Word-Conditioned 3D American Sign Language Motion Generation

Lu Dong **Xiao Wang** **Ifeoma Nwogu**
University at Buffalo, SUNY
{ludong, xwang277, inwogu}@buffalo.edu

Abstract

Sign words are the building blocks of any sign language. In this work, we present wSignGen, a word-conditioned 3D American Sign Language (ASL) generation model dedicated to synthesizing realistic and grammatically accurate motion sequences for sign words. Our approach leverages a transformer-based diffusion model, trained on a curated dataset of 3D motion meshes from word-level ASL videos. By integrating CLIP, wSignGen offers two advantages: image-based generation, which is particularly useful for children learning sign language but not yet able to read, and the ability to generalize to unseen synonyms. Experiments demonstrate that wSignGen significantly outperforms the baseline model in the task of sign word generation. Moreover, human evaluation experiments show that wSignGen can generate high-quality, grammatically correct ASL signs effectively conveyed through 3D avatars.

1 Introduction

Sign languages are natural languages with their own grammatical structures and lexicons, primarily used by Deaf and Hard-of-Hearing (DHH) communities (Müller et al., 2023). While most current sign language research focuses on sign-to-text translation (Müller et al.; Moryossef et al., 2021; Ye et al., 2023; Zhang and Duh, 2023), we address the reverse problem with wSignGen [1]– a model dedicated to sign language synthesis and production, converting text into corresponding signing motions.

Many sign synthesis works in the literature (more on the sign synthesis approaches are discussed in Section 2) aim to directly learn the mappings between continuous visual signs and written language, via neural machine translation (NMT). But we propose a different approach where the NMT would occur between the textual words and the corresponding ASL gloss[2]. The resulting gloss sentences can then readily be furnished with ASL word-sign motions, the focus of this work.

Current continuous sign language production methods only perform close-set generation and cannot handle words that were not previously seen; whereas, we empirically observe that new ASL signers can construct a large number signed phrases based only on a fixed number of signs, further motivating the importance of word sign synthesis.

But word sign synthesis can be a highly challenging task as ASL consists of manual (hands-only) and nonmanual (non-hands, such as mouthing cues, facial expressions, etc) cues. In this work, we focus on the manual cues which are expressed with four grammatical parameters: handshape, palm orientation, location, and movement (Tennant and Brown, 1998). Also, different signers may perform the same signs with different movement amplitudes and forces, and these variations should be considered in a signing avatar.

CLIP (Radford et al., 2021) pre-trains an image encoder and a text encoder to predict which images correspond to specific text pairs in the dataset. Instead of using one-hot encoding for the sign word, we leverage the CLIP text encoder as a conditional input. This approach offers two key advantages: (1) image-based generation, where the image embedding connects to the most semantically similar word, which can be particularly useful for children who cannot read yet but are learning sign language; and (2) generalization to unseen synonyms. Since many words with similar meanings (e.g., 'table' in the dataset and 'desk' not in the dataset) share the same sign motion, CLIP's semantically aligned text embeddings are well-suited to handle such synonyms.

[1] The code and data are available on the project page.

[2] ASL Gloss is the written transcription of the ASL signing motions. It is a writing tool used to connect the ordered components of a sign sequence to their semantic meanings.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 9993–9999
November 12-16, 2024 ©2024 Association for Computational Linguistics

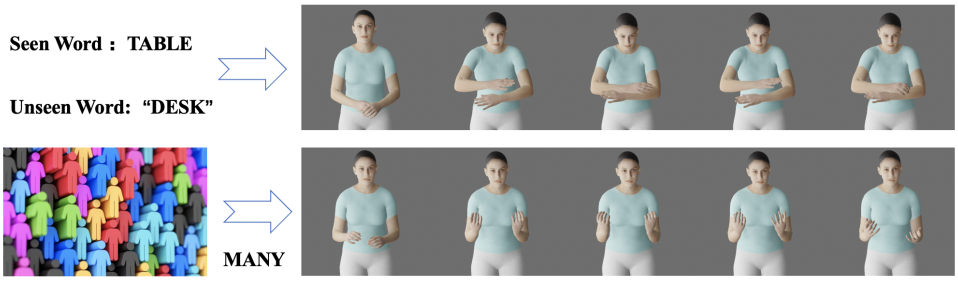

Figure 1: Top row: given a text word in dataset, wSignGen can generate a 3D signing avatar corresponding to that word. Also, wSignGen supports unseen but semantically close words. Bottom row: given a semantically meaningful image, the closest associated word is identified *via CLIP* and wSignGen generates the corresponding sign.

Our contributions are threefold: First, we introduce a novel approach for word-conditioned 3D ASL motion generation, synthesizing more accurate, realistic, and expressive sign language motions. Second, we present wSignGen, a model that combines CLIP with a diffusion-based approach using a classifier-free strategy, offering two advantages: image-based generation and the ability to generalize to unseen synonyms. Third, extensive experiments show that wSignGen consistently outperforms baseline models on standard generative metrics and receives highly positive feedback from human evaluators.

2 Related Work

Sign Language Synthesis Sign Language Synthesis (/Production) aims to convert text input into grammatically correct sign sequence motions. Researchers have explored various seq2seq models (Koller et al., 2015; Stoll et al., 2018; Saunders et al., 2020b, 2021a; Hwang et al., 2021), and multi-channel approaches (Saunders et al., 2020a, 2021a), to synthesize pixel-level videos (Stoll et al., 2018, 2020) or produce sign language skeletal sequences (Zelinka et al., 2019). Other works have relied solely on the regression of 2D/3D joint positions (Saunders et al., 2021b) with OpenPose (Cao et al., 2017) or post-regression SMPLX (Pavlakos et al., 2019) fitting to elevate 2D joints to 3D features (Stoll et al., 2022). Reconstructing 3D avatars (Dong et al., 2024a; Xu et al., 2024; Dong et al., 2024b) further advances recent studies. We go a step further by utilizing a diffusion model to directly synthesize 3D features, thereby preserving lifelike diversity and authenticity in our avatars.

Motion generative models Numerous studies on human motion generation have achieved significant success using the variational autoencoder (VAE) framework. ACTOR (Petrovich et al., 2021) involves a transformer-based VAE with numerical labels as the condition. Subsequent works in language-guided generation include (Zhang et al., 2017; Li et al., 2020b; Nam et al., 2018; Ramesh et al., 2021). The CLIP model (Radford et al., 2021) has been incorporated into various motion generation techniques, such as its combination with an Autoencoder (AE) in (Tevet et al., 2022a), a VAE in (Petrovich et al., 2022; Guo et al., 2022a; Zhai et al., 2023), and a VQ-VAE in TM2T (Guo et al., 2022b). Additionally, pre-trained language models have been employed to further enhance semantic complexity in motion generation (Jiang et al., 2023; Zhang et al., 2023).

Unlike these VAE models, diffusion models have been shown to generate samples that more closely follow the distribution of their training samples. For conditioned generation, researchers investigated classifier-guided diffusion (Dhariwal and Nichol, 2021; Nichol et al., 2021) and classifier-free guidance (Ho and Salimans, 2022; Shan et al., 2024) technologies in succession. The MDM architecture (Tevet et al., 2022b) utilized a classifier-free guidance diffusion architecture to predict human motion in a similar structure as employed by wSignGen.

3 Methodology

Problem formulation Our goal is to generate 3D SMPLX-based motion sequences that match the meanings of sign language words, based on either input words or images, as illustrated in the Figure 2. The generative process learns clean motion patterns through a diffusion model, trained with classifier-free guidance by randomly masking the condition c. At each step, a transformer encoder is employed to learn the spatial and temporal relationships from the sign motion sequence.

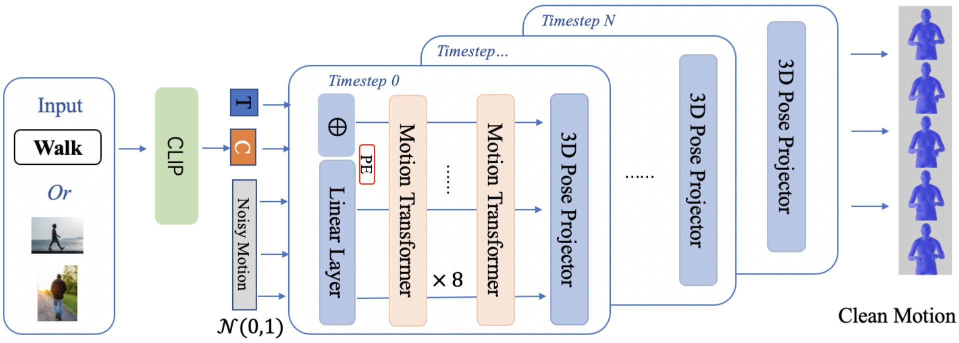

Figure 2: Overview of the wSignGen architecture.Given timestep T, Condition c, Nosy Motion sampled from $\mathcal{N}(0,1)$ distribution, the wSignGen generative process learn clean motion through a diffusion model. At each step, timestep T, Condition c is projected together into a token Z. This token, combined with the output from a linear layer, is passed through position encoding PE, and then sent into the transformer encoder blocks. In the final phase, the motion output passes through a 3D Pose Projector to obtain the SMPLX rotation features.

3.1 Conditioned Word or Image

The input modality can be either text or image, and we leverage CLIP to align their semantics. Using the CLIP text encoder, we obtain text embeddings E_{text}, so that for a given word w, the conditioning embedding is set as $c = E_{\text{text}}(w)$. Similarly, since CLIP supports zero-shot alignment between text and images, for a given image i, we extract its encoding $E_{\text{image}}(i)$ and align it to obtain the conditional embedding $c = E_{\text{text}}(w_i)$, where w_i represents the word associated with image i through the pre-trained CLIP model.

3.2 Diffusion Process

Our learning approach incorporates a diffusion model, to generate samples from isotropic Gaussian noise by iteratively removing the noise at each step. There are three major sub-processes: the forward process, the reverse process, and the sampling procedure. For the sign motion X_0, our final goal is to model a distribution $X_0 \sim q(X_0)$.

The forward diffusion process follows a Markov chain over T steps, evolving by gradually injecting noise into the samples until the distribution of X_T approximates $\mathcal{N}(0,1)$. Formally, this process is denoted as:

$$q(X_t|X_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}X_{t-1}, (1-\alpha_t)\mathrm{I}) \tag{1}$$

where $\alpha_t \in (0,1)$ are constant hyper-parameters. From here on we use X_t to denote the full sequence at noising step t. The reversed diffusion is a gradually denoised motion X_t based on the condition. In our context, it models the distribution $p(X_0|c)$, where c are conditional embeddings from CLIP. Following previous works in motion diffusion, instead of predicting noise, we predict the motion signal itself i.e. $\hat{X}_0 = G(X_t, t, c)$ with the objective shown in Equation 2:

$$\mathcal{L}_{base} = \mathrm{E}_{X_0 \sim q(X_0|c), t\in[1,T]}\left[\|X_0 - G_c(X_t,t)\|_2^2\right] \tag{2}$$

Training losses For regularization, following the method by Tevet et al. (2022b), we utilize a velocity loss $\mathcal{L}_{vel}$ which considers physical characteristics and mitigates the emergence of artifacts. The velocity loss and the combined losses are shown in Equations 3 and 4, respectively.

$$\mathcal{L}_{vel} = \frac{1}{N-1}\sum_{i=1}^{N-1}\left\|(x_0^{i+1} - x_0^i) - (\hat{x}_0^{i+1} - \hat{x}_0^i)\right\|_2^2 \tag{3}$$

$$\mathcal{L} = \mathcal{L}_{base} + \mathcal{L}_{vel} \tag{4}$$

Sampling Sampling here involves the iterative prediction of the clean sample $\hat{X}_0 = G_c(X_t,t)$ and noising it back to X_{t-1}, repeated T times back to X_0. We train our model G with classifier-free guidance. G learns both the conditioned and the unconditioned distributions by randomly setting $c = \emptyset$ (null condition) for 10% of the samples, such that $G_\emptyset(X_t,t)$ approximates $p(X_0)$.

$$G(X_t,t) = sG_c(X_t,t) - (s-1)G_\emptyset(X_t,t) \tag{5}$$

During sampling, we can control the trade-off between diversity and fidelity by adjusting the hyperparameter s.

4 Experiments

Dataset We curated a 3D SMPLX-based dataset from the sign recognition video dataset, WLASL(Li et al., 2020a). We sorted the dataset by sample size and then selected the top 103 words

Table 1: **Comparison of CVAE Baseline and our Diffusion Model** We compare a motion generation baseline algorithm with our proposed method using the curated datasets. Notation Keys: $\rightarrow$: implies that motions are better when the metric is closer to those computed for GT^{train} and GT^{test}; "Acc.": accuracy; "Div.": diversity; "Mul.": multimodality; "Gen": Generation. Gen^{train} and Gen^{test} are generated from the same model, and we report them separately to compare with the original training and testing data distribution on FID, Div., and Mul. metrics.

ASL3D$_S$	**Acc. $\uparrow$**	**FID $\downarrow$**	**Div.$\rightarrow$**	**Mul.$\rightarrow$**	**ASL3D**	**Acc.$\uparrow$**	**FID$\downarrow$**	**Div.$\rightarrow$**	**Mul.$\rightarrow$**
Original Data (no Generative Process)									
GT^{train}	1.0	-	30.001	9.921	GT^{train}	1.0	-	34.565	13.256
GT^{test}	0.897	-	26.252	11.180	GT^{test}	0.765	-	30.599	12.289
CVAE Baseline (ACTOR$^+$)									
Gen^{train}	0.884	75.243	24.566	8.250	Gen^{train}	0.515	126.830	25.732	16.500
Gen^{test}	-	65.285	24.187	6.600	Gen^{test}	-	100.147	25.393	12.289
wSignGen (Our Model)									
Gen^{train}	1.0	5.348	29.592	8.855	Gen^{train}	1.0	7.339	33.927	11.417
Gen^{test}	-	40.834	29.278	6.494	Gen^{test}	-	37.873	33.608	8.538

(with more than 18 samples per word) for this work. We further selected a subset of 30 words denoted as ASL3D$_S$ for our scalability evaluation; Next, we used the Hand4whole model (Moon et al., 2022) to extract SMPLX features, creating our SMPLX-based ASL3D Dataset.

Evaluation Metrics In accordance with previous motion generation approaches in the literature, (Guo et al., 2020; Petrovich et al., 2021; Tevet et al., 2022b), we evaluate wSignGen based recognition accuracy, Fréchet Inception Distance (FID), variation of motion across all words (diversity), and per-word motion variation (multimodality). To obtain the accuracy metrics, using our training data split, we trained a sign classifier over STGCN features (Yu et al., 2017). Training and test accuracies on the original data are shown in the first part of Table 1.

4.1 Word-Based Sign Synthesis

To demonstrate the efficacy of wSignGen, we trained a conditional VAE (CVAE), widely used in motion synthesis, as a baseline to compare with the performance of wSignGen. This VAE was a variation of ACTOR (Petrovich et al., 2021) trained our 3D sign motion data. To align with our objectives, we modified the condition vector to use CLIP embeddings rather than the previous one-hot action labels and trained the model with standard reconstruction and KL divergence losses. We denote this modified model as ACTOR$^+$. We quantitatively compared the generative capabilities of wSignGen to the CVAE baseline, as shown in Table 1. As can be seen, wSignGen far outperforms the CVAE in all four metrics. For accuracy, we report only one value as both the train and test samples were generated from the same model. Notably, wSignGen achieves an accuracy score of **1.00** both for ASL3D$_S$ and the larger ASL3D datasets, thus indicating its excellent scalability. The remaining metrics further confirm the high quality of the generated data.

4.2 Image-Based Sign Synthesis

Image-based sign synthesis involves generating the sign motion corresponding to the word-level semantic meaning of an input image. During the training phase, we assumed that the corresponding semantics (from the input image) was precisely identified. During the evaluation phase, we randomly selected five images from Google Image Search for each of the words in ASL3D$_S$ ($N = 150$). Note that the quality of the results here depends heavily on how well the images being presented are resolved to their correct semantics. After presenting the 150 images to wSignGen, the resulting accuracy was **86%**, and this included errors from CLIP occasionally mislabeling the image.

4.3 Human Evaluation

As recommended by Fox et al. (2023), we evaluated the acceptability of our generated avatar by having seven subjects (five hearing and two Deaf or Hard-of-Hearing (DHH) individuals) participate in a human evaluation test. Specifically, we began by presenting them with two original human-signing videos of a word. Next, avatar videos of the same word, generated by wSignGen were shown to them.

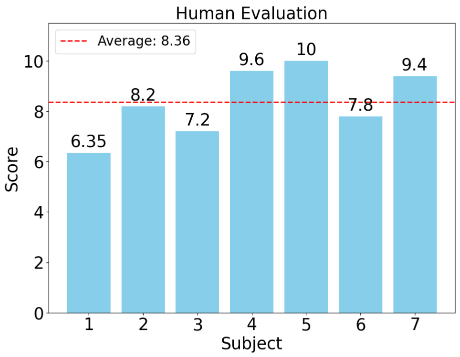

Figure 3: Human Evaluation Results

They were instructed to provide ratings on a scale of 1 to 10 on evaluation criteria which included acceptability and the correctness of their observations as related to hand shape, palm orientation, location, and movement. The average score of each tester is shown in Figure 3. The overall average scores from the experiment was **8.36 out of 10**, showing that the generated signs were received quite favorably.

5 Conclusion

We introduce the wSignGen model, representing a significant milestone in the realm of word-conditioned 3D sign language motion generation. By leveraging the CLIP model to align word semantics and a diffusion model to learn clean signing motions, wSignGen demonstrates notable advancements over existing methodologies. Extensive experiments showcase our work has excellent scalability and can generalize to unseen words.

6 Limitations

Currently, we are working with 1,547 videos (corresponding to 103 words) due to the limited number of signing videos that have a sufficient number of training examples; but this can be readily remedied by collecting more word-level data.

Also, in order to extend this work to the translation of continuous text phrases, it is important to first have a text-to-gloss translation model, although in its current form, wSignGen can still be successfully used to teach ASL word signs.

It can be challenging to estimate the 3D pose from some hand and finger motions, especially for signs that require fine granularity and this can lead to inaccurate word representation during training.

Lastly, ASL consists of both manual (related to hands) and nonmanual (facial expressions, mouth movements, body postures, etc) components, but in this work we explicitly model only the manual aspect of ASL, while the nonmanuals are generated only by side-effect from the training data. In the future, we will explicitly model and incorporate these nonmanuals at the word level.

7 Ethical Consideration

The dataset used for this work was curated from a publicly available dataset, WLASL, which has been made public since 2020. We assume that the participants in the original dataset collection gave their voluntary consent for participation. WLASL is licensed under the Computational Use of Data Agreement (C-UDA). The data is intended for academic and computational use only. This work strictly adheres to those licensing constraints.

In our work, confidentiality is maintained because we do not plan to share the faces or other identifying features of the participants in their original form; rather, we will only show the avatars that were learned from the movements of a large number of diverse signers.

Their is little potential for harm in this work although it must be acknowledged that words might be mistranslated by the model resulting in inaccurate communication between non-signing and Deaf and Hard-of-Hearing (DHH) individuals who rely on such a model (but this work is still in the early stages).

Acknowledgments

This work is based upon work supported under the National AI Research Institutes program by the National Science Foundation (NSF) and the Institute of Education Sciences (IES), U.S. Department of Education, through Award # 2229873 and NSF Award # 2223507. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF, the IES, or the U.S. Department of Education.

References

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Lu Dong, Lipisha Chaudhary, Fei Xu, Xiao Wang, Mason Lary, and Ifeoma Nwogu. 2024a. Signavatar: Sign language 3d motion reconstruction and generation. *arXiv preprint arXiv:2405.07974*.

Lu Dong, Xiao Wang, Srirangaraj Setlur, Venu Govindaraju, and Ifeoma Nwogu. 2024b. Ig3d: Integrating 3d face representations in facial expression inference. *arXiv preprint arXiv:2408.16907*.

Neil Fox, Bencie Woll, and Kearsy Cormier. 2023. Best practices for sign language technology research. *Universal Access in the Information Society*, pages 1–9.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022a. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161.

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. 2022b. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *ECCV*.

Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. 2020. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Euijun Hwang, Jung-Ho Kim, and Jong C Park. 2021. Non-autoregressive sign language production with gaussian space. In *British Machine Vision Conference*.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2023. Motiongpt: Human motion as a foreign language. *arXiv preprint arXiv:2306.14795*.

Oscar Koller, Jens Forster, and Hermann Ney. 2015. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020a. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020b. Unicoder-vl: A universal encoder for vision and language by cross-modal pretraining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344.

Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. 2022. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2308–2317.

Amit Moryossef, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgoz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Muller, and Sarah Ebling. 2021. Evaluating the immediate applicability of pose estimation for sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3434–3440.

Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-Bonet, Anne Göhring, Roman Grundkiewicz, et al. Findings of the second wmt shared task on sign language translation.

Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-Bonet, Anne Göhring, Roman Grundkiewicz, et al. 2023. Findings of the second wmt shared task on sign language translation (wmt-slt23). Association for Computational Linguistics.

Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. 2018. Text-adaptive generative adversarial networks: manipulating images with natural language. volume 31.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985.

Mathis Petrovich, Michael J Black, and Gül Varol. 2021. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995.

Mathis Petrovich, Michael J Black, and Gül Varol. 2022. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020a. Adversarial training for multi-channel sign language production. *arXiv preprint arXiv:2008.12405*.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020b. Progressive transformers for end-to-end sign language production. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 687–705. Springer.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021a. Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks. *International journal of computer vision*, 129(7):2113–2135.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021b. Mixed signals: Sign language production via a mixture of motion primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1919–1929.

Mengyi Shan, Lu Dong, Yutao Han, Yuan Yao, Tao Liu, Ifeoma Nwogu, Guo-Jun Qi, and Mitch Hill. 2024. Towards open domain text-driven synthesis of multi-person motions. *arXiv preprint arXiv:2405.18483*.

Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 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.

Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2sign: towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128(4):891–908.

Stephanie Stoll, Armin Mustafa, and Jean-Yves Guillemaut. 2022. There and back again: 3d sign language generation from text using back-translation. In *2022 International Conference on 3D Vision (3DV)*, pages 187–196. IEEE.

Richard A Tennant and Marianne Gluszak Brown. 1998. *The American sign language handshape dictionary*. Gallaudet University Press.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. 2022a. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2022b. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*.

Fei Xu, Lipisha Chaudhary, Lu Dong, Srirangaraj Setlur, Venu Govindaraju, and Ifeoma Nwogu. 2024. A comparative study of video-based human representations for american sign language alphabet generation. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE.

Jinhui Ye, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Hui Xiong. 2023. Cross-modality data augmentation for end-to-end sign language translation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13558–13571, Singapore. Association for Computational Linguistics.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. 2017. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*.

Jan Zelinka, Jakub Kanis, and Petr Salajka. 2019. Nn-based czech sign language synthesis. In *Speech and Computer: 21st International Conference, SPECOM 2019, Istanbul, Turkey, August 20–25, 2019, Proceedings 21*, pages 559–568. Springer.

Yuanhao Zhai, Mingzhen Huang, Tianyu Luan, Lu Dong, Ifeoma Nwogu, Siwei Lyu, David Doermann, and Junsong Yuan. 2023. Language-guided human motion synthesis with atomic actions. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5262–5271.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 2023. T2m-gpt: Generating human motion from textual descriptions with discrete representations. *arXiv preprint arXiv:2301.06052*.

Xuan Zhang and Kevin Duh. 2023. Handshape-aware sign language recognition: Extended datasets and exploration of handshape-inclusive methods. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.