SignX: The Foundation Model for Sign Recognition

Sen Fang[1], Chunyu Sui[2], Hongwei Ye[3], Carol Neidle[4], Dimitris N. Metaxas[1]
[1] Rutgers University, [2] Columbia University, [3] Max-Planck Institute, [4] Boston University

Abstract

The complexity of sign language data processing brings many challenges. The current approach to recognition of ASL signs aims to translate RGB sign language videos through pose information into English-based ID glosses, which serve to uniquely identify ASL signs. Note that there is no shared convention for assigning such glosses to ASL signs, so it is essential that the same glossing conventions are used for all of the data in the datasets that are employed. This paper proposes SignX, a foundation model framework for sign recognition. It is a concise yet powerful framework applicable to multiple human activity recognition scenarios. First, we developed a Pose2Gloss component based on an inverse diffusion model, which contains a multi-track pose fusion layer that unifies five of the most powerful pose information sources—SMPLer-X, DWPose, Mediapipe, PrimeDepth, and Sapiens Segmentation—into a single latent pose representation. Second, we trained a Video2Pose module based on ViT that can directly convert raw video into signer pose representation. Through this 2-stage training framework, we enable sign language recognition models to be compatible with existing pose formats, laying the foundation for the common pose estimation necessary for sign recognition. Experimental results show that SignX can recognize signs from sign language video, producing predicted gloss representations with greater accuracy than has been reported in prior work.

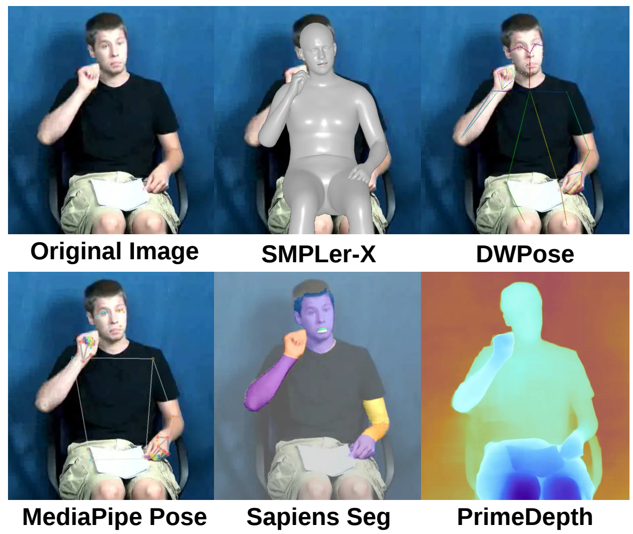

Figure 1. SMPLer-X [2] can provide accurate 3D human body model parameters; DWPose [85] focuses on real-time 2D keypoint detection; Mediapipe [51] provides lightweight but efficient 3D pose prediction; PrimeDepth [90] can obtain scene depth information; while Sapiens segmentation [36] provides fine-grained human body part segmentation results. These methods each have their own characteristics, providing rich feature representations for sign language recognition.

1. Introduction

American Sign Language (ASL) is the primary means of communication for the Deaf community in the US and parts of Canada. According to the World Health Organization, there are approximately 466 million people with hearing disabilities worldwide, of whom over 70 million use sign language for daily communication [35]. Sign language recognition aims to automatically interpret sign language videos into text or Glosses (represented as capitalized text, which serve as unique identifiers of signs). This has important social value in facilitating barrier-free communication between deaf and hearing individuals [5, 15, 69, 88].

However, sign language (SL) recognition faces technical challenges: (1) Signed languages are complex multimodal systems, involving hand and arm movements as well as facial expressions and body postures, with complex temporal dependencies among these channels [1, 76, 78]. (2) Existing SL datasets are limited in scale and lack uniform preprocessing format, with complicated processing pipelines, which severely constrains the application of deep learning methods [15, 21, 61]. (3) Given the lack of a unified pose representation, it is difficult to integrate the innovative advantages of different types of models, limiting model comparison and performance improvement [22, 23, 40].

To address these challenges, we developed a Pose2Gloss module based on the inverse diffusion model [30], which unifies information from 5 of the most powerful pose estimation models—SMPLer-X [2], DWPose [85], Mediapipe [51], PrimeDepth [90], and Sapiens segmentation [36]—

into a multi-track pose fusion layer [6, 60]. We learn the Pose2Gloss process in the U-Net of the diffusion model through unified pose representation. The first three methods are primary, while the other 2 are auxiliary learning. Through this approach, we learn not only hand movements, but also more complex body information, such as finger overlapping, palm tilting, or vertical translation, with the goal of fully capturing a person's pose [54, 64, 69, 84, 97].

Moreover, most existing models have several limitations: (1) Most are based on pose data learning, which is inconvenient in practical applications and requires assistance from other pose information extraction models [80, 81]. (2) Furthermore, models capable of real-time translation have low accuracy, while highly accurate models have slow output speeds [66]. (3) Even purely vision-based models are limited by high computational costs, and their attention mechanisms cannot support higher performance ceilings for longer and more accurate translations [18, 41, 44, 92].

To address these challenges, this paper proposes SignX, a foundational model framework for sign recognition. SignX adopts a 2-stage training architecture, achieving, for the first time, unified processing from raw sign language videos to multiple heterogeneous pose information and then to Gloss representations [8, 67, 91].

The backbone of the first training stage is an inverse diffusion model [30], which trains the Pose2Gloss module. The core of this module maps multi-track pose information to a unified latent representation space [5, 12, 13, 49]. Through carefully designed diffusion processes, the module can learn the intrinsic correlations among different types of pose information, where multiple information representations represent a single pose and generate one pose representation. This unified feature representation not only addresses the problem of heterogeneous information integration, but also facilitates the subsequent expansion of the model [15, 40, 86, 95].

The second training stage is a Video2Pose module based on a Vision Transformer [19]. Since raw video is typically used as input in SL recognition applications, we freeze the Pose2Gloss model and then train ViT to directly convert video sequences into unified pose representations [1, 12, 13]. This module effectively captures dynamic information in videos through spatiotemporal attention mechanisms and enhances feature representation accuracy through multi-scale feature aggregation. This end-to-end processing approach greatly simplifies the pose representation process that is necessary for SL recognition.

Through this 2-stage training framework, SignX achieves powerful functionality while maintaining architectural simplicity: it can handle various existing pose formats, while directly processing raw video input. Experimental results show that SignX has achieved good performance on our ASL datasets, surpassing existing methods in both accuracy and robustness [11, 24, 28, 42].

The main contributions of this paper can be summarized in the following four points:

- We propose SignX, a 2-stage foundational model for sign recognition that can uniformly process multiple heterogeneous types of pose information; it is capable of recognizing both sign videos and pose files, demonstrating strong practical application value.
- We design a Pose2Gloss module based on inverse diffusion, featuring a multi-track pose fusion layer that achieves, for the first time, effective integration of 5 mainstream pose information types, expanding the model's compatibility with different types of pose information through innovative feature representation.
- We develop a Video2Pose module based on improved ViT, achieving direct conversion from raw video to unified representation in an end-to-end manner, significantly enhancing the practical value that was previously underaddressed in existing work.
- Through our carefully designed 2-stage training framework, we achieve powerful performance while maintaining model simplicity. SignX can recognize signs from SL video and generate accurate Gloss representations, thereby establishing a new baseline in continuous SL recognition.

2. Related Work

2.1. Sign Language Recognition

Sign language recognition is an important research direction, at the intersection of computer vision and natural language processing. Early works mainly relied on hand-crafted features and rules to recognize small sign language vocabularies [25, 33]. Subsequently, methods based on the use of machine learning (HMMs) and 3D human pose and motion analytics were used to recognize ASL [53, 77]. With the development of deep learning, neural network-based methods began to dominate this field. The earliest deep learning methods used CNN [29] to extract spatial features from video frames and then used RNN [70] to model temporal relationships [18, 43]. Although these methods were simple and direct, they struggled to capture fine-grained movements in sign language.

Subsequently, researchers began to introduce human pose estimation technology [9, 16, 34, 39, 55, 87], converting sign language videos into skeleton sequences for recognition, which significantly improved the models' ability to perceive action details [64, 65]. With the advancement of computer vision technology, various powerful pose estimation methods [2, 85] have emerged. However, because of their different data formats and representation methods, how to effectively integrate this heterogeneous information has become an urgent problem to solve [22, 23, 32, 69]. Notably, lexical units (signs) play an important role in SL understanding. Understanding also requires taking account of the complex syntax

Dataset	Duration(h)			Vocabulary(K)			Output Type	Year	Domain
	train	val	test	train	val	test			
KETI [39]	20.05	2.24	5.70	←	0.49	→	Spoken Text	2019	Emergency situations
PHOENIX-2014T [3]	9.2	0.6	0.7	2	0.9	1	Spoken Text, Gloss	2018	Weather Forecast
CSL Daily [96]	20.62	1.24	1.41	2	1.3	1.3	Spoken Text, Gloss	2021	Daily life
OpenASL [68]	←	288	→	←	33	→	Spoken Text	2022	Youtube (news + vlogs)
How2Sign [20]	69.6	3.9	5.6	15.6	3.2	3.6	Spoken Text	2021	Instructional
ASLLRP [58]	←	80	→	←	2.1	→	Gloss	2022	Comprehensive

Table 1. **Comparison of information and translation type on mainstream datasets:** The top 5 rows primarily focus on spoken word text. However, our main focus is on recognition of ASL signs (identified by Gloss ID labels) from poses and sign language videos. ← → indicate that in some cases only statistics on the whole dataset are provided.

of these languages. This involves not only linear sign order, but also interactions with non-manual signals, which play an essential grammatical role in signed languages. Gloss-based representations like "I-GO-STORE" can help, in part, to model the unique linguistic rules of a given signed language [5, 15, 46, 91, 92].

2.2. Pose Estimation

Pose estimation technology has made significant progress in recent years. SMPLer-X [2] achieves precise modeling of the body, hands, and face by providing complete 3D human body model parameter estimation [6, 48]. DWPose [85] focuses on 2D keypoint detection, achieving real-time performance while ensuring stability [51, 80]. Google's MediaPipe [51] provides lightweight but accurate 3D pose prediction, making real-time applications possible [8, 51].

Beyond traditional pose estimation, emerging technologies continue to appear. PrimeDepth [90], specifically designed for depth estimation, can provide 3D structural information about scenes [7, 60]. Sapiens segmentation [36] provides a new perspective for interpreting signers' fine-grained actions through precise human body part segmentation [26, 66]. These advances provide rich feature representation methods for SL recognition, offering good extensions to the information dimensions of pose estimation [41, 72].

2.3. Diffusion Models

Diffusion models have achieved breakthroughs in the generative field in recent years. From the initial DDPM [30] to Stable Diffusion [62], diffusion models have demonstrated powerful image generation capabilities [67, 81]. Researchers have also begun applying diffusion to discrete text sequence generation, with promising results [1, 40]. Particularly in cross-modal conversion tasks, diffusion models have shown excellent performance, providing important references for converting SL videos to text [17, 64].

The success of diffusion models lies in their unique noise addition and denoising process, which enables models to learn the inherent structure of data. Through gradual denoising, models can generate high-quality samples [69, 80]. This characteristic makes diffusion models particularly suitable for handling feature conversion problems in SL recognition, especially when processing continuously changing pose sequences [5, 21].

2.4. Vision Transformer

Since the development of ViT [19], the use of Transformers in vision tasks has become increasingly widespread. Compared to traditional architectures, Transformers can better model long-range dependencies through self-attention mechanisms [32, 65]. The original ViT first applied pure Transformer structure to image classification, pioneering a new paradigm of vision Transformers. Subsequent improvements, such as Swin Transformer and other models, further enhanced model performance in vision tasks by introducing local attention mechanisms [15, 40].

In the field of video event recognition, researchers have introduced spatiotemporal attention mechanisms, enabling ViT to better process video sequence data [1, 24]. This advance is particularly important for SL recognition, as SL videos contain rich temporal information. Through well-designed spatiotemporal attention mechanisms, models can effectively capture temporal dependencies in sequences of signing [42, 66].

3. Methodology

3.1. Data Construction and Processing

Dataset Preparation. Our experiments utilize a linguistically annotated ASL dataset of continuous signing: the ASLLRP SignStream® 3 Corpus [58]. This dataset contains over 80 hours of American Sign Language (ASL) videos with synchronized front view, side view, and facial close-up recordings. These videos have been linguistically annotated using SignStream® [57] software (see https://www.bu.edu/asllrp/SignStream/3/), providing detailed manual Gloss labels, sign types, handshape informa-

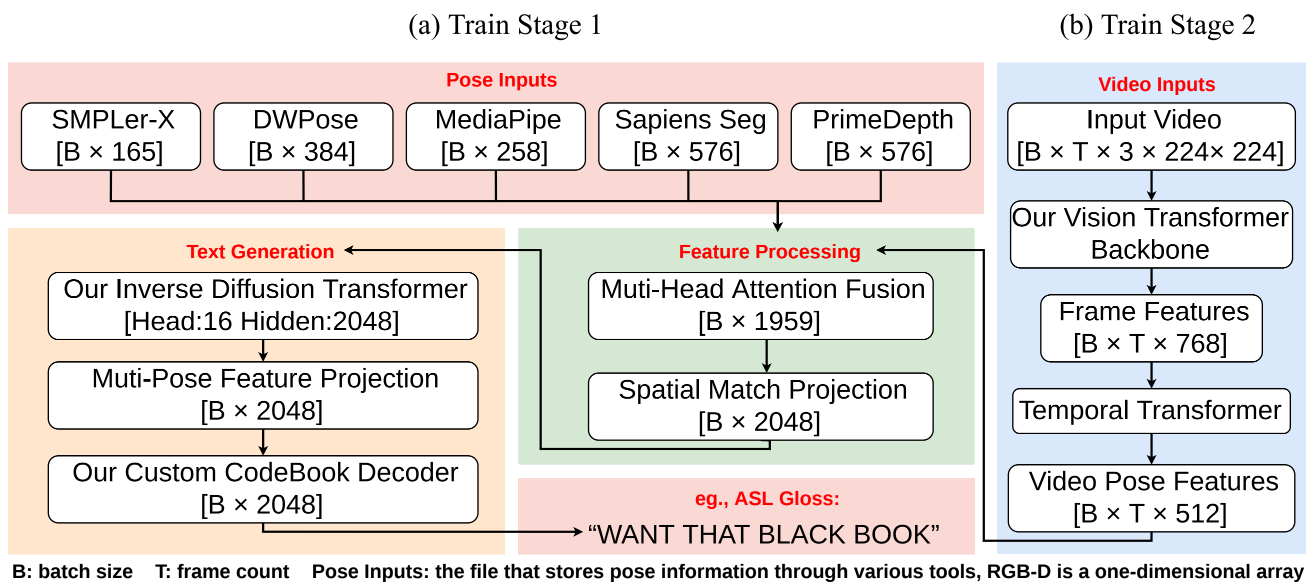

Figure 2. **Overview of the Pose2Gloss and Video2Gloss training pipeline:** (a) Stage 1 (Pose2Gloss) processes diverse pose features from SMPLer-X, DWPose, MediaPipe, Sapiens Segmentation, and PrimeDepth through a multi-head attention fusion and spatial match projection, employing an inverse diffusion transformer with a custom CodeBook decoder to generate textual descriptions. (b) Stage 2 (Video2Pose) extends this approach by utilizing a vision transformer backbone and temporal transformer to extract and process video frame features, enabling comprehensive translation of visual and pose information into coherent textual narratives.

tion, and non-manual grammatical markers. The dataset comprises 2,127 utterances with 17,522 sign tokens, plus an additional 542 sentences from DawnSignPress.

For video processing, we apply 5 pose extraction models: SMPLer-X [2] for accurate 3D human body parameters, DWPose [85] for real-time 2D keypoint detection, MediaPipe [51] for lightweight 3D pose prediction, PrimeDepth [90] for scene depth information, and Sapiens segmentation [36] for fine-grained body part segmentation. Unlike conventional approaches, our preprocessing strategy retains some incomplete or slightly occluded video samples to enhance model robustness to quality variations in real-world scenarios. Videos are standardized to 30fps and processed at uniform spatial resolution. We extract multimodal pose features from all videos to form a unified representation space.

Data Construction. To address the challenge of integrating heterogeneous pose information for SL recognition, we first construct a comprehensive processing pipeline that standardizes multiple pose formats. For each input video $V \in \mathbb{R}^{T\times H\times W\times 3}$, where T represents the number of frames and H, W represent height and width, respectively, we extract sequential pose representations for each pose type: $P_{\text{DWPose}} \in \mathbb{R}^{T\times 384}$, $P_{\text{MediaPipe}} \in \mathbb{R}^{T\times 258}$, $P_{\text{SMPLer-X}} \in \mathbb{R}^{T\times 165}$, $P_{\text{PrimeDepth}} \in \mathbb{R}^{T\times 576}$, $P_{\text{Sapiens}} \in \mathbb{R}^{T\times 576}$. The dimensionality of each pose type reflects its specific information structure. For instance, DWPose contains 18 body keypoints, 21 keypoints for each hand, and 68 facial keypoints, along with their respective confidence scores, totaling 384 dimensions per frame.

As shown in Fig. 2, we flatten the information from each frame into a 1-dimensional list, then concatenate 5 types of poses to form an input of length 1959. In this way, we obtain rich pose information for each frame. This standardized structure not only prevents different types of pose information from interfering with each other, but also makes heterogeneous types of pose information mutually compatible.

3.2. Inverse Diffusion for Multi-Track Pose2Gloss

To convert multi-track pose information into text effectively, we develop an inverse diffusion-based approach that learns the mapping from the unified pose representation to natural language descriptions. Our model consists of several key components: a pose encoder for multimodal fusion, a UNet backbone for the diffusion process, and a text decoder for generating the final output.

Data Encoding and Projection The pose encoder E_{pose} processes the 5 types of pose information with type-specific encoders and fuses them through a multi-head attention mechanism:

$$f_i = E_i(P_i), \tag{1}$$

where E_i is the encoder for pose type i, and $f_i \in \mathbb{R}^{d_h}$ is the encoded feature with hidden dimension d_h. The encoded features from different pose types are then fused through a multi-head attention layer:

$$f_{\text{fused}} = \text{MultiHeadAttention}(f_1, f_2, ..., f_5) \quad (2)$$

This fusion approach preserves the unique characteristics of each pose type, while allowing information sharing among them [40, 47, 93, 93].

Multimodal Pose Fusion Layer The multimodal pose fusion layer is the core component of our Pose2Gloss module, enabling effective integration of heterogeneous pose information. For each encoded pose feature f_i, we project it to a unified representation space through a dimension matching layer:

$$z_i = \text{PadMatch}(f_i) \in \mathbb{R}^{2048} \quad (3)$$

where z_i is the noisy representation of input position i; the projected features are transformed into a unified feature representation that preserves the relationships among pose modalities while normalizing their dimensions. The diffusion process operates on this unified representation by gradually adding and then removing noise:

$$z_t = \sqrt{\alpha_t} z_0 + \sqrt{1 - \alpha_t} \epsilon \quad (4)$$

where z_t is the noisy representation at timestep t, z_0 is the original representation, α_t is the noise schedule, and $\epsilon \sim \mathcal{N}(0, I)$. During training, the model learns to predict the noise ϵ from z_t, while during inference, it performs the reverse process to recover the original representation from random noise. This approach allows the model to learn the intricate connections among different pose types and create a unified representation that captures the essence of the sign language movements [13, 49, 69].

3.3. ViT for Video2Pose

To directly process raw video inputs, we develop a Video2Pose module based on Vision Transformer (ViT) [19]. Given an input video $V \in \mathbb{R}^{T \times H \times W \times 3}$, the Video2Pose module extracts frame-level features and models their temporal relationships:

$$V_{\text{frame}} = v_1, v_2, ..., v_T = \text{ViT}_{\text{spatial}}(V) \quad (5)$$

where $v_t \in \mathbb{R}^{d_v}$ represents the feature of frame t extracted by the spatial ViT. To capture temporal dependencies, we employ a temporal attention mechanism:

$$V_{\text{temp}} = \text{TemporalAttention}(V_{\text{frame}}) \quad (6)$$

The Video2Pose module then projects these temporal features to the same format as the 5 pose types:

$$\hat{P}_i = \text{Projection}i(V\text{temp}) \quad (7)$$

where $\hat{P}_i$ is the predicted pose of type i. This end-to-end approach eliminates the need for separate pose extraction models during deployment, greatly simplifying the application process [10, 66, 74, 82, 83].

4. Two-Stage Training Methodology

Sign recognition presents significant challenges in multimodal representation learning, particularly in converting continuous visual gestures into discrete textual descriptions [14, 15, 23, 37, 63]. We propose a 2-stage training approach that systematically learns pose-to-text and video-to-pose representations.

4.1. Stage I: Pose2Gloss Module Training

Noise Prediction Loss Drawing from diffusion models [30], we introduce a noise prediction loss L_{noise} that operates on pose representations. The loss measures the discrepancy between predicted and actual noise across different timesteps, where $z_0 \in \mathbb{R}^D$ represents the original pose representation, ϵ is Gaussian noise, t is time step, and $\hat{\epsilon}_\theta$ is a neural network predicting noise:

$$L_{\text{noise}} = \mathbb{E}z_0, \epsilon, t[||\epsilon - \hat{\epsilon}\theta(z_t, t)||^2] \quad (8)$$

Text Generation Loss Following sequence-to-sequence learning principles [71], we employ a cross-entropy text generation loss [52] that measures the probabilistic divergence between predicted and ground truth token sequences:

$$L_{\text{text}} = -\sum_{i=1}^{N} y_{\text{true},i} \log(y_{\text{pred},i}) \quad (9)$$

where N denotes sequence length, $y_{\text{true},i}$ represents ground truth tokens, and $y_{\text{pred},i}$ represents predicted token probabilities.

Word Matching Loss To further improve vocabulary precision and text coherence, we introduce a word matching loss $\mathcal{L}_{\text{word}}$ that directly compares predicted and ground-truth token embeddings using cosine similarity. This loss is defined as:

$$\mathcal{L}_{\text{word}} = \frac{1}{B} \sum_{b=1}^{B} (1 - \text{sim}(e_{\text{pred},b}, e_{\text{true},b})), \quad (10)$$

where B is the batch size, $e_{\text{pred},b}$ and $e_{\text{true},b}$ represent the embeddings of the predicted and ground-truth tokens for batch sample b, respectively, and $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity function. By minimizing this loss, the model learns to align the predicted token embeddings with their true counterparts, enhancing the semantic accuracy and fluency of the generated text. This approach proves particularly effective in addressing over-long or incoherent outputs observed with the text generation loss alone.

Contrastive Learning Loss Inspired by recent multimodal representation learning [10, 50, 94], we introduce a contrastive alignment loss that encourages semantic proximity between pose and text embeddings:

$$L_{\text{contrast}} = -\log \frac{\exp(sim(z_{\text{pose}}, e_{\text{text}})/\tau)}{\sum_j \exp(sim(z_{\text{pose}}, e^j_{\text{text}})/\tau)}. \quad (11)$$

The loss leverages cosine similarity between pose embeddings z_{pose} and text embeddings e_{text}, with τ as a temperature scaling parameter to control embedding separation.

Composite Loss Formulation The total loss combines these components with learnable weighting coefficients:

$$L_{\text{total}} = \lambda_{\text{noise}} L_{\text{noise}} + \lambda_{\text{text}} L_{\text{text}} + \lambda_{\text{word}} L_{\text{word}} + \lambda_{\text{contrast}} L_{\text{contrast}}. \tag{12}$$

4.2. Stage II: Video2Pose Module Training

Following feature transfer learning principles [89], we freeze the Pose2Gloss module and train the Video2Pose component by minimizing the reconstruction error:

$$L_{\text{video2pose}} = \sum_i ||\hat{P}_i - P_i||^2, \tag{13}$$

where $\hat{P}_i$ represents predicted pose representations and P_i denotes ground truth pose representations. This 2-stage approach enables specialized learning for each module, creating a robust sign recognition system that effectively bridges visual and textual domains.

4.3. Implementation Details

For Stages I and II, we use the linguistically annotated ASLLRP SignStream® 3 Corpus of continuous signing [58] as primary datasets for American Sign Language (ASL) recognition. During the development of SignX, we initially identified that noise in pose features from diverse sources—SMPLer-X, DWPose, MediaPipe, PrimeDepth, and Sapiens Segmentation—significantly impacted the performance of our SL recognition system. To address this, we prioritized optimizing noise reduction and reconstruction in the Pose2Gloss module (Stage I), focusing on a noise prediction loss $\mathcal{L}_{\text{noise}} = \mathbb{E}_{z_0,\epsilon,t}[||\epsilon - \hat{\epsilon}_\theta(z_t, t)||^2]$, where $z_0 \in \mathbb{R}^D$ is the original pose representation, z_t is the noisy version at timestep t, ϵ is Gaussian noise, and $\hat{\epsilon}_\theta$ is the neural network prediction of noise. However, we observed that relying solely on noise optimization did not sufficiently address text generation quality.

Next, we explored a text generation loss $\mathcal{L}_{\text{text}} = -\sum_{i=1}^{N} y_{\text{true},i} \log(y_{\text{pred},i})$, where N denotes sequence length, $y_{\text{true},i}$ represents ground-truth tokens, and $y_{\text{pred},i}$ represents predicted token probabilities. While this approach improved basic text output, it frequently generated over-long and incoherent text, limiting practical utility. To refine this, we investigated an n-gram consistency loss with an F1-based metric, $\mathcal{L}_{\text{ngram}}$, designed to balance precision and recall of n-gram matches in generated text. However, we realized that this implicit matching approach was less effective than directly focusing on the vocabulary and explicit token alignment of the generated text.

We therefore introduced a word matching loss $\mathcal{L}_{\text{word}}$, which measures cosine similarity between predicted and ground-truth token embeddings, encouraging precise vocabulary generation. This loss, defined as $\mathcal{L}_{\text{word}} = \frac{1}{B}\sum_{b=1}^{B}(1 - \text{sim}(e_{\text{pred},b}, e_{\text{true},b}))$—where B is the batch size, $e_{\text{pred},b}$ and $e_{\text{true},b}$ are the predicted and true token embeddings, respectively, and sim is cosine similarity—significantly improved text coherence and vocabulary accuracy. Incorporation of a contrastive learning loss $\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(\text{sim}(z_{\text{pose}}, e_{\text{text}})/\tau)}{\sum_j \exp(\text{sim}(z_{\text{pose}}, e^j_{\text{text}})/\tau)}$, where z_{pose} and e_{text} are pose and text embeddings, and τ is a temperature parameter, encouraging semantic proximity between multimodal representations, further enhanced semantic fluency.

Config	Stage I (Pose2Gloss)	Stage II (Video2Pose)
optimizer	AdamW	AdamW
base learning rate	1×10^{-3}	1×10^{-3}
weight decay	0.01	0.01
optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.999$	$\beta_1 = 0.9, \beta_2 = 0.999$
learning rate schedule	Cosine decay (min = $0.05 \times lr$)	Cosine decay (min = $0.01 \times lr$)
warmup steps	Dynamic (2%-10% of total steps)	10% of total steps
training epochs	500–1000	300
batch size	32	32
gradient accumulation	4	0
gradient clipping	7.0	1.0
label smoothing	0.1	N/A
teacher forcing ratio	Decay from 0.5 to 0	N/A
noise loss weight	Decay from 1 to 0	N/A
text loss weight	1	1
word match weight	1	N/A
ngram loss weight	0.6	N/A
contrast loss weight	1	N/A
max sequence length	64	64
temperature	0.7 (inference)	0.7 (inference)
beam size	5 (inference)	5 (inference)
length penalty	0.7 (inference)	0.7 (inference)
top-k	50 (sampling)	50 (sampling)
top-p	0.92 (sampling)	N/A
repetition penalty	2.5 (beam search)	N/A

Table 2. **Comprehensive training configurations and hyperparameters:** The table details optimizer settings, learning schedules, training parameters, and inference configurations for both stages.

For Stage II (Video2Pose), we froze the Pose2Gloss parameters trained in Stage I and developed a Vision Transformer-based model to directly convert raw video inputs into the unified pose representation, minimizing the reconstruction loss $\mathcal{L}_{\text{video2pose}} = \sum_i ||\hat{P}_i - P_i||^2$, where $\hat{P}_i$ represents predicted pose representations and P_i denotes ground-truth poses. This 2-stage approach, leveraging the pre-trained Pose2Gloss module, ensures robust video-to-pose conversion while maintaining eval efficiency.

These optimizations, combined with gradient accumulation (batch size of 32 for Stage I and 4 for Stage II), a cosine decay learning rate schedule with a minimum learning rate ratio of 0.05 (resulting in a decay factor of $\cos(\pi \cdot \text{progress})$, where progress is the normalized training progress after warmup), and the AdamW optimizer with a learning rate of 1×10^{-3}, weight decay of 0.01, and momentum parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ (as detailed in Table 2), resolved initial training instability and enhanced model performance.

	Pose2Gloss			Video2Gloss		
	BLEU	**Top-5**	**Best**	**BLEU**	**Top-5**	**Best**
Ours.	10.60	23.42	76.13	12.01	21.86	84.67

Table 3. Best scores on the 2000-gloss ASLLRP dataset [58].

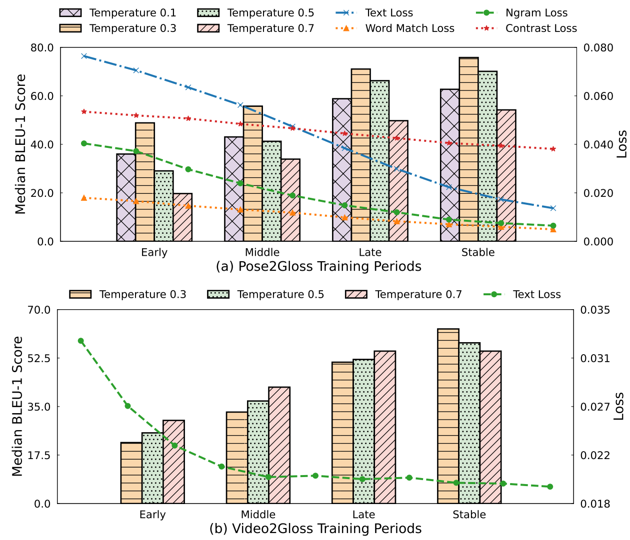

Figure 3. **Training Efficiency & Ablation Analysis:** We compare the effects of different training periods (*i.e.*, (a) and (b) in the figure) and temperature settings in the 2-stage training process. The different periods are divided and determined based on every quarter of the total number of epochs.

5. Experiments

Here we evaluate our 2-stage SL recognition framework, consisting of Pose2Gloss and Video2Gloss modules, using the ASLLRP [58] dataset. Our experiments assess sign recognition accuracy as well as training efficiency, with comparisons to baseline methods where applicable. For sign recognition experiments, we used the ASLLRP dataset, which contains over 2,100 annotated video utterances. From this, we designated 100 videos as the test set. This consistent split across tasks enables fair performance assessment and analysis.

5.1. Evaluation Metrics

For sign recognition (Pose2Gloss and Video2Gloss), we measure: (1) **BLEU**, the standard BLEU score [59] (ranging from 0 to 100), which calculates n-gram overlap between generated and ground-truth gloss sequences; (2) **Top-n Accuracy**, the percentage of samples where the ground-truth gloss is among the top-n predictions; (3) **Best Score**, the highest BLEU-1 score achieved across all test samples, highlighting peak performance; and (4) **Word Error Rate (WER)**, the minimum number of substitutions [38], insertions, and deletions required to transform the predicted gloss sequence into the ground-truth sequence, divided by the length of the ground-truth sequence, expressed as a percentage (lower WER indicates better performance). All metrics are computed using NLTK's BLEU implementation with a brevity penalty, averaged over 100 test samples.

5.2. Experiments on Sign Recognition

Pose2Gloss & Video2Gloss Recognition Study. We evaluate Pose2Gloss and Video2Gloss on the ASLLRP dataset, focusing on the 100-video test set. Pose2Gloss generates gloss sequences from pre-extracted pose features, while Video2Gloss processes raw video inputs before gloss generation. Table 3 summarizes our results. Pose2Gloss achieves a BLEU of 10.60, Top-5 Accuracy of 23.42, and a Best Score of 76.13, indicating moderate n-gram overlap with strong peak performance on well-aligned samples. Video2Gloss improves to a BLEU of 12.01, with a Top-5 Accuracy of 21.86 and a Best Score of 84.67, suggesting that video context enhances gloss recovery, though Top-5 Accuracy drops because of pose prediction complexity.

Training Efficiency Study. The training efficiency of the Pose2Gloss module is illustrated in Fig. 3, which tracks the progression of 4 loss components—Text Loss, Word Match Loss, Ngram Loss, and Contrast Loss—throughout the training process. The line plots reveal a consistent decline in Text Loss and Word Match Loss over time, reflecting improved token prediction and vocabulary precision within the Pose2Gloss module. Ngram Loss plateaus after an initial decrease, suggesting stabilized n-gram consistency, whereas Contrast Loss shows a slight downward trend, potentially indicating challenges in maintaining pose-text alignment as training advances. These patterns highlight the effectiveness of the composite loss design in Stage I, particularly the interplay between text generation and word matching losses, while suggesting that further adjustment of the contrastive loss may be needed to sustain multimodal coherence.

Ablation Study. We conducted an ablation study to evaluate the impact of temperature settings on gloss generation accuracy within the Pose2Gloss module in Stage I. Fig. 3 illustrates the BLEU scores across four temperature settings (0.1, 0.3, 0.5, and 0.7) at early, middle, late, and stable training periods. The bar plots demonstrate that a temperature of 0.3 consistently yields the highest performance, peaking at the stable phase, while temperatures of 0.1, 0.5, and 0.7 achieve progressively lower scores. This indicates that a moderate temperature optimally enhances gloss generation accuracy in Stage I. However, we can also observe that in the second stage of training, lower temperatures actually yield better results. These findings suggest that temperature plays a critical role in balancing the trade-off between generation precision and training stability, warranting further exploration to refine this parameter.

Comparison with Previous Work. As shown in Table 5, we tested our model's performance on different vocabulary sizes, using 2 different metrics. SignX outperforms previous

Video2Gloss
Generated Text: part:indef ENJOY CHAT FRIEND WITH SECOND-IN-LIST ns-fs-JESSICA SAME OTHER WITH CHAT EACH+ONE IX-1p *Ground Truth:* DURING/WHILE IX-1p ENJOY CHAT WITH ns-fs-DONNA AND SECOND-IN-LIST ns-fs-JESSICA IX-1p SAME ENJOY CHAT WITH EACH+ONE OTHER SAME IX-1p
Generated Text: DATE/DESSERT #IF FAVORITE/PREFER PIE *Ground Truth:* DATE/DESSERT POSS-1p FAVORITE/PREFER PIE
Pose2Gloss
Generated Text: ns-fs-MARY VEGETABLE QMwg #NO IX HAVE #BBQ ONE ONE-MONTH PAST HAVE *Ground Truth:* IX ns-fs-MARY VEGETABLE QMwg #NO IX HAVE #BBQ ONE ONE-MONTH PAST HAVE
Generated Text: #IF POSS-1p SISTER IX GET HEADBAND_2 FOR BIRTHDAY IX SISTER GUARANTEE FUTURE DECORATE WITH HEADBAND_2 *Ground Truth:* #IF POSS-1p SISTER IX GET HEADBAND_2 FOR POSS BIRTHDAY IX SISTER GUARANTEE FUTURE DECORATE WITH BAKING-SPRINKLES

Table 4. Qualitative Evaluation of Video2Gloss and Pose2Gloss Outputs on the ASLLRP Dataset.

Gloss Size:	100	300		1000	2000
Method:	**Top-1% Accuracy ↑**		Method:	**WER score ↓**	
POSE-GRU [79]	46.51	33.68	GEU [73]	49.9	-
POSE-TGCN [79]	55.43	38.32	SLT [16]	24.5	32.0
GCN-BERT [75]	60.15	42.18	CorrNet [31]	20.5	30.1
SPOTER [1]	63.18	43.78	GPGN [27]	20.5	30.0
SignX (Ours)	68.02	62.15	SignX (Ours)	18.6	24.3

Table 5. Performance comparison using Word Error Rate (WER; lower is better) and Top-1% Accuracy (higher is better) for sign recognition. These two metrics each have their own distinct calculation methods and should not be conflated.

models across all vocabulary sizes. For smaller vocabularies (100-300 glosses), SignX achieves top-1% accuracy of 68.02% and 62.15%, surpassing all the other methods, as shown in the Table. With larger vocabularies (1000-2000 glosses), SignX achieves WER scores of 18.6% and 24.3%, also better than all other previous models. Much of the previous work presented in the Table [1, 75, 79] was evaluated on the WLASL, with comparable vocabulary size (100/300 glosses) [45]. (Note, however, that there are issues with inconsistent Gloss labeling in the WLASL data [56], which will have a negative impact on the results.) Our model outperforms those and others using different datasets e.g, [96] (CSL-Daily, 200 glosses) and [4] (RWTH-2014T,1000 glosses).

5.3. Qualitative Experiments

Recognition Effect Analysis. Qualitatively, Pose2Gloss performs well on high-frequency signs, but struggles with noisy inputs (where images are incorrectly recognized), producing Gloss strings that may contain errors (e.g., "FUTURE DECORATE WITH HEADBAND_2" vs. source "FUTURE DECORATE WITH BAKING-SPRINKLES"). Video2Gloss often generates extraneous elements in Gloss sequences (e.g., producing "ENJOY CHAT FRIEND WITH" when the source is "ENJOY CHAT WITH"). Such errors may be attributable to recognition inaccuracies and/or the model's tendency to favor high-frequency signs, particularly in experimental settings, where there is a bias toward commonly occurring signs. In our qualitative evaluation, we found that beam search works better than greedy decoding, and can balance diversity and accuracy.

6. Discussion

Societal Impacts. Our model is crucial for practical applications, sign language research, and scalability. (1) By learning representations of multiple poses, our model can be used in rougher real-world scenarios, which is critical for the widespread popularization and application of sign language recognition. (2) Since our model can accommodate mainstream sign language data without extremely fine-grained processing, facilitates transfer learning, enables model comparisons, and promotes research on less-studied languages. (3) Our model can be used in a lightweight manner, and stage 2 training is based on original videos, which provides our model with strong extensibility.

Limitations. Our model is challenging to train because of its basis in diffusion models. We require very cumbersome multiple training settings and various initialization techniques to formally enter stage 2. This may raise the training threshold considerably.

Future Work. Generating appropriate glosses from a continuous video is an extremely difficult task. There are too many aspects to focus on, and real-world situations are far too complex. Various parameters and constraints need to be carefully considered in order to achieve a decent score. However, this approach lacks elegance. In fact, sign language recognition might be even more challenging than generating full-body poses from text—a field where countless researchers have struggled to make significant progress. Overall, for tasks like Sign Language Translation (SLT) or Sign Language Recognition (SLR), reinforcement learning might offer a more elegant and concise path forward, something future sign language researchers could seriously explore.

7. Conclusion

In summary, we propose the sign language recognition framework named SignX, using inverse diffusion [30] models and ViT [19] as the backbone, and separately training Pose2Gloss and Video2Pose modules. We adopt a 2-stage training approach to bridge the significant differences between these 2 modules, and employ techniques such as multi-pose projection fusion, contrastive learning, multi-loss learning, and custom text decoder with codebook to refine this newly implemented baseline. We validate our performance on mainstream-scale datasets.

References

[1] Matyáš Boháček and Marek Hrúz. Sign pose-based transformer for word-level sign language recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 182–191, 2022. 1, 2, 3, 8

[2] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. Smpler-x: Scaling up expressive human pose and shape estimation, 2023. 1, 2, 3, 4

[3] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793, 2018. 3

[4] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793, 2018. 8

[5] Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural Sign Language Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3

[7] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[8] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody Dance Now. In *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, 2019. 2, 3

[9] James Charles, Tomas Pfister, Mark Everingham, and Andrew Zisserman. Automatic and Efficient Human Pose Estimation for Sign Language Videos. *International Journal of Computer Vision (IJCV)*, 2014. 2

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 5

[11] Xuanting Chen, Junjie Ye, Can Zu, Nuo Xu, Rui Zheng, Minlong Peng, Jie Zhou, Tao Gui, Qi Zhang, and Xuanjing Huang. How robust is gpt-3.5 to predecessors? a comprehensive study on language understanding tasks, 2023. 2

[12] Hung-Yueh Chiang, Yen-Liang Lin, Yueh-Cheng Liu, and Winston H Hsu. A unified point-based framework for 3d segmentation. In *3DV*, 2019. 2

[13] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021. 2, 5

[14] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *CVPR*, pages 7784–7793, 2018. 5

[15] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 3, 5

[16] Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 8

[17] Helen Cooper and Richard Bowden. Large Lexicon Detection of Sign Language. In *International Workshop on Human-Computer Interaction*, 2007. 3

[18] Runpeng Cui, Hu Liu, and Changshui Zhang. Recurrent Convolutional Neural Networks for Continuous Sign Language Recognition by Staged Optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2, 3, 5, 8

[20] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[21] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2Sign: A Large-Scale Multimodal Dataset for Continuous American Sign Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 3

[22] Sen Fang, Chunyu Sui, Yanghao Zhou, Xuedong Zhang, Hongbin Zhong, Minyu Zhao, Yapeng Tian, and Chen Chen. Signdiff: Diffusion models for american sign language production, 2024. 1, 2

[23] Sen Fang, Lei Wang, Ce Zheng, Chunyu Sui, Mingyu Zhao, Yapeng Tian, and Chen Chen. Signllm: Sign language production large language models, 2024. 1, 2, 5

[24] Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus H Piater, and Hermann Ney. RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2012. 2, 3

[25] Kirsti Grobel and Marcell Assan. Isolated Sign Language Recognition using Hidden Markov Models. In *IEEE International Conference on Systems, Man, and Cybernetics*, 1997. 2

[26] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7297–7306, 2018. 3

[27] Leming Guo, Wanli Xue, Bo Liu, Kaihua Zhang, Tiantian Yuan, and Dimitris Metaxas. Gloss prior guided visual feature learning for continuous sign language recognition. *IEEE Transactions on Image Processing*, 33:3486–3495, 2024. 8

[28] Mutian He, Yan Deng, and Lei He. Robust Sequence-to-Sequence Acoustic Modeling with Stepwise Monotonic Attention for Neural TTS. *arXiv e-prints*, 2019. 2

[29] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 2

[30] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 1, 2, 3, 5, 8

[31] Lianyu Hu, Liqing Gao, Zekang Liu, and Wei Feng. Continuous sign language recognition with correlation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 8

[32] Wencan Huang, Wenwen Pan, Zhou Zhao, and Qi Tian. Towards Fast and High-Quality Sign Language Production. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 2, 3

[33] Timor Kadir, Richard Bowden, Eng-Jon Ong, and Andrew Zisserman. Minimal Training, Large Lexicon, Unconstrained Sign Language Recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2004. 2

[34] Slava M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Trans. Acoust. Speech Signal Process.*, 35(3):400–401, 1987. 2

[35] Muhammad Ajmal Khan, Hina Fazil, and Ayesha Wajihullah. Barriers faced by individuals with hearing impairment in accessing information at public places: A qualitative inquiry. *Al-Mahdi Research Journal (MRJ)*, 5(4):582–593, 2024. 1

[36] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *arXiv preprint arXiv:2408.12569*, 2024. 1, 3, 4

[37] Youngmin Kim, Minji Kwak, Dain Lee, Yeongeun Kim, and Hyeongboo Baek. Keypoint based sign language translation without glosses. *arXiv preprint arXiv:2204.10511*, 2022. 5

[38] Dietrich Klakow and Jochen Peters. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38 (1-2):19–28, 2002. 7

[39] Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. Neural sign language translation based on human keypoint estimation. *Applied sciences*, 9(13):2683, 2019. 2, 3

[40] Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. Neural Sign Language Translation based on Human Keypoint Estimation. *Applied Sciences*, 2019. 1, 2, 3, 5

[41] Oscar Koller. Quantitative Survey of the State of the Art in Sign Language Recognition. *arXiv preprint arXiv:2008.09918*, 2020. 2, 3

[42] Oscar Koller, Jens Forster, and Hermann Ney. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125, 2015. 2, 3

[43] Oscar Koller, Jens Forster, and Hermann Ney. Continuous Sign Language Recognition: Towards Large Vocabulary Statistical Recognition Systems Handling Multiple Signers. *Computer Vision and Image Understanding (CVIU)*, 2015. 2

[44] Jolanta Lapiak. Gloss: transcription symbols. `https://www.handspeak.com/learn/index.php?id=3`. Accessed: 2019-08-20. 2

[45] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1459–1469, 2020. 8

[46] Scott K Liddell et al. *Grammar, gesture, and meaning in American Sign Language*. Cambridge University Press, 2003. 3

[47] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022. 5

[48] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. In *ACM TOG*, 2015. 3

[49] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. UNIFIED-IO: A Unified Model for Vision, Language, and Multi-modal Tasks. 2023. 2, 5

[50] Yujie Lu, Pan Lu, Zhiyu Chen, Wanrong Zhu, Xin Eric Wang, and William Yang Wang. Multimodal procedural planning via dual text-image prompting. *CoRR*, abs/2305.01795, 2023. 5

[51] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 1, 3, 4

[52] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pages 23803–23828. PMLR, 2023. 5

[53] D. Metaxas. Deformable model and hmm-based tracking, analysis and recognition of gestures and faces. In *Proceedings International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems. In Conjunction with ICCV'99 (Cat. No.PR00378)*, pages 136–140, 1999. 2

[54] Amit Moryossef. Complete toolkit for working with poses. `https://github.com/AmitMY/pose-format/`, 2022. 2

[55] Amit Moryossef, Ioannis Tsochantaridis, Roee Yosef Aharoni, Sarah Ebling, and Srini Narayanan. Real-time sign language detection using human pose estimation. 2020. https://www.slrtp.com/. 2

[56] Carol Neidle, Augustine Opoku, Carey Ballard, Konstantinos M. Dafnis, Evgenia Chroni, and Dimitris Metaxas. Resources for Computer-Based Sign Recognition from Video, and the Criticality of Consistency of Gloss Labeling across Multiple Large ASL Video Corpora. In *10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources. LREC, Marseille, France, June 2022*, pages 165–172, . 8

[57] Carol Neidle, Augustine Opoku, Gregory Dimitriadis, and Dimitris Metaxas. NEW Shared & Interconnected ASL Resources: SignStream® 3 Software; DAI 2 for Web Access to Linguistically Annotated Video Corpora; and a Sign Bank. In *Language Resources and Evaluation. 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community. Miyazaki, Japan, 2018-05-12*, pages 147–154, . 3

[58] Carol Neidle, Augustine Opoku, and Dimitris Metaxas. ASL Video Corpora & Sign Bank: Resources Available through the American Sign Language Linguistic Research Project (ASLLRP), 2022. 3, 6, 7

[59] 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*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. 7

[60] Shi Qiu, Saeed Anwar, and Nick Barnes. Dense-resolution network for point cloud classification and segmentation. In *WACV*, pages 3813–3822, 2021. 2, 3

[61] Manny Rayner, Pierrette Bouillon, Sarah Ebling, Johanna Gerlach, Irene Strasly, and Nikos Tsourakis. An open web platform for rule-based speech-to-sign translation. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 162–168, 2016. 1

[62] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3

[63] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Everybody sign now: Translating spoken language to photo realistic sign language video. *arXiv preprint arXiv:2011.09846*, 2020. 5

[64] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Progressive Transformers for End-to-End Sign Language Production. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[65] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Continuous 3D Multi-Channel Sign Language Production via Progressive Transformers and Mixture Density Networks. *International Journal of Computer Vision (IJCV)*, 2021. 2, 3

[66] Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. Mixed SIGNals: Sign Language Production via a Mixture of Motion Primitives. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 5

[67] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Skeletal Graph Self-Attention: Embedding a Skeleton Inductive Bias into Sign Language Production. *arXiv preprint arXiv:2112.05277*, 2021. 2, 3

[68] Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. Open-domain sign language translation learned from online video. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. 3

[69] Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. Text2Sign: Towards Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. *International Journal of Computer Vision (IJCV)*, 2020. 1, 2, 3, 5

[70] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014. 2

[71] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014. 5

[72] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015. 3

[73] Shengeng Tang, Dan Guo, Richang Hong, and Meng Wang. Graph-based multimodal sequential embedding for sign language translation. *IEEE Transactions on Multimedia*, 24: 4433–4445, 2022. 8

[74] Laia Tarrés, Gerard I. Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró i Nieto. Sign language translation from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2023. 5

[75] Anirudh Tunga, Sai Vidyaranya Nuthalapati, and Juan Wachs. Pose-based sign language recognition using gcn and bert. In *2021 IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 31–40, 2021. 8

[76] Špela Vintar, Boštjan Jerko, and Marjetka Kulovec. Compiling the slovene sign language corpus. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon. Language Resources and Evaluation Conference (LREC)*, pages 159–162, 2012. 1

[77] C. Vogler and D. Metaxas. Asl recognition based on a coupling between hmms and 3d motion analysis. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 363–369, 1998. 2

[78] U. Von Agris and K.-F. Kraiss. Signum database: Video corpus for signer-independent continuous sign language recognition. In *Workshop on Representation and Processing of Sign Languages*, pages 243–246, 2010. 1

[79] Julia Walczynska. *HandTalk: American sign language recognition by 3D-CNNs*. PhD thesis, 2022. 8

[80] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-Video Synthesis. In *Advances in Neural Information Processing Systems (NIPS)*, 2018. 2, 3

[81] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3

[82] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and

efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090, 2021. 5

[83] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022. 5

[84] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose Guided Human Video Generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[85] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 1, 2, 3, 4

[86] Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning. *CoRR*, 2023. 2

[87] Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. Mlslt: Towards multilingual sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5109–5119, 2022. 2

[88] Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. Including signed languages in natural language processing. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021. 1

[89] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks?, 2014. 6

[90] Denis Zavadski, Damjan Kalšan, and Carsten Rother. Primedepth: Efficient monocular depth estimation with a stable diffusion preimage, 2024. 1, 3, 4

[91] Jan Zelinka and Jakub Kanis. Neural sign language synthesis: Words are our glosses. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 3395–3403, 2020. 2, 3

[92] Jan Zelinka and Jakub Kanis. Neural Sign Language Synthesis: Words Are Our Glosses. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020. 2, 3

[93] Feihu Zhang, Jin Fang, Benjamin Wah, and Philip Torr. Deep fusionnet for point cloud semantic segmentation. In *ECCV*, 2020. 5

[94] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *CoRR*, abs/2302.00923, 2023. 5

[95] Hengshuang Zhao, Jiaya Jia, and Vladlen Koltun. Exploring self-attention for image recognition. In *CVPR*, 2020. 2

[96] Hao Zhou, Wen gang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325, 2021. 3, 8

[97] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive Pose Attention Transfer for Person Image Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2