

Towards Italian Sign Language Generation for digital humans

Emanuele Colonna^{1,*}, Alessandro Arezzo², Domenico Roberto¹, David Landi², Felice Vitulano², Gennaro Vessio¹ and Giovanna Castellano¹

¹Department of Computer Science, University of Bari Aldo Moro, Italy

²QuestIT S.r.l., Siena, Italy

Abstract

In the rapidly evolving field of human-computer interaction, the need for inclusive and accessible communication methods has become increasingly vital. This paper introduces an early exploration of Text-to-LIS, a new model designed to generate contextually accurate Italian Sign Language (LIS) gestures for digital humans. Our approach addresses the importance of non-verbal communication in virtual environments, focusing on enhancing interaction for the deaf and hard-of-hearing community. The core contribution of this work is developing an iterative framework that leverages a comprehensive multimodal dataset, integrating textual and audio inputs with visual data. Utilizing state-of-the-art deep learning algorithms and advanced human pose estimation techniques, the framework enables the progressive refinement of generated gestures, ensuring realism and contextual relevance. The potential applications of the Text-to-LIS model are wide-ranging, from improving accessibility in digital environments to supporting educational tools and promoting LIS in the digital age. The code is publicly available at: <https://github.com/CarpiDiem98/text-to-lis/>.

Keywords

Sign language generation, Human pose estimation, Digital humans, Inclusive technology

1. Introduction

The advancement of graphics and robotics technology has significantly contributed to the rise of virtual and socially intelligent agents, making them increasingly popular for human interaction. This progress has enabled the development of artificial agents with either virtual or physical embodiments, such as avatars or robots, capable of interacting with humans across diverse settings. Among these, digital humans are particularly impactful, replicating human form and behavior within virtual environments [2].

A key component of effective interaction with digital humans is nonverbal communication, which includes facial expressions, gestures, and body language [3]. Gestures, especially co-speech gestures that accompany verbal communication, enhance these agents' realism and engagement. However, automatically generating natural and synchronized gestures remains a significant challenge due to the complexity and diversity of human nonverbal communication [4].

In this context, sign languages such as Italian Sign Language (LIS) introduce an even more complex dimension of nonverbal communication. Sign languages are not simply gestures but fully developed languages that serve as the primary means of communication for the deaf and hard-of-hearing community. This paper addresses the challenge of generating realistic LIS gestures for digital human agents, recognizing sign languages' critical role in communication and the unique needs of the deaf community.

Specifically, we propose a novel approach that employs an iterative refinement process, training a model on a comprehensive dataset of text and image pairs representing LIS signs (Fig. 1). Our approach integrates textual descriptions and visual data to generate accurate and expressive LIS gestures. The

NL4AI 2024: Eighth Workshop on Natural Language for Artificial Intelligence, November 26-27th, 2024, Bolzano, Italy [1]

*Corresponding author.

✉ emanuele.colonna@uniba.it (E. Colonna); arezzo@quest-it.com (A. Arezzo); d.roberto8@studenti.uniba.it (D. Roberto); d.landi@quest-it.com (D. Landi); felice.vitulano@quest-it.com (F. Vitulano); gennaro.vessio@uniba.it (G. Vessio); giovanna.castellano@uniba.it (G. Castellano)

🆔 0009-0009-0932-3424 (E. Colonna); 0009-0002-8896-7840 (D. Roberto); 0009-0006-6642-1918 (D. Landi); 0000-0002-0883-2691 (G. Vessio); 0000-0002-6489-8628 (G. Castellano)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

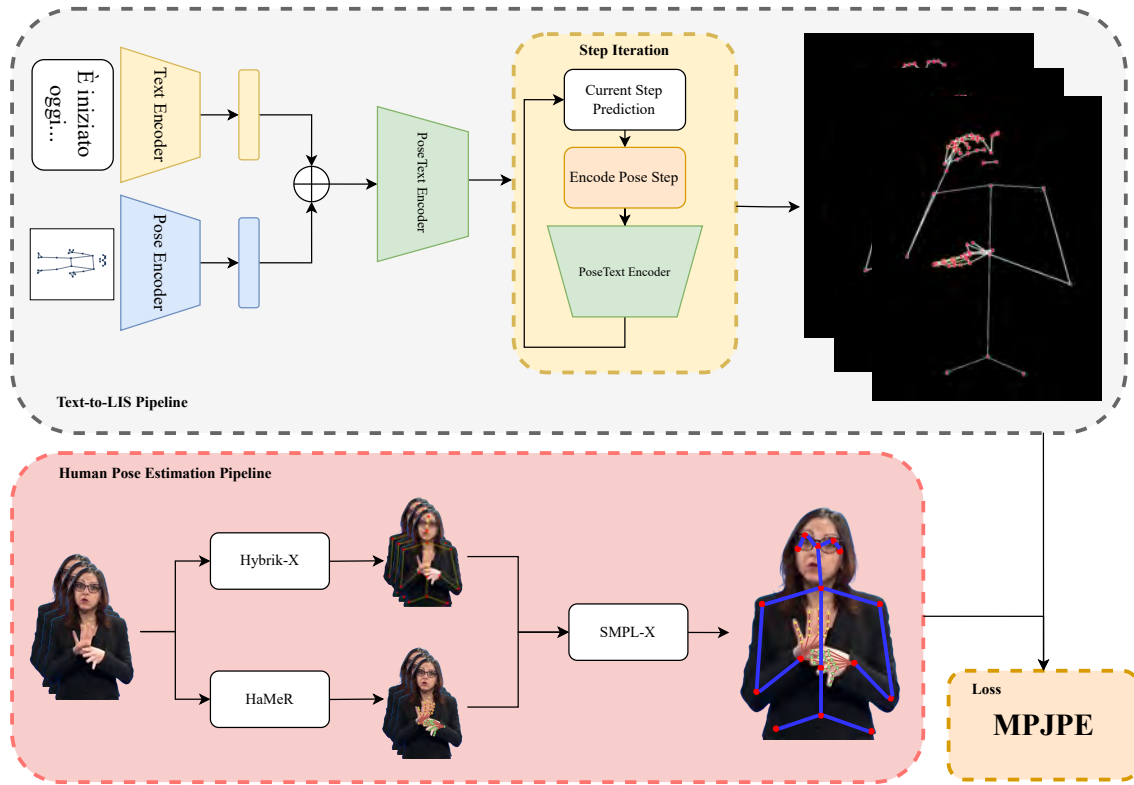


Figure 1: Pipeline of the Text-to-LIS framework. The red section shows the human pose estimation system for generating pseudo-ground truth. The gray section depicts the Text-to-LIS model for generating LIS motion from text. The orange section indicates the metrics for calculating the loss and improving pose quality and accuracy.

model has five main parts: a text encoder, which uses Transformers; a pose encoder, which handles poses; a pose-text encoder, which combines the two; a step encoder, which makes refinements at each step; and a projection module, which produces the final poses. This model captures linguistic and visual aspects of LIS signs. The iterative process begins with an initial generic pose and progresses through multiple steps. Advanced human pose estimation techniques serve as the ground truth for our dataset, allowing for the precise capture and translation of human body movements into 3D animations for virtual models. We present a robust solution for creating natural and coherent LIS gestures by combining textual and visual data.

By advancing technology for LIS gesture generation, we aim to achieve several important goals. First, we seek to improve accessibility by generating accurate and natural LIS gestures, which can enhance communication tools for the deaf community, making digital content and interactions more accessible. Additionally, our work promotes LIS, a minority language facing challenges in preservation and promotion, by contributing to its digital representation and documentation, thus supporting its importance in the digital era. Another significant aim is to enhance education; accurate LIS gesture generation can serve as a valuable resource for educational tools, helping deaf individuals learn written Italian and aiding hearing individuals in acquiring LIS. Moreover, as virtual and augmented reality technologies become prominent, it is crucial to ensure that LIS users can fully participate in these digital environments, fostering inclusivity. Lastly, our model and dataset offer valuable resources for linguistic research, particularly for scholars studying the structure and patterns of LIS, thereby contributing to the broader understanding of sign languages.

The rest of this paper is structured as follows. Section 2 reviews the existing literature. Section 3 introduces the proposed dataset. Section 4 details the proposed Text-to-LIS model. Section 5 presents

preliminary results and discusses future work directions.

2. Related work

Our Text-to-LIS model builds on several areas of research, including pose extraction, sign language datasets, gesture generation, and Italian Sign Language research. This section provides an overview of the relevant work in these fields.

2.1. Pose extraction

Pose extraction is essential for creating realistic digital humans, as it captures and translates human body movements into 3D animations. Using computer vision techniques, pose estimation methods infer human poses from images without requiring markers. These methods are typically categorized into whole-body and single-part estimations, each with specific challenges.

Several models have emerged to reconstruct human body posture from a single image. PIXIE [5] generates complete 3D models even with challenging poses or incomplete body information. Hand4Whole [6] simultaneously estimates both the full-body and hand poses, outperforming prior methods such as FrankMocap [7] and PIXIE. PyMAF-X [8] improves accuracy and speed, estimating SMPL-X parameters with detailed joint rotation and depth information. SMPL-X (Skinned Multi-Person Linear model with eXpressive hands and face) is a comprehensive 3D human body model that integrates detailed representations of the body, face, and hands [9]. SMPL-X parameters are the values used to configure this model, including joint angles, body shape coefficients, and facial expression parameters.

Hand pose estimation has also advanced significantly. Early approaches like those by Baek et al. [10] and Boukhayma et al. [11] used parametric hand models such as MANO [12] to match hand shapes to images. Later methods, such as the already mentioned PyMAF-X [8], moved away from predefined models, directly predicting the 3D shape of the hand point by point, allowing for greater detail and flexibility.

2.2. Sign language datasets

While several datasets exist for various sign languages, there remains a need for more extensive and diverse resources, especially for the Italian language. Notable sign language datasets include:

- RWTH-Phoenix-2014T: A German Sign Language (DGS) dataset with approximately 11 hours of content [13].
- Boston104: An American Sign Language (ASL) dataset with about 9 hours of video [14].
- How2Sign: A large-scale multimodal ASL dataset with 79 hours of content [15].
- TGLIS-227: A LIS dataset with approximately 19 hours of video [16].

Other LIS datasets, such as those in [17, 18], are private or partially accessible. Our work aims to complement and extend these existing resources by providing a novel and comprehensive multimodal dataset for LIS, including video, audio, text, and extracted key points.

2.3. Gesture generation

Recent advancements in gesture generation have focused on creating more natural and context-aware movements. Yoon et al. [3] proposed generating speech gestures using trimodal context, incorporating text, audio, and speaker identity. Their approach highlights the importance of considering multiple modalities for realistic gesture synthesis. Similarly, Yang et al. [4] introduced DiffuseStyleGesture, a diffusion-based model for generating stylized co-speech gestures, demonstrating the potential of advanced generative models for creating diverse and expressive movements.

In the context of sign language generation, Shi et al. [19] developed an open-domain sign language translation system learned from online videos, showcasing the feasibility of generating sign language

from large-scale web data. Our Text-to-LIS model builds on these advancements by incorporating iterative refinement and utilizing textual and visual information to generate accurate and expressive LIS gestures.

2.4. Italian Sign Language research

Research on LIS is growing, but there is still a need for more comprehensive studies and resources. Marchisio et al. [17] introduced deep learning techniques with data augmentation for LIS recognition. Fagiani et al. [18] contributed by creating a new LIS database, adding to the resources available for LIS research. Bertoldi et al. [16] developed a large-scale Italian-LIS parallel corpus, which has been valuable for machine translation and linguistic studies. However, their work primarily focused on text-based representations rather than visual gesture generation.

Our research extends these efforts by creating a more comprehensive LIS dataset and developing a model specifically designed for generating realistic LIS gestures from textual input. This work bridges the gap between textual representations and visual sign language production, contributing to computational linguistics and assistive technology.

3. Proposed dataset

Our proposed dataset is a comprehensive, multimodal collection designed to advance research and application development in Italian Sign Language. It addresses the scarcity of publicly available LIS data and supports various applications, including human movement analysis, nonverbal communication recognition, and understanding human behavior in digital environments.

The dataset includes approximately 37 hours of LIS content:

- Video: High-quality video recordings of signers performing LIS during TV news broadcasts, segmented to align with spoken phrases.
- Audio: Corresponding audio recordings, including the signer’s voice and ambient news sounds.
- Text: Transcriptions of the spoken content, initially generated using Whisper [20] and manually corrected for accuracy.
- Key points: Body and hand joint positions, stored in pickle file format for each frame of the videos.

The segmented videos were generated based on transcriptions produced by Whisper. To streamline the automated process, no preprocessing was applied to the transcription output. As noted qualitatively, glossary extraction techniques, common in many datasets, were not applied as they can potentially decrease the deaf community’s understanding of the movement. In this dataset, a whole sentence is considered text.

We utilized a fully automated web scraping mechanism to gather LIS news broadcast videos from multiple platforms, primarily YouTube, while ensuring compliance with privacy regulations. This approach allowed us to collect diverse signers and contexts, enhancing the dataset’s diversity and representativeness. For key point extraction, we employed two state-of-the-art techniques:

- Hybrik-X [21]: Known for its accuracy and robustness, Hybrik-X is optimized for real-time execution on mobile devices and performs well in high-detail scenarios.
- HaMeR [22]: HaMeR reconstructs a 3D hand mesh from a single RGB image, utilizing a Vision Transformer (ViT) [23] for detailed hand pose estimation.

The extracted key points were normalized using the SMPL-X model [9], ensuring consistency between the body and hand models. Figure 2 shows an overview of the multiple modalities collected in our dataset.

Compared to current state-of-the-art sign language datasets, our LIS dataset stands out in its multimodal nature and substantial duration (see Table 1). While other LIS datasets exist, they are often either private or limited in accessibility [17, 18]. We aim to continually expand this dataset to enhance its utility for LIS research.



Figure 2: Overview of our LIS dataset, comprising approximately 37 hours of LIS videos, including multiple modalities such as video, audio, text, and key points.

Table 1

Overview of publicly available sign language datasets, including ours.

Dataset	Language	Duration (h)	Modalities					
			Multiview	Transcription	Gloss	Pose	Depth	Speech
RWTH-Phoenix-2014T	DGS	11	✗	✓	✗	✓	✓	✗
Boston104	ASL	≈ 9	✗	✓	✓	✗	✗	✗
How2Sign	ASL	79	✓	✓	✓	✓	✓	✓
TGLIS-227	LIS	≈ 19	✗	✓	✗	✓	✗	✓
LIS (ours)	LIS	≈ 37	✗	✓	✗	✓	✗	✓

4. Proposed method

This section presents our Text-to-LIS model for automatic gesture generation based on textual descriptions. Our approach builds on the work of Zhang et al. [24], employing an iterative refinement process to generate a sequence of poses from textual input generated by automatic transcription [20]. The key innovation of our method lies in its ability to progressively enhance pose quality through multiple refinement steps, leveraging both textual and positional information.

4.1. Model architecture

The core components of our Text-to-LIS model, shown in Fig. 1, include:

- **Text encoder:** A Transformer-based encoder that processes text embeddings to generate a dense representation of the input text. It uses multi-head attention mechanisms and feed-forward neural networks to capture the contextual relationships between tokens. The text encoder receives the corresponding phrase of the LIS gesture as its input.
- **Pose encoder:** A Transformer-based encoder designed to handle the sequence of poses. This encoder applies attention mechanisms to the current state of the gesture (which, in the initial iteration, is a generic starting pose) to represent the directional matrices.
- **Pose-text encoder:** This component combines and processes the joint information from text and pose data.
- **Step encoder:** A small neural network representing the current iterative process step. It refines this representation with embedding layers, integrating information from previous steps to inform subsequent pose adjustments.
- **Projection module:** This module transforms hidden representations into final poses, mapping the refined poses back into the appropriate output space.

The process is iterated, with each iteration taking the output from the previous step as its new input. This iterative approach attempts to translate sentences into fluid and accurate movements effectively.

4.2. Iterative refinement process

The iterative refinement process is the heart of our Text-to-LIS model. It works similarly to an artist creating a painting, starting with a rough sketch and gradually refining it through multiple steps until a detailed work of art emerges. The process begins with a textual input (a description of a gesture in LIS) and an initial generic pose, which serves as a foundation. From this starting point, the model iterates through a series of refinements, progressively improving the pose.

At each iteration, the text and current pose are processed by their respective encoders, allowing the model to “understand” both the description and the current pose. The step encoder keeps track of the progress made so far, integrating information from previous refinements. Based on this understanding, the model outputs an improved pose version. This process repeats over several iterations, with each cycle producing a more accurate and detailed representation of the LIS gesture described in the text. This gradual refinement allows the model to capture subtle nuances and correct errors step by step, leading to more natural and expressive gesture generation.

4.3. Training procedure

Training our Text-to-LIS model involves strategies to facilitate effective learning, creativity, and precision. Two key techniques used during training are:

- **Teacher forcing [25]:** Similar to guiding an apprentice artist, this technique alternates between allowing the model to make its predictions and providing the correct pose for the next step. This approach enables the model to learn from independent attempts and supervised guidance, improving its ability to generate accurate poses.
- **Controlled noise injection:** To improve robustness and flexibility, we introduce random variations (or “noise”) into the poses during training. This involves adding small Gaussian noise to joint positions. This is akin to practicing under different conditions—such as using different brushes or lighting in art—which helps prevent overfitting and encourages the model to learn the underlying structure of LIS gestures.

5. Preliminary results and future work

We conducted exploratory experiments training the model for 200 epochs with a batch size of 16. Our analyses were performed on a subset of the dataset, consisting of approximately six thousand videos, each with an average duration of ten seconds. Table 2 summarizes the hyperparameters used during model training and evaluation.

Two loss functions were employed to evaluate the model’s performance: the MPJPE (Mean per Joint Position Error) and a refined loss specifically designed to account for the model’s confidence in each predicted pose point. To define this loss function, first, the squared error between the ground truth pose P_i^j and the predicted pose \hat{P}_i^j is calculated:

$$E_i^j = \|\hat{P}_i^j - P_i^j\|^2. \quad (1)$$

This error is then weighted by a confidence vector C_i^j that represents the model’s certainty about each predicted joint position, leading to the loss function:

$$L_i^j = C_i^j \|\hat{P}_i^j - P_i^j\|^2. \quad (2)$$

This loss function enables the model to prioritize joints with higher confidence while assigning less weight to uncertain predictions. Finally, the mean weighted error is calculated and normalized, yielding the final refined loss:

$$L_{\text{refined}} = \frac{1}{N \cdot J} \sum_{i=1}^N \sum_{j=1}^J L_i^j \cdot \log(S + 1), \quad (3)$$

Table 2

Hyperparameters used in model training.

Category	Hyperparameter	Value
Generic	Seed	42
	Batch size	16
Sequence	Max sequence size	10000
	Noise epsilon	1e-4
	Sequence length weight in loss calculation	2e-5
Model	Dimension of hidden encoder	128
	# Text encoder layers	2
	# Pose encoder layers	4
	# Pose refinement steps	10
	Encoder feed-forward size	2048
Optimizer	Adam learning rate	1e-3

Table 3

Comparison of the number of refinement steps and the quality of the generated poses.

# Steps	Refined loss (Train)	MPJPE (Test)
1	0.07	0.20
10	0.12	0.10

where N represents the number of the samples in the batch multiplied by the number of the joints in each pose J . A key feature of this loss is the normalization based on the number of model steps S , computed with the logarithmic function $\log(S + 1)$. The MPJPE is a widely used metric for assessing the accuracy of 3D pose estimation. It quantifies the average discrepancy between predicted and actual joint positions across all samples:

$$\text{MPJPE} = \frac{1}{N \cdot J} \sum_{i=1}^N \sum_{j=1}^J \|\hat{P}_i^j - P_i^j\|_2, \quad (4)$$

where \hat{P}_i^j represents the predicted 3D coordinate for joint j of sample i , P_i^j is the corresponding ground truth, N is the number of samples, and J is the number of joints per sample.

We conducted experiments comparing different configurations to determine the optimal number of refinement steps. As shown in Table 3, increasing the number of refinement steps significantly improves the quality of the generated poses. The improvement was most pronounced up to ten refinement passes, after which further increases produced diminishing returns and significantly increased the generation time. Specifically, with ten refinement passes, the optimal balance between the generated poses' accuracy and the model's computational demands was observed. Each refined step took approximately a few seconds when training the model on a GeForce RTX 4090 graphics card.

The preliminary results demonstrate the effectiveness of the Text-to-LIS model in generating realistic LIS poses from textual descriptions. The model's iterative refinement approach produces high-quality poses, as evidenced by qualitative evaluation. These results (Fig. 3) indicate the model's potential as a valuable tool for enhancing digital human interactions, virtual reality environments, and nonverbal communication systems.

While the results are promising, several avenues for further research and development remain. Expanding the dataset with more diverse signers, gestures, and contexts is essential to improve the model's generalization capabilities. On the technical side, investigating advanced attention mechanisms and temporal modules may help the model better capture long-term dependencies and subtle nuances in gestures. Real-time sign language generation is another critical goal for practical applications,

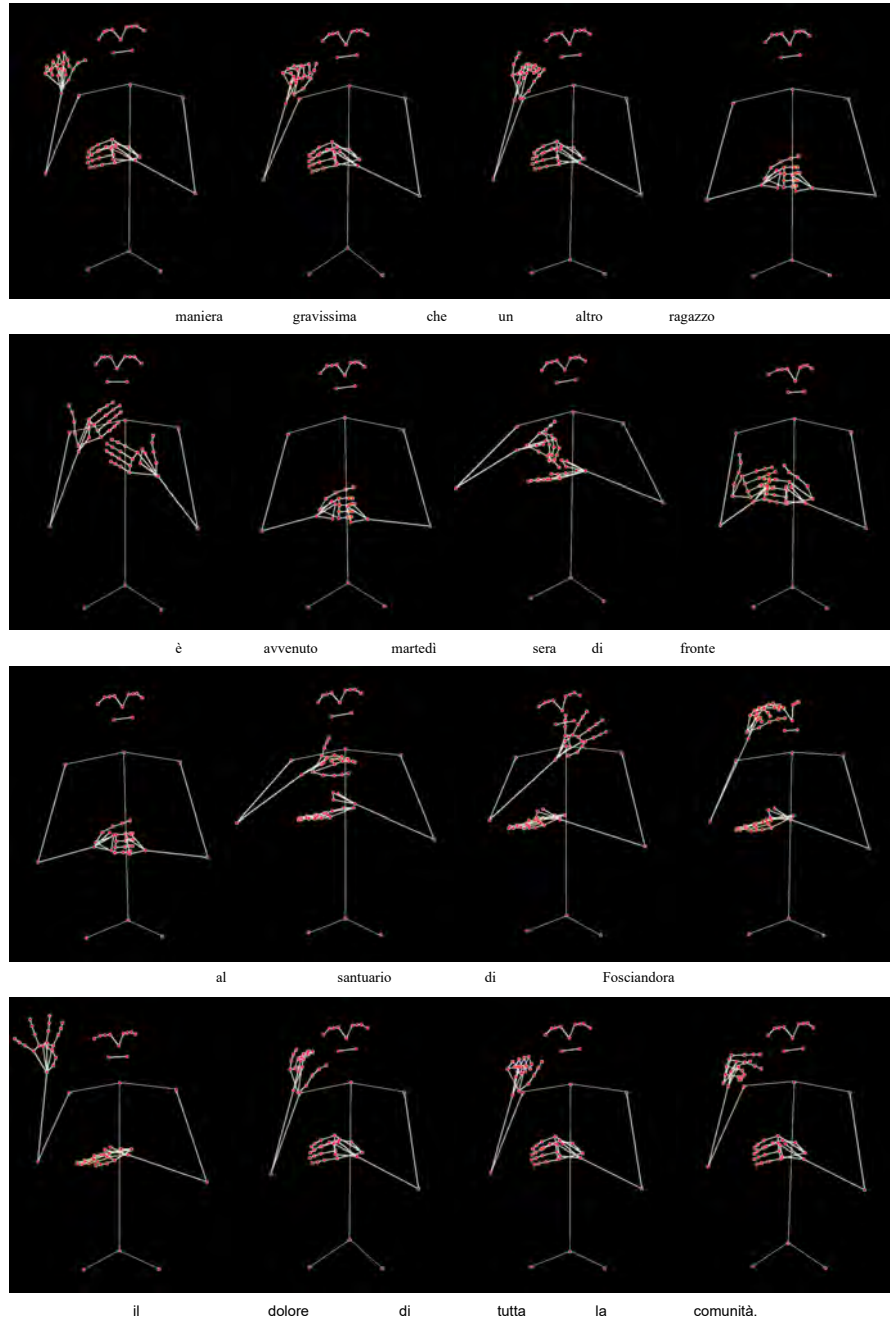


Figure 3: The example illustrates the generation of LIS poses from textual input. The input is taken from the text “*in maniera gravissima un altro ragazzo, è avvenuto martedì sera di fronte al santuario di Fosciandora, il dolore di tutta la comunità*” (in english: “*very serious way another boy, happened Tuesday night in front of the sanctuary of Fosciandora, the pain of the whole community*”). The Text-to-LIS model was employed to generate the LIS poses.

and techniques like model pruning and quantization could reduce computational complexity without sacrificing accuracy. Given that sign language communication is inherently multimodal, future work should also focus on integrating hand gestures, facial expressions, and body language into a unified model to generate more natural and expressive LIS gestures. Moreover, although the current focus is on LIS, the techniques developed in this research could be adapted to other sign languages or nonverbal communication systems, broadening the scope and impact of this work.

Finally, collaboration with the deaf community, linguists, and technologists will be essential to ensure that our advancements are both technically sound and socially impactful.

Acknowledgments

This research was supported by a PhD fellowship awarded to Emanuele Colonna, funded under the Italian National Recovery and Resilience Plan (D.M. n. 117/23), Mission 4, Component 2, Investment 3.3. The PhD project, titled “Study of AI Techniques for Efficient Generation of Digital Humans and 3D Environments” (CUP H91I23000690007), is co-funded by QuestIT S.r.l. Additionally, this research was partially supported by the UNIBA-MAML (Microsoft Azure Machine Learning) agreement.

References

- [1] G. Bonetta, C. D. Hromei, L. Siciliani, M. A. Stranisci, Preface to the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI 2024) co-located with 23th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2024), 2024.
- [2] M.-Y. Yin, J.-G. Li, A systematic review on digital human models in assembly process planning, *The International Journal of Advanced Manufacturing Technology* 125 (2023) 1037–1059.
- [3] Y. Yoon, B. Cha, J.-H. Lee, M. Jang, J. Lee, J. Kim, G. Lee, Speech gesture generation from the trimodal context of text, audio, and speaker identity, *ACM Transactions on Graphics (TOG)* 39 (2020) 1–16.
- [4] S. Yang, Z. Wu, M. Li, Z. Zhang, L. Hao, W. Bao, M. Cheng, L. Xiao, Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models, in: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23, International Joint Conferences on Artificial Intelligence Organization, 2023, pp. 5860–5868. URL: <https://doi.org/10.24963/ijcai.2023/650>. doi:10.24963/ijcai.2023/650.
- [5] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, M. J. Black, Collaborative Regression of Expressive Bodies using Moderation, in: International Conference on 3D Vision (3DV), 2021.
- [6] G. Moon, H. Choi, K. M. Lee, Accurate 3D Hand Pose Estimation for Whole-Body 3D Human Mesh Estimation, in: Computer Vision and Pattern Recognition Workshop (CVPRW), 2022.
- [7] Y. Rong, T. Shiratori, H. Joo, FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration, in: IEEE International Conference on Computer Vision Workshops, 2021.
- [8] H. Zhang, Y. Tian, Y. Zhang, M. Li, L. An, Z. Sun, Y. Liu, Pymaf-x: Towards well-aligned full-body model regression from monocular images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
- [9] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, M. J. Black, Expressive Body Capture: 3D Hands, Face, and Body from a Single Image, in: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2019.
- [10] S. Baek, K. I. Kim, T.-K. Kim, Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1067–1076.
- [11] A. Boukhayma, R. d. Bem, P. H. Torr, 3d hand shape and pose from images in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10843–10852.
- [12] J. Romero, D. Tzionas, M. J. Black, Embodied hands: modeling and capturing hands and bodies together, *ACM Trans. Graph.* 36 (2017). URL: <https://doi.org/10.1145/3130800.3130883>. doi:10.1145/3130800.3130883.
- [13] J. Forster, C. Schmidt, O. Koller, M. Bellgardt, H. Ney, Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather, in: LREC, 2014, pp. 1911–1916.

- [14] P. Dreuw, D. Rybach, T. Deselaers, M. Zahedi, H. Ney, Speech recognition techniques for a sign language recognition system, *hand 60* (2007) 80.
- [15] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, X. Giro-i Nieto, How2sign: a large-scale multimodal dataset for continuous american sign language, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2735–2744.
- [16] N. Bertoldi, G. Tiotto, P. Prinetto, E. Piccolo, F. Nunnari, V. Lombardo, A. Mazzei, R. Damiano, L. Lesmo, A. Principe, On the creation and the annotation of a large-scale italian-lis parallel corpus, *International Conference on Language Resources and Evaluation* (2010) 19–22.
- [17] M. Marchisio, A. Mazzei, D. Sammaruga, Introducing Deep Learning with Data Augmentation and Corpus Construction for LIS, in: *Italian Conference on Computational Linguistics*, 2023. URL: <https://api.semanticscholar.org/CorpusID:266726316>.
- [18] M. Fagiani, S. Squartini, E. Principi, F. Piazza, A New Italian Sign Language Database, 2012. doi:10.1007/978-3-642-31561-9_18.
- [19] B. Shi, D. Brentari, G. Shakhnarovich, K. Livescu, Open-Domain Sign Language Translation Learned from Online Video, in: *EMNLP*, 2022.
- [20] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via Large-Scale Weak Supervision, 2022.
- [21] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, C. Lu, Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3383–3393.
- [22] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, J. Malik, Reconstructing hands in 3d with transformers, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9826–9836.
- [23] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, X. Zhai, An image is worth 16x16 words: Transformers for image recognition at scale, 2021.
- [24] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, Z. Liu, Motiondiffuse: Text-driven human motion generation with diffusion model, *arXiv preprint arXiv:2208.15001* (2022).
- [25] S. Bengio, O. Vinyals, N. Jaitly, N. Shazeer, Scheduled sampling for sequence prediction with recurrent Neural networks, in: *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, NIPS'15*, MIT Press, Cambridge, MA, USA, 2015, p. 1171–1179.