RuSignBot: Russian Sign Language Synthesis via Customized MimicMotion

Daria Bondarenko[1,5], Emilia Bojarskaja[1,5], Maxim Novopoltsev[2], Aleksandr Tulenkov[2], Ruslan Murtazin[2], Iuliia Zemtsova[2], Ilya Makarov[3,4,5], Andrey Savchenko[4,6]

[1]AI Talent Hub; [2]Sber AI; [3]AI Research Institute, Moscow, Russia;
[4]Research Center of the Artificial Intelligence Institute, Innopolis University, Innopolis, Russia;
[5]ITMO University, Saint Petersburg, Russia; [6]Sber AI Lab, Moscow, Russia
dnbondarenko@itmo.ru, eeboiarskaia@itmo.ru, iamakarov@hse.ru, avsavchenko@hse.ru

Abstract

The hearing-impaired community is often underserved due to barriers such as a lack of linguistically appropriate services, especially in non-native English-speaking countries. This paper presents RuSignBot, a novel isolated word-to-sign generation for the Russian language based on an adapted MimicMotion architecture. To enhance the realism and expressiveness of output videos, we introduce a domain adaptation strategy on a large-scale sign language corpus. Quantitative evaluation demonstrates that our fine-tuned model achieves superior performance in standard full-reference metrics, specifically SSIM and PSNR, compared to its base version. Furthermore, we propose a human-centric Sign Understandability Score and conduct a user study with fluent signers. The results confirm that the generated signs are recognized with high accuracy, underscoring the model's communicative efficacy. To facilitate practical application, we integrate the model into a Telegram-based application that converts user-input text into animated sign language videos. The system supports both default and user-defined avatars, highlighting its potential for real-world deployment in assistive technology contexts.

Code and pre-trained models —
https://github.com/ds-hub-sochi/mimic_text2video

Introduction

The hearing-impaired community is often underserved due to barriers such as a lack of linguistically appropriate services (McKee et al. 2022; Kaur et al. 2024). Sign language serves as a primary linguistic medium for hearing-impaired communities worldwide (Novopoltsev et al. 2024; Wong, Camgoz, and Bowden 2024). Yet, technologies for automated sign language generation (SLG) remain limited in accessibility and linguistic expressiveness, especially for specific languages with limited labeled data available for model training. Two distinct methodological paradigms currently dominate the field of SLG, each presenting significant limitations. The first approach employs neural end-to-end models, often based on sequence-to-sequence or diffusion architectures, to generate pose sequences or videos directly from input text or glosses (Fang et al. 2024). While promising, these methods are highly data-intensive and frequently fail to accurately articulate linguistically complex manual parameters, such as precise handshapes and movements. The second paradigm comprises pose-guided systems, which leverage human pose keypoints to drive the video generation process from a source image (Fang et al. 2024). Although effective in controlling global body motion, these systems are critically limited by the inherent inaccuracies of pose estimation algorithms. Errors in extracting fine-grained hand configurations and facial landmarks directly propagate through the generation process, resulting in output videos that often lack the articulatory precision required for natural and intelligible sign language communication.

A significant portion of this underserved community uses American Sign Language (ASL), which is a distinct language, and many hearing-impaired individuals face difficulties in accessing healthcare information and services that are not provided in a culturally or linguistically appropriate manner, especially in non-native English-speaking countries (Murphy and Dodd 2010). In this paper, we introduce RuSignBot, a novel approach for generating isolated signs in Russian Sign Language (RSL). Our core contribution is a domain adaptation of the MimicMotion model (Zhang et al. 2024) for SLG, leveraging its confidence-aware pose guidance mechanism to prioritize reliable keypoints for precise articulation and reduced motion artifacts. We quantitatively validate our approach using standard full-reference metrics (e.g., PSNR, SSIM), demonstrating clear gains over the baseline model. We also introduce a human-centric evaluation methodology to assess the naturalness and intelligibility of synthesized signs. The implemented system operates by retrieving video exemplars via sign gloss from a database and using their extracted ground-truth pose keypoints to guide the synthesis of a photo-realistic signing avatar. Finally, we demonstrate the practical viability of this technology through a fully-functional Telegram bot, providing an accessible interface for text-to-sign video generation with support for both a default avatar and custom user uploads.

Related Work

Rule-based systems and animated avatars initially addressed SLG (Cox et al. 2002; Braffort et al. 2016). These methods leveraged predefined linguistic rules and glossaries to ensure grammatical correctness but often produced rigid and unnatural animations, failing to capture essential nuances, such as

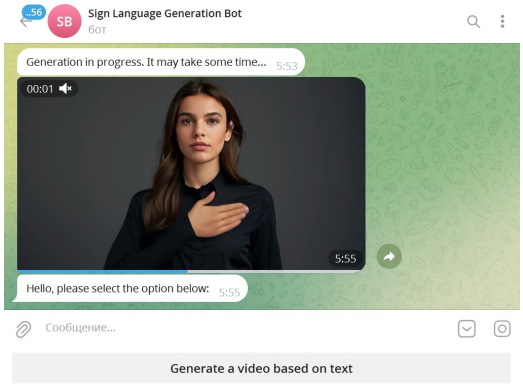


Figure 1: RuSignBot user interface

non-manual signals (Kipp et al. 2011).

The field has since shifted to data-driven neural approaches, enabling the learning of complex grammatical structures and generating more fluid motion (Saunders, Camgoz, and Bowden 2020c; Zelinka and Kanis 2020; Inan et al. 2022). However, these methods are hindered by the scarcity of large parallel datasets, and their output is limited to pose data, which lacks the visual detail required for fully comprehensible communication.

To overcome this visual deficit, recent work has integrated generative video models (Saunders, Camgoz, and Bowden 2020a; Xiao, Qin, and Yin 2020; Ventura, Duarte, and Giró-i Nieto 2020). Pose-conditioned GANs have been employed to synthesize photo-realistic signers, significantly enhancing intelligibility and user acceptance over prior methods (Saunders, Camgoz, and Bowden 2020b; Stoll et al. 2018, 2020). However, a fundamental limitation persists: these methods often result in unnatural articulations or artifacts.

More recently, diffusion models have demonstrated superior performance in generating high-fidelity, temporally coherent videos (Tripathy, Kannala, and Rahtu 2021; Mallya et al. 2020). The MimicMotion model, in particular, has shown notable success in generating videos of human actions with impressive detail and accuracy (Zhang et al. 2024). Despite these advancements, a critical challenge remains: they often fail to capture the fast, small movements typical of sign language.

Building upon these advancements, we propose a novel approach based on the MimicMotion framework (Zhang et al. 2024) to ensure high visual fidelity and linguistic accuracy. This is achieved by explicitly modeling the intricate synchrony of sign language through specialized pre-training on a large, domain-specific corpus of sign language.

Proposed Approach

To provide an accessible and user-friendly interface for sign language generation, we designed a modular pipeline that transforms natural language text into a video of an avatar performing the corresponding signs in RSL. This complete pipeline has been integrated into a practical Telegram bot application (Fig. 1). Our system (Fig. 2) consists of three core modules that operate sequentially to achieve this goal.

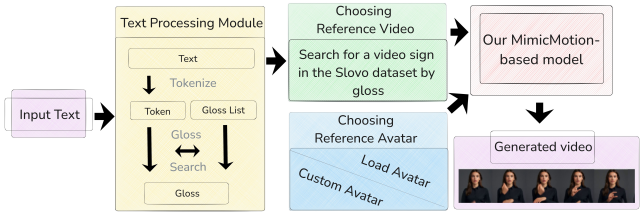


Figure 2: Sign Language Generation System Architecture

1. Input Text Preprocessing Module. Our generative model requires precise pose sequences; however, text input remains the most natural and accessible method for users. This module bridges that gap by converting an input word in either Russian or English into a standardized gloss label from our RSL database, which is derived from the Slovo dataset (Kapitanov et al. 2023). The processing begins with lemmatizing the input word, followed by an attempt to match it directly against the database dictionary. For out-of-vocabulary words, the module conducts an embedding-based similarity search. It encodes the lemma using a multilingual Sentence Transformer and retrieves the nearest gloss via a FAISS index of precomputed gloss embeddings.

2. Gloss-Video Matcher. The primary function of this module is to provide the ground-truth visual sign language data necessary to guide the subsequent video generation process. Utilizing the gloss label obtained from the first module, this component is responsible for retrieving a corresponding example of the sign from a curated database of sign language videos. To address out-of-vocabulary words or unsuccessful gloss matches, the linguistic module employs a fallback strategy that returns the sign corresponding to the dictionary word with the closest semantic meaning, identified through an embedding search.

3. Sign Language Generation Module is the core generative component of our system. It takes the retrieved sign video and a user-defined avatar image as input. The module first extracts a sequence of human pose keypoints (e.g., hand, body, face landmarks) from the retrieved video. This pose sequence, along with the target avatar, is then fed into our adapted MimicMotion model (Zhang et al. 2024). The model, which we have specifically trained for this task, generates a novel video of the chosen avatar performing the sign (Fig. 3). The output preserves the precise articulation from the retrieved pose data while rendering it seamlessly with the user's preferred avatar, striking a balance between high visual quality and low inference latency for interactive use.

The backend of our system is implemented in PyTorch with GPU acceleration support via CUDA to handle the computational demands of the diffusion model. The system generates videos at a resolution of 576px and 15 FPS, providing a user-friendly and responsive text-to-sign interaction experience. The current average latency from the query to the generated video is between 1 and 1.5 minutes.

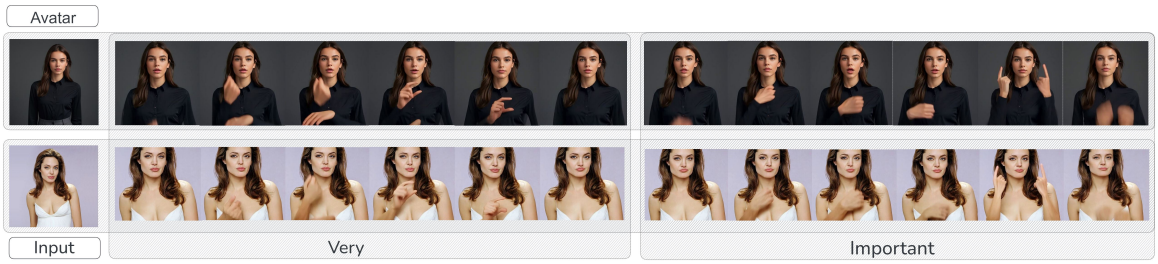


Figure 3: Sample frames generated by the proposed system.

Our Sign Language Generation Model

MimicMotion (Zhang et al. 2024) is a pose-guided human video generation framework that incorporates confidence-aware pose guidance. It utilizes confidence maps to prioritize reliable keypoints during synthesis, thereby enhancing spatial accuracy and temporal consistency. These properties make MimicMotion especially suitable for generating high-quality sign language videos, where clarity of hand shape and fluidity of motion are essential. However, the base MimicMotion algorithm is not explicitly designed for sign language, necessitating domain-specific training to capture the fine-grained articulations required in SLG.

Data

We use the validation subset of the How2Sign dataset (Duarte et al. 2021), specifically its green-screen studio component. This data consists of 1,741 frontal-view RGB video clips (1280x720 resolution, 30 FPS, $\sim$ 5.4 seconds), each paired with English sentences and 2D body pose keypoints. The dataset features 11 signers, including deaf or hard-of-hearing individuals who use ASL as their primary language, thereby ensuring linguistic authenticity and natural signing production. While the dataset provides keypoints, we opted for enhanced accuracy by re-extracting them using the DWPose model (Yang et al. 2023), as proposed in (Zhang et al. 2024).

Training

The proposed framework utilizes frozen, pre-trained components from the Stable Video Diffusion model (Blattmann et al. 2023). Specifically, it employs the Variational Autoencoder (VAE) encoder and decoder, which encode frames into a latent space and reconstruct them with temporal consistency. The primary learnable component of the system is a spatiotemporal U-Net (Ronneberger, Fischer, and Brox 2015), which performs denoising in latent space and adapts to new motion patterns based on pose sequences and reference images. Additionally, PoseNet (Kendall, Grimes, and Cipolla 2015) encodes both pose and confidence data to guide motion synthesis.

The input of the training pipeline (Fig. 4) is RGB frames, reference images, and pose sequences. The model learns to synthesize videos that faithfully replicate motion trajectories while maintaining visual appearance. The loss function is based on per-pixel MSE, regionally weighted by the confidence of keypoints. Additional scaling of the loss is applied to hand regions to improve articulation fidelity.

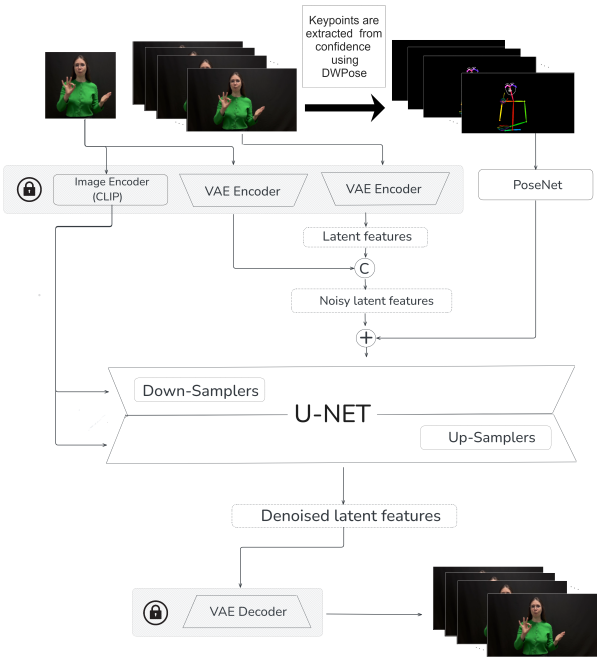


Figure 4: Proposed SLG model of RuSignBot

To evaluate training quality, we conducted a visual assessment on the Slovo dataset (Kapitanov et al. 2023) for RSL. As shown in Fig. 5, we observed a degradation in quality in later epochs, particularly in motion realism and the naturalness of the facial expressions exhibited by the generated character. Hence, we limited the number of training epochs.

Experiments

To evaluate our SLG model, we computed two key metrics, SSIM and PSNR, using a sample of 100 randomly selected, non-repeating glosses from the Slovo dataset's glossary. These metrics provide critical insights into the perceptual quality and structural integrity of generated sign language videos. PSNR focuses on pixel-level fidelity, and SSIM evaluates structural similarity. The results indicate that domain adaptation is effective in achieving a high-quality model for SLG (Table 1).

The MimicMotion model was selected as the foundation

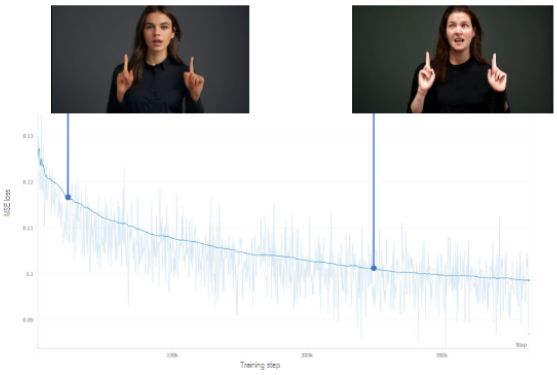


Figure 5: Training curve for our model

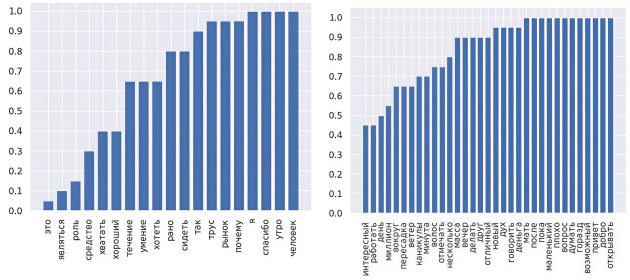
Table 1: SSIM, PSNR, and APE results for the MimicMotion model with and without domain-specific training

Model	SSIM ↑	PSNR ↑	APE ↓
Original MimicMotion	0.7597	19.7309	0.4723 ± 0.07
SLG MimicMotion (Ours)	**0.8169**	**21.1261**	$\mathbf{0.4567 \pm 0.05}$

for our approach due to its proven ability to replicate reference movements with high fidelity. We quantified this capability using the keypoint-based Average Position Error (APE) metric, conducting evaluations both before and after a specific pre-training phase. This metric is used to assess the fidelity of generated 2D poses by quantifying the average Euclidean distance between the predicted and ground-truth joint positions across all frames. The analysis was performed on the same 100 samples, similarly to other metrics reported in Table 1. The resulting APE scores for the original MimicMotion model and our proposed model were found to be similar (Table 1). This outcome demonstrates that the proposed pre-training successfully preserved the model's core capacity for high-fidelity motion generation and even slightly enhanced its movement accuracy.

The primary objective of sign language generation is to produce semantically accurate and intelligible animations. To move beyond kinematic metrics and directly assess functional performance, we introduce the "Sign Understandability Score" (SUS) metric. This metric quantifies the core communicative efficacy of a generated sign by measuring its recognition rate (Makarov et al. 2019; Savchenko 2012; Savchenko and Belova 2015), evaluating whether the intended meaning is successfully conveyed.

We evaluated a set of 50 generated sign videos (across two packs of 20 and 30). Each video was presented to 20 distinct signers of RSL (a total of 40 unique participants). Participants provided free-form text responses describing the recognized sign. These responses were normalized using a predefined lemmatization dictionary (e.g., collapsing "go", "walking" to "walk"; "cowardly" to "be afraid") to account for semantic equivalence and lexical variation. The per-sign accuracy was calculated as the proportion of correct identifications after normalization. The final SUS metric is the mean accuracy across all 50 signs, representing the average probability that a generated sign will be correctly understood, as shown in Fig. 6.



a) A pack of 20 words b) A pack of 30 words

Figure 6: Distribution of the SUS metric for each word in the estimated sample

Table 2: SUS, SSIM, and PSNR results for our system in a set of 20 and 30 generated sign videos pack

Samples	**SUS**	**SSIM**	**PSNR**
A pack of 20 words	0.668	0.772	20.767
A pack of 30 words	0.845	0.837	22.247
Mean for two packs	0.7565	0.8045	21.907

The values of metrics SUS, SSIM, and PSNR for each word pack, along with their averages, are presented in Table 2. Analysis of these data reveals a correlation between the standard full-reference metrics and the proposed human-centric SUS metric.

Conclusion and Future Work

This study presents a practical word-to-sign synthesis system for Russian Sign Language using a customized MimicMotion framework. Experiments have shown that our model, fine-tuned using a domain-specific approach, achieves superior performance in quantitative metrics (PSIM and SSIM) compared to its base version. Furthermore, our human-centric evaluation demonstrated a high level of understandability, confirming that native RSL users accurately recognize the generated signs.

RSL synthesis algorithms have a wide range of applications, primarily in enhancing accessibility for the Deaf and hard-of-hearing community. In our Telegram-based RuSignBot, we demonstrated the potential of using this technology for intuitive, user-driven sign language learning and communication, supporting both default and custom avatars.

Future research will pursue two primary directions. First, we aim to significantly expand the system's lexicon by augmenting both the training dataset and the gloss database. Second, we will focus on advancing from generating isolated signs to modeling fluent, continuous sign language sequences. Furthermore, the proposed pipeline is not limited to RSL and can be adapted for other sign languages, provided the necessary pose and video data are available. Another promising direction is to optimize the model for faster inference, making it suitable for real-time applications in a broader range of devices.

Acknowledgments

The work of Ilya Makarov and Andrey Savchenko was supported by the Ministry of Economic Development of the Russian Federation (agreement No. 139-10-2025-034 dd. 19.06.2025, IGK 000000C313925P4D0002).

References

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Braffort, A.; Filhol, M.; Delorme, M.; Bolot, L.; Choisier, A.; and Verrecchia, C. 2016. KAZOO: a sign language generation platform based on production rules. *Universal Access in the Information Society*, 15(4): 541–550.

Cox, S.; Lincoln, M.; Tryggvason, J.; Nakisa, M.; Wells, M.; Tutt, M.; and Abbott, S. 2002. Tessa, a system to aid communication with deaf people. In *Proceedings of the fifth international ACM conference on Assistive technologies*, 205–212.

Duarte, A.; Palaskar, S.; Ventura, L.; Ghadiyaram, D.; DeHaan, K.; Metze, F.; Torres, J.; and Giro-i Nieto, X. 2021. How2sign: a large-scale multimodal dataset for continuous american sign language. In *CVPR*, 2735–2744.

Fang, S.; Chen, C.; Wang, L.; Zheng, C.; Sui, C.; and Tian, Y. 2024. SignLLM: Sign Language Production Large Language Models. *arXiv preprint arXiv:2405.10718*.

Inan, M.; Zhong, Y.; Hassan, S.; Quandt, L.; and Alikhani, M. 2022. Modeling intensification for sign language generation: A computational approach. *arXiv preprint arXiv:2203.09679*.

Kapitanov, A.; Karina, K.; Nagaev, A.; and Elizaveta, P. 2023. Slovo: Russian sign language dataset. In *International Conference on Computer Vision Systems*, 63–73. Springer.

Kaur, B.; Chaudhary, A.; Bano, S.; Yashmita; Reddy, S.; and Anand, R. 2024. Fostering inclusivity through effective communication: Real-time sign language to speech conversion system for the deaf and hard-of-hearing community. *Multimedia Tools and Applications*, 83(15): 45859–45880.

Kendall, A.; Grimes, M.; and Cipolla, R. 2015. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, 2938–2946.

Kipp, M.; Nguyen, Q.; Heloir, A.; and Matthes, S. 2011. Assessing the deaf user perspective on sign language avatars. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, 107–114.

Makarov, I.; Veldyaykin, N.; Chertkov, M.; and Pokoev, A. 2019. American and Russian sign language dactyl recognition. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, 204–210.

Mallya, A.; Wang, T.-C.; Sapra, K.; and Liu, M.-Y. 2020. World-consistent video-to-video synthesis. In *European Conference on Computer Vision*, 359–378. Springer.

McKee, M.; James, T. G.; Helm, K. V.; Marzolf, B.; Chung, D. H.; Williams, J.; and Zazove, P. 2022. Reframing our health care system for patients with hearing loss. *Journal of Speech, Language, and Hearing Research*, 65(10): 3633–3645.

Murphy, J.; and Dodd, B. 2010. A diagnostic challenge: Language difficulties and hearing impairment in a secondary-school student from a non-English-speaking background. *Child Language Teaching and Therapy*, 26(3): 207–220.

Novopoltsev, M.; Tulenkov, A.; Murtazin, R.; Akhidov, R.; Zemtsova, I.; Bojarskaja, E.; Bondarenko, D.; Savchenko, A.; and Makarov, I. 2024. Video-based learning of sign languages: one pre-train to fit them all. In *International Conference on Data Mining Workshops (ICDMW)*, 899–902. IEEE.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Saunders, B.; Camgoz, N. C.; and Bowden, R. 2020a. Adversarial training for multi-channel sign language production. *arXiv preprint arXiv:2008.12405*.

Saunders, B.; Camgoz, N. C.; and Bowden, R. 2020b. Everybody sign now: Translating spoken language to photo realistic sign language video. *arXiv preprint arXiv:2011.09846*.

Saunders, B.; Camgoz, N. C.; and Bowden, R. 2020c. Progressive transformers for end-to-end sign language production. In *European Conference on Computer Vision*, 687–705. Springer.

Savchenko, A. V. 2012. Adaptive video image recognition system using a committee machine. *Optical Memory and Neural Networks*, 21(4): 219–226.

Savchenko, A. V.; and Belova, N. S. 2015. Statistical testing of segment homogeneity in classification of piecewise-regular objects. *International Journal of Applied Mathematics and Computer Science*, 25(4).

Stoll, S.; Camgöz, N. C.; Hadfield, S.; and Bowden, R. 2018. Sign language production using neural machine translation and generative adversarial networks. In *Proceedings of the 29th British Machine Vision Conference (BMVC 2018)*. British Machine Vision Association.

Stoll, S.; Camgoz, N. C.; Hadfield, S.; and Bowden, R. 2020. Text2Sign: towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128(4): 891–908.

Tripathy, S.; Kannala, J.; and Rahtu, E. 2021. Facegan: Facial attribute controllable reenactment gan. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1329–1338.

Ventura, L.; Duarte, A.; and Giró-i Nieto, X. 2020. Can everybody sign now? exploring sign language video generation from 2d poses. *arXiv preprint arXiv:2012.10941*.

Wong, R.; Camgoz, N. C.; and Bowden, R. 2024. Sign2GPT: Leveraging large language models for gloss-free sign language translation. *arXiv preprint arXiv:2405.04164*.

Xiao, Q.; Qin, M.; and Yin, Y. 2020. Skeleton-based Chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. *Neural networks*, 125: 41–55.

Yang, Z.; Zeng, A.; Yuan, C.; and Li, Y. 2023. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4210–4220.

Zelinka, J.; and Kanis, J. 2020. Neural sign language synthesis: Words are our glosses. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 3395–3403.

Zhang, Y.; Gu, J.; Wang, L.-W.; Wang, H.; Cheng, J.; Zhu, Y.; and Zou, F. 2024. MimicMotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*.