SynSign: Data Synthesis for Sign Language Translation with Controllable Semantic and Appearance Variation

Xiaoyun Zheng[1,2,3], Yuegan Yi[1,2,3], Ruiquan Zhang[1,2,3], Yidong Chen[1,2,3]*
[1]School of Informatics, Xiamen University, Xiamen, China.
[2]Key Lab of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian–Taiwan (XMU), Xiamen, China.
[3]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, China.
xyzheng@stu.xmu.edu.cn ydchen@xmu.edu.cn

Abstract

Sign Language Translation (SLT) is constrained by the high cost and limited diversity of real-world training data. Although existing efforts attempt to mitigate this through the construction of new datasets or data augmentation, physical recording constraints and shallow feature transformations still hinder substantial expansion in both semantic coverage and visual diversity. To address these limitations, we propose SynSign, a data synthesis framework that combines Large Language Model (LLM)-based gloss generation with 3D motion reconstruction and controllable rendering. SynSign combines Retrieval-Augmented Generation (RAG) with vocabulary-guided generation to produce valid gloss sequences, which are then transformed into continuous avatar signing through motion retrieval, blending, retargeting, and diverse visual rendering. On CSL-Daily, pretraining with 10,129 SynSign videos achieves performance comparable to WLASL pretraining while using approximately 48% as many video sequences. Combining the two data sources further yields the best overall performance. Moreover, across two signer-independent splits, SynSign pretraining consistently outperforms WLASL pretraining, suggesting its potential to improve cross-signer transfer.

1 Introduction

Sign Language Translation (SLT) aims to translate continuous visual signals into spoken-language text. This task remains challenging due to the complex correspondence between spatiotemporal motions and discrete linguistic representations. Existing studies have primarily addressed this challenge through advances in model architectures and learning strategies (Lai et al., 2026; Fu et al., 2025; Rao et al., 2026; Chen et al., 2025). Despite these advances, from a data perspective, further development remains constrained by the limited scale and diversity of available training data.

Existing sign language datasets are costly to collect and annotate, and often provide limited semantic and visual variation. Controlled recording environments restrict variation in signer appearance, clothing, and background (Shen et al., 2024; Zhang et al., 2024). Meanwhile, limited corpus coverage constrains the range of linguistic expressions represented in the data. Previous studies have expanded data through television broadcasts, 3D avatars, or visual and textual augmentation (Shen et al., 2023; Zuo et al., 2024; Ye et al., 2023a,b). However, these approaches either remain dependent on existing datasets or provide limited control over semantic content and visual appearance simultaneously.

To address these limitations, we propose SynSign, a sign language data synthesis framework that integrates a Large Language Model (LLM) with 3D motion reconstruction and controllable rendering. SynSign expands training data along two complementary dimensions: semantic content and visual appearance. First, its semantic expansion module retrieves candidate glosses from a predefined vocabulary and uses them to guide LLM-based gloss sequence generation. The corresponding isolated-sign videos are then reconstructed as 3D motions and temporally blended to form continuous sentence-level sequences. Second, its appearance control module retargets the synthesized motions to different virtual avatars and renders them with varied backgrounds. We evaluate SynSign on CSL-Daily (Zhou et al., 2021). Despite containing fewer video sequences, SynSign pretraining achieves translation performance comparable to WLASL pretraining. Moreover, under two signer-independent splits, SynSign pretraining consistently yields numerical improvements over WLASL pretraining, suggesting its potential for

*Corresponding Author.

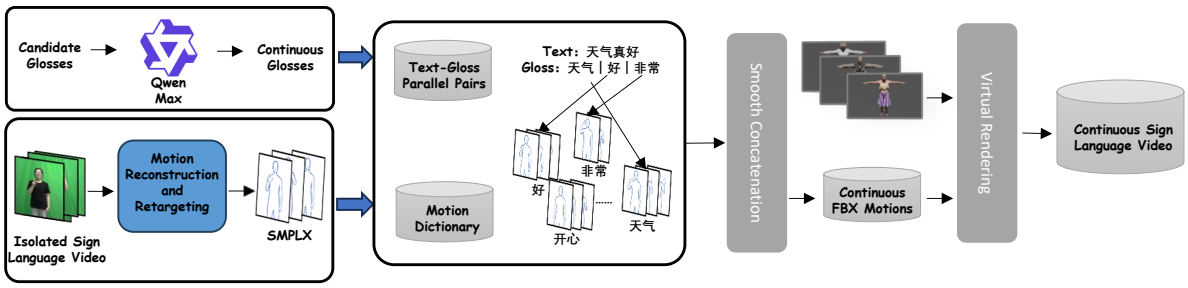


Figure 1: Overview of SynSign. Given an input sentence, RAG retrieves semantically related candidates from the gloss vocabulary to guide LLM-based gloss sequence generation. The corresponding isolated-sign videos are reconstructed as 3D motions, temporally blended into sentence-level sequences, retargeted to virtual avatars, and rendered with varied avatar identities and backgrounds.

transfer to unseen signers.

In summary, our contributions are:

- We propose SynSign, a sign language data synthesis framework that combines semantic expansion with controllable 3D rendering to generate sentence-level sign language videos with varied semantic content and visual appearance.
- We develop a synthesis pipeline that combines Retrieval-Augmented Generation (RAG) (Lewis et al., 2020)-based candidate retrieval and vocabulary-guided gloss generation with 3D motion reconstruction, temporal blending, avatar retargeting, and controllable rendering.
- Experiments on CSL-Daily show that SynSign provides a competitive source of sign-language pretraining data. It achieves performance comparable to WLASL pretraining with fewer video sequences, while producing consistent numerical improvements under two signer-independent splits.

2 Related Work

Sign language research relies primarily on 2D video datasets for isolated or continuous recognition and translation (Zhou et al., 2021; Jin et al., 2025; Li et al., 2020; Camgoz et al., 2018). Although recent datasets introduce more signers, backgrounds, or 3D avatars (Shen et al., 2023; Zuo et al., 2024), scaling real-world data remains costly because of recording and annotation requirements. Synthetic data therefore provides an alternative for increasing visual diversity without additional physical recording. Existing augmentation methods mainly modify visual or textual data. Visual approaches use cropping, temporal resampling, color transformation, or motion-level manipulation (Chen et al., 2022a; Zhou et al., 2023; Wei and Chen, 2023), while textual approaches employ LLM rewriting or back-translation (Shen et al., 2024; Ye et al., 2023a; Guo et al., 2024). Cross-modal methods further mix representations or synthesize videos by concatenating retrieved sign segments (Ye et al., 2023b; Zhou et al., 2021). These approaches can enlarge the training set but remain largely tied to existing samples and provide limited joint control over semantic content and appearance. SynSign instead combines semantic generation with controllable 3D rendering.

3 Method

As shown in Figure 1, SynSign comprises two components: a semantic expansion module that generates gloss sequences and constructs the corresponding sentence-level 3D sign motions, and an appearance control module that retargets these motions to different virtual avatars and renders them in varied environments.

3.1 Semantic Expansion Module

3.1.1 Gloss Sequence Generation

To construct the synthetic corpus, we first use an LLM to generate Chinese source sentences from the semantic domains summarized in Table 1. Given a gloss dictionary $\mathcal{G} = \{g_1, \ldots, g_M\}$, each generated sentence $X = (x_1, \ldots, x_L)$ is converted into a gloss sequence $\hat{Y} = (\hat{y}_1, \ldots, \hat{y}_K)$, where $\hat{y}_j \in \mathcal{G}$. To reduce out-of-vocabulary (OOV) hallucinations during free-form LLM generation, we use an automated pipeline that combines RAG-based candidate gloss retrieval, vocabulary-guided generation, and post-generation validation, as shown in

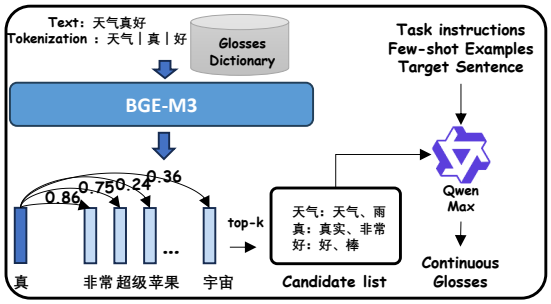


Figure 2: Gloss sequence generation via RAG-based semantic retrieval and vocabulary-guided generation.

Figure 2.

RAG-based semantic candidate retrieval. BGE-M3 (Chen et al., 2024) encodes each input token and gloss into $\boldsymbol{e}_{x_i}$ and $\boldsymbol{e}_{g_j}$. For each token, the top-k gloss candidates are

$$C_i = \underset{g_j \in \mathcal{G}}{\arg\,\text{top-k}} \cos(\boldsymbol{e}_{x_i}, \boldsymbol{e}_{g_j}), \tag{1}$$

and the sentence-level candidate set is $C_X = \bigcup_{i=1}^{L} C_i$, with $|C_X| \leq Lk$.

Vocabulary-guided generation. The candidate set C_X is included in a structured few-shot prompt for Qwen (Bai et al., 2023), together with task instructions, text-to-gloss examples, and the target sentence. Qwen is instructed to select glosses only from C_X. Generated glosses are subsequently validated against $\mathcal{G}$, and any OOV output is replaced by the closest valid gloss according to embedding similarity.

3.1.2 3D Motion Construction

For each gloss, we retrieve the corresponding isolated-sign video from National-CSL (Jin et al., 2025) and apply SMPLer-X (Cai et al., 2023) to recover frame-wise SMPL-X (Pavlakos et al., 2019) parameters $(\beta, \theta_t, \psi_t)$. Here, $\beta \in \mathbb{R}^{10}$ represents body shape, $\theta_t \in \mathbb{R}^{156}$ represents 3D pose rotations, and $\psi_t \in \mathbb{R}^{10}$ represents facial expression coefficients. These parameters provide a unified 3D representation of body and fine-grained hand motions required for subsequent motion composition and retargeting.

To construct sentence-level motion, isolated-sign SMPL-X clips are concatenated according to the generated gloss order. Direct concatenation often produces discontinuities at word boundaries. Therefore, we introduce a fixed-length overlap blending strategy.

Statistic	Value
Synthesized videos	10,129
Average glosses per sequence	7.09
Average gloss characters per sequence	14.63
Average text characters per sentence	21.25
Average video length	371 frames
Minimum video length	101 frames
Maximum video length	769 frames
Example semantic domains	Daily life, weather, transportation, healthcare, and family
Virtual avatar identities	1,000
Background settings	Indoor and outdoor scenes

Table 1: Statistics of the SynSign synthetic dataset.

For two adjacent clips A and B, we use a transition window of $W = 20$ frames, consisting of the last W frames of A and the first W frames of B. For $i = 0, \ldots, W-1$, the normalized progress and smoothstep weight are:

$$t_i = \frac{i+1}{W+1}, \qquad w_i = 3t_i^2 - 2t_i^3. \tag{2}$$

Because SMPL-X contains both rotational and non-rotational parameters, we apply interpolation according to their geometric properties. Let $q_{i,j}^A$ and $q_{i,j}^B$ denote the rotation quaternions of joint j in the paired transition frames from clips A and B, and let y_i^A and y_i^B denote the corresponding non-rotational parameters. Rotational parameters are interpolated with spherical linear interpolation (Slerp), while non-rotational parameters use linear interpolation (Lerp):

$$\begin{aligned} q_{i,j}^{\text{blend}} &= \text{Slerp}(q_{i,j}^A, q_{i,j}^B; w_i), \\ y_i^{\text{blend}} &= (1 - w_i) y_i^A + w_i y_i^B. \end{aligned} \tag{3}$$

The blended transition is then concatenated with the non-overlapping regions of the two clips. This training-free operation directly manipulates SMPL-X parameters to smooth transitions between adjacent sign clips.

3.2 Appearance Control Module

Sentence-level SMPL-X motions are retargeted to FBX avatars with different body proportions using a hierarchical skeleton mapping inspired by SynBody (Yang et al., 2023). SMPL-X rotations are converted to quaternions and mapped to corresponding FBX joints, while the torso, arms, and fingers are adjusted hierarchically according to the



Figure 3: Comparison of real videos from National-CSL and synthetic samples from SynSign.

target skeleton. Global joint positions are reconstructed through forward kinematics. This retargeting process separates motion content from avatar appearance, enabling a sentence-level SMPL-X motion to be instantiated across virtual characters with different body proportions. The shared source motion preserves the temporal organization of the sign sequence while allowing visual appearance to vary across rendered samples. The retargeted motions are then rendered in Blender using 1,000 virtual avatar assets sourced from SynBody and randomly sampled environment maps. Videos are rendered with a fixed frontal camera using Cycles at 256×256 resolution and 25 frames per second (fps). The resulting corpus statistics are summarized in Table 1. Examples of the rendered SynSign videos are shown in Figure 3.

4 Experiments

4.1 Experimental Settings

Datasets and Evaluation Metrics. We use three public sign language datasets. CSL-Daily (Zhou et al., 2021), a continuous Chinese Sign Language dataset containing 20,654 videos from 10 signers, is used for downstream SLT evaluation. National-CSL (Jin et al., 2025), an isolated Chinese Sign Language dataset containing approximately 67,000 videos covering 6,707 high-frequency glosses, provides the source videos for constructing 3D motion assets. WLASL (Li et al., 2020), an isolated American Sign Language dataset containing 21,083 videos from 119 signers, serves as the real-data baseline for sign-language-domain pretraining of the S3D (Xie et al., 2018) backbone. We report BLEU-1 to BLEU-4 (Papineni et al., 2002) and ROUGE-L (Lin, 2004) as evaluation metrics.

Experimental Design. The HyperSign (Zhang et al., 2025) setup uses a two-stream architecture consisting of an RGB stream and a keypoint stream, with keypoints pre-extracted using HRNet (Wang et al., 2020). Both streams are retained in all experiments unless otherwise specified. Following MMTLB (Chen et al., 2022a), we compare four S3D pretraining settings: (1) Kinetics-400 initialization only; (2) WLASL isolated-sign classification pretraining; (3) SynSign sequence-level pretraining with Connectionist Temporal Classification (CTC) loss; and (4) joint pretraining on WLASL and SynSign, where mini-batches from the two datasets are alternated during training. The shared S3D backbone is optimized with the classification loss for WLASL batches and the CTC loss for SynSign batches. All four settings are subsequently fine-tuned and evaluated on CSL-Daily.

Implementation Details. Across all settings, the S3D backbone is initialized with Kinetics-400 weights and pretrained for 30 epochs using a learning rate of 10^{-4} with cosine annealing and a batch size of 32. We set $k = 3$ for candidate retrieval,

Method	Pretraining Data	B1	B2	B3	B4	R
Baseline: Without In-Domain Pretraining						
HyperSign (Zhang et al., 2025)	—	58.25	45.31	36.09	29.07	57.31
Real Data Pretraining						
MMTLB (Chen et al., 2022a)	WLASL	53.31	40.41	30.87	23.92	53.25
TS-SLT (Chen et al., 2022b)	WLASL	55.44	42.49	32.87	25.79	55.72
CV-SLT (Zhao et al., 2024)	WLASL	58.29	45.15	35.77	28.94	57.06
HyperSign	WLASL	58.96	45.93	36.53	29.55	57.89
Synthetic and Joint Pretraining						
HyperSign	SynSign	58.77	45.81	36.44	29.49	57.78
HyperSign	**WLASL+SynSign**	**59.02**	**45.96**	**36.55**	**29.55**	**57.92**

Table 2: Impact of sign language pretraining strategies on CSL-Daily test set translation performance.

Component	Variant	B4	Δ
Full model	—	**29.49**	—
Gloss gen.	w/o RAG	29.27	-0.22
Motion concat.	w/o smoothing	29.14	-0.35
	w/o Lerp	29.34	-0.15
Identity div.	—	29.19	-0.30
Background div.	—	29.41	-0.08

Table 3: Ablation study of key components in SynSign.

yielding approximately 22 candidate glosses per sentence. All experiments run on a single RTX 3090 GPU.

4.2 Effect of Pretraining on SLT Performance

Table 2 compares different pretraining strategies. Relative to Kinetics-400 initialization alone, WLASL pretraining improves all translation metrics, indicating that sign-language-specific visual pretraining provides a more suitable initialization for downstream SLT.

Using 10,129 synthetic continuous Chinese Sign Language videos, SynSign pretraining achieves results close to WLASL across all translation metrics while using approximately 48% as many video sequences. The comparable downstream performance suggests that the synthesized motion and appearance variations provide task-relevant signals that can support SLT backbone pretraining.

Combining WLASL and SynSign yields the strongest overall performance. The two data sources provide different training signals: WLASL contributes real-world visual and motion characteristics, whereas SynSign provides Chinese Sign Language sequences together with controllable appearance diversity. Their combination therefore suggests complementary value for pretraining.

4.3 Ablation Study

Table 3 evaluates the contributions of the main components of SynSign. For semantic expansion, removing RAG-based candidate retrieval leads to lower translation performance, highlighting the importance of semantic guidance before LLM generation. By narrowing the candidate space to glosses related to the input sentence, retrieval supports more effective gloss sequence construction. For motion construction, removing transition smoothing leads to lower translation performance, highlighting the importance of temporal continuity when composing continuous motions from isolated-sign clips. Removing the Lerp component also degrades performance, indicating that interpolation of non-rotational parameters contributes to sentence-level motion construction. Reducing the 1,000 virtual identities to a single fixed avatar decreases pretraining performance, suggesting a potential benefit from avatar variation. Similarly, replacing diverse environments with a single solid-color background produces a smaller performance decrease, indicating that background variation also contributes to pretraining performance.

4.4 Signer-Independent Evaluation

Because the standard CSL-Daily split contains overlapping signer identities, we construct two signer-independent splits. Split A holds out signers 1–3, yielding 13,294/3,327/4,033 training/validation/test videos; Split B holds out signers 4–6, yielding 13,348/3,342/3,964 videos. Table 4 details the performance under this protocol. Adding the keypoint stream increases BLEU-4 from 7.24 to 20.01 on Split A and from 18.99 to 27.02 on Split B. These large improvements across both splits indicate that structural motion cues are

Method	Pretraining Data	Split A					Split B				
		B1	**B2**	**B3**	**B4**	**R**	**B1**	**B2**	**B3**	**B4**	**R**
HyperSign w/o Keypoint	WLASL	27.43	16.78	10.76	7.24	27.96	45.85	33.72	25.03	18.99	46.64
HyperSign	WLASL	47.66	35.35	26.34	20.01	48.23	55.98	43.07	33.74	27.02	55.91
HyperSign	**SynSign**	**48.29**	**35.93**	**26.95**	**20.54**	**48.83**	**56.61**	**43.85**	**34.55**	**27.75**	**56.51**

Table 4: Comparison of translation performance on CSL-Daily under the signer-independent protocol.

particularly important when signer identities do not overlap between training and testing. Under the same two-stream architecture, SynSign pretraining yields modest BLEU-4 gains over WLASL pretraining, from 20.01 to 20.54 on Split A and from 27.02 to 27.75 on Split B, with the same numerical trend across all other reported metrics. These results suggest that SynSign, with its expanded appearance variation, provides useful pretraining signals for cross-signer transfer in SLT.

5 Conclusion

We present SynSign, a controllable sign language data synthesis framework combining RAG-based semantic retrieval and vocabulary-guided gloss generation with 3D motion reconstruction and controllable rendering. On CSL-Daily, SynSign pretraining achieves performance close to WLASL using fewer video sequences, while joint pretraining yields the strongest overall results. Under two signer-independent splits, SynSign consistently obtains higher scores than WLASL pretraining. These results support the feasibility of using controllable synthetic data for SLT pretraining and motivate further study of semantic and appearance diversity.

Limitations

SynSign constructs sentence-level motions from isolated-sign clips; transition smoothing reduces boundary discontinuities but cannot reproduce all coarticulation and sentence-level non-manual cues of natural signing. Rendered videos also retain a domain gap from real recordings, and the current pipeline uses a fixed frontal camera rather than broader viewpoint or conversational settings.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62476232 and in part by First Batch of Projects for the 2025 "Intergovernmental International Science, Technology and Innovation Cooperation" of the National Key Research and Development Program of China under Grant 2025YFE0121700.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. 2023. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36:11454–11468.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5).

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022a. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5120–5130.

Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. 2022b. Two-stream network for sign language recognition and translation. *Advances in neural information processing systems*, 35:17043–17056.

Zhigang Chen, Benjia Zhou, Yiqing Huang, Jun Wan, Yibo Hu, Hailin Shi, Yanyan Liang, Zhen Lei, and Du Zhang. 2025. C 2 rl: Content and context representation learning for gloss-free sign language translation and retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(9):8533–8544.

Biao Fu, Liang Zhang, Peigen Ye, Pei Yu, Cong Hu, Xiaodong Shi, and Yidong Chen. 2025. Improving end-to-end sign language translation via multi-level contrastive learning. *IEEE transactions on audio, speech and language processing*, 33:1230–1242.

Zhengsheng Guo, Zhiwei He, Wenxiang Jiao, Xing Wang, Rui Wang, Kehai Chen, Zhaopeng Tu, Yong Xu, and Min Zhang. 2024. Unsupervised sign language translation and generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14041–14055.

Peng Jin, Hongkai Li, Jun Yang, Yazhou Ren, Yuhao Li, Lilan Zhou, Jin Liu, Mei Zhang, Xiaorong Pu, and Siyuan Jing. 2025. A large dataset covering the chinese national sign language for dual-view isolated sign language recognition. *Scientific Data*, 12(1):660.

Chang Hao Lai, Rui Zhao, Xuewen Zhong, Jinsong Su, and Yidong Chen. 2026. Selective contrastive learning for gloss free sign language translation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 45648–45660.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. 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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

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.

Zhi Rao, Yucheng Zhou, Benjia Zhou, Yiqing Huang, Sergio Escalera, and Jun Wan. 2026. Rvlf: A reinforcing vision-language framework for gloss-free sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9237–9247.

Xin Shen, Lei Shen, Shaozu Yuan, Heming Du, Haiyang Sun, and Xin Yu. 2024. Diverse sign language translation. *arXiv preprint arXiv:2410.19586*.

Xin Shen, Shaozu Yuan, Hongwei Sheng, Heming Du, and Xin Yu. 2023. Auslan-daily: Australian sign language translation for daily communication and news. *Advances in Neural Information Processing Systems*, 36:80455–80469.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364.

Fangyun Wei and Yutong Chen. 2023. Improving continuous sign language recognition with cross-lingual signs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23612–23621.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321.

Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, et al. 2023. Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20282–20292.

Jinhui Ye, Wenxiang Jiao, Xing Wang, and Zhaopeng Tu. 2023a. Scaling back-translation with domain text generation for sign language gloss translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 463–476.

Jinhui Ye, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Hui Xiong. 2023b. Cross-modality data augmentation for end-to-end sign language translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13558–13571.

Biao Zhang, Garrett Tanzer, and Orhan Firat. 2024. Scaling sign language translation. *Advances in neural information processing systems*, 37:114018–114047.

Ruiquan Zhang, Rui Zhao, Zhicong Wu, Liang Zhang, Haoqi Zhang, and Yidong Chen. 2025. Dynamic feature fusion for sign language translation using hypernetworks. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6227–6239.

Rui Zhao, Liang Zhang, Biao Fu, Cong Hu, Jinsong Su, and Yidong Chen. 2024. Conditional variational autoencoder for sign language translation with cross-modal alignment. In *Proceedings of the aaai conference on artificial intelligence*, volume 38, pages 19643–19651.

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pretraining.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20871–20881.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1316–1325.

Ronglai Zuo, Fangyun Wei, Zenggui Chen, Brian Mak, Jiaolong Yang, and Xin Tong. 2024. A simple baseline for spoken language to sign language translation with 3d avatars. In *European Conference on Computer Vision*, pages 36–54. Springer.