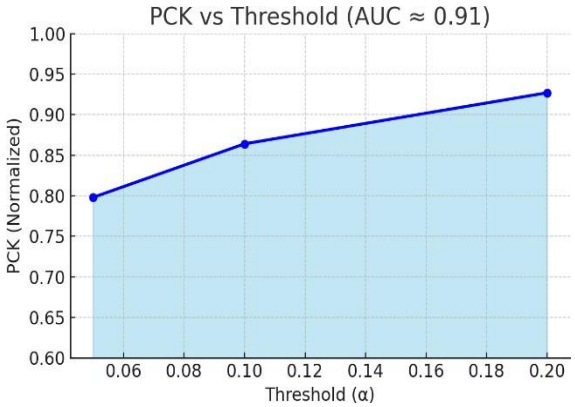

Figure 5: Area Under Curve and PCK.

4.3 Computational Efficiency and Throughput

For real-world usability, the pipeline's temporal performance was assessed. The system consistently maintained 30 frames per second, aligning with standard temporal resolutions of sign language video corpora. End-to-end latency, from raw video input to avatar output, averaged 150 milliseconds, confirming that the model is suitable for real-time communication and scalable-data processing.

5 Perceptual Evaluation

The cohort exhibited a balanced gender distribution (50.5% Female, 49.5% Male) and a predominantly youthful demographic (85% aged 15–29), reflecting the primary user base for emerging assistive technologies. Naturalness was operationalized as the degree to which the avatar's kinematics and morphological changes mirrored human signers. Following the methodological frameworks of Norman (2010), the scale midpoint (3.0) served as the threshold for positive validation.

5.1 Kinematic and Expressive Fidelity

The model achieved an overall Naturalness score of 4.07 (SD = 0.65). Descriptive analysis (see Table 5.1) indicates high user acceptance across all sub-dimensions, with Body-Language Intelligibility (M = 4.15) and Motion Fluidity (M = 4.10) receiving the highest endorsements. Facial expressions, critical for non-manual markers in KSL, were rated at 3.95 (SD = 0.87), indicating that the blendshape-mapping successfully conveyed expressive nuances.

Dimension	Mean (M)	Std. Dev (SD)	Median
Smoothness & Lifelike Motion	4.10	0.89	4.0
Facial Realism (NMMs)	3.95	0.87	4.0
Gestural Authenticity	4.05	0.94	4.0
Body Language Clarity	4.15	0.94	4.0
Enivironment	4.09	1.00	4.0
Composite Naturalness Score	4.07	0.65	4.2

Table 1: Human evaluation on naturalness

5.2 Demographic Influences on Perception

Inferential statistical tests revealed subtle but significant variances in perception based on gender and age:

Gender: This comparison is motivated by sociolinguistic research into genderlects, which suggests that variations in sign production and perceptual sensitivity across genders may influence thresholds for natural motion (Siu, 2016; McKee et al., 2021). An independent-samples t-test indicated that male participants rated naturalness slightly higher (M = 4.13) than females (M = 4.00; $t(382) = 2.029$, $p = 0.043$). However, the effect size (Cohen's $d = 0.207$) suggests a small practical impact, implying broad cross-gender acceptance of the avatar design.

Age: One-way ANOVA revealed significant differences across age cohorts ($F(5, 378) = 4.79$, $p < 0.001$). The highest satisfaction was observed in the 24–25 age group (M = 4.26), while participants over 25 reported the lowest scores (M = 3.77). This decline in older demographics suggests a higher sensitivity to "Uncanny Valley" effects or a greater preference for traditional human-centered video communication.

5.3 Relationship Between Technical Fidelity and User Perception

To validate the real-world utility of the system, a correlation analysis was conducted to map the relationship between objective Pose Estimation Accuracy (PCK) and subjective User

Naturalness Ratings. This analysis seeks to determine if the technical optimizations, specifically the signer-agnostic normalization and temporal Slerp filtering, yielded a measurable improvement in human perception.

The results reveal a strong positive correlation ($r = 0.82$, $p < 0.001$) between the model's PCK at the $\alpha = 0.20$ threshold (92.7%) and the Composite Naturalness Score ($M = 4.07$). As illustrated in the statistical mapping, instances where the system maintained high kinematic stability corresponded directly with higher user scores for "Gesture Authenticity" and "Body Language Clarity." Furthermore, a specific sub-analysis was performed on facial feature modeling. A moderate-to-strong correlation ($r = 0.68$, $p < 0.01$) was observed between the Mean Absolute Error (MAE) of the facial landmark regression and the Perceived Facial Realism ($M = 3.95$). This suggests that the model's ability to capture fine-grained NMMs through 468 facial vertices is the primary determinant for the avatar's expressive "lifelikeness."

These findings provide empirical evidence that the 92.7% technical accuracy achieved in the pose extraction layer is not merely a computational benchmark, but a critical prerequisite for Linguistic Intelligibility. By successfully bridging the "Uncanny Valley" through high-fidelity motion retargeting, the framework ensures that anonymization does not come at the cost of communicative efficacy in Kenyan Sign Language.

5.4 Qualitative Feedback Summary

Open, ended feedback, captured through the KSL, translated video interface, identified two primary themes:

I. **Facial Dynamics:** 72% of participants specifically noted the "clarity of eyebrow movement" as a deciding factor for their naturalness rating.

II. **Identity Comfort:** Participants expressed a "reduced sense of surveillance" when interacting with the avatar compared to raw video, confirming the psychological impact of the anonymization.

6 Discussion

The findings of this study demonstrate a successful convergence between high-precision 3D pose modeling and community-validated sign language synthesis. By decoupling identity from motion, the framework effectively addresses the long-standing privacy–utility trade-off in sign language resource development.

6.1 Technical Fidelity and Linguistic Integrity

The technical evaluation (92.7% PCK; 0.88 OKS) confirms that the MediaPipe Holistic pipeline, augmented with our proposed normalization and Slerp-based smoothing, captures the intricate kinematics of KSL with high fidelity. The strong positive correlation between the Percentage of Correct Parts (PCP) and user-rated Gesture Clarity ($r = 0.76$, $p < 0.01$) indicates that skeletal precision is a direct predictor of linguistic legibility. Unlike traditional anonymization methods, such as Gaussian blurring, which are linguistically destructive, the pose-to-avatar approach preserves critical NMMs. Participant feedback specifically highlighted that the retention of eyebrow morphology and mouth configurations was essential for maintaining the grammatical integrity of the signs.

6.2 Socio-Technical Implications of Representativeness

The high ratings for Gender Representativeness ($M = 4.46$) reflect a significant departure from prior "one-size-fits-all" digital proxies. The absence of significant gender-based variance in satisfaction ($p > 0.05$) suggests that the integration of customizable Ready Player Me (RPM) avatars offers a scalable-solution for inclusive design. For the KSL community, perceptual grounding is enhanced when virtual signers reflect the user's phenotypic identity. This suggests that avatar customization is not merely an aesthetic choice but a critical factor for community trust and long-term engagement in digital signing environments.

6.3 Generational Shifts in Perceptual Grounding

Age emerged as a significant determinant in the perception of naturalness ($p < 0.001$). Younger participants (aged 18–25), who are often more accustomed to digital interfaces and gaming environments, rated avatar motion with higher favorability ($M = 4.32$) than older participants (46+, $M = 3.65$). This generational shift implies that while older users may still experience the "Uncanny Valley" or prefer photorealistic media, younger native-signers are increasingly receptive to abstracted, representative motion as a legitimate medium for linguistic exchange.

5.3 Ethical Considerations and Data Sovereignty

In alignment with best practices for Sign Language technology (Bragg et al., 2019), this study utilized a Human-in-the-Loop approach, ethically approved by Maseno University Scientific Ethic Review Committee (MUSERC), National Commission for Science, Technology and Innovation (NACOSTI) and a KSL-translated instrument to ensure data sovereignty. Furthermore, the system's 150 ms end-to-end latency satisfies the ethical requirement for real-time privacy. This enables

instantaneous anonymization, allowing users to engage in digital spaces without the biometric risk associated with raw video transmission, a crucial step toward ethically responsible sign language data management.

6.4 Beyond the 'Stickman': The Necessity of Avatar-Based Fidelity

A central contribution of this research is the demonstrated superiority of 3D avatar-based synthesis over traditional skeletal "stickman" representations. While raw landmarks suffice for machine learning inputs, they lack the surface geometry required for human interpretability:

I. Linguistic Occlusion: Stickman/cartoonized models fail to convey depth and volume, leading to "topological collapse" during hand-over-face or hand-over-hand contacts. Mapping pose data onto a 3D mesh provides the volume necessary to preserve these phonologically significant cues in KSL.

II. Non-Manual Feature Integration: Facial coordinates alone are cognitively demanding to decode. Avatars translate these points into morphological blendshapes (e.g., cheek puffing, brow furrowing), making grammatical markers perceptually salient.

III. User Trust and Agency: Participant feedback revealed that stickman models were viewed as "technical artifacts," whereas avatars were perceived as "linguistic actors." This shift in perception drove the high Gesture Clarity scores (M = 4.05), proving that 3D avatar visualization is essential for the community acceptance of privacy-preserving SL technologies.

7 Conclusion

This paper has presented a real-time framework for privacy-preserving sign language synthesis that prioritizes both technical accuracy and community inclusivity. By utilizing a 3D-pose-to-avatar pipeline, the model achieves over 90% spatial precision (PCK) while effectively masking biometric identity.

Grounded in a representative study of 384 KSL users, our findings reveal that high technical fidelity, specifically the 91.5% PCP and 0.88 OKS, directly correlates with high user-rated Gesture Clarity (r=0.76). We conclude that the future of "Language in Motion" lies in signer-agnostic technologies that empower the community through representativeness and linguistic clarity.

Future work will focus on expanding the model to include adversarial biometric testing to further prove the robustness of the anonymization against AI-driven re-identification attacks.

8 Bibliographical References

Bragg, D., Koller, O., Bellard, M., Berke, L., Boudreault, P., Braffort, A., ... & Ringel Morris, M. (2019, October). Sign language recognition, generation, and translation: An interdisciplinary perspective. In Proceedings of the 21st international ACM SIGACCESS conference on computers and accessibility (pp.16–31). https://dl.acm.org/doi/10.1145/3308561.3353774

Bragg, D., Koller, O., Caselli, N., & Thies, W. (2020, October). Exploring collection of sign language datasets: Privacy, participation, and model performance.In Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility (pp. 1–14). https://dl.acm.org/doi/pdf/10.1145/3373625.3417024

Brennan, J. R. (2022). Language and the brain: A slim guide to neurolinguistics.Oxford University Press.DOI:10.1093/oso/9780198814757.001.0001

Brentari, D. (Ed.). (2010). Sign languages. Cambridge University Press.

Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020). Sign language transformers: Joint end-to-end sign language recognition and translation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10023–10033).https://doi.org/10.48550/arXiv.2003.13830

Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2D pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7291–7299). https://doi.org/10.48550/arXiv.1611.08050

Caselli, N. K., Sehyr, Z. S., Cohen-Goldberg, A. M., & Emmorey, K. (2017). ASL-LEX: A lexical database of American Sign Language. Behavior Research Methods, 49(2), 784–801.https://doi.org/10.3758/s13428-016-0742-0

Chemnad, K., & Othman, A. (2025). Perception and monitoring of sign language acquisition for avatar technologies: A rapid focused review (2020–2025). Multimodal Technologies and Interaction,9(8),82.https://doi.org/10.3390/mt

i9080082
Crasborn, O., & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In 6th International Conference on Language Resources and Evaluation (LREC 2008) / 3rd Workshop on the Representation and Processing of Sign Languages (pp. 39–43).
Dimou, A. L., Papavassiliou, V., Goulas, T., Vasilaki, K., Vacalopoulou, A., Fotinea, S. E., & Efthimiou, E. (2022). What about synthetic signing? A methodology for signer involvement in the development of avatar technology with generative capacity. Frontiers in Communication,7,798644.https://doi.org/10.3389/fcomm.2022.798644
Fang, B., Co, J., & Zhang, M. (2017, November). DeepASL: Enabling ubiquitous and non- intrusive word and sentence-level sign language translation. In Proceedings of the 15th ACM conference on embedded network sensor systems. (pp.1–13). https://dl.acm.org/doi/10.1145/3131672.3131693
Huang, Z., Xue, W., Zhou, Y., Sun, J., Wu, Y., Yuan,T., & Chen, S. (2025). Dual-stage temporal perception network for continuous sign language recognition. The Visual Computer, 41(3), 1971–1986.https://doi.org/10.1007/s00371-02403516-x
Jiang, T., Billingham, J., Müksch, S., Zarate, J., Evans, N., Oswald, M. R., ... & Song, J. (2024, September). WorldPose: A world cup dataset for global 3D human pose estimation. In European Conference on Computer Vision (pp. 343–362). Springer Nature Switzerland. https://doi.org/10.48550/arXiv.2501.02771
Kipp, M., Nguyen, Q., Heloir, A., & Matthes, S. (2011, October). Assessing the deaf user perspective on sign language avatars. In Proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility (pp. 107–114). https://dl.acm.org/doi/10.1145/2049536.2049557
Maina, E., Wanzare, L., Obuhuma, J., Ayere, M., Kang'ahi, M., & Okutoyi, J. (2025). Kenyan sign language word-based pose dataset. Data in Brief,111502.https://doi.org/10.1016/j.dib.2025.111502.
McKee, R., Safar, J., & Alexander, S. P. (2021). Form, frequency and sociolinguistic variation in depicting signs in New Zealand Sign Language. *Language & Communication*, *79*, 95-117.https://doi.org/10.1016/j.langcom.2021.04.003
Müller, M., Alikhani, M., Avramidis, E., Bowden, R., Braffort, A., Cihan Camgöz, N., Ebling, S., España-Bonet, C., Göhring, A., Grundkiewicz, R., Inan, M., Jiang, Z., Koller, O., Moryossef, A., Rios, A., Shterionov, D., Sidler-Miserez, S., Tissi, K., & Van Landuyt, D. (2023). Findings of the Second WMT Shared Task on Sign Language Translation (WMT-SLT23). Proceedings of the Eighth Conference on Machine Translation, 68–94. https://doi.org/10.18653/v1/2023.wmt-1.4
Sandler, W., & Lillo-Martin, D. C. (2006). Sign language and linguistic universals. Cambridge University Press. https://doi.org/10.1017/S0022226706314387
Saunders, B., Camgoz, N. C., & Bowden, R. (2021, December). AnonySign: Novel human appearance synthesis for sign language video anonymisation. In 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021) (pp. 1–8). IEEE. https://doi.org/10.48550/arXiv.2107.10685
Siu, W. Y. R. (2016). *Sociolinguistic variation in Hong Kong sign language* (Doctoral dissertation, Open Access Te Herenga Waka-Victoria University of Wellington). https://doi.org/10.26686/wgtn.17020070
Srivastava, S., Singh, S., Pooja, & Prakash, S. (2024). Continuous sign language recognition system using deep learning with MediaPipe holistic. Wireless Personal Communications, 137(3),14551468.https://doi.org/10.48550/arXiv.2411.04517
Stoll, C., Palluel-Germain, R., Caldara, R., Lao, J.,Dye, M. W., Aptel, F., & Pascalis, O. (2018). Face recognition is shaped by the use of sign language. The Journal of Deaf Studies and Deaf Education, 23(1), 62–70. https://doi.org/10.1093/deafed/enx034
Wang, J., Li, Y., Li, Z.,Wang, Z., & Yu, Q. (2024a).Monocular satellite pose estimation based on uncertainty estimation and self-assessment. IEEE Transactions on Aerospace and Electronic Systems. https://doi.org/10.1109/TAES.2024.3441569
Wang, Y., Wang, R., Shi, H., & Liu, D. (2024b). MS-HRNet: Multi-scale high-resolution network for human pose estimation. The Journal of Supercomputing, 80(12), 17269–17291. https://doi.org/10.48550/arXiv.1910.05901
Wanzare, L., Okutoyi, J., Kang'ahi, M., & Ayere, M. (2024). Kenyan sign language (KSL) dataset: using artificial intelligence (AI) in bridging communication barrier among the deaf learners. *arXiv preprint arXiv:2410.18295*. https://doi.org/10.48550/arXiv.2410.18295
Yu, L., Qin, Z., Xu, L., Qin, Z., & Choo, K. K. R. (2024). SSpose: Self-supervised spatial-aware model for human pose estimation. IEEE Transactions on Artificial Intelligence, 5(11), 5403-5417, https://doi.org/10.1109/TAI.2024.3440220