SIGN-SALD: A SKELETON-AWARE LATENT DIFFUSION MODEL FOR TEXT-DRIVEN SIGN LANGUAGE PRODUCTION

Jiayu Shen[*], *Kalin Stefanov*[†], *Lay-Ki Soon*[*], *Vee Yee Chong*[‡], *and KokSheik Wong*[*]

[*]School of Information Technology, Monash University Malaysia, Malaysia
[†]Faculty of Information Technology, Monash University, Australia
[‡]Jeffrey Cheah School of Medicine & Health Sciences, Monash University, Malaysia
{jiayu.shen, kalin.stefanov, soon.layki, anthonyalexanderveeyee.chong, wong.koksheik}@monash.edu

ABSTRACT

Denoising diffusion models have shown great promise for text-driven sign language production. However, existing approaches often oversimplify the representations of skeletal joints, temporal frames, and textual inputs, limiting their ability to capture modality-specific information and cross-modal dependencies. To address this issue, we propose Sign-SALD, a skeleton-aware latent diffusion method that explicitly models interactions among temporal, spatial, and semantic modalities. Specifically, a skeleton-aware Variational Autoencoder constructs a latent space that decouples spatial and temporal structures while preserving skeleton-specific dependencies. Within this latent space, a diffusion model employs a sequential attention mechanism to progressively integrate these modalities, enabling coherent sign production. Experiments on the How2Sign benchmark demonstrate that Sign-SALD outperforms existing methods in both motion quality and semantic consistency.

Index Terms— Sign language production, Diffusion model, Variational autoencoder

1. INTRODUCTION

Text-driven sign language production (SLP) aims to generate semantically aligned sign pose sequences from textual descriptions. This task presents significant challenges due to the complexity of cross-modal semantic mapping and the intricate spatiotemporal structure of signs. Generative models have been increasingly used for SLP. Transformer-based approaches [1] emphasize temporal dynamics but under-utilize spatial correlations among joints, while GAN-based methods [2] enable photorealistic video synthesis, but overlook fine-grained semantics by compressing sequences into single embeddings. Diffusion models [3] improve stability and long-sequence generation, but discretization of continuous motion causes information loss and jitter. Despite these advances, existing approaches rely on oversimplified representations, which limits the modeling of spatial, temporal, and semantic structures, highlighting the need for structured representations that explicitly model spatial joint relationships, temporal joint dependencies, and textual semantics.

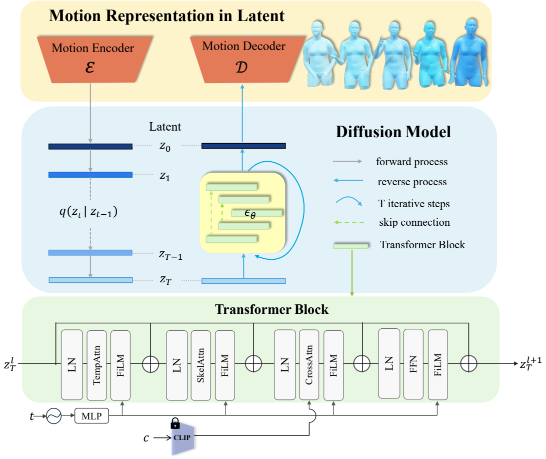

Fig. 1: Overview of the proposed Sign-SALD method. Skeleton-aware VAE (top) reconstructs sign pose sequences to learn skeletal-temporal representations and latent diffusion model (middle) learns text-conditioned denoiser (bottom) to generate continuous motions.

To address this gap, we introduce Sign-SALD (see Fig. 1), a skeleton-aware latent diffusion model that operates in a skeletal-temporally structured latent space to explicitly model interactions among skeletal joints, temporal frames, and textual inputs. We first employ a skeleton-aware variational autoencoder (VAE) [4] to construct a sign pose latent space, where the skeletal-temporal convolution layers decouple spatial and temporal dimensions while enabling information exchange between adjacent joints and frames. Skeletal-temporal pooling layers then aggregate these features into a compact representation, reducing dimensionality and computational cost during diffusion sampling. Within this latent space, a diffusion model with a sequential attention architecture progressively integrates temporal dynamics, skeletal topology, and textual semantics to structure high-quality sign language production. Our contributions are:

- We propose Sign-SALD, a skeleton-aware latent diffusion model for text-driven sign language production within a skeletal-temporally structured latent space.
- We introduce a sequential attention mechanism that leverages a structured latent space to progressively integrate temporal, spatial, and semantic information, capturing the intricate inter-relationships essential for coherent sign language production.
- We evaluate Sign-SALD on the How2Sign [5] benchmark and demonstrate competitive performance over existing methods in terms of sign naturalness, spatial accuracy, and semantic consistency.

2. METHOD

The goal is to generate sign pose sequences $\mathbf{x}_{1:N}$ from text prompts c, where N denotes the number of frames. To capture both spatial skeletal topology and temporal dynamics, we construct skeleton-aware latent representations (see Sec. 2.1). Within this latent space, a diffusion model with a sequential attention architecture progressively integrates temporal cues and textual semantics and generates coherent sign motions (see Sec. 2.2). The overall method is illustrated in Fig. 1.

2.1. Skeleton-Aware Representation Learning

To model skeleton-aware pose features that explicitly capture joint-level dynamics, we introduce a VAE with skeletal-temporal convolution and pooling operations [6], as illustrated in Fig.2. In contrast to conventional spatial-temporal approaches [7, 8] that rely on oversimplified representations of skeletal structures, our skeletal-temporal method specifically models the structured relationships between anatomical joints as the primary spatial component, moving beyond oversimplified single-vector pose representations.

We employ skeletal-temporal convolution (STConv), which decouples joint and temporal dimensions while enabling structured information exchange between adjacent joints and frames. This decomposition facilitates fine-grained modeling of joint-level dynamics by explicitly respecting skeletal topology, but it also introduces structural complexity through a dedicated joint axis. Compared to vector-based pose encoding, the resulting skeleton-aware representation expands the latent space dimensionality. This decomposition facilitates fine-grained modeling while maintaining computational efficiency through our STPool design.

To address this dimensionality challenge while preserving skeletal structure, we introduce skeletal-temporal pooling (STPool) layers in the encoder to compress the representation into a compact skeleton-aware latent space. Correspondingly, skeletal-temporal unpooling (STUnpool) layers in the decoder restore the compressed features back to the full skeletal-temporal resolution.

Encoder. The encoder processes sign pose sequences $\mathbf{x}_{1:N}$ by first decomposing each pose into joint-wise features through joint-specific MLPs, resulting in $\mathbf{h} \in \mathbb{R}^{N \times J \times D}$, where J denotes the number of joints and D the feature dimensions. Skeletal and temporal dependencies are then integrated via STConv layers:

$$\tilde{\mathbf{h}} = F_{ST}(\mathbf{h}) = F_{skel}(\mathbf{h}) \oplus F_{temp}(\mathbf{h}), \tag{1}$$

where $F_{skel}(\cdot)$ is a graph convolutional network (GCN) [7] operating in the joint dimension and capturing skeletal topology and inter-joint relationships, $F_{temp}(\cdot)$ is a 1D temporal convolutional network [9] modeling frame-wise dynamics, and $\oplus$ denotes element-wise addition. This factorization enables independent but coordinated processing of the skeletal structure and temporal evolution.

To mitigate dimensionality issues, we use STPool layers:

$$\hat{\mathbf{h}} = P_{ST}(\tilde{\mathbf{h}}) = P_{temp}(P_{skel}(\tilde{\mathbf{h}})), \tag{2}$$

where $P_{skel}(\cdot)$ aggregates the features of spatially adjacent joints while preserving the underlying skeletal topology, and $P_{temp}(\cdot)$ performs temporal downsampling. These operations are commutative because of their orthogonal dimensions. The skeletal pooling reduces the joint space to a set of atomic joints that maintain representational sufficiency.

This process yields a compact latent representation:

$$\mathbf{z} \in \mathbb{R}^{N' \times J' \times D}, \quad \text{where } N' < N \text{ and } J' < J. \tag{3}$$

We retain $J' = 7$ atomic joints (root, spine, head, shoulders, and hands) as primary kinetic chain anchors for upper-body articulation: intermediate joints (e.g., elbows, wrists) are biomechanically constrained by these anchors and reconstructable via STUnpool, as validated empirically in Sec. 3.

Decoder. The decoder reconstructs sign pose sequences from the compact latent representation $\mathbf{z}$ through a symmetric architecture. STUnpool layers restore both skeletal and temporal resolution, while STConv layers restore skeletal-temporal dependencies. Joint-wise MLPs then reconstruct the final pose sequence $\hat{\mathbf{x}}_{1:N}$.

Training. The VAE is trained using a composite objective:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{m}} + \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}}, \tag{4}$$

where $\mathcal{L}_{\text{m}}$, $\mathcal{L}_{\text{pos}}$, and $\mathcal{L}_{\text{vel}}$ are L1 reconstruction losses for sign pose sequences, joint positions, and velocities, respectively, while $\mathcal{L}_{\text{KL}}$ is the KL divergence regularization term that promotes structured latent representations.

2.2. Skeleton-Aware Denoising

Architecture. Given the skeleton-aware pose latent vector $\mathbf{z}_0 \in \mathbb{R}^{N' \times J' \times D}$ from the encoder, we adopt a transformer-based diffusion denoiser that sequentially applies temporal attention (TempAttn), skeletal attention (SkelAttn), and

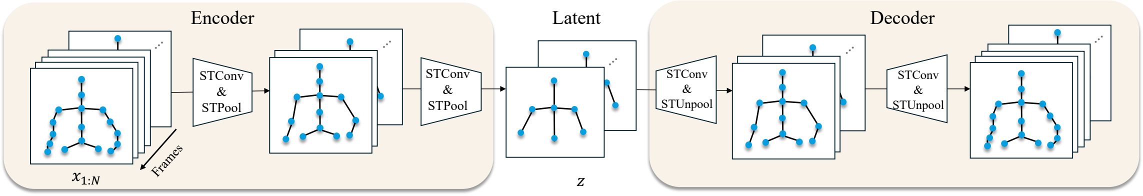

Fig. 2: Architecture of the skeletal-temporal VAE model. The encoder maps sign pose sequences into a skeletal-temporal latent space, and the decoder restores the skeletal-temporal latent representations into sign pose sequences.

cross-attention (CrossAttn). This systematic decomposition explicitly models structured motion patterns by encoding domain knowledge of temporal coherence, skeletal constraints, and semantic correspondence into the architectural design. Within each transformer layer, the latent representation $\mathbf{z}_t^{(l)}$ at timestep t is refined through a sequential composition of temporal attention, skeletal attention, cross-attention, and a feedforward network:

$$\mathbf{z}_t^{(l,k+1)} = \mathbf{z}_t^{(l,k)} + \mathrm{FiLM}\left(A_k(\mathrm{LN}(\mathbf{z}_t^{(l,k)})), t\right), \tag{5}$$

where $k \in \{1, 2, 3, 4\}$, A_1: TempAttn, A_2: SkelAttn, $A_3(\cdot)$: CrossAttn($\cdot$, CLIP(c)), and A_4: FFN.

The temporal attention module first captures dependencies across frames, effectively modeling motion dynamics and ensuring temporal coherence. Building on this temporal foundation, the skeletal attention module encodes spatial relationships among joints within each frame, thereby enforcing anatomical consistency and preserving skeletal topology. Subsequently, a cross-attention mechanism facilitates fine-grained semantic interaction between the textual features extracted from a frozen CLIP text encoder [10] and skeletal-temporal structures. Finally, a feedforward network applies nonlinear transformations to refine and integrate the aggregated information from the preceding attention modules.

To maintain stability across diffusion steps, each module is equipped with residual connections, layer normalization [11], and FiLM modulation [12], where:

$$\mathrm{FiLM}(\mathbf{h}, t) = \gamma_t \odot \mathbf{h} + \beta_t, \quad \gamma_t, \beta_t \in \mathbb{R}^D, \tag{6}$$

and the learned timestep embeddings are broadcast over temporal (N') and joint (J') dimensions. In addition, to preserve fine-grained skeletal-temporal information throughout the denoiser, we incorporate long skip connections [13] linking non-adjacent layers ($l \to l+\tau$ with $\tau > 1$) via learnable linear projections, which facilitate stable gradient flow and long-range information propagation. This formulation embodies the sequential orchestration of temporal, skeletal, and cross-modal processing, enabling explicit modeling of structured motion patterns through principled architectural design.

Training and Inference. The denoiser is trained within the diffusion model. The forward process is defined as:

$$q(\mathbf{z}_t \mid \mathbf{z}_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}\, \mathbf{z}_{t-1}, (1-\alpha_t)\mathbf{I}), \tag{7}$$

where $\mathbf{z}_t$ denotes the noisy latent at timestep t, which is subsequently processed by the denoiser through layers $\mathbf{z}_t^{(1)}, \ldots, \mathbf{z}_t^{(L)}$. The model is optimized to predict the injected noise with the objective:

$$\mathcal{L} = \mathbb{E}_{\substack{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}),\\ t\sim\mathrm{Uniform}(1,T),\, c}} \left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, c)\right\|_2^2, \tag{8}$$

where $\boldsymbol{\epsilon} \in \mathbb{R}^{N' \times J' \times D}$ is the ground truth noise and $\boldsymbol{\epsilon}_\theta$ is the model prediction [14]. During inference, DDIM sampling [15] produces a clean latent $\hat{\mathbf{z}}_0$, which is then decoded by the skeleton-aware decoder into a sign pose sequence.

3. EXPERIMENTS

We evaluate Sign-SALD on the SMPL-X [16] poses of the How2Sign dataset [5], a benchmark released in Neural Sign Actors and subsequently reused in later works, ensuring reliability and comparability. The dataset contains ~35k sequences, spanning 79 hours of ASL videos with a 16k-word vocabulary. Sign-SALD is implemented in PyTorch. The VAE is trained for 50 epochs with a batch size of 128, while the diffusion denoiser ϵ_θ is trained for 500 epochs with a batch size of 64. Both models use the Adam optimizer [17] with a learning rate of 1×10^{-3}. The diffusion process uses a linear noise schedule with 1000 steps during training, and inference relies on DDIM sampling with 50 steps for a balance between efficiency and quality. All experiments are conducted on a single NVIDIA A100 GPU.

Following [5], we adopt a back-translation evaluation, where generated pose sequences are translated back into text and compared with references using BLEU-4 [18] and ROUGE [19] to assess semantic consistency. For motion-level evaluation, we report the Mean Per Joint Position Error (MPJPE) and the Fréchet Inception Distance (FID) [20] between generated and ground truth sequences. MPJPE measures the average Euclidean distance between corresponding joints, commonly used in human motion prediction tasks [21, 22], while FID evaluates the distributional similarity of generated and real pose sequences.

Table 1 records the results for Sign-SALD and five baselines [1, 23, 2, 24, 5], using official implementations when available and otherwise following prior evaluations for fair comparison on How2Sign. Sign-SALD achieves the best

Table 1: Results on the How2Sign dataset.

Methods	BLEU-4 ↑	ROUGE ↑	FID ↓	MPJPE ↓
PT [1]	2.75	29.87	4.71	70.06
NSLP-G [23]	5.75	31.98	4.45	63.25
Adv.Trai [2]	6.21	32.33	3.98	65.25
MotionGPT [24]	9.38	35.16	2.71	42.53
NSA [5]	13.12	47.55	1.56	35.87
Sign-SALD (Ours)	**13.79**	**48.06**	**1.32**	**31.25**

Table 2: Comparison of VAE architectures with parameter counts.

Methods	FID ↓	MPJPE ↓	Params (M) ↓
MotionGPT [24]	0.75	0.335	11.29
Parco [25]	0.63	0.316	5.37
Sign-SALD (Ours)	**0.58**	**0.209**	**2.81**

Table 3: Ablation results on the VAE and denoiser.

Methods	BLEU-4 ↑	FID ↓
w/o ST-Latent	13.07	2.76
w/o CrossAttn	12.50	3.18
w/o TempAttn	13.61	3.29
w/o SkelAttn	13.72	2.45
J’=5,7,9	13.11 / 13.79 / 12.43	2.45 / 1.32 / 1.29
w/o ST-Latent+CrossAttn	12.08	3.65
Full model	**13.79**	**1.32**

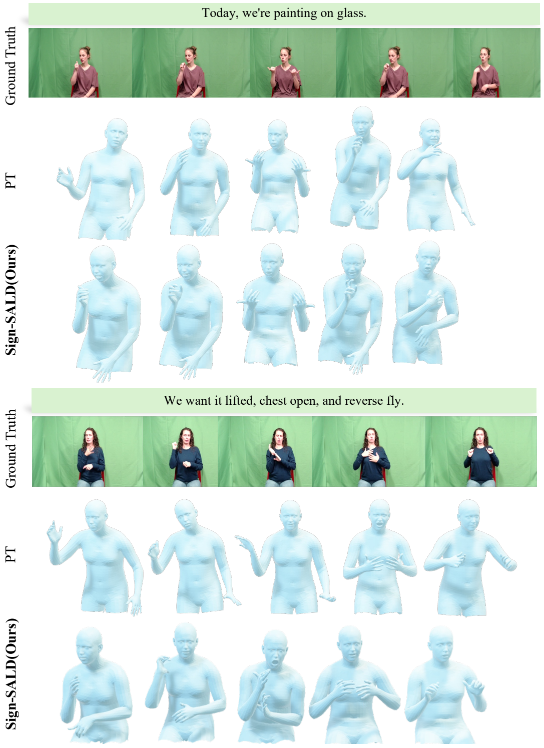
Today, we're painting on glass.
Ground Truth
PT
Sign-SALD(Ours)
We want it lifted, chest open, and reverse fly.
Ground Truth
PT
Sign-SALD(Ours)

Fig. 3: Qualitative results of PT and Sign-SALD models.

scores for BLEU-4 (13.79) and ROUGE (48.06). Compared to the best-performing baseline NSA [5], Sign-SALD achieves gains on motion quality metrics, reducing FID from 1.56 to 1.32 and MPJPE from 35.87 to 31.25. These consistent improvements across semantic and motion metrics suggest that explicitly modeling skeletal-temporal-textual interactions produces higher quality sign language poses while maintaining greater semantic consistency with the textual inputs. In addition to the quantitative results, Fig. 3 illustrates that the sign poses generated by Sign-SALD are more plausible and dynamic than those of PT [1], particularly in hand details and limb movements, highlighting its advantage in capturing both joint positions and bone orientations.

To contextualize the efficiency of our latent representation, Table 2 compares reconstruction quality against two representative VAE architectures: the VQ-VAE in MotionGPT [24] and the part-aware VQ-VAE in Parco [25]. We note that this comparison involves both architectural differences (skeleton-aware vs. holistic/part-based) and representation type (continuous vs. discrete); the contribution of the skeleton-aware design is isolated in Table 3 (“w/o ST-Latent”), which compares against a standard continuous VAE and shows substantial degradation (FID: 1.32→2.76). Our approach achieves strong reconstruction with only 2.81M parameters, indicating that the skeleton-aware design provides an efficient foundation for diffusion-based generation.

We further test i) the skeletal-temporal latent (ST-Latent) by replacing it with a standard VAE, ii) the cross-attention mechanism by substituting it with self-attention over concatenated sentence-level CLIP and sign pose features, iii) temporal and skeletal attention modules separately (w/o TempAttn, w/o SkelAttn), and iv) different numbers of atomic joints $J' = 5, 7, 9$. As shown in Table 3, removing ST-Latent increases FID from 1.32 to 2.76; replacing cross-attention reduces BLEU-4 by 9.35%. Combined removal causes most severe degradation. Temporal attention contributes more (3.29 vs 2.45 FID increase), with $J' = 7$ providing the best efficiency–performance balance, as $J' = 5$ loses semantic detail and $J' = 9$ adds redundancy. These ablations collectively demonstrate the necessity of each component for consistent sign sequence generation.

4. CONCLUSION

We presented Sign-SALD, a diffusion-based framework for text-driven sign language production that explicitly models interactions among skeletal joints, temporal dynamics, and textual inputs. The method integrates a skeletal-temporal VAE, which decouples spatial and temporal dimensions to learn structured motion representations, with a sequential attention diffusion model that ensures semantic coherence and motion consistency in the latent space. Experimental results demonstrate that Sign-SALD consistently outperforms strong baselines across multiple evaluation metrics. Future work will investigate long-sequence generation by modeling longer-range dependencies and broaden evaluation as more standardized SMPL-X datasets become available.

5. REFERENCES

[1] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden, "Progressive transformers for end-to-end sign language production," in *European Conference on Computer Vision*. Springer, 2020, pp. 687–705.

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

[3] Sen Fang, Chunyu Sui, Yanghao Zhou, Xuedong Zhang, Hongbin Zhong, Yapeng Tian, and Chen Chen, "Signdiff: Diffusion model for american sign language production," in *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2025, pp. 1–11.

[4] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[5] Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou, "Neural sign actors: A diffusion model for 3d sign language production from text," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1985–1995.

[6] Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen, "Skeleton-aware networks for deep motion retargeting," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 62–1, 2020.

[7] Sijie Yan, Yuanjun Xiong, and Dahua Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, 2018, vol. 32.

[8] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 183–192.

[9] Shaojie Bai, J Zico Kolter, and Vladlen Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

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

[11] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[12] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, 2018, vol. 32.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[15] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[16] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black, "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*, 2019, pp. 10975–10985.

[17] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[19] Chin-Yew Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[21] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan, "Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13232–13242.

[22] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang, "Deep kinematics analysis for monocular 3d human pose estimation," in *Proceedings of the IEEE/CVF Conference on computer vision and Pattern recognition*, 2020, pp. 899–908.

[23] Eui Jun Hwang, Jung-Ho Kim, and Jong C Park, "Non-autoregressive sign language production with gaussian space.," in *BMVC*, 2021, p. 197.

[24] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen, "Motiongpt: Human motion as a foreign language," *Advances in Neural Information Processing Systems*, vol. 36, pp. 20067–20079, 2023.

[25] Qiran Zou, Shangyuan Yuan, Shian Du, Yu Wang, Chang Liu, Yi Xu, Jie Chen, and Xiangyang Ji, "Parco: Part-coordinating text-to-motion synthesis," in *European Conference on Computer Vision*. Springer, 2024, pp. 126–143.