SignGen: End-to-End Sign Language Video Generation with Latent Diffusion

Fan Qi[1], Yu Duan[1], Huaiwen Zhang[2⋆], and Changsheng Xu[3]

[1] Tianjin University of Technology, Tianjin, China
[2] College of Computer Science, Inner Mongolia University, China
[3] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China
fanqi@email.tjut.edu.cn; dyiwork@stud.tjut.edu.cn;
huaiwen.zhang@imu.edu.cn; csxu@nlpr.ia.ac.cn;

Abstract. The seamless transformation of textual input into natural and expressive sign language holds profound societal significance. Sign language is not solely about hand gestures. It encompasses vital facial expressions and mouth movements essential for nuanced communication. Achieving both semantic precision and emotional resonance in text-to-sign language translation is of paramount importance. Our work pioneers direct end-to-end translation of text into sign language videos, encompassing a realistic representation of the entire body and facial expressions. We go beyond traditional diffusion models by tailoring the multi-modal conditions for sign language videos. Additionally, our modified motion-aware sign generation framework enhances alignment between text and visual cues in sign language, further improving the quality of the generated sign language videos. Extensive experiments show that our approach significantly outperforms the state-of-the-art approaches in terms of semantic consistency, naturalness, and expressiveness, presenting benchmark quantitative results on the RWTH-2014, RWTH-2014-T, WLASL, CSL-Daily, and AUTSL. Our code is available at https://github.com/mingtiannihao/SignGen.

Keywords: Sign Language Generation · Video Generation · Diffusion

1 Introduction

Approximately 466 million people worldwide experience hearing impairments [12]. It is imperative for this global deaf community to have equal access to information. This involves two primary components: to be understood and to understand. While significant strides have been made in the latter, known as sign language recognition [9, 53, 64], in recent years, the former, termed Sign Language Generation (SLG), remains a formidable challenge. Given the considerable expenses tied to employing human translators for the deaf community, the urgency

⋆ Corresponding author

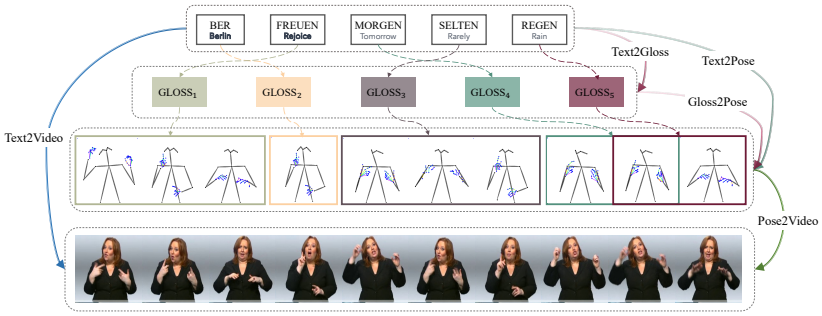

Fig. 1: The sign language video generation approaches. Given a spoken language sentence, existing methods generate sign language poses or videos by Text2Gloss2Pose [44], Text2Pose [44, 46], Gloss2Pose [51], Text2Pose2Video [17, 42, 58], and Text2Gloss2Pose2Video [50, 57]. In this paper, we present `SignGen`, a Text2Video method that streamlines the process by directly translating text into sign language videos, bypassing intermediate stages for increased efficiency and coherence.

for a robust machine translation solution to cater to their daily communication needs cannot be overstated.

Sign language generation approaches typically involve preprocessing input text for morphosyntactic analysis to categorize words into types like nouns, verbs, and particles. These words are then transformed into symbolic sequences by machine translation models and subsequently rendered as symbolic glosses, video representations, or animated avatars [4, 10, 25, 42, 44, 48, 50, 51, 58]. However, translating to sign glosses frequently results in misalignment with spoken language, loss of contextual nuances, and unnatural translations. Moreover, creating realistic avatars is hindered by high data collection and annotation costs, scalability issues, and the need for expert validation to ensure accuracy. Fig.1 illustrates the existing approaches in SLG, highlighting the lack of end-to-end solutions in the established pipeline. The conversion from text to "gloss" often results in a significant loss of context, hindering the accurate conveyance of intended meanings. Besides, generating realistic human videos from "gloss" poses additional challenges, such as ensuring video quality and achieving naturalness in the synthesized content. Compounding these issues is the problem of error propagation, where inaccuracies at each stage magnify risks in subsequent steps, potentially leading to compromised overall outcomes.

Actually, sign language translation extends beyond hand gestures. It incorporates crucial facial expressions and mouth movements, integral to conveying nuanced communication. While existing approaches focus on maintaining only the semantic content, they often neglect the emotional subtleties inherent in sign language. Similar to how the hearing population discerns emotions from facial cues, the deaf community intricately integrates emotions into sign language [63]. In the text to sign video generation task, capturing both precise meaning (**semantic clarity**) and emotional intent (**emotional resonance**) is essential. An ideal sign language translation seamlessly amalgamates these

aspects. Recently, diffusion models [8, 13, 15, 61, 71] have exhibited significant potential in visual generation, outperforming traditional generative adversarial networks in image synthesis and video generation. The stepwise process of diffusion models offers precise control over generation, making them ideal for blending gestures with facial and mouth cues.

In this paper, we introduce `SignGen`, an enhanced diffusion model designed to improve text-to-video generation for sign language translation. We utilize a multi-modal condition fusion module that takes into account both text conditions and motion conditions. Within motion conditions, we harness three critical features: optical flow, pose, and depth, tailored specifically for sign videos. These features are intelligently fused in both temporal and spatial dimensions, enabling us to capture the nuanced movements inherent in sign language expression. Efficiency and consistency in video generation are central to our approach. To achieve this, we develop a motion-aware sign video generation structure. It not only ensures semantic clarity but also prioritizes emotional resonance in the resulting sign language videos. For **semantic clarity**, our model incorporates text into the conditional encoder, providing guidance throughout the generation process. Additionally, we introduce a novel spatio-temporal-semantic attention mechanism that enhances the semantic aspects of the generated content, ensuring faithful conveyance of intended meaning. For **emotional resonance**, our approach places a strong emphasis on accurate facial features detection within the pose features. These facial features play a pivotal role in conveying emotions in sign language. To maintain the fluency and naturalness of emotional expressions, we employ a U-Net-like architecture coupled with a unique design for fully semantic-frame interaction. Our comprehensive approach addresses both semantic clarity and emotional resonance, resulting in generated sign language videos that faithfully convey meaning while embodying emotional depth and authenticity.

The main contributions of this work can be summarized as follows:

- Our work pioneers directly translating text to sign videos, featuring a realistic representation of the entire body and facial expressions.
- We customize a diffusion model that goes beyond conventional approaches by incorporating multi-modal conditions for sign video. To ensure both consistency and efficiency in video generation, we devise an appropriate U-Net-like architecture complemented by motion-aware attention blocks.
- We evaluate our model with five widely used sign language generation datasets. Comprehensive evaluations firmly establish our framework as the leading solution for sign language generation to date.

2 Related work

2.1 Sign Language Production

To bridge the communication gap between hearing and deaf and hard-of-hearing individuals, researchers have explored sign avatars, such as Tessa [10], dictasign [14], Sign3D [20] to perform sign language. These are 3D animated models

that display signed conversations, replicating the motions of fingers, hands, facial gestures, and the body. However, they've faced criticism for appearing unnatural and missing non-manual information like eye gaze and facial expressions. This led to avatars based on motion capture data, which are more realistic but limited due to data collection costs. Recently, Saunders et al. [44] propose a progressive transformer-based SLP model that translates spoken language sentences into 3D sign pose sequences. Text2Gloss2Pose [44] uses a Symbolic Transformer for translating from source text to target gloss sequences. A Progressive Transformer translates from the symbolic domains of gloss or text into continuous sign pose sequences. Gloss2Pose [51] employs transformer encoders to yield pose predictions, with these encoders formulating representations that assimilate both the spatial and temporal dimensions of text and corresponding poses. Text2Gloss2Pose2Video [57] initially translates spoken language sentences into sign gloss sequences through an encoder-decoder network. Following this, a data-driven approach is applied to map the gloss sequences onto skeletal sequences, which then conditions a generative model that synthesizes realistic sign language video sequences. Recently, Saunders et al. [50] employ an encoder-decoder transformer to translate text to gloss and introduce a frame selection network that refines the temporal coherence of the sign sequences. The process culminates with a pose-conditioned human synthesis model that produces photo-realistic sign language videos directly from the skeletal poses.

2.2 Video Generation

The realm of automatic image and video generation has been revolutionized by deep learning, with several neural network-based architectures emerging as frontrunners. For example, Van den Oord et al. [33] bring PixelRNNs to the fore, focusing on the sequential creation of image pixels. The combination of **VAEs** and **GANs**, as StyleGAN [24], harnesses both stability and discriminative capabilities, making it particularly suitable for tasks like generating visuals of individuals and, notably, individuals performing sign language [3, 43, 44, 52, 57, 60]. While these methodologies have ushered in advancements, challenges like mode collapse, non-convergence, and instability persist. However, innovative solutions, including the design of appropriate network architecture and the fine-tuning of objective functions, are continuously proposed to mitigate these issues. Furthermore, the advent of video **Diffusion** models [8, 13, 15, 61, 71] has opened new avenues in the field of video generation, promising more realistic and coherent video sequences. While several recent studies [2, 17, 66] have explored the application of the diffusion model in generating poses and gestures, they fall short in producing realistic sign language videos, as shown in Fig.1 Text2Pose. A significant challenge is that when converting pose data to a digital human model, important contextual details, particularly facial expressions, are lost. In our work, we propose a novel end-to-end diffusion-based framework specifically designed for text-to-sign language video generation.

3 Method

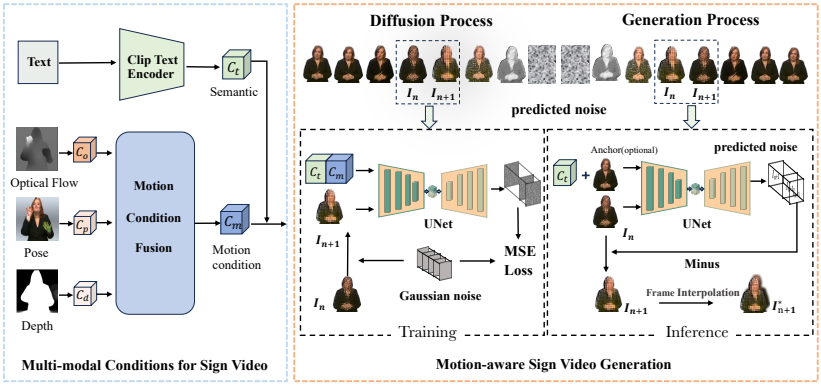

Fig. 2: The Overview of SignGen. First, a video is decomposed into two types of conditions, including text conditions and motion conditions. Then, we feed these conditions into the motion condition fusion module or the CLIP module to embed control signals. Finally, the resulting conditions are leveraged to jointly guide VLDMs for denoising, and then are input to a video frame interpolation model.

In this section, we will comprehensively present SignGen to showcase how it can end-to-end generate sign language video that strictly aligns the meaning of a given text. Firstly, we briefly introduce Video synthesis Latent Diffusion Models (VLDMs) and the guidance directions upon which SignGen is designed. Subsequently, we delve into the details of SignGen's architecture, including the multi-modal conditions for sign video and motion-aware sign video generation, as illustrated in Fig.2.

3.1 Video synthesis with Latent Diffusion Models

Latent Diffusion Models (LDMs) [39], an efficient variant of the diffusion models [21], leverage the diffusion process within the latent space instead of the traditional RGB space. The core of LDMs lies in two primary components. The first component is an encoder, denoted as $\mathcal{E}$, which compresses an image $\boldsymbol{x}$ into a latent code $\boldsymbol{z} = \mathcal{E}(\boldsymbol{x})$. This is complemented by a decoder, responsible for reconstructing the image from the latent code, approximating $\boldsymbol{x} \approx \mathcal{D}(\boldsymbol{z})$. The second component involves learning the distribution of these latent image codes, $\boldsymbol{z}_0 \sim \mathbb{P}_{data}(\boldsymbol{z}_0)$, following the Denoising Diffusion Probabilistic Models (DDPM) framework [21], which includes both forward and reverse processes. In our work, we utilize a modified LDM tailored for sign video generation, operating within the latent space to ensure enhanced local fidelity and preservation of the visual manifold's integrity.

Represent Video in Latent Space. To efficiently process video data, we follow LDMs by introducing a pre-trained encoder [16] to project the given video $\boldsymbol{x} \in \mathbb{R}^{F\times H\times W\times 3}$ into a more compact latent space representation $\boldsymbol{z} = \mathcal{E}(\boldsymbol{x})$, where $\boldsymbol{z} \in \mathbb{R}^{F\times h\times w\times c}$. A corresponding decoder $\mathcal{D}$ then maps these latent representations back to the original pixel space, resulting in $\tilde{\boldsymbol{x}} = \mathcal{D}(\boldsymbol{z})$. We have chosen specific dimensions with $channels = 8$ and a downsampling ratio of $H/h = W/w = 32$ for efficient processing. This setup strikes a balance between reducing data size for computational efficiency and retaining sufficient detail for accurate reconstruction.

Diffusion models in Latent Space. To accurately model the video distribution $\mathbb{P}(x)$, diffusion models [21,55] focus on denoising normally-distributed noise to reconstruct realistic visual content. This process effectively reverses a Markov Chain of length T, typically set to 1000 steps for optimal balance between quality and computational tractability. In our approach, the reverse diffusion process commences with the injection of noise into the latent representation $\boldsymbol{z}$, forming a noise-corrupted version $\boldsymbol{z}_t$ [39]. Following this, a denoising function $\boldsymbol{\epsilon}_\theta(\cdot,\cdot,t)$ is applied to $\boldsymbol{z}_t$ along with the selected conditions $\boldsymbol{c}$, iteratively refining the latent representation across the sequence $t \in \{1,...,T\}$. The optimized objective can be formulated as:

$$\mathcal{L}_{VLDM} = \mathbb{E}_{\mathcal{E}(x),\boldsymbol{\epsilon}\in\mathcal{N}(0,1),\boldsymbol{c},t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t,\boldsymbol{c},t)\|_2^2\right]. \tag{1}$$

To exploit the inductive bias of locality and temporal inductive bias of sequentiality during denoising, we instantiate $\boldsymbol{\epsilon}_\theta(\cdot,\cdot,t)$ as a U-Net augmented with temporal convolution and cross-attention mechanism following [1,22,40].

Conditional Data Sampling. Classifier-free guidance is most widely employed in recent works [32,36,41] for conditional data sampling from a diffusion model, where the predicted noise is adjusted via:

$$\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t,\mathbf{c}) = \omega\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,\mathbf{c}) + (1-\omega)\boldsymbol{\epsilon}_\theta(\mathbf{x}_t), \tag{2}$$

where $\mathbf{x}_t = a_t\mathbf{x}_0 + \sigma_t\boldsymbol{\epsilon}$, and ω is a guidance weight. DDIM [56] is often adopted to speed up the sampling process of diffusion models.

3.2 Multi-modal Conditions for Sign Video

We decompose sign language videos into two distinct modalities of conditions, namely text and motion. Specifically, the text condition specifies the semantic content of sign video, the motion information consists of optical flow, pose, and depth, which can jointly determine the spatial and temporal patterns in the sign language video.

Text Condition. To capture the essence of videos from textual descriptions, focusing on their visual content and motion, we leverage the OpenCLIP ViT-H/14 [35] text encoder to extract semantic embeddings from text, effectively linking language to visual elements.

Motion Condition. To accomplish finer control along the temporal dimension, we introduce the motion condition, which consists of three temporal information:

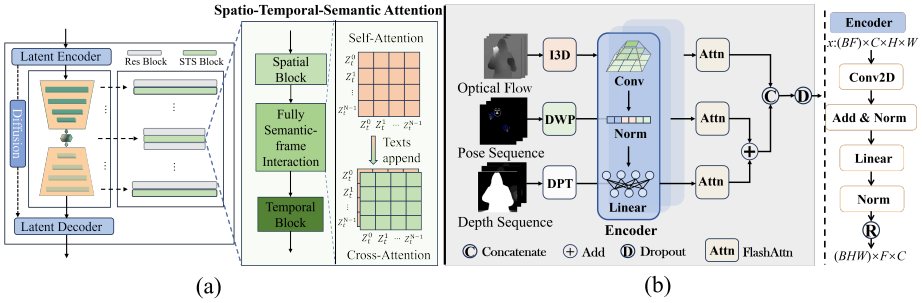

Fig. 3: (a) Motion-aware Sign Video Generation and (b) Motion Condition Fusion Module.

• *Optical Flow*: Optical flow is a valuable representation for analyzing motion in video data as it encodes the direction and magnitude of motion over time at a pixel-level granularity. We extract this information using a pre-trained I3D model [7].

• *Depth Sequence*: Depth maps capture the intrinsic 3D geometry of scenes and provide per-pixel depth values. This additional structure from depth enables more accurate reconstruction of spatial configurations between objects. We extract the depth information using a pre-trained DPT [37] model.

• *Pose Sequence*: To enable finer-grained spatial control, we utilize pose information as an additional conditioning input. Pose maps effectively capture key articulations of objects at the pixel-level, especially the delicate features of faces and hands. We extract pose representations from input frames using a pre-trained image encoder from DWPose [67].

To fuse these multi-modal cues, we propose a Motion Condition Fusion (MCF) module as in Fig.3. It takes as input the extracted features from the optical flow, depth, and pose. Rather than simply concatenating the features, the encoder learns to combine the complementary information through a series of fusion operations. Precisely, a lightweight 2D convolutional network to extract local spatial features, Batch normalization and linear projection are then applied to regularize and project the convolutional features into a lower-dimensional embedding space, capturing the localized appearance. To model long-range temporal dependencies lacking in per-frame representations, we introduce a FlashAttention [11] module incorporating a non-recurrent self-attentive mechanism. For textual conditions provided as embeddings, cross-attention is used to inject the guidance. By explicitly embedding motion cues via motion condition fusion, our method facilitates a unified conditioning interface to enhance temporal coherence for diverse inputs. After processing all conditions, their outputs are fused via element-wise addition and concatenated with the target latent code z_t as control signals.

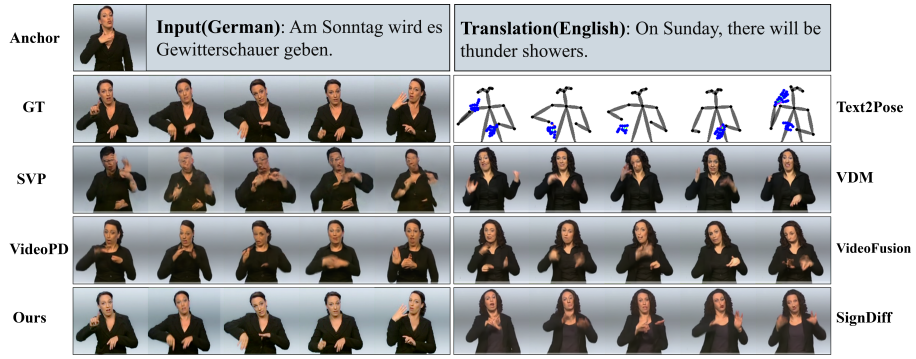

Fig. 4: The key frames generated by various baseline models with the same "anchor" image and corresponding text.

3.3 Motion-aware Sign Video Generation

As illustrated in Fig.3, we design a novel structure termed motion-aware sign video generation towards consistent and efficient video generation. Firstly, we improve the 2D-Unet architecture backbones [39] by reforming the input representation. Specifically, we integrate the spatial and temporal dimensions into a single channel, allowing the convolutional operations to capture the spatiotemporal relations better and ensure appearance consistency with less quality degradation. Secondly, we adopt a spatio-temporal-semantic attention module. It models the dependencies across frames as well as the correlations between visual content and linguistic semantics. Finally, we incorporate a fully semantic-frames attention mechanism to align better the semantics conveyed in text with the visual patterns in sign videos. It helps generate videos with more faithful reproduction of linguistic and semantic information.

Spatio-Temporal-Semantic Attention. Modeling the temporal coherence across frames and establishing precise correspondence between video frames and text is crucial for sign language video generation. Specifically, sign motions at semantic switches and certain actions like greetings exhibit consistent patterns across sentences due to similar sentence structures. Additionally, variations in hand shape, size, and linguistic abilities across individuals necessitate considering relationships between frames as well as alignment between frames and text. We propose the Spatio-Temporal-Semantic Attention to the standard U-Net model by incorporating various building blocks to better capture spatio-temporal dependencies. A series of structural blocks in the downsampling and upsampling path are presented, forming a nested Res-Spatial-Temporal-Res structure. Moreover, we utilize the Fully Semantic-frame Interaction to maintain visual coherence across frames and interaction to jointly establish frame-text correspondence during reconstruction.

Fully Semantic-frame Interaction. Leveraging the controllability of MCF, motion sequences could provide coarse-level consistency in structure. Nonetheless, even using the same initial noise, individually producing all frames will lead

to drastic inconsistencies in appearance. To keep the video appearance coherent, we concatenate all video frames to become a "large image", so that their content could be shared via inter-frame interaction. Considering that self-attention in SD [39] is driven by appearance similarities [65], we propose to enhance the holistic coherency by adding attention-based Fully Semantic-frame Interaction. As shown in Fig.3, it extends self-attention by adding interaction across all frames: $\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d}}) \cdot \boldsymbol{V}$, where $\boldsymbol{Q} = \boldsymbol{W}^Q \boldsymbol{z}_t$, $\boldsymbol{K} = \boldsymbol{W}^K \boldsymbol{z}_t$, $\boldsymbol{V} = \boldsymbol{W}^V \boldsymbol{z}_t$, $\boldsymbol{z}_t = \{\boldsymbol{z}_t^i\}_{i=0}^{N-1}$ denotes all latent frames at timestep t, while $\boldsymbol{W}^Q$, $\boldsymbol{W}^K$, and $\boldsymbol{W}^V$ project $\boldsymbol{z}_t$ into query, key, and value, respectively.

We propose a two-step approach to effectively fusing linguistic information from input text with visual patterns in sign videos. First, self-attention is applied in all video frames. Then, we perform cross-attention between encoded text features and video features to guide the generation. Specifically, we modify the standard cross-attention module by defining: Q as the encoded text features. K and V are separately defined as the concatenated features of video and text, where the weights for video and text features set to 0.4 and 0.6, respectively. With this design, the encoder can learn the correspondence between video and text. During decoding, the generation is constrained by video structure and guided by text simultaneously, leading to higher quality and more semantically accurate results.

3.4 Video Frame Interpolation

While the Motion-aware Sign Video Generation and Motion Condition Fusion Module effectively preserve the realistic representation of the entire body and facial expressions, the generated videos may still exhibit frame-level flickering issues [30]. To address frame flickering, we introduce a video frame interpolation model [38] that refines pixel space $\bar{\boldsymbol{x}}$ for smooth transitions between frames. Due to space constraints, more experimental details can be found in the supplementary materials.

3.5 Training and inference

As a conditional video diffusion model, our method takes prompt text c_t, an anchor image I, and motion condition c_m (optional) as input to predict the noise at the current step. The anchor image provides prior knowledge (i.e., signer, pose, and depth) for generation. Without an anchor image, our model could run as well, but the signer will be generated randomly and the semantic consistency will be reduced, as shown in Fig.6(a). The motion condition c_m could consist of optical flow c_o, pose c_p, and depth c_d information. During the training, we adhere to Composer [23], using a probability of 0.1 to keep all conditions, a probability of 0.1 to discard all conditions, and an independent probability of 0.5 to keep or discard a specific condition. Therefore, in inference, even if certain conditions are absent, we can still generate high-quality videos with the text and anchor image as input. More details can be found in Fig.7.

Table 1: Back translation results for the Text2Pose task.

Method	Extra Data	DEV					TEST				
		ROUGE-L↑	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	ROUGE-L↑	BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑
ProTran [45]	n/a	34.05	33.12	20.71	14.71	11.43	31.07	29.74	17.62	12.53	9.68
MCG [47]	n/a	33.68	31.84	20.58	15.61	12.65	32.74	30.93	18.99	13.72	10.81
MOMP [49]	n/a	37.76	35.23	23.49	17.50	14.03	36.77	35.89	23.27	16.86	13.30
FS-NET [50]	dictionary	40.94	-	-	-	19.14	40.60	-	-	-	18.78
`SignGen`	n/a	45.00	43.71	31.56	25.63	20.21	44.56	45.18	30.68	24.31	19.71

Table 2: Video generation results on different datasets for 10 predicted key frames, given inputs of text and anchor image (signer+pose+depth).

Dataset	**Method**	VDM [22]	SVP [27]	VideoLDM [5]	VideoPD [68]	VideoFusion [31]	SignDiff [17]	`SignGen`
AUTSL	FVD↓	1638	2512	1836	1213	1678	1169	**556**
	SSIM↑	0.49	0.12	0.43	0.29	0.39	0.33	**0.93**
	PSNR↑	13.37	8.36	11.29	18.21	12.57	18.82	**27.53**
	LPIPS↓	0.28	0.79	0.29	0.82	0.31	0.52	**0.03**
RWTH-2014	FVD↓	1289	1317	1301	1132	1032	930	**579**
	SSIM↑	0.67	0.59	0.59	0.69	0.72	0.63	**0.73**
	PSNR↑	15.58	13.27	12.30	19.27	14.53	16.22	**20.22**
	LPIPS↓	0.16	-	0.23	0.31	0.18	0.21	**0.03**
RWTH-2014T	FVD↓	1445	1300	1267	659	896	661	**640**
	SSIM↑	0.60	0.51	0.61	0.71	0.67	0.70	**0.89**
	PSNR↑	14.11	11.20	14.21	20.31	14.80	21.23	**24.84**
	LPIPS↓	0.22	-	0.21	0.16	0.17	0.12	**0.01**
CSL-Daily	FVD↓	1763	2411	1931	1682	1802	1326	**424**
	SSIM↑	0.21	0.09	0.23	0.11	0.27	0.25	**0.93**
	PSNR↑	8.23	7.61	9.27	13.56	11.34	18.92	**34.92**
	LPIPS↓	0.95	0.84	0.83	0.67	0.56	0.54	**0.08**
WLASL	FVD↓	1782	2386	1842	1217	1324	972	**493**
	SSIM↑	0.19	0.11	0.27	0.32	0.32	0.57	**0.91**
	PSNR↑	8.11	9.73	10.56	15.76	15.24	17.63	**32.72**
	LPIPS↓	0.92	0.83	0.39	0.46	0.31	0.27	**0.11**

4 Experiment

4.1 Experimental Settings

Datasets. To demonstrate the robustness of the model across various backgrounds and signers. We conducted experiments on five distinct datasets. The RWTH-2014 [26] and RWTH-2014-T [19] datasets feature video recordings from German weather forecasts, containing 6,842 sentences and 1,295 sign words by 9 signers. The AUTSL [54] dataset is a comprehensive multi-modal collection featuring 36,302 isolated Turkish sign language video samples of 226 signs by 43 signers, set against 20 diverse backgrounds. The CSL-Daily [72] is a comprehensive Chinese sign language dataset for continuous sign language translation, encompassing daily life scenarios such as travel, shopping, and medical care, and includes spoken language translations and gloss-level annotations from 10 signers. The WLASL [28] (Word-Level American Sign Language) video dataset contains more than 2000 words performed by over 100 signers.

Baselines. We evaluate our model against six publicly available approaches, distinguishing between methods based on their foundational concepts. Among these, one approach [27] is based on the idea of GAN. The remaining five [5, 17,

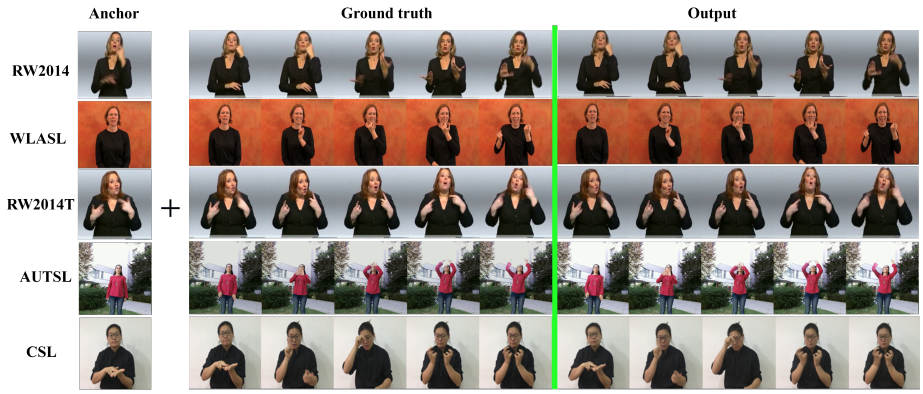

Fig. 5: We randomly showcase the outcomes generated by an "anchor" image across five distinct datasets, illustrating the versatility and consistency of `SignGen` in diverse settings.

Table 3: Ablation study results showcasing different conditions with provided "anchor" image on the RWTH-2014 dataset.

Conditions			FVD	SSIM	PSNR	LPIPS	Dev		Test	
Optical-Flow	Pose	Depth					ROUGE-L	BLEU-4	ROUGE-L	BLEU-4
✓			726	0.49	14.20	0.33	39.55	16.34	36.35	12.45
✓	✓		628	0.60	16.25	0.15	42.37	17.91	39.79	15.13
✓	✓	✓	**579**	**0.73**	**20.22**	**0.03**	**45.00**	**20.21**	**45.18**	**19.71**

22, 31, 68] are primarily based on diffusion processes, demonstrating the recent shift towards diffusion-based frameworks in video synthesis. Specifically, we include SignDiff [17], a novel approach that utilizes a ControlNet [69] architecture, highlighting its unique position in the context of controlled generation.

Evaluation Metrics. We utilize three types of metrics to evaluate `SignGen`: *i)* To evaluate the semantic consistency between the generated sign language video and the original text input, we follow ProTran [44] to use the back translation metric to evaluate the accuracy of SLG, which uses a pre-trained SLT model [6] to translate sign language back to text and then calculates the BLEU [34] and ROUGE-L [29] scores between the generated text and the original text. *ii)* To evaluate the video quality of generated sign language content, we utilize PSNR [18], SSIM [62], and LPIPS [70]. These metrics quantify the similarity and structural integrity between the generated video and ground truth, emphasizing the precise reconstruction of motion details. *iii)* To evaluate the diversity of generated sign video, we use the FVD [59] to gauge the diversity in generated sequences, ensuring they reflect the variable temporal patterns seen in natural sign languages.

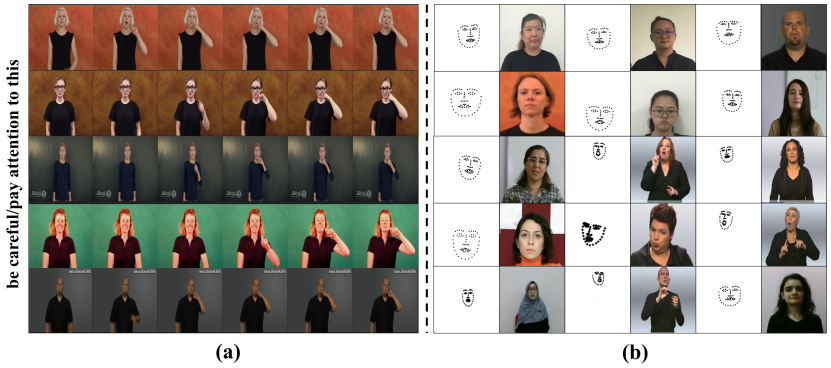

Fig. 6: (a) The generated video of `SignGen` with only **`text`** input is used for inference. (b) The detailed facial expressions generated when facial key points are provided.

4.2 Qualitative and quantitative comparisons

To qualitatively assess our approach against leading contemporary approaches, we present keyframes from sign language sequences synthesized using various visual cues in Fig.4. As illustrated, our model accurately renders natural finger and palm distributions and captures the nuanced movements characteristic of sign language, including subtle facial features like eyebrow movements, mouth shapes, and head poses, indicating our approach's capability to produce temporally coherent sequences that effectively capture the dynamics of sign language. In contrast, baseline approaches often produce ambiguous hand shapes and lack detailed articulation.

Semantic Consistency. In line with prior studies, we evaluate our Text2Pose task and compare it with state-of-the-art approaches, treating the generated video as a skeleton for back translation metrics (Tab.1). Our approach significantly surpasses the continuous SLG method [46] and outperforms the FS-NET [50], which relies on additional high-quality isolated signs from sign language lexicons. Compared with previous work [44, 48] that employs a cascaded pipeline from Text2Gloss and Gloss2Pose, our method achieves better quantitative results. We attribute the improvement to the avoidance of middle information loss. The Spatio-Temporal-Semantic Attention and Fully Semantic-frame Interaction further ensure the semantic consistency between generated video and input text.

Video Quality. As illustrated in Fig.4 and Fig.5, our method outperforms other sign language generation techniques in quantitative metrics. Specifically, the groundtruth exhibits motion blurs in hand regions. Competing baselines either fail to generate clear hands or introduce artifacts. In contrast, our approach consistently generates sharp, accurate hand poses and motions. The proposed method significantly surpasses six baseline approaches in PSNR and SSIM metrics, as detailed in Tab.2, indicating superior video quality in our generated sign language. Furthermore, with notably lower LPIPS and higher PSNR scores

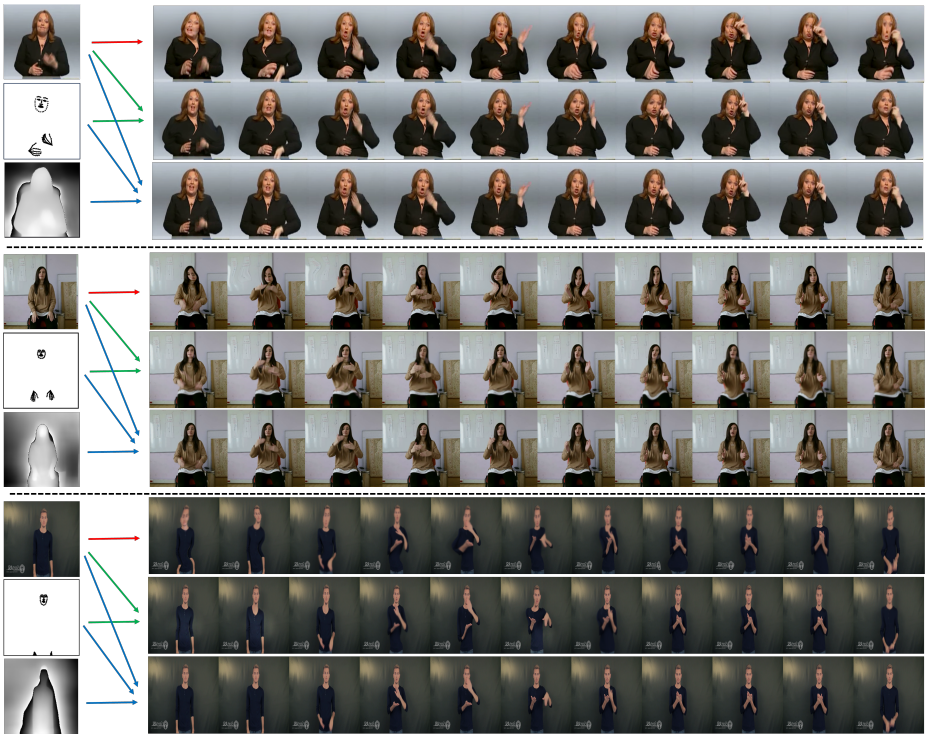

Fig. 7: We present the generated results under three different input conditions for the model: (i) text + anchor, (ii) text + anchor + pose, and (iii) text + anchor + pose + depth.

compared to the other six baselines, our approach demonstrates greater semantic consistency with the ground truth, effectively capturing the intricate content structure of sign language. This superior performance in both visual quality and semantic fidelity will further bridge the communication between signers and non-signers. As demonstrated in Fig.6(b), by extracting facial poses, our approach is guided to focus more on emotional expressions and subtle facial changes during training, thereby enabling it to generate complex and nuanced **facial expressions** during inference.

Generation Diversity. Diversity metrics, particularly FVD, provide meaningful insights into generation quality when corroborated by PSNR/SSIM scores, as evidenced in Tab.2. Our method excels in both quality (PSNR, SSIM) and FVD, maintaining FVD within an optimal range, signifying robust dynamic diversity, and effective one-to-many mapping in sign language generation. In the absence of an "anchor" image, inputting identical text results in the generation of distinct signers for each instance as in Fig.6(a). There is a noticeable decline in the semantic consistency of the generated content.

Table 4: Ablation study results on the RWTH-2014 dataset.

Metrics	FVD	SSIM	PSNR	LPIPS	Dev		Test	
					ROUGE-L	BLEU-4	ROUGE-L	BLEU-4
w/o MCSV	1014	0.54	10.20	0.53	36.55	13.34	36.35	12.45
w/o MSVG	631	0.61	19.11	0.15	41.23	17.32	40.02	13.65
`SignGen`	**579**	**0.73**	**20.22**	**0.03**	**45.00**	**20.21**	**45.18**	**19.71**

4.3 Ablation Study

Effect of the Proposed Modules. Tab.4 shows the efficacy of our proposed modules in enhancing sign language video generation. The Multi-modal Conditions for Sign Video module, integrating text, optical flow, depth, and pose inputs, offers a richer understanding of cross-modal semantics. This leads to a more accurate grounding of text in the video generation process. The Motion-aware Sign Video Generation module, deviating from conventional U-Net architecture, employs optical flow and 3D pose data to model motion dynamics. This focus on temporal aspects, rather than just frame appearance, results in more semantically aligned signing video, closely matching the text descriptions. In essence, the integration of multi-modal information and emphasis on motion dynamics address the challenge of semantic ambiguity in translation. These combined innovations enhance the quality of text-to-video grounding, ensuring the generated content faithfully mirrors the complexity of sign language.
Effect of Different Conditions. To further analyze the impact of additional modalities, we conduct an ablation study where pose, optical flow, depth, and text features are progressively incorporated into the multi-modal encoder. The results in Tab.3 validate that introducing more visual modalities significantly improves generation quality over a text-only baseline. Fig.7 clearly illustrates the effect on video generation when both pose and depth conditions are provided. The stark contrast between the two underscores the essential role of multi-modality in enhancing generation quality. In summary, this ablation study presents empirical evidence that multi-modal modeling is important for sign language generation through the collective extraction of semantic meanings encoded across visual and linguistic data.

5 Conclusion

In this paper, we introduce `SignGen`, an innovative diffusion model for natural sign language generation from text. Our approach has demonstrated superior performance on benchmark datasets in terms of semantic consistency, naturalness, and expressiveness. While `SignGen` offers promising results, it is essential to acknowledge the need to accommodate these dialectical variations. Current models may not cover all dialects, potentially leading to inaccuracies or limited applicability in specific communities. As we look ahead, our future work will focus on resolving this challenge and ensuring that sign language generation becomes more inclusive and accessible across linguistic variations.

Acknowledgements

This work was supported by NSFC (No.62206200, 62206137, 62036012, 62376196, U23A20387), and Tianjin Natural Science Foundation (No.22JCQNJC00940, 22JCYBJC00030).

References

1. Text to video synthesis in modelscope (2023), `https://modelscope.cn/models/damo/text-to-video-synthesis/summary`
2. Ao, T., Zhang, Z., Liu, L.: Gesturediffuclip: Gesture diffusion model with clip latents. ACM Trans. Graph. **42**(4) (2023)
3. Arkushin, R.S., Moryossef, A., Fried, O.: Ham2pose: Animating sign language notation into pose sequences pp. 21046–21056 (2023)
4. Bangham, J.A., Cox, S., Elliott, R., Glauert, J.R., Marshall, I., Rankov, S., Wells, M.: Virtual signing: Capture, animation, storage and transmission-an overview of the visicast project. In: IEE Seminar on speech and language processing for disabled and elderly people (Ref. No. 2000/025). pp. 6–1. IET (2000)
5. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models pp. 22563–22575 (2023)
6. Camgoz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation (2020)
7. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset (2018)
8. Chen, W., Wu, J., Xie, P., Wu, H., Li, J., Xia, X., Xiao, X., Lin, L.: Control-a-video: Controllable text-to-video generation with diffusion models. arXiv preprint arXiv:2305.13840 (2023)
9. Cooper, H., Holt, B., Bowden, R.: Sign language recognition. In: Visual Analysis of Humans: Looking at People, pp. 539–562. Springer (2011)
10. Cox, S., Lincoln, M., Tryggvason, J., Nakisa, M., Wells, M., Tutt, M., Abbott, S.: Tessa, a system to aid communication with deaf people. In: ASSETS. pp. 205–212. ACM (2002)
11. Dao, T., Fu, D.Y., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness (2022)
12. Doe, J.: Example website title. `https://www.who.int/health-topics/hearing-loss` (2022), accessed: 2023-10-26
13. Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., Giro-i Nieto, X.: How2sign: a large-scale multimodal dataset for continuous american sign language pp. 2735–2744 (2021)
14. Efthimiou, E., Fotinea, S., Hanke, T., Glauert, J.R.W., Bowden, R., Braffort, A., Collet, C., Maragos, P., Lefebvre-Albaret, F.: The dicta-sign wiki: Enabling web communication for the deaf. In: ICCHP (2). Lecture Notes in Computer Science, vol. 7383, pp. 205–212. Springer (2012)
15. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models pp. 7346–7356 (2023)
16. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR. pp. 12873–12883 (2021)

17. Fang, S., Sui, C., Zhang, X., Tian, Y.: Signdiff: Learning diffusion models for american sign language production. arXiv preprint arXiv:2308.16082 (2023)
18. Fardo, F.A., Conforto, V.H., de Oliveira, F.C., Rodrigues, P.S.: A formal evaluation of PSNR as quality measurement parameter for image segmentation algorithms (2016)
19. Forster, J., Schmidt, C., Koller, O., Bellgardt, M., Ney, H.: Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-weather (2014)
20. Gibet, S., Lefebvre-Albaret, F., Hamon, L., Brun, R., Turki, A.: Interactive editing in french sign language dedicated to virtual signers: Requirements and challenges. Universal Access in the Information Society **15**, 525–539 (2016)
21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020)
22. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. arXiv preprint arXiv:2204.03458 (2022)
23. Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: Creative and controllable image synthesis with composable conditions. arXiv preprint arXiv:2302.09778 (2023)
24. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
25. Kipp, M., Heloir, A., Nguyen, Q.: Sign language avatars: Animation and comprehensibility. In: Intelligent Virtual Agents: 10th International Conference, IVA 2011, Reykjavik, Iceland, September 15-17, 2011. Proceedings 11. pp. 113–126. Springer (2011)
26. Koller, O., Forster, J., Ney, H.: Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. Computer Vision and Image Understanding **141**, 108–125 (2015)
27. Lee, A.X., Zhang, R., Ebert, F., Abbeel, P., Finn, C., Levine, S.: Stochastic adversarial video prediction. arXiv preprint arXiv:1804.01523 (2018)
28. Li, D., Rodriguez, C., Yu, X., Li, H.: 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. pp. 1459–1469 (2020)
29. Lin, C.Y., Och, F.J.: Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics (2004)
30. Liu, Y., Zhao, H., Chan, K.C.K., Wang, X., Loy, C.C., Qiao, Y., Dong, C.: Temporally consistent video colorization with deep feature propagation and self-regularization learning (2021)
31. Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., Tan, T.: Videofusion: Decomposed diffusion models for high-quality video generation pp. 10209–10218 (2023)
32. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In: International Conference on Machine Learning (2021)
33. Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al.: Conditional image generation with pixelcnn decoders. Advances in neural information processing systems **29** (2016)
34. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation (2002)

35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
36. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. ArXiv **abs/2204.06125** (2022)
37. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction (2021)
38. Reda, F., Kontkanen, J., Tabellion, E., Sun, D., Pantofaru, C., Curless, B.: Film: Frame interpolation for large motion pp. 250–266 (2022)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
40. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)
41. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, S.K.S., Ayan, B.K., Mahdavi, S.S., Lopes, R.G., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. ArXiv **abs/2205.11487** (2022)
42. Saunders, B., Camgöz, N.C., Bowden, R.: Everybody sign now: Translating spoken language to photo realistic sign language video. CoRR **abs/2011.09846** (2020)
43. Saunders, B., Camgoz, N.C., Bowden, R.: Everybody sign now: Translating spoken language to photo realistic sign language video. arXiv preprint arXiv:2011.09846 (2020)
44. Saunders, B., Camgoz, N.C., Bowden, R.: Progressive transformers for end-to-end sign language production. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 687–705. Springer (2020)
45. Saunders, B., Camgöz, N.C., Bowden, R.: Progressive transformers for end-to-end sign language production. CoRR **abs/2004.14874** (2020)
46. Saunders, B., Camgoz, N.C., Bowden, R.: Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks. International journal of computer vision **129**(7), 2113–2135 (2021)
47. Saunders, B., Camgöz, N.C., Bowden, R.: Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks. CoRR **abs/2103.06982** (2021)
48. Saunders, B., Camgoz, N.C., Bowden, R.: Mixed signals: Sign language production via a mixture of motion primitives. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1919–1929 (2021)
49. Saunders, B., Camgöz, N.C., Bowden, R.: Mixed signals: Sign language production via a mixture of motion primitives. CoRR **abs/2107.11317** (2021)
50. Saunders, B., Camgoz, N.C., Bowden, R.: Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5141–5151 (2022)
51. Shalev-Arkushin, R., Moryossef, A., Fried, O.: Ham2pose: Animating sign language notation into pose sequences. In: CVPR. pp. 21046–21056. IEEE (2023)
52. Siarohin, A., Sangineto, E., Lathuiliere, S., Sebe, N.: Deformable gans for pose-based human image generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3408–3416 (2018)
53. Sincan, O.M., Junior, J., Jacques, C., Escalera, S., Keles, H.Y.: Chalearn lap large scale signer independent isolated sign language recognition challenge: Design, re-

sults and future research. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3472–3481 (2021)
54. Sincan, O.M., Keles, H.Y.: Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. IEEE Access **8**, 181340–181355 (2020). https://doi.org/10.1109/ACCESS.2020.3028072
55. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICCV. pp. 2256–2265 (2015)
56. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. ArXiv **abs/2010.02502** (2020)
57. Stoll, S., Camgöz, N.C., Hadfield, S., Bowden, R.: Sign language production using neural machine translation and generative adversarial networks. In: Proceedings of the 29th British Machine Vision Conference (BMVC 2018). British Machine Vision Association (2018)
58. Stoll, S., Camgöz, N.C., Hadfield, S., Bowden, R.: Text2sign: Towards sign language production using neural machine translation and generative adversarial networks. Int. J. Comput. Vis. **128**(4), 891–908 (2020)
59. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & (2018)
60. Vasani, N., Autee, P., Kalyani, S., Karani, R.: Generation of indian sign language by sentence processing and generative adversarial networks. In: 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). pp. 1250–1255. IEEE (2020)
61. Voleti, V., Jolicoeur-Martineau, A., Pal, C.: Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. Advances in Neural Information Processing Systems **35**, 23371–23385 (2022)
62. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity (2004)
63. Weast, T.P.: Questions in American Sign Language: A quantitative analysis of raised and lowered eyebrows. The University of Texas at Arlington (2008)
64. Wei, F., Chen, Y.: Improving continuous sign language recognition with cross-lingual signs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23612–23621 (2023)
65. Wu, J.Z., Ge, Y., Wang, X., Lei, W., Gu, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. arXiv preprint arXiv:2212.11565 (2022)
66. Xie, P., Zhang, Q., Li, Z., Tang, H., Du, Y., Hu, X.: Vector quantized diffusion model with codeunet for text-to-sign pose sequences generation. arXiv preprint arXiv:2208.09141 (2022)
67. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation (2023)
68. Yu, S., Sohn, K., Kim, S., Shin, J.: Video probabilistic diffusion models in projected latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18456–18466 (2023)
69. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
70. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric (2018)
71. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. arXiv preprint arXiv:2305.13077 (2023)

72. Zhou, H., Zhou, W., Qi, W., Pu, J., Li, H.: Improving sign language translation with monolingual data by sign back-translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1316–1325 (2021)