MaDiS: Taming Masked Diffusion Language Models for Sign Language Generation

Ronglai Zuo, Rolandos Alexandros Potamias, Qi Sun, Evangelos Ververas, Jiankang Deng, Stefanos Zafeiriou

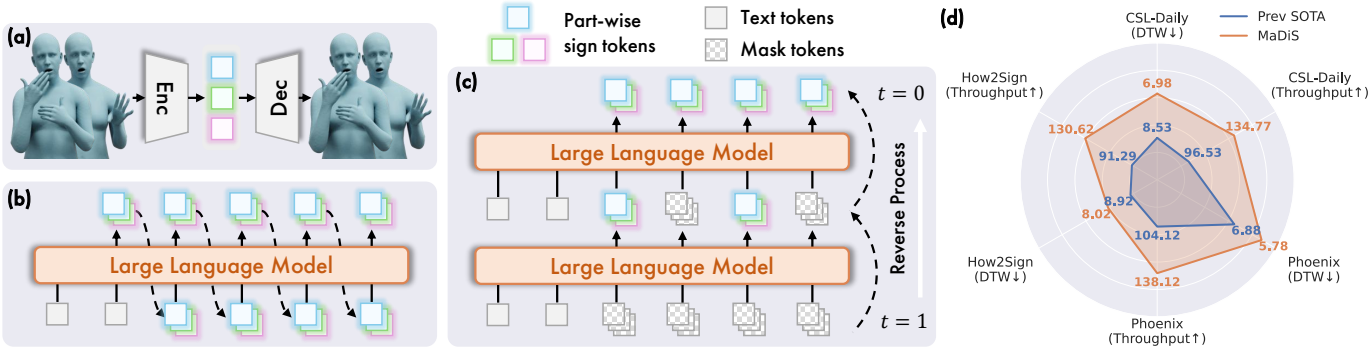

Fig. 1. We propose MaDiS, a novel sign language generation approach built upon masked diffusion language models (MDLMs) [1]. (a) A sign tokenizer discretizes continuous sign motions into part-wise tokens [2]. (b) Conventional autoregressive language models generate tokens in a left-to-right manner, limiting utilization of contexts and inference efficiency. (c) The emerging MDLMs model token distributions with bidirectional contexts and enable parallel multi-token sampling during inference. (d) MaDiS achieves SOTA performance across multiple benchmarks [3]–[5] while delivering a 40% higher throughput.

***Abstract*—Sign language generation (SLG) aims to translate written texts into expressive sign motions, bridging communication barriers for the Deaf and Hard-of-Hearing communities. Recent studies formulate SLG within the language modeling framework using autoregressive language models, which suffer from unidirectional context modeling and slow token-by-token inference. To address these limitations, we present MaDiS, a masked-diffusion-based language model for SLG that captures bidirectional dependencies and supports efficient parallel multi-token generation. We further introduce a tri-level cross-modal pretraining scheme that jointly learns from token-, latent-, and 3D physical-space objectives to leverage complementary, multi-level sign representations. To accelerate model convergence in the fine-tuning stage, we design a novel unmasking strategy with temporal checkpoints, which restructures generation in a coarse-to-fine manner and reduces the combinatorial complexity of unmasking orders by over 10^{41} times. In addition, a mixture-of-parts embedding layer is developed to effectively fuse information stored in different part-wise sign tokens through a learnable gate and well-optimized codebooks. Extensive experiments on CSL-Daily, Phoenix-2014T, and How2Sign demonstrate that MaDiS achieves superior performance across multiple metrics, including DTW error and two new complementary metrics, SiBLEU and SiCLIP, while delivering a 40% higher throughput. Code, models, and demos are available on the project page: https://2000zrl.github.io/madis/.**

***Index Terms*—Sign Language Generation, Diffusion Language Model, Sign Language Pretraining**

R. Zuo, R.A. Potamias, Q. Sun, E. Ververas, J. Deng, and S. Zafeiriou are with Imperial College London, London SW7 2AZ, United Kingdom. E-mail: {r.zuo; r.potamias; qi.sun24; e.ververas16; j.deng16; s.zafeiriou}@imperial.ac.uk.

Corresponding author: Jiankang Deng

I. Introduction

Sign language is the primary communication mode for Deaf and Hard-of-Hearing (DHH) individuals, yet in predominantly oral societies they are often encouraged to rely on text instead of their preferred sign language [6]. Research in sign language translation (SLT, sign-to-text) and generation (SLG, text-to-sign) aims to improve the accessibility of sign language, thereby mitigating the communication gap between the hearing and DHH communities. While SLT has seen significant progress [7]–[15], SLG remains relatively under-explored.

Early SLG approaches typically rely on glosses, the written form of signs, as an intermediate representation and formulate SLG as a text-to-gloss-to-sign pipeline [16]–[19]. While glosses encode rich linguistic information and provide explicit guidance for sign generation, the performance of gloss-based methods is constrained by the quality of the text-to-gloss module: since gloss annotations are expensive, gloss-based datasets are often limited in scale, leading to degraded text-to-gloss performance. For example, Gemini-generated glosses exhibit an error rate of over 60% [12]. In addition, such cascaded pipelines incur higher inference latency.

Given these limitations, recent efforts have increasingly adopted gloss-free SLG, directly generating signs from text. Typical works [20]–[23] formulate the task as a visual content generation problem, using either GANs [17] or diffusion models [24]. Recently, motivated by the linguistic character of sign language and the generalizability of large language models (LLMs), a new research direction has emerged that models SLG through tokenization (Fig. 1a) and autoregressive language models (ARLMs) [2], [25]–[28] (Fig. 1b). However, ARLMs have two main limitations: their left-to-right

generation order restricts the use of bidirectional contexts and their one-token-at-a-time decoding introduces an inference bottleneck.

To address such problems, we propose MaDiS (Fig. 1c), a novel SLG approach built upon an emerging and promising language modeling paradigm, the masked diffusion language model (MDLM) [1], [29]–[31]. Unlike conventional ARLMs, MDLMs are based on masked diffusion models [32]–[34], which introduce a forward data masking process and train a mask predictor to approximate the reverse denoising process. This mechanism enables MDLMs to construct model distributions with bidirectional dependencies, allowing them to capture richer contextual information. During the reverse process, multiple tokens can be sampled simultaneously, thereby improving inference efficiency.

Pretraining plays a crucial role in sign language understanding models [7], [35]–[39], yet pretraining for SLG models remains largely unexplored. MDLMs are typically pretrained in the token space using a mask-then-predict framework [1], where a subset of tokens is randomly masked and subsequently predicted by the model. Although such token-space objective is well-established in natural language processing, recent advances in self-supervised learning [40]–[42] have demonstrated that latent-space pretraining is often more effective in a wide range of vision tasks, including image and video understanding [41], [43] and world modeling [44], [45]. Moreover, sign language, as a visual language, conveys most of its semantics through body and hand motions in the 3D physical space. Motivated by the advantages of different representation spaces, we propose a tri-level cross-modal pretraining method in which the model learns to predict masked text and sign tokens, latent features, and 3D sign motions simultaneously. This challenging and comprehensive pretraining objective encourages the model to fully exploit complementary sign representations and yields significant performance improvements in the downstream SLG task.

After pretraining, we perform supervised fine-tuning conditioned on unmasked text tokens. The original MDLM can be viewed as an any-order ARLM [46]. Due to its unconstrained confidence-based unmasking strategy, the number of possible unmasking orders is enormous. For example, generating 100 tokens in 25 steps results in approximately 10^{123} possible orders. Although this property enables MDLMs to perform well in data-constrained scenarios, it also leads to significantly slower convergence, often requiring more training epochs [47]. To address this issue, we propose a novel unmasking strategy with temporal checkpoints. Inspired by the effectiveness of VAR [48], which generates images progressively from low to high resolution, we insert checkpoints composed of uniformly distributed tokens at noise levels 0.75 and 0.5, corresponding to temporal resolutions of $T/4$ and $T/2$. This design decomposes generation in a coarse-to-fine manner and dramatically prunes unmasking orders by over 10^{41} times, thereby accelerating convergence during training. Furthermore, to better integrate part-wise sign tokens [2], we propose a plug-and-play mixture-of-parts (MoP) embedding layer. Unlike the previous SOTA method [2] that simply averages token embeddings from different body parts, our MoP embedding layer is built upon well-optimized VQ-VAE codebooks, and part-wise embeddings are dynamically fused through a learnable gate. In summary, our contributions are as follows:

- We propose MaDiS, the first approach to introduce MDLMs to gloss-free SLG, enabling bidirectional context modeling and parallel multi-token generation.
- A tri-level cross-modal pretraining framework is developed that jointly learns from token, latent, and 3D physical spaces, enabling the model to exploit the benefits from complementary, multi-level sign representations.
- To accelerate training convergence, we design a novel token unmasking strategy with temporal checkpoints, which significantly reduces unmasking order complexity through a coarse-to-fine decomposition. To enhance the integration of part-wise sign tokens, we introduce an MoP embedding layer that is conditioned on VQ-VAE codebooks and controlled by a learnable gate.
- Extensive experiments on widely adopted benchmarks [3]–[5] demonstrate that MaDiS achieves SOTA performance across multiple metrics, including DTW error and two new complementary metrics, SiBLEU and SiCLIP, while enhancing inference throughput by 40% (Fig. 1d).

II. Related Work

A. Sign Language Generation

Sign language generation (SLG), also known as sign language production, aims to translate texts into sign language outputs represented in various forms, such as skeletons [49]–[56], RGB videos [17], [22], [23], [57]–[59], and 3D motions [2], [16], [21], [25], [26], [60], [61]. Together with sign language translation (SLT), SLG enables bidirectional interaction between hearing and Deaf or Hard-of-Hearing individuals, thereby helping to mitigate the communication barriers. Although real-person videos are more realistic, developing end-to-end sign video generation models is challenging due to the high dimensionality of RGB video data. Therefore, we focus on generating sign motions, which can serve as an intermediate representation for subsequent sign video synthesis [20], [62], [63] or be used to control humanoid robots [64], [65].

One line of SLG research relies on glosses, either as intermediate representations [16]–[19], [53], [58], [66], [67] or as additional supervision [49]. For example, Signing-at-Scale [17] and Spoken2Sign [16] adopt a language model for text-to-gloss translation, use the predicted glosses to retrieve sign language dictionaries in the form of videos or motions, and learn a connector to model co-articulations between adjacent signs. However, such two-stage pipelines inevitably introduce an information bottleneck. To mitigate this issue, MoMP [49] uses glosses only as auxiliary supervision for the text-to-gloss module, rather than as a mandatory intermediate representation.

Although glosses provide useful linguistic guidance, their annotations are expensive, resulting in small-scale gloss-based datasets and consequently degraded text-to-gloss performance [12]. Recent works have therefore shifted towards gloss-free methods, *i.e.*, generating signs directly from texts in an end-to-end manner. One line of work formulates SLG as a visual

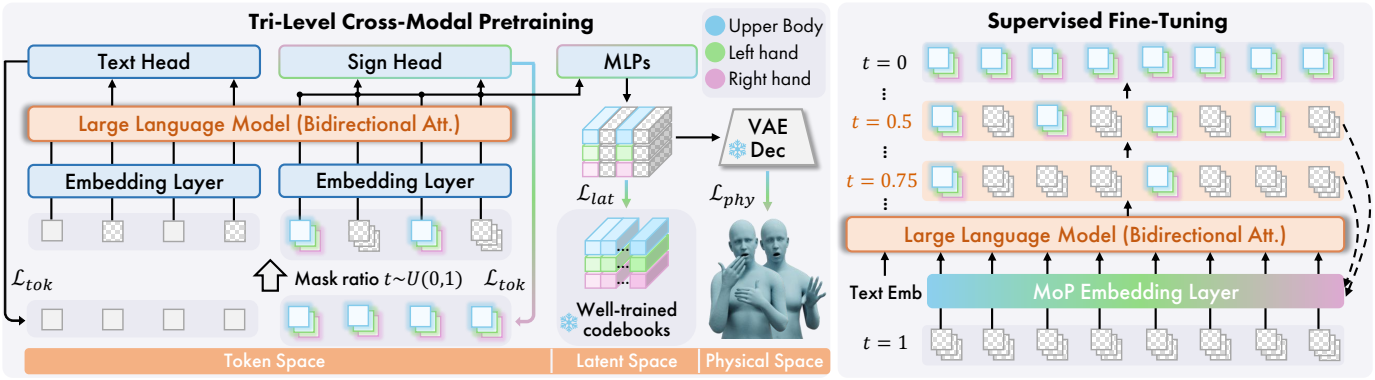

Fig. 2. MaDiS is built upon the emerging masked diffusion language model (MDLM), implemented by modifying a standard decoder-only LLM with bidirectional attention. The MDLM is first pretrained with three objectives ($\mathcal{L}_{tok}$, $\mathcal{L}_{lat}$, $\mathcal{L}_{phy}$) from the token, latent, and 3D physical spaces, respectively. We then fine-tune the model conditioned on text inputs using the proposed temporal-checkpoint unmasking strategy and a dedicated mixture-of-parts sign embedding layer.

generation problem [21], [23], [56], [59], [61], where text features are used as conditions for visual generative models. Motivated by the linguistic nature of sign language, another line of work [2], [25], [26], [28] formulates SLG within a language modeling framework by combining a discrete sign tokenizer [2] with an autoregressive language model (ARLM). Due to the difficulty of gloss-free SLG, some methods further introduce additional information, such as external sign dictionaries [2] or key frames [61].

In this work, we follow the language-model-based paradigm and study gloss-free SLG without relying on any additional information, which represents a more practical and challenging setting. Unlike previous methods based on well-established ARLMs, which are inherently limited by their left-to-right generation process, we explore emerging masked diffusion language models for SLG, which model token distributions bidirectionally and enable efficient parallel generation.

B. Masked Diffusion Language Model

Current LLMs are mostly built upon ARLMs. Although they demonstrate strong capability, their left-to-right generation order limits both efficiency and context modeling. Recently, masked diffusion language models (MDLMs) [1], [31], [68], [69] have emerged as a new language modeling paradigm. Built upon masked diffusion models [32]–[34], MDLMs define a forward data-masking process and train a mask predictor to approximate the reverse denoising process, which enables bidirectional context modeling and allows multiple tokens to be sampled in a single generation step. MDLMs have been successfully applied to diverse domains, including multimodal understanding [29], [30], protein design [70], [71], and robotic manipulation [72], [73]. A related line of work in human motion generation [74]–[77] adopts a similar masked generation approach [78], but their training objectives are determined heuristically and lack the rigorous theoretical foundation of MDLMs. In this work, we present the first effort to introduce MDLMs into SLG and propose a temporal-checkpoint unmasking strategy to improve training efficiency.

C. Sign Language Pretraining

Pretraining plays a crucial role in sign language understanding tasks. Early works [9], [79], [80] use isolated or continuous sign language recognition as pretext tasks and observe the transferability of the learned features to the translation task. Recent works can be broadly categorized into two groups based on their dependence on text labels.

For text-based methods, one line of work [36], [38], [39], [81]–[83] pretrains models by learning sign-text alignment in a contrastive manner. For example, VAP [39] learns frame-level alignments between sign videos and texts, and transfers the learned features to translation and retrieval tasks; MMSLT [81] instead learns sentence-level alignments and uses them as a regularizer for an off-the-shelf multimodal large language model. Another line of work [7], [84] directly adopts SLT as the pretext task and pretrains models with a next-token-prediction loss.

Text-free methods can better exploit unlabeled data. Typical works [35], [37], [85]–[88] pretrain models by reconstructing sign language inputs in either video or pose format. For example, SignBERT+ [35] performs pretraining by reconstructing masked sign poses, while SHuBERT [88] provides a different perspective by predicting cluster assignments instead of raw sign language inputs. Beyond reconstruction, SignDINO [89] learns sign-aware representations from global video frames by self-distillation with different data views.

However, an important oversight is that all these works focus solely on understanding tasks. Our work represents the first attempt to pretrain sign language generation models through a cross-modal and multi-level strategy that jointly predicts sign and text tokens, latent features, and 3D sign motions, thereby injecting sign-aware knowledge into MDLMs.

III. Methodology

An overview of MaDiS is illustrated in Fig. 2. In the first stage, both text and sign tokens are randomly masked according to a masking ratio t, and the MDLM is pretrained with objectives defined across multiple representation spaces (token, latent, and physical) to exploit the benefits from multi-level sign representations. In the subsequent fine-tuning stage,

the MDLM keeps text tokens unmasked and learns to generate sign tokens using the proposed unmasking with temporal checkpoints (UTC) strategy and the mixture-of-parts (MoP) embedding layer.

A. Preliminaries

1) Sign Tokenizer: We adopt the open-sourced decoupled tokenizer [2] to discretize sign motions, which can be reconstructed with high accuracy. It consists of three parallel VQ-VAEs [90] responsible for different body parts. Each VQ-VAE includes an encoder $\mathcal{E}_p$, a decoder $\mathcal{D}_p$, and a codebook $\mathcal{C}_p = \{\mathbf{c}_p^i\}_{i=1}^{N_c}$, where $p \in \{b, l, r\}$ represents the upper body (including face), left hand, and right hand, respectively, and N_c is the codebook size. Using this tokenizer, a T-frame sign motion sequence $\mathbf{S} \in \mathbb{R}^{T \times d_s}$ is decomposed into three part-wise token sequences, $x_p = \{x_p^i\}_{i=1}^{L_s}$. Here, $d_s = 133$ denotes the number of SMPL-X parameters [91], comprising 11 upper-body joints, 30 hand joints, and 10 facial expression parameters; $x_p^i \in \{1, \ldots, N_c\}$ denotes the code index of the i-th token, and L_s denotes the token sequence length.

2) Masked Diffusion Language Model: MDLMs [1], [31] introduce a new language modeling paradigm based on forward-reverse discrete diffusion. In the forward process, the original token sequence x_0 is corrupted with mask tokens $|\mathrm{M}|$ according to the time (noise level) $t \in [0, 1]$. The conditional distribution of the masked sequence x_t is defined as $q_{t|0}(x_t|x_0) = \prod_{i=1}^{L} q_{t|0}(x_t^i|x_0^i)$, where

$$q_{t|0}\big(x_t^i \mid x_0^i\big) = \begin{cases} 1-t, & x_t^i = x_0^i, \\ t, & x_t^i = |\mathrm{M}|. \end{cases} \tag{1}$$

The reverse process aims to generate meaningful tokens from the all-masked sequence x_1. The conditional distribution is defined as $q_{s|t}(x_s|x_t) = \prod_{i=1}^{L} q_{s|t}(x_s^i|x_t)$, where $0 \le s < t \le 1$, and

$$q_{s|t}(x_s^i|x_t) = \begin{cases} 1, & x_s^i = x_t^i \neq |\mathrm{M}|, \\ \frac{s}{t}, & x_s^i = x_t^i = |\mathrm{M}|, \\ (1-\frac{s}{t}) q_{0|t}(x_s^i|x_t), & x_s^i \neq |\mathrm{M}|, x_t^i = |\mathrm{M}|, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

A proof in [34] shows that $q_{0|t}(x_s^i|x_t) = p_{\text{data}}(x_0^i|x_t^{\mathrm{UM}})$, where x_t^{UM} denotes the unmasked tokens in x_t that are identical to those in x_0. This implies that the time condition t can be omitted from the model inputs, leading to a simplified implementation of MDLM in which the model only needs to predict masked tokens conditioned on the unmasked ones.

B. Tri-Level Cross-Modal Pretraining

Existing works [92], [93] reveal that the knowledge of LMs mainly comes from pretraining. As shown in Fig. 2, we pretrain the MDLM at three spaces, enabling it to absorb knowledge from sign representations of different levels.

1) Token Space: At the first level, we pretrain the MDLM by predicting masked tokens. Following [1], at each training iteration, we uniformly sample a noise level (mask ratio) $t \sim U(0, 1)$. Denoting text tokens as $x_e = \{x_e^i\}_{i=1}^{L_e}$ and an end-of-sentence token as $|\mathrm{eos}|$, the input sequence is formed by concatenating text tokens, $|\mathrm{eos}|$, and sign tokens: $x_0 = \mathrm{cat}(x_e, |\mathrm{eos}|, x_{\{p\}}, |\mathrm{eos}|) = \{x_0^i\}_{i=1}^{L_e+L_s+2}$. For simplicity, we use $\{p\}$ to denote all part-wise sign tokens, since they are masked and unmasked simultaneously once their index is selected. Each index i is selected with a probability of t, and its corresponding token is masked to obtain x_t. The token-space loss is defined as a cross-entropy loss [1]:

$$\mathcal{L}_{tok} = -\frac{1}{t} \sum_{i \in I_m} \log p_\theta(x_0^i \mid x_t), \tag{3}$$

where I_m denote the set of masked indices, and $p_\theta(\cdot)$ denotes the token probabilities predicted by the model.

2) Latent Space: Inspired by the effectiveness of joint-embedding predictive architectures (JEPAs) [40], [94] in numerous vision tasks [41], [44], [45], we extend this idea to MDLM pretraining. Unlike the original JEPAs that train an additional target encoder to extract embeddings, we directly leverage the well-trained codebooks $\mathcal{C}_p$, whose embeddings can accurately reconstruct sign motions via the decoders, as the prediction targets to simplify model design.

More specifically, let the last hidden states of the MDLM corresponding to the sign tokens be $\mathbf{H} \in \mathbb{R}^{L_s \times d}$. We use three MLPs to map $\mathbf{H}$ into the spaces of the codebooks: $\mathbf{H}_p = \mathrm{MLP}_p(\mathbf{H}) \in \mathbb{R}^{L_s \times d_c}$, where d_c denotes the dimension of each code embedding. Let I_s denote the set of masked sign-token indices. For each $i \in I_s$, the target is the corresponding code embedding $\mathbf{c}_p^{n_i}$, where $1 \le n_i \le N_c$ denotes the code index of the i-th token. Finally, the latent-space pretraining loss is defined as the smoothed L1 loss ($\|\cdot\|_1$) between the mapped hidden states and their target code embeddings:

$$\mathcal{L}_{lat} = \sum_{i \in I_s} \sum_{p} \|\mathbf{H}_p^i - \mathbf{c}_p^{n_i}\|_1. \tag{4}$$

3) Physical Space: Inspired by the visual-language nature of sign languages, we incorporate 3D sign motion reconstruction as an additional pretraining objective in the physical space. More specifically, we reuse the well-trained VAE decoder $\mathcal{D}_p$, which is frozen and used to reconstruct sign motions from the mapped hidden states $\mathbf{H}_p$. Denoting the ground-truth part-wise sign motions as $\mathbf{S}_p$, the objective function in the physical space is defined as:

$$\mathcal{L}_{phy} = \sum_{p} \|\mathcal{D}_p(\mathbf{H}_p) - \mathbf{S}_p\|_1. \tag{5}$$

Unlike existing works [35], [36], [85] that mask and reconstruct sign poses, we mask tokens but reconstruct poses (motions), which is more challenging and encourages the model to learn more grounded sign representations. The overall pretraining loss is defined as:

$$\mathcal{L}_{pre} = \mathcal{L}_{tok} + \mathcal{L}_{lat} + \mathcal{L}_{phy}. \tag{6}$$

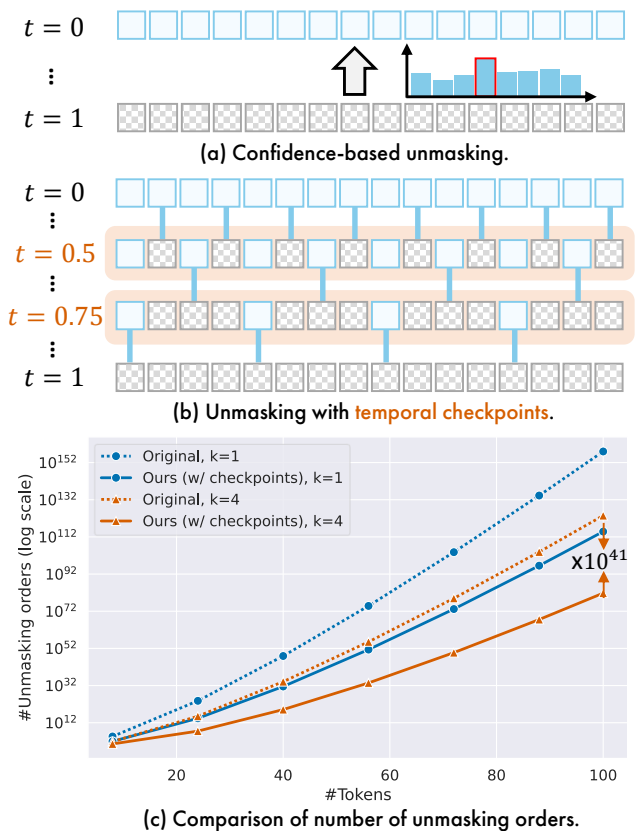

Fig. 3. (a) Vanilla MDLMs adopt an unconstrained confidence-based unmasking strategy [1]. (b) We insert temporal checkpoints at noise levels 0.75 and 0.5 to reduce the complexity of unmasking orders. (c) The original unmasking strategy allows arbitrary unmasking orders (about 10^{123} for generating 100 tokens), while our method reduces this complexity by over 10^{41} times.

Algorithm 1 Adding Noise for UTC

Input: Token sequence $x_0 \in \mathbb{N}^{B\times L}$; noise level $t \in [0,1]^B$; mask token $|\mathrm{M}|$
Output: Noisy token sequence x_t
Define temporal grids $\mathcal{Q} = \{i \mid i \bmod 4 = 0\}$, $\mathcal{H} = \{i \mid i \bmod 2 = 0\}$, and $\mathcal{O} = \{0,\dots,L-1\} \setminus \mathcal{H}$
Initialize the kept mask $\mathbf{K} \leftarrow \mathrm{False}^{B\times L}$
for $b = 0,\dots,B-1$ **do**
  Case 1: randomly keep quarter-grid tokens.
  if $t_b \geq 0.75$ **then**
    $q \leftarrow (t_b - 0.75)/0.25$
    Sample $\mathbf{r} \in \{0,1\}^L$ with $r_i \sim \mathrm{Bernoulli}(q)$
    $\mathbf{K}_{b,i} \leftarrow \mathbf{1}_{\{i\in\mathcal{Q}\}}(1-r_i)$
  Case 2: fix quarter-grid tokens and sample the remaining half-grid tokens.
  else if $0.5 \leq t_b < 0.75$ **then**
    $q \leftarrow (t_b - 0.5)/0.25$
    Sample $\mathbf{r} \in \{0,1\}^L$ with $r_i \sim \mathrm{Bernoulli}(q)$
    $\mathbf{K}_{b,i} \leftarrow \mathbf{1}_{\{i\in\mathcal{Q}\}} \vee (\mathbf{1}_{\{i\in\mathcal{H}\setminus\mathcal{Q}\}}(1-r_i))$
  Case 3: fix half-grid tokens and sample the remaining tokens.
  else
    $q \leftarrow t_b/0.5$
    Sample $\mathbf{r} \in \{0,1\}^L$ with $r_i \sim \mathrm{Bernoulli}(q)$
    $\mathbf{K}_{b,i} \leftarrow \mathbf{1}_{\{i\in\mathcal{H}\}} \vee (\mathbf{1}_{\{i\in\mathcal{O}\}}(1-r_i))$
  end if
end for
$x_t \leftarrow x_0$
$x_t[\mathbf{K} = 0] \leftarrow |\mathrm{M}|$
return x_t

C. Supervised Fine-Tuning

After the pretraining stage, the MDLM is expected to acquire basic sign language knowledge and adapt to the diffusion process with bidirectional contexts. We then fine-tune the entire model while keeping the text tokens unmasked.

1) Unmasking with Temporal Checkpoints (UTC): During the reverse process, the original MDLM [1] adopts a simple confidence-based unmasking strategy (Fig. 3a). Specifically, at each step, the model unmasks k tokens with the highest confidence, defined as the maximum of their categorical probabilities. However, this unconstrained strategy leads to highly diverse unmasking orders. Denoting the number of generated tokens as M, the total number of possible unmasking orders is:

$$N_u = \binom{M}{k}\binom{M-k}{k}\cdots\binom{k}{k} = \frac{M!}{(k!)^{M/k}}. \tag{7}$$

When $M = 100$ and $k = 4$, this number can achieve 2.9×10^{123}. Although such diversity can enhance model performance, it also results in slow convergence [47].

Inspired by VAR [48], which adopts a coarse-to-fine generation scheme with progressively enhanced resolution, we propose a novel unmasking strategy with temporal checkpoints (UTC), as shown in Fig. 3b. Specifically, we divide the reverse process into three stages at $t = 0.75$ and $t = 0.5$, where $1/4$ and $1/2$ of the tokens have been unmasked, respectively. Both pivot states are enforced to contain uniformly distributed tokens, while a standard confidence-based unmasking strategy is applied within each stage[1]. Consequently, the number of possible orders under our UTC is substantially reduced to:

$$N_u^{\mathrm{UTC}} = \frac{(M/4)!}{(k!)^{M/4k}} \cdot \frac{(M/4)!}{(k!)^{M/4k}} \cdot \frac{(M/2)!}{(k!)^{M/2k}}. \tag{8}$$

When $M = 100$ and $k = 4$, UTC reduces the number of orders by $10^{41}\times$ compared to vanilla unmasking strategy (Fig. 3c). Moreover, the resulting generation process follows a coarse-to-fine structure, with both factors jointly easing model learning and accelerating convergence.

Algorithm 1 describes the noise-adding process used for UTC during training. Instead of randomly masking tokens over the whole sequence, we constrain the noisy states according to three temporal stages: when $t \geq 0.75$, only quarter-grid tokens are allowed to remain unmasked; when $0.5 \leq t < 0.75$, the quarter-grid tokens are fixed and the remaining half-grid tokens are randomly sampled; when $t < 0.5$, all half-grid tokens are fixed and the remaining tokens can be sampled. In each stage, the random sampling follows a Bernoulli distribution. This design ensures that the noisy sequence follows the same

[1]When the target number of unmasked tokens at a pivot state is not divisible by k, the final iteration of that stage unmasks all remaining tokens, which may be fewer than k.

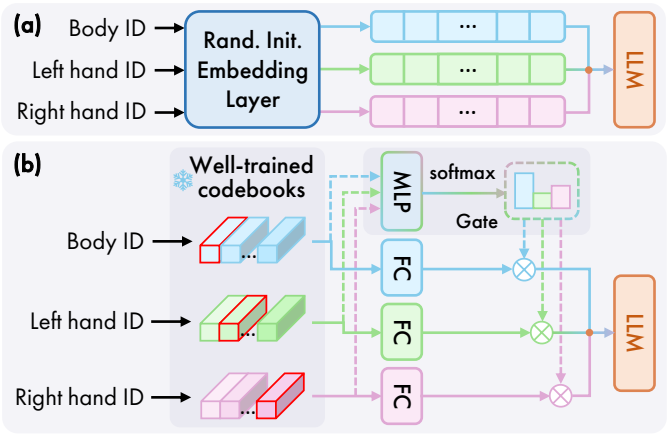

Fig. 4. (a) A conventional sign embedding layer is randomly initialized and simply averages the embeddings of part-wise tokens [2]. (b) The proposed mixture-of-parts embedding layer leverages optimized VAE codebooks and uses a learnable gate to control contributions from different body parts.

coarse-to-fine structure as the UTC reverse process, thereby ensuring training-inference consistency.

2) Mixture-of-Parts (MoP) Embedding Layer: In contrast to the pretraining stage, where different part-wise tokens are treated equally, we aim to more discriminatively capture the relations among body parts during fine-tuning with text conditions. To this end, our MoP embedding layer is built upon the well-trained VQ-VAE codebooks and employs a learnable gate to control the contributions from different parts. For instance, since a large number of signs are produced with a single hand [95], [96], it would be beneficial to assign distinct weights to each hand.

As shown in Fig. 4b, given the part IDs (i_p), the part-wise code embeddings are passed through fully-connected (FC) layers to match the model dimension of MDLM: $\mathbf{E}_p = \mathrm{FC}_p(\mathbf{c}_p^{i_p})$. We then use an MLP followed by a softmax layer to predict the gating weights $\mathbf{G}$, which adaptively control the contribution of each body part. The final sign embedding, denoted as $\mathbf{E}$, is obtained as:

$$\begin{aligned} \mathbf{G} &= \{g_p\} = \mathrm{softmax}(\mathrm{MLP}(\mathrm{cat}(\{\mathbf{c}_p^{i_p}\}))), \\ \mathbf{E} &= \sum_p g_p \cdot \mathbf{E}_p. \end{aligned} \tag{9}$$

Similar to [39], we also observe that continuing to use pretraining objectives during fine-tuning is beneficial. Therefore, the fine-tuning loss is defined as:

$$\mathcal{L}_{sft} = \mathcal{L}_{tok} + \alpha(\mathcal{L}_{lat} + \mathcal{L}_{phy}), \tag{10}$$

where $\alpha = 0.5$ is determined through a sensitivity analysis.

D. Complementary Metrics: SiBLEU and SiCLIP

Existing SLG models [2], [16], [21], [51] are typically evaluated using dynamic time warping (DTW) error [51] and back-translation (BT) [18]. However, DTW error may over-optimistically reflect model performance since it measures the minimum distance between generated and ground-truth sequences. Moreover, its computation is inefficient due to the reliance on dynamic programming. On the other hand, BT evaluates performance in the text space, which overlooks the visual and multi-cue nature of sign language, and the translation process itself introduces unavoidable errors [97]. To address these issues, we propose two new complementary evaluation metrics.

1) SiBLEU – Straightforward and Efficient Evaluation in the Token Space: For LM-based SLG approaches, signs are tokenized, which enables a straightforward evaluation by computing BLEU scores [98] over generated tokens against ground-truth tokens. We term this the SiBLEU score. Specifically, denoting the generated and ground-truth token sequences as (y'_b, y'_l, y'_r) and (y_b, y_l, y_r) for the body, left hand, and right hand, respectively, the SiBLEU scores for the body and hands are computed as:

$$\begin{aligned} &\mathrm{SiBLEU\_Body} = \mathrm{BLEU}(y'_b, y_b), \\ &\mathrm{SiBLEU\_Hands} = (\mathrm{BLEU}(y'_l, y_l) + \mathrm{BLEU}(y'_r, y_r))/2. \end{aligned} \tag{11}$$

A higher SiBLEU indicates greater overlap with the ground-truth tokens.

2) SiCLIP – Evaluation in a Joint Embedding Space: In related domains such as human motion generation [99], [100] and text-to-video generation [101]–[103], CLIP [104] is widely used to evaluate the alignment between prompts (texts) and generations in a joint-embedding space. Motivated by this, we train a CLIP model over ground-truth sign-text pairs in the training sets, and adopt retrieval-based metrics [60], [105], [106] for evaluating the semantic alignment between texts and generated signs.

Specifically, sign motions are encoded using part-wise sign encoders, each consisting of a stack of 1D ResNet blocks [2]. The resulting features are concatenated and passed through an MLP to obtain sign embeddings. For text, we adopt a multilingual language model [107] as the encoder. We empirically find that its first 14 layers are sufficient to achieve high retrieval performance; therefore, the remaining layers are discarded for efficiency. The resulting hidden states are then projected into the shared embedding space through another MLP. The overall model is trained with a fine-grained contrastive loss [108], which implicitly learns precise sign-to-word alignments.

IV. Experiments

A. Datasets and Evaluation Metrics

1) Datasets: We evaluate MaDiS on three widely-adopted datasets: CSL-Daily [4], Phoenix-2014T [5], and How2Sign [3], which contain 20K/8K/35K video-text pairs for Chinese, German, and American sign languages, respectively. We use the curated SMPL-X poses provided in [2] as sign motion representations.

2) Evaluation Metrics: Due to the length difference between generated signs and ground truth, we follow the previous work [2] to use dynamic time warping over joint position errors (DTW-JPE) as the major metric to measure sequence-level distances. Note that we do not compute Procrustes Alignment (PA) since it will align rotation before computing joint errors, leading to overoptimistic performance [109]. As

Fig. 5. Qualitative comparisons of generated signs between our proposed method, MaDiS, with the SOTA method, SOKE [2], on the test sets of CSL-Daily (left), Phoenix-2014T (middle), and How2Sign (right). Video demonstrations are available on the project page.

complementary metrics, we report SiBLEU to assess token-level precision, and sentence-level sign-to-text retrieval performance, measured by Recall@1 (R@1) [60], [108], using SiCLIP as the joint feature extractor for signs and texts.

B. Implementation Details

We instantiate the MDLM with Qwen3-0.6B-Base [107], a leading open-source LLM supporting 119 languages. To enable bidirectional modeling, we replace the causal attention mask in each transformer layer with a non-causal one [1]. During pretraining, the model is trained on the combined training sets of all three datasets to maximize data utilization [35], [84]. In the fine-tuning stage, it is trained on each dataset independently. We empirically set $k = 4$ and $M = 100$, which corresponds to 400 frames after decoding due to the 4x upsampling layer in the VAE decoder. This length is sufficient to cover about 98% of training videos, and over-length videos are uniformly sampled to fit within 400 frames. When generating sign tokens, we adopt multi-head decoding [2] to predict all three part-wise tokens simultaneously, and tokens appearing

TABLE I
HYPER-PARAMETERS OF MLPS.

Module	Latent Map.	Gating Net.	Motion Enc.	Text Enc.
d_{in}	1024	1536	1536	1024
d_m	3072	1024	3072	3072
d_{out}	512	3	512	512

TABLE II
DETAILS OF COMPUTATIONAL COSTS, MEASURED ON NVIDIA GH200 GPUS. *DENOTES TIME COSTS ON CSL-DAILY/PHOENIX-2014T/HOW2SIGN, RESPECTIVELY.

Stage	Memory Cost	Time Cost
Pretraining	96GB	144 GPUh
Supervised fine-tuning	96GB	40/15/70 GPUh*
Inference	8GB	0.76 sec/video

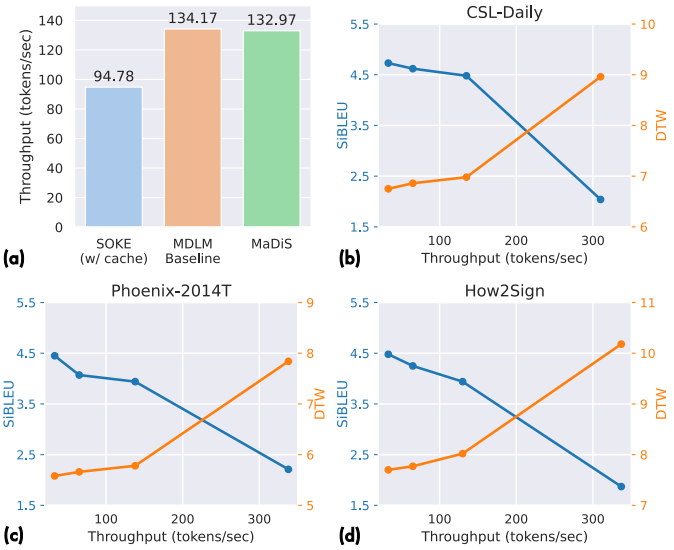

Fig. 6. (a) Comparison of throughput measured in output tokens per second. (b-d) Efficiency-quality trade-off on three datasets. The generation length is fixed to 100 tokens, and the number of sampling steps is set to 100, 50, 25 (default), and 10, corresponding to decoding 1, 2, 4, and 10 tokens per step, respectively. DTW and SiBLEU refer to the average on body and hands.

after the end-of-sentence token are discarded [1]. The initial learning rate is set to $2e-4$, and we adopt the AdamW optimizer [110] with a cosine learning rate scheduler and a batch size of 64 per GPU for both stages. The models are trained for 200 and 150 epochs in the pretraining and fine-tuning stages, respectively, using 4x NVIDIA GH200 GPUs.

MLPs are extensively used in pretraining, the MoP embedding layer, and SiCLIP. Each MLP consists of two linear layers with a SiLU [111] activation function in between. As shown in Tab. I, let d_{in}, d_m, and d_{out} denote the dimensions of the input, intermediate, and output features, respectively.

- For MLPs used in latent-space pretraining, $d_{in} = 1024$ to match the hidden state dimension, $d_m = 3072$, and $d_{out} = 512$ to match the code embedding dimension.
- For the MLP used in the gating network of the MoP embedding layer, $d_{in} = 512 \times 3$, $d_m = 1024$, and $d_{out} = 3$, corresponding to the three body parts.
- For the MLP appended to the motion encoder in SiCLIP, $d_{in} = 512 \times 3$, $d_m = 3072$, and $d_{out} = 512$. For the MLP appended to the text encoder in SiCLIP, $d_{in} = 1024$, $d_m = 3072$, and $d_{out} = 512$.

We further report the computational costs of different stages in Tab. II. The model is first pretrained on the combination of the three datasets, requiring 144 GPU hours. Supervised fine-tuning is then performed independently on each dataset; for example, fine-tuning on CSL-Daily requires only 40 GPU hours. During inference, MaDiS is highly efficient owing to the parallel reverse process of MDLMs, requiring only 0.76 seconds on average to generate a video. In addition, the relatively small model size (0.6B parameters) results in a low GPU memory footprint of 8GB.

C. Comparison with State-of-the-Art Methods

1) Qualitative Comparison: We first conduct qualitative comparisons with the previous SOTA method, SOKE [2], which proposes a simple autoregressive SLG model without dedicated pretraining and can only capture unidirectional contexts due to the nature of ARLMs. In contrast, our MaDiS is pretrained across multiple sign representation spaces and benefits from an MDLM backbone that enables bidirectional contextual modeling. As shown in Fig. 5, our model produces more precise and natural sign motions, demonstrating better generalization across diverse hand shapes, palm orientations, single- and double-handed signs, and finger details. Video demos are available on the project page.

2) Quantitative Comparison: As detailed in Tab. III, we conduct a quantitative comparison with SOTA SLG methods. Since we introduce two new evaluation metrics, we carefully reproduce previous works using their open-source codes [2], [18], [75], [112], [113] or our own reimplementations [21], [25]. MoMask++ [75] is a leading human motion generation method based on a masked generation approach [78]. However, its masking schedule is heuristically determined without a theoretical foundation, and its multi-scale tokenization introduces error accumulation during inference, leading to inferior performance than SOKE. For fairness, we also report the performance of SOKE when using Qwen3-0.6B as the LM, instead of the originally adopted mBART-Large [114]. Although both LMs have a similar size (0.6B), we observe that employing Qwen3 leads to slightly better performance, which can be attributed to the larger pretraining corpus of the Qwen3 series.

Furthermore, we report the performance of an MDLM baseline obtained by directly training an MDLM on sign language datasets, without any of our proposed techniques (pretraining, UTC, or MoP). Notably, this baseline performs on par with SOKE, which utilizes external sign dictionaries, while offering 40% higher throughput (Fig. 6a). This highlights the potential of MDLMs for SLG. Finally, our proposed MaDiS achieves new SOTA results across the test sets of all three datasets and all evaluation metrics.

3) Efficiency: As shown in Fig. 6a, across all test samples of the three datasets, our method achieves a throughput of

TABLE III
COMPARISON WITH SOTA SIGN LANGUAGE GENERATION METHODS. NOTE THAT SINCE SIBLEU SCORES ARE COMPUTED OVER SIGN TOKENS, WE REPORT "N/A" FOR TOKENIZER-FREE METHODS [18], [21], [112]. SIMILARLY, FOR METHODS [25], [75], [113] THAT EMPLOY A SINGLE TOKENIZER FOR WHOLE-BODY MOTIONS, THE SIBLEU-HANDS RESULTS ARE ALSO MARKED AS "N/A". *DTW ERRORS ARE REPORTED BY [2]. †DENOTES METHODS USING EXTERNAL SIGN DICTIONARIES.

Method	CSL-Daily					Phoenix-2014T					How2Sign				
	DTW-JPE↓		SiBLEU↑		SiCLIP↑	DTW-JPE↓		SiBLEU↑		SiCLIP↑	DTW-JPE↓		SiBLEU↑		SiCLIP↑
	Body	Hands	Body	Hands	R@1	Body	Hands	Body	Hands	R@1	Body	Hands	Body	Hands	R@1
Prog. Trans.* [18]	16.30	32.63	N/A		13.65	15.01	31.77	N/A		14.03	14.74	30.17	N/A		9.88
Text2Mesh* [112]	13.76	30.37	N/A		15.84	14.04	31.64	N/A		12.08	15.50	32.97	N/A		9.30
T2S-GPT* [25]	12.32	15.43	2.11	N/A	20.03	11.65	19.09	1.93	N/A	22.10	12.65	18.44	2.13	N/A	12.58
NSA [21]	–	–	N/A		–	–	–	N/A		–	9.16	18.51	N/A		15.04
S-MotionGPT* [113]	11.58	11.31	2.26	N/A	20.37	10.42	9.08	1.97	N/A	32.42	12.41	13.74	2.32	N/A	12.77
MoMask++ [75]	8.93	10.57	1.53	N/A	22.62	8.16	9.96	1.49	N/A	37.88	9.06	11.34	1.58	N/A	20.39
SOKE (mBART)† [2]	7.38	9.68	2.59	2.14	23.64	6.04	7.72	2.71	1.85	34.30	7.75	10.08	2.52	1.74	16.12
SOKE (Qwen3)† [2]	7.29	9.59	2.71	2.27	25.05	6.09	7.65	2.38	1.98	38.53	7.18	10.09	2.57	1.97	18.48
MDLM baseline	7.12	9.57	3.17	2.55	29.17	5.98	7.58	2.36	2.14	44.25	7.02	10.16	2.73	1.79	25.78
MaDiS	**5.99**	**7.96**	**4.87**	**4.08**	**38.61**	**4.97**	**6.59**	**4.54**	**3.33**	**57.94**	**6.59**	**9.45**	**4.41**	**3.46**	**34.27**

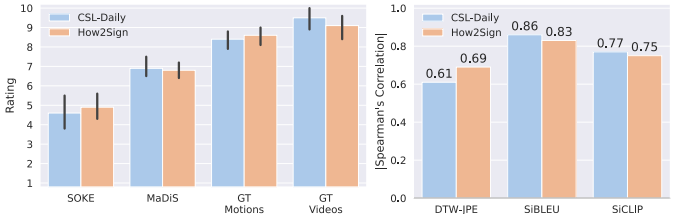

Fig. 7. Left: User study with professional CSL/ASL signers. We report average ratings of the previous SOTA method, SOKE [2], our proposed MaDiS, ground-truth motions, and ground-truth RGB videos. Error bars indicate 95% confidence intervals. Right: Absolute values of Spearman's rank correlation coefficient [115] between CSL/ASL signers' ratings and metric scores. DTW and SiBLEU refer to the average on body and hands.

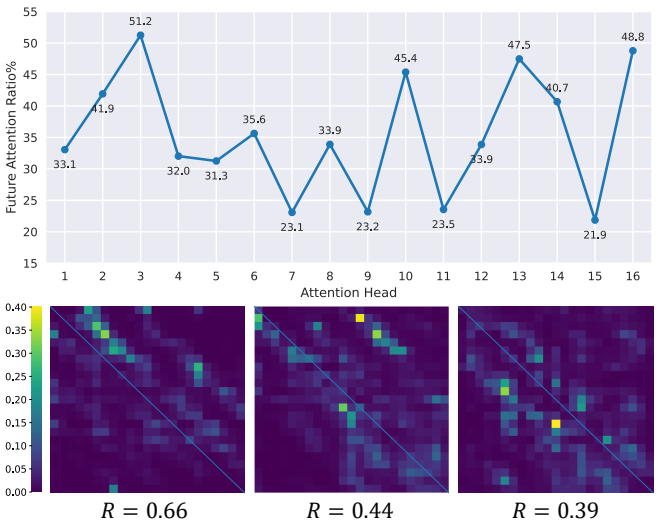

Fig. 8. Top: Future-attention ratios (R) of the 16 attention heads in the penultimate Transformer layer. Bottom: A case study visualizing the attention maps of the three heads with the highest future-attention ratios.

132 tokens/sec, representing a 40% improvement over SOKE with KV caching. This high throughput enables the generation of a 400-frame (100-token) video in only 0.76s, thereby satisfying real-time requirements. This improvement stems from the highly efficient parallel token generation enabled by the reverse process in MDLMs.

In Fig. 6b-d, we observe that MaDiS provides a flexible trade-off between generation quality and efficiency, where quality is measured by DTW and SiBLEU scores. When the number of sampling steps is set to 100, corresponding to sampling one token at a time, the model achieves the highest generation quality but the lowest throughput. In contrast, with only 10 sampling steps, inference becomes extremely fast (exceeding 300 tokens/sec) at the cost of degraded performance. Using 25 sampling steps yields a balance between generation quality and efficiency.

D. Human Evaluation with Professional Signers

To assess the intelligibility of the generated signs, we conducted a user study involving 5 professional CSL signers and 3 professional ASL signers, each with 4 to 13 years of signing experience. Participants were asked to rate the alignment between generated signs and text annotations on a scale from 1 to 10, where higher scores indicate higher sign accuracy and better semantic conveyance. Ground-truth motions and RGB videos were also included for reference. Specifically, we provided 15 generated signs from SOKE and from our proposed MaDiS, together with the corresponding ground-truth motions and videos. The order of all samples was randomly shuffled to avoid potential bias. All participants were compensated above the local average hourly rate.

As shown in the left part of Fig. 7, on CSL-Daily, our method achieves an average rating of 6.9, substantially outperforming SOKE, which receives an average rating of 4.6. The ground-truth motions obtain an average rating of 8.4, further demonstrating the feasibility of using motions as effective sign language representations. Similar trends are observed on How2Sign, demonstrating the generalizability of our method across different sign languages.

Furthermore, we report the absolute values of Spearman's rank correlation coefficient [115] to assess the agreement

TABLE IV
ABLATION STUDY FOR THE KEY TECHNIQUES.

Method	DTW-JPE↓		SiBLEU↑		SiCLIP↑
	Body	Hands	Body	Hands	R@1
MaDiS	**5.99**	**7.96**	**4.87**	**4.08**	**38.61**
w/o Pretraining	6.93	9.27	3.48	2.84	31.49
w/o UTC	6.61	9.05	3.87	3.11	32.81
w/o MoP	6.35	8.50	4.06	3.42	35.95

TABLE V
ABLATION STUDY FOR THE TRI-LEVEL CROSS-MODAL PRETRAINING.

Token		Latent	Physical	DTW-JPE↓		SiBLEU↑		SiCLIP↑
Sign	Text			Body	Hands	Body	Hands	R@1
				6.93	9.27	3.48	2.84	31.49
✓				6.81	8.90	3.76	2.97	33.01
✓	✓			6.41	8.64	4.18	3.24	34.76
✓	✓	✓		6.22	8.39	4.45	3.67	35.96
✓	✓	✓	✓	**5.99**	**7.96**	**4.87**	**4.08**	**38.61**

between automatic metrics and human judgments. A value closer to 1 indicates stronger monotonic agreement between an automatic metric and signers' ratings, while a value closer to 0 indicates weaker agreement. As shown in the right part of Fig. 7, SiBLEU and SiCLIP exhibit stronger agreement, whereas DTW, based on minimum-distance alignment, shows relatively weaker agreement.

E. Ablation Studies

We conduct all ablation studies on CSL-Daily unless otherwise specified.

1) Usage of Bidirectional Contexts: To verify that the MDLM effectively models bidirectional contexts, we visualize attention maps from the penultimate Transformer layer at the final sampling step, which captures rich contextual interactions while being less influenced by the task-specific output projection of the final layer [116], [117].

As shown in the top part of Fig. 8, we compute the future-attention ratio for each attention head as:

$$R = \frac{\sum_{i,j>i} \mathbf{A}_{i,j}}{\sum_{i,j} \mathbf{A}_{i,j}}, \tag{12}$$

where $\mathbf{A}$ denotes the attention map. The ratios vary substantially across the 16 attention heads, ranging from 21.9% to 51.2%, with an average of 35.4%. Heads 3, 13, and 16 exhibit the highest ratios, indicating that they allocate nearly half of their attention to future tokens, *i.e.*, the upper-triangular region, and effectively capture bidirectional contextual dependencies. As a case study, we further visualize the attention maps of these three heads for a representative test sample in the bottom part of Fig. 8.

2) Contribution of Key Techniques: Tab. IV validates the effectiveness of the key techniques in MaDiS. Removing the tri-level cross-modal pretraining leads to the largest performance degradation, increasing DTW-JPE from 5.99/7.96 to 6.93/9.27 for body and hands, respectively, while reducing

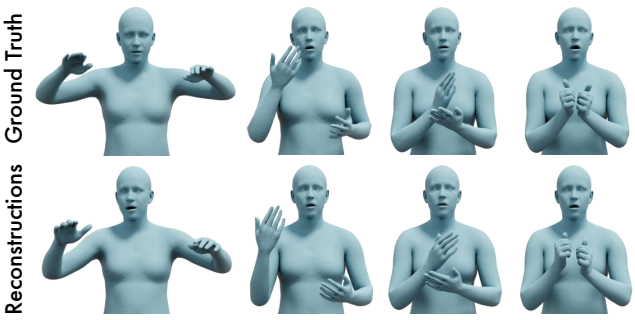

Fig. 9. Visualization of reconstructed signs from mask tokens.

TABLE VI
ABLATION STUDY FOR PRETRAINING DATA.

Pretraining data	DTW-JPE↓		SiBLEU↑		SiCLIP↑
	Body	Hands	Body	Hands	R@1
No pretraining	6.93	9.27	3.48	2.84	31.49
CSL	6.33	8.40	3.96	3.74	35.51
CSL+Phoenix	6.21	8.25	4.19	3.92	36.83
CSL+Phoenix+How2Sign	**5.99**	**7.96**	**4.87**	**4.08**	**38.61**

SiBLEU and SiCLIP R@1 from 4.87/4.08 and 38.61 to 3.48/2.84 and 31.49. This demonstrates that pretraining across complementary spaces is crucial for learning effective sign language representations. Disabling the proposed unmasking with temporal checkpoints (UTC) also consistently worsens all metrics, validating that the coarse-to-fine unmasking strategy improves generation quality. In addition, removing the mixture-of-parts (MoP) embedding layer results in inferior performance, showing the benefit of dynamically fusing part-wise sign tokens instead of treating body and hand tokens equally. Overall, these results confirm that each technique contributes positively to MaDiS.

3) Tri-Level Cross-Modal Pretraining: We pretrain the model across three spaces (token, latent, and physical) to fully exploit the benefits from multi-level sign representations. Results in Tab. V validate the effectiveness of each pretraining level. First, at the token level, masking both sign and text tokens yields better performance than masking sign tokens alone, demonstrating the benefit of cross-modal training. This observation is consistent with previous findings on alignment-based pretraining in sign language translation [38], [39], [118]. Second, predicting latent features proves effective in reducing DTW errors and enhancing sentence-level semantic alignment. Finally, incorporating physical-space training by reconstructing sign motions leads to the best overall performance. This design also enables the pretrained model to learn physically grounded representations, allowing it to accurately reconstruct sign motions from masked tokens (Fig. 9). Quantitatively, when $t = 0.25$ (*i.e.*, masking 25% of tokens), the reconstructed signs achieve mean position errors of 24.79/5.64mm on body/hand joints, comparable to the performance of the sign tokenizer [2].

In the pretraining stage, we use the combined training sets of all three datasets to maximize data utilization. Although

TABLE VII
ABLATION STUDY FOR UNMASKING WITH TEMPORAL CHECKPOINTS.

Checkpoint (distribution, noise level)	DTW-JPE↓		SiBLEU↑		SiCLIP↑
	Body	Hands	Body	Hands	R@1
N/A	6.61	9.05	3.87	3.11	32.81
Uniform, $t = 0.75$	6.33	8.73	4.15	3.52	34.09
Uniform, $t = 0.5$	6.18	8.23	4.63	3.94	36.73
Uniform, $t = 0.75/0.5$	**5.99**	**7.96**	4.87	**4.08**	**38.61**
Uniform, $t = 0.875/0.75/0.5$	6.07	8.14	**4.92**	3.97	38.04
Contiguous, $t = 0.75/0.5$	6.28	8.44	4.31	3.80	34.85

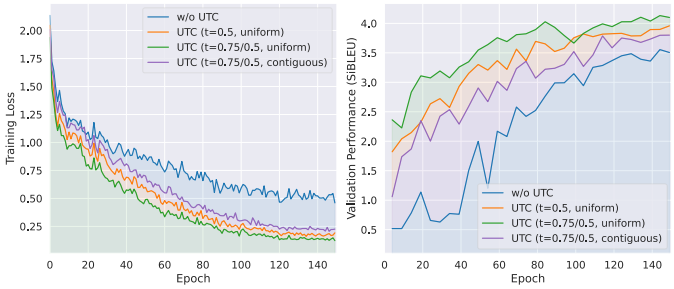

Fig. 10. Visualization of training loss (left) and validation performance (average of SiBLEU-Body and SiBLEU-Hands, right).

these datasets come from different languages, pretraining on multilingual data has been shown to improve performance in sign language recognition and translation [35], [84]. As shown in Tab. VI, we study the effect of different pretraining data sources on downstream SLG performance on CSL-Daily. The results show that pretraining on the same data source leads to significant performance gains, while further incorporating data from other languages continues to improve performance. This also suggests the scalability of our model.

4) Unmasking with Temporal Checkpoints (UTC): To accelerate model convergence, we propose a novel unmasking strategy with temporal checkpoints, which substantially reduces the complexity of unmasking orders in a coarse-to-fine manner. As shown in Tab. VII, inserting checkpoints composed of uniformly distributed tokens at either $t = 0.75$ ($T/4$ temporal resolution) or $t = 0.5$ ($T/2$ temporal resolution) yields superior performance over the no-checkpoint setting. Inserting checkpoints at both $t = 0.75$ and $t = 0.5$ achieves the best performance. Adding an additional checkpoint at $t = 0.875$ ($T/8$ temporal resolution) yields comparable results but provides no clear performance gain.

We further construct a baseline that retains only the first $1/4$ or $1/2$ contiguous tokens at the pivot states. While this design also reduces the order complexity, it does not explicitly model the coarse-to-fine prior. Although this baseline outperforms the no-checkpoint setting, the best performance is achieved when both checkpoints are applied with uniformly distributed tokens, indicating the benefits of both the reduced complexity and the coarse-to-fine design. Finally, we visualize the training loss ($\mathcal{L}_{tok}$) and validation performance in Fig. 10. The curves demonstrate that our UTC leads to faster convergence, which is reflected by lower training loss and higher validation performance at the same epoch.

TABLE VIII
ABLATION STUDY FOR THE MIXTURE-OF-PARTS EMBEDDING LAYER.

Condition	Fusion Strategy	DTW-JPE↓		SiBLEU↑		SiCLIP↑
		Body	Hands	Body	Hands	R@1
N/A	Average [2]	6.35	8.50	4.06	3.42	35.95
Codebook	Average	6.19	8.25	4.45	3.73	36.98
	MLP	6.22	8.29	4.19	3.86	37.02
	Transformer	6.17	8.13	4.33	3.62	36.49
	Sparse MoE (Top-1)	6.48	8.74	3.76	3.27	35.65
	Sparse MoE (Top-2)	6.29	8.41	3.91	3.50	36.87
	Dense MoE (ours)	**5.99**	**7.96**	**4.87**	**4.08**	**38.61**

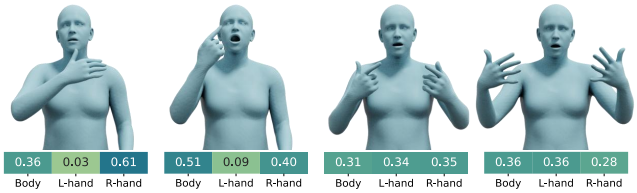

Fig. 11. Visualization of the learned gating weights $\mathbf{G}$.

5) Mixture-of-Parts (MoP) Embedding Layer: Our MoP sign embedding layer dynamically fuses token embeddings from different body parts via a learnable gate. As shown in Tab. VIII, MoP significantly outperforms the prior design that simply averages randomly initialized part-wise token embeddings [2]. When conditioned on well-trained codebooks, we evaluate three standard fusion strategies: 1) averaging, 2) an MLP applied to the concatenation of part-wise code embeddings, and 3) a two-layer Transformer equipped with a $|\text{CLS}|$ token. Results show that all conditioned fusion strategies outperform the non-conditioned baseline [2], validating the effectiveness of codebook conditioning.

Our MoP is built upon a dense mixture-of-experts (MoE) formulation [119], where all experts, corresponding to different body parts, are activated for each input and their contributions are adaptively weighted by the gating network. As reported in Tab. VIII, we further compare the dense formulation with sparse MoE variants that retain only the experts associated with the top-1 or top-2 gate activations. Although these sparse variants reduce the number of active experts, they consistently yield inferior performance to the default dense setting. This suggests that discarding experts with lower gate weights may remove useful co-occurrence cues, whereas dense fusion provides a more comprehensive representation of part-wise sign information.

To further analyze the learned gating behavior, we visualize MoP gate weights for four example signs in Fig. 11. For single-handed signs (left two examples), the MoP layer assigns higher weights to the active hand, indicating that it effectively captures hand dominance. In contrast, for two-handed signs, the layer learns comparable weights for both hands, reflecting balanced contributions during generation. Quantitatively, we identify the signing hand using a simple rule based on wrist position relative to the spine-01 joint [91], and accordingly classify signs as single- or two-handed. Tab. IX reports the correlation between hand dominance and MoP gate weights

TABLE IX
DISTRIBUTION OF GATE WEIGHTS ACROSS DIFFERENT SIGN CATEGORIES.

Part \ Category	Left-Handed	Right-Handed	Two-Handed
Body	0.41	0.47	0.32
Left Hand	0.47	0.08	0.31
Right Hand	0.12	0.45	0.37

TABLE X
SENSITIVITY ANALYSIS FOR THE LOSS WEIGHT (α).

Weight (α)	DTW-JPE↓		SiBLEU↑		SiCLIP↑
	Body	Hands	Body	Hands	R@1
0.0	6.13	8.04	4.39	3.64	36.99
0.2	6.05	7.98	4.81	**4.15**	38.07
0.5	**5.99**	**7.96**	**4.87**	4.08	**38.61**
0.7	6.02	8.01	4.65	3.99	37.90
1.0	6.03	7.97	4.43	3.83	37.23

(body/left hand/right hand), indicating that the learned gates align with interpretable physical structure.

6) Choice of Loss Weight: Following [39], we retain the pretraining objectives during the fine-tuning stage and introduce a hyperparameter, α, to weight $\mathcal{L}_{lat}$ and $\mathcal{L}_{phy}$ (Eq. (10)). The sensitivity of model performance to α is shown in Tab. X. Based on these results, we set $\alpha = 0.5$, which achieves the best performance among all the values.

7) SiCLIP: To better evaluate the semantic alignment between texts and generated signs, we follow established practices in related fields [99], [100] and train a CLIP model on ground-truth sign motions and their corresponding texts, referred to as SiCLIP. We assess the quality of the learned joint embedding space using the sign-to-text retrieval task, where higher retrieval accuracy indicates stronger semantic alignment. As shown in Tab. XI, using only a standard contrastive loss [106] results in relatively low retrieval performance. Introducing decoupled sign encoders [2], which model different body parts separately, yields noticeable improvements. Most of the gains come from the fine-grained contrastive loss [108], which implicitly learns more precise word-level alignments between sign motions and texts. The final retrieval performance is comparable to that of a leading motion retrieval model [105], validating both the quality of the learned embedding space and the reliability of SiCLIP as an SLG evaluator.

V. DISCUSSION

A. Broader Impacts

Sign language is the primary communication method for Deaf and Hard-of-Hearing (DHH) communities. However, DHH individuals are often encouraged to communicate with hearing people through text rather than their preferred sign language [6]. SLG research therefore has great potential to improve sign language accessibility and foster a more inclusive society. Our proposed MaDiS advances the state of the art in SLG in both generation quality and inference speed, laying the groundwork for practical, real-time communication systems between DHH and hearing individuals.

TABLE XI
ABLATION STUDY FOR THE SICLIP.

Method	CSL-Daily↑		Phoenix↑		How2Sign↑	
	R@1	R@5	R@1	R@5	R@1	R@5
Baseline (CLIP) [106]	41.07	81.97	58.41	88.01	9.27	28.47
+ Decoupled encoders	45.07	85.88	61.84	90.03	12.13	38.48
+ Fine-grained loss	**59.10**	**95.66**	**68.85**	**95.64**	**41.33**	**88.43**

B. Limitations and Future Work

In this work, we focus on a relatively small diffusion language model with 0.6B parameters. The performance is expected to improve further with larger model sizes and larger-scale datasets. We leave this direction for future work, subject to the availability of additional computational resources. In addition, since the adopted tokenizer [2] operates solely on motions, intra-sign variance may be introduced; that is, varied motions of the same sign may be mapped to different tokens. This factor may lead to lower SiBLEU scores and an underestimation of model performance. Therefore, in this study, we use SiBLEU in token space together with other metrics, such as DTW-JPE in physical space, to provide a more comprehensive evaluation. Future work should explore more robust tokenizers with reduced sensitivity to intra-sign variation. Finally, while we have demonstrated the generalizability of our method across three sign languages, there are over 200 distinct sign languages worldwide. In future work, we will investigate the performance of our method on a broader range of sign languages.

VI. CONCLUSION

We present MaDiS, a novel SLG approach built upon MDLMs that enable bidirectional context modeling and efficient generation. To effectively inject sign language knowledge into MDLMs, we design a tri-level cross-modal pretraining framework that jointly learns from token, latent, and 3D physical spaces. To improve training efficiency of MDLMs, we simplify unmasking orders by proposing a novel unmasking strategy with temporal checkpoints. Moreover, a mixture-of-parts sign embedding layer is introduced to enhance part-wise token integration through gated, codebook-conditioned fusion. Experiments on standard SLG benchmarks demonstrate that MaDiS achieves SOTA performance, while improving inference throughput by 40%.

ACKNOWLEDGMENTS

S. Zafeiriou and part of the research were funded by the EPSRC Fellowship DEFORM (EP/S010203/1), EPSRC Project GNOMON (EP/X011364/1) and Turing AI Fellowship (EP/Z534699/1). R.A. Potamias and R. Zuo were supported by EPSRC Project GNOMON (EP/X011364/1). J. Deng was supported by the NVIDIA Academic Grant. We would also

like to thank Yuecong Min for coordinating the user study, and the anonymous signers who participated in it.

The authors acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR). Isambard-AI [120] is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023].

REFERENCES

[1] S. Nie, F. Zhu, Z. You, X. Zhang, J. Ou, J. Hu, J. Zhou, Y. Lin, J.-R. Wen, and C. Li, "Large language diffusion models," *NeurIPS*, 2025.
[2] R. Zuo, R. A. Potamias, E. Ververas, J. Deng, and S. Zafeiriou, "Signs as tokens: A retrieval-enhanced multilingual sign language generator," in *ICCV*, 2025.
[3] A. Duarte, S. Palaskar, L. Ventura, D. Ghadiyaram, K. DeHaan, F. Metze, J. Torres, and X. Giro-i Nieto, "How2sign: a large-scale multimodal dataset for continuous american sign language," in *CVPR*, 2021, pp. 2735–2744.
[4] H. Zhou, W. Zhou, W. Qi, J. Pu, and H. Li, "Improving sign language translation with monolingual data by sign back-translation," in *CVPR*, 2021.
[5] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," in *CVPR*, 2018.
[6] K. Yin, A. Moryossef, J. Hochgesang, Y. Goldberg, and M. Alikhani, "Including signed languages in natural language processing," in *ACL*, 2021, pp. 7347–7360.
[7] Z. Li, W. Zhou, W. Zhao, K. Wu, H. Hu, and H. Li, "Uni-sign: Toward unified sign language understanding at scale," in *ICLR*, 2025.
[8] J. Gong, L. G. Foo, Y. He, H. Rahmani, and J. Liu, "Llms are good sign language translators," in *CVPR*, 2024, pp. 18 362–18 372.
[9] Y. Chen, R. Zuo, F. Wei, Y. Wu, S. Liu, and B. Mak, "Two-stream network for sign language recognition and translation," in *NeurIPS*, 2022.
[10] Y. Jang, H. Raajesh, L. Momeni, G. Varol, and A. Zisserman, "Lost in translation, found in context: Sign language translation with contextual cues," in *CVPR*, 2025.
[11] R. Wong, N. C. Camgoz, and R. Bowden, "Sign2gpt: Leveraging large language models for gloss-free sign language translation," in *ICLR*, 2024.
[12] J. Guo, P. Li, and T. Cohn, "Bridging sign and spoken languages: Pseudo gloss generation for sign language translation," *NeurIPS*, 2025.
[13] E. Fish and R. Bowden, "Geo-sign: Hyperbolic contrastive regularisation for geometrically aware sign language translation," *NeurIPS*, 2025.
[14] S. Gan, Y. Yin, Z. Jiang, L. Xie, S. Lu, and H. Wen, "Mixsigngraph: A sign sequence is worth mixed graphs of nodes," in *NeurIPS*, 2025.
[15] H. Yao, W. Zhou, H. Feng, H. Hu, H. Zhou, and H. Li, "Sign language translation with iterative prototype," in *ICCV*, 2023, pp. 15 592–15 601.
[16] R. Zuo, F. Wei, Z. Chen, B. Mak, J. Yang, and X. Tong, "A simple baseline for spoken language to sign language translation with 3d avatars," in *ECCV*, 2024.
[17] B. Saunders, N. C. Camgoz, and R. Bowden, "Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production," in *CVPR*, 2022, pp. 5141–5151.
[18] B. Saunders, N. C. Camgoz, and R. Bowden, "Progressive transformers for end-to-end sign language production," in *ECCV*, 2020, pp. 687–705.
[19] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden, "Text2sign: towards sign language production using neural machine translation and generative adversarial networks," *IJCV*, vol. 128, no. 4, pp. 891–908, 2020.
[20] T. Shi, L. Hu, F. Shang, L. Gao, and W. Feng, "Greg: Geometry-aware region refinement for sign language video generation," in *ICCV*, 2025.
[21] V. Baltatzis, R. A. Potamias, E. Ververas, G. Sun, J. Deng, and S. Zafeiriou, "Neural sign actors: A diffusion model for 3d sign language production from text," in *CVPR*, 2024, pp. 1985–1995.
[22] S. Fang, C. Sui, Y. Zhou, X. Zhang, H. Zhong, Y. Tian, and C. Chen, "Signdiff: Diffusion model for american sign language production," in *FG*, 2025.
[23] C. Wang, Z. Deng, Z. Jiang, F. Shen, Y. Yin, S. Gan, Z. Cheng, S. Ge, and Q. Gu, "Advanced sign language video generation with compressed and quantized multi-condition tokenization," in *NeurIPS*, 2025.
[24] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.
[25] A. Yin, H. Li, K. Shen, S. Tang, and Y. Zhuang, "T2S-GPT: Dynamic vector quantization for autoregressive sign language production from text," in *ACL*, 2024.
[26] Z. Yu, S. Huang, Y. Cheng, and T. Birdal, "Signavatars: A large-scale 3d sign language holistic motion dataset and benchmark," in *ECCV*, 2024, pp. 1–19.
[27] S. Fang, L. Wang, C. Zheng, Y. Tian, and C. Chen, "Signllm: Sign languages production large language models," in *ICCVW*, 2025.
[28] L. Dong, L. Chaudhary, F. Xu, X. Wang, M. Lary, and I. Nwogu, "Signavatar: Sign language 3d motion reconstruction and generation," in *FG*, 2024.
[29] L. Yang, Y. Tian, B. Li, X. Zhang, K. Shen, Y. Tong, and M. Wang, "Mmada: Multimodal large diffusion language models," in *NeurIPS*, 2025.
[30] Z. You, S. Nie, X. Zhang, J. Hu, J. Zhou, Z. Lu, J.-R. Wen, and C. Li, "Llada-v: Large language diffusion models with visual instruction tuning," in *CVPR*, 2026.
[31] S. Nie, F. Zhu, C. Du, T. Pang, Q. Liu, G. Zeng, M. Lin, and C. Li, "Scaling up masked diffusion models on text," in *ICLR*, 2025.
[32] J. Shi, K. Han, Z. Wang, A. Doucet, and M. Titsias, "Simplified and generalized masked diffusion for discrete data," *NeurIPS*, 2024.
[33] S. Sahoo, M. Arriola, Y. Schiff, A. Gokaslan, E. Marroquin, J. Chiu, A. Rush, and V. Kuleshov, "Simple and effective masked diffusion language models," *NeurIPS*, 2024.
[34] J. Ou, S. Nie, K. Xue, F. Zhu, J. Sun, Z. Li, and C. Li, "Your absorbing discrete diffusion secretly models the conditional distributions of clean data," *ICLR*, 2025.
[35] H. Hu, W. Zhao, W. Zhou, and H. Li, "SignBERT+: Hand-model-aware self-supervised pre-training for sign language understanding," *TPAMI*, 2023.
[36] W. Zhou, W. Zhao, H. Hu, Z. Li, and H. Li, "Scaling up multimodal pre-training for sign language understanding," *TPAMI*, 2025.
[37] R. Wong, N. C. Camgoz, and R. Bowden, "Signrep: Enhancing self-supervised sign representations," in *ICCV*, 2025.
[38] B. Zhou, Z. Chen, A. Clapés, J. Wan, Y. Liang, S. Escalera, Z. Lei, and D. Zhang, "Gloss-free sign language translation: Improving from visual-language pretraining," in *ICCV*, 2023, pp. 20 871–20 881.
[39] P. Jiao, Y. Min, and X. Chen, "Visual alignment pre-training for sign language translation," in *ECCV*, 2024, pp. 349–367.
[40] Y. LeCun, "A path towards autonomous machine intelligence," *Open Review*, vol. 62, no. 1, pp. 1–62, 2022.
[41] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas, "Revisiting feature prediction for learning visual representations from video," *TMLR*, 2024.
[42] F. Baldassarre, M. Szafraniec, B. Terver, V. Khalidov, F. Massa, Y. LeCun, P. Labatut, M. Seitzer, and P. Bojanowski, "Back to the features: Dino as a foundation for video world models," *arXiv preprint arXiv:2507.19468*, 2025.
[43] S. Mo and S. Tong, "Connecting joint-embedding predictive architecture with contrastive self-supervised learning," *NeurIPS*, 2024.
[44] G. Zhou, H. Pan, Y. LeCun, and L. Pinto, "Dino-wm: World models on pre-trained visual features enable zero-shot planning," in *ICML*, 2025.
[45] M. Assran, A. Bardes, D. Fan, Q. Garrido, R. Howes, M. Muckley, A. Rizvi, C. Roberts, K. Sinha, A. Zholus *et al.*, "V-jepa 2: Self-supervised video models enable understanding, prediction and planning," *arXiv preprint arXiv:2506.09985*, 2025.
[46] Y. Song, Z. Zhang, C. Luo, P. Gao, F. Xia, H. Luo, Z. Li, Y. Yang, H. Yu, X. Qu *et al.*, "Seed diffusion: A large-scale diffusion language model with high-speed inference," *arXiv preprint arXiv:2508.02193*, 2025.
[47] M. Prabhudesai, M. Wu, A. Zadeh, K. Fragkiadaki, and D. Pathak, "Diffusion beats autoregressive in data-constrained settings," *NeurIPS*, 2025.
[48] K. Tian, Y. Jiang, Z. Yuan, B. Peng, and L. Wang, "Visual autoregressive modeling: Scalable image generation via next-scale prediction," *NeurIPS*, 2024.
[49] B. Saunders, N. C. Camgoz, and R. Bowden, "Mixed signals: Sign language production via a mixture of motion primitives," in *ICCV*, 2021, pp. 1919–1929.
[50] S. Tang, J. He, L. Cheng, J. Wu, D. Guo, and R. Hong, "Discrete to continuous: Generating smooth transition poses from sign language observations," in *CVPR*, 2025.

[51] R. S. Arkushin, A. Moryossef, and O. Fried, "Ham2pose: Animating sign language notation into pose sequences," in *CVPR*, 2023, pp. 21 046–21 056.

[52] P. Xie, Q. Zhang, P. Taiying, H. Tang, Y. Du, and Z. Li, "G2p-ddm: Generating sign pose sequence from gloss sequence with discrete diffusion model," in *AAAI*, 2024, pp. 6234–6242.

[53] S. Tang, R. Hong, D. Guo, and M. Wang, "Gloss semantic-enhanced network with online back-translation for sign language production," in *MM*, 2022, pp. 5630–5638.

[54] X. Liu, S. Gan, Y. Yin, B. Guo, Z. Jiang, S. Meng, L. Xie, and S. Lu, "Signpr: A progressive vector-quantized diffusion framework for sign language production," in *CVPR*, 2026.

[55] Y. Yu, S. Liu, Y. Feng, Z. Jin, Y. Jiang, and M. Xu, "Focal-general diffusion model with semantic consistent guidance for sign language production," in *CVPR*, 2026.

[56] M. Ye, X. Ye, and S. Manoharan, "Hybrid autoregressive-diffusion model for real-time sign language production," in *ACL*, 2026.

[57] F. Qi, Y. Duan, C. Xu, and H. Zhang, "Signgen: End-to-end sign language video generation with latent diffusion," in *ECCV*, 2024.

[58] S. Fang, Y. Feng, H. Zhong, Y. Zhang, and D. N. Metaxas, "Stable signer: Hierarchical sign language generative model," in *ACL*, 2026.

[59] F. Qi, Y. Duan, H. Zhang, and C. Xu, "Signmod: Sign language video generation via mixture of diffusion," *TPAMI*, 2026.

[60] L. Bensabath, M. Petrovich, and G. Varol, "Text-driven 3d hand motion generation from sign language data," in *CVPR*, 2026.

[61] J. Low, A. Symeonidis-Herzig, M. Ivashechkin, O. M. Sincan, and R. Bowden, "Signspark: Efficient multilingual sign language production via sparse keyframe learning," in *ECCV*, 2026.

[62] T. Shi, L. Hu, F. Shang, J. Feng, P. Liu, and W. Feng, "Pose-guided fine-grained sign language video generation," in *ECCV*, 2024, pp. 392–409.

[63] X. Wang, S. Tang, L. Cheng, F. Li, S. Wang, and R. Hong, "Signaligner: Harmonizing complementary pose modalities for coherent sign language generation," *arXiv preprint arXiv:2506.11621*, 2025.

[64] G. Qiao, S. Lin, R. Zuo, Z. Wu, K. Jia, and G. Liu, "Signbot: Learning human-to-humanoid sign language interaction," in *ICRA*, 2026.

[65] N. Khan, B. Wu, R. Shi, B. Yen, T. Ashizawa, C. T. Ishi, T. Minato, and K. Nakadai, "From sign language generation to humanoid execution: Vision-language guided retargeting with collision mitigation," in *IROS*, 2026.

[66] W. Huang, Z. Zhao, J. He, and M. Zhang, "Dualsign: Semi-supervised sign language production with balanced multi-modal multi-task dual transformation," in *MM*, 2022, pp. 5486–5495.

[67] T. Lee, H. Nam, G. Moon, and K. M. Lee, "Signer: Temporally grounded sign language generation via time-resolved conditioning," in *ECCV*, 2026.

[68] T. Bie, M. Cao, K. Chen, L. Du, M. Gong, Z. Gong, Y. Gu, J. Hu, Z. Huang, Z. Lan *et al.*, "Llada2. 0: Scaling up diffusion language models to 100b," *arXiv preprint arXiv:2512.15745*, 2025.

[69] S. Gong, S. Agarwal, Y. Zhang, J. Ye, L. Zheng, M. Li, C. An, P. Zhao, W. Bi, J. Han *et al.*, "Scaling diffusion language models via adaptation from autoregressive models," in *ICLR*, 2025.

[70] X. Wang, Z. Zheng, F. Ye, D. Xue, S. Huang, and Q. Gu, "Diffusion language models are versatile protein learners," in *ICML*, 2024.

[71] J. Yin, C. Zha, W. He, C. Xu, and X. Gao, "Cfp-gen: Combinatorial functional protein generation via diffusion language models," in *ICML*, 2025.

[72] Y. Wen, H. Li, K. Gu, Y. Zhao, T. Wang, and X. Sun, "Llada-vla: Vision language diffusion action models," in *IROS*, 2026.

[73] J. Chen, W. Song, P. Ding, Z. Zhou, H. Zhao, F. Tang, D. Wang, and H. Li, "Unified diffusion VLA: Vision-language-action model via joint discrete denosing diffusion process," in *ICLR*, 2026.

[74] C. Guo, Y. Mu, M. G. Javed, S. Wang, and L. Cheng, "Momask: Generative masked modeling of 3d human motions," in *CVPR*, 2024, pp. 1900–1910.

[75] C. Guo, I. Hwang, J. Wang, and B. Zhou, "Snapmogen: Human motion generation from expressive texts," in *NeurIPS*, 2025.

[76] E. Pinyoanuntapong, M. Saleem, K. Karunratanakul, P. Wang, H. Xue, C. Chen, C. Guo, J. Cao, J. Ren, and S. Tulyakov, "Maskcontrol: Spatio-temporal control for masked motion synthesis," in *ICCV*, 2025.

[77] Y. Xia, Q. Zhan, X. Luo, X. Shi, and Y. Wang, "Signmask: Structure-aware masked modeling for holistic 3d sign language production," *ACM TOMM*, vol. 22, no. 1, pp. 1–28, 2026.

[78] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman, "Maskgit: Masked generative image transformer," in *CVPR*, 2022.

[79] D. Li, C. Xu, X. Yu, K. Zhang, B. Swift, H. Suominen, and H. Li, "TSPNet: Hierarchical feature learning via temporal semantic pyramid for sign language translation," in *NeurIPS*, 2020.

[80] Y. Chen, F. Wei, X. Sun, Z. Wu, and S. Lin, "A simple multi-modality transfer learning baseline for sign language translation," in *CVPR*, 2022, pp. 5120–5130.

[81] J. Kim, H. Jeon, J. Bae, and H. Y. Kim, "Leveraging the power of mllms for gloss-free sign language translation," in *ICCV*, 2025.

[82] J. Ye, X. Wang, W. Jiao, J. Liang, and H. Xiong, "Improving gloss-free sign language translation by reducing representation density," *NeurIPS*, 2024.

[83] Z. Chen, B. Zhou, Y. Huang, J. Wan, Y. Hu, H. Shi, Y. Liang, Z. Lei, and D. Zhang, "C2rl: Content and context representation learning for gloss-free sign language translation and retrieval," *IEEE TCSVT*, 2025.

[84] B. Zhang, G. Tanzer, and O. Firat, "Scaling sign language translation," *NeurIPS*, 2024.

[85] P. Rust, B. Shi, S. Wang, N. C. Camgoz, and J. Maillard, "Towards privacy-aware sign language translation at scale," in *ACL*, 2024, pp. 8624–8641.

[86] W. Zhao, H. Hu, W. Zhou, J. Shi, and H. Li, "BEST: BERT pre-training for sign language recognition with coupling tokenization," in *AAAI*, 2023.

[87] W. Zhao, H. Hu, W. Zhou, Y. Mao, M. Wang, and H. Li, "Masa: Motion-aware masked autoencoder with semantic alignment for sign language recognition," *IEEE TCSVT*, 2024.

[88] S. Gueuwou, X. Du, G. Shakhnarovich, K. Livescu, and A. H. Liu, "Shubert: Self-supervised sign language representation learning via multi-stream cluster prediction," in *ACL*, 2025.

[89] S. Gan, X. Liu, Y. Yin, N. Liu, K. Liu, D. Tuerdaken, Z. Jiang, L. Xie, S. Lu, and H. Wen, "Learning effective sign features without text for gloss-free sign language translation," in *CVPR*, 2026.

[90] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *NeurIPS*, vol. 30, 2017.

[91] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3D hands, face, and body from a single image," in *CVPR*, 2019, pp. 10 975–10 985.

[92] H. Chang, J. Park, S. Ye, S. Yang, Y. Seo, D.-S. Chang, and M. Seo, "How do large language models acquire factual knowledge during pretraining?" *NeurIPS*, 2024.

[93] K. Tirumala, A. Markosyan, L. Zettlemoyer, and A. Aghajanyan, "Memorization without overfitting: Analyzing the training dynamics of large language models," *NeurIPS*, 2022.

[94] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-supervised learning from images with a joint-embedding predictive architecture," in *CVPR*, 2023.

[95] R. Battison, *Lexical borrowing in American sign language.* ERIC, 1978.

[96] M.-P. Forte, P. Kulits, C.-H. P. Huang, V. Choutas, D. Tzionas, K. J. Kuchenbecker, and M. J. Black, "Reconstructing signing avatars from video using linguistic priors," in *CVPR*, 2023, pp. 12 791–12 801.

[97] S. Imai, M. İnan, A. Sicilia, and M. Alikhani, "Silverscore: Semantically-aware embeddings for sign language generation evaluation," in *International Conference on Recent Advances in Natural Language Processing*, 2025.

[98] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.

[99] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or, "Motionclip: Exposing human motion generation to clip space," in *ECCV*, 2022.

[100] Z. Meng, Y. Xie, X. Peng, Z. Han, and H. Jiang, "Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression," in *CVPR*, 2025.

[101] Y. Zhang, Y. Wei, D. Jiang, X. Zhang, W. Zuo, and Q. Tian, "Controlvideo: Training-free controllable text-to-video generation," in *ICLR*, 2024.

[102] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou, "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation," in *ICCV*, 2023.

[103] Z. Xing, Q. Dai, H. Hu, Z. Wu, and Y.-G. Jiang, "Simda: Simple diffusion adapter for efficient video generation," in *CVPR*, 2024.

[104] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[105] M. Petrovich, M. J. Black, and G. Varol, "Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis," in *ICCV*, 2023, pp. 9488–9497.
[106] Z. Jiang, G. Sant, A. Moryossef, M. Müller, R. Sennrich, and S. Ebling, "Signclip: Connecting text and sign language by contrastive learning," in *EMNLP*, 2024.
[107] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.
[108] Y. Cheng, F. Wei, J. Bao, D. Chen, and W. Zhang, "Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning," in *CVPR*, 2023.
[109] M. J. Black, P. Patel, J. Tesch, and J. Yang, "Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion," in *CVPR*, 2023.
[110] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.
[111] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural networks*, vol. 107, pp. 3–11, 2018.
[112] S. Stoll, A. Mustafa, and J.-Y. Guillemaut, "There and back again: 3d sign language generation from text using back-translation," in *3DV*, 2022, pp. 187–196.
[113] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen, "Motiongpt: Human motion as a foreign language," *NeurIPS*, vol. 36, pp. 20 067–20 079, 2023.
[114] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *TACL*, vol. 8, pp. 726–742, 2020.
[115] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.
[116] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does bert look at? an analysis of bert's attention," in *ACL workshop BlackboxNLP*, 2019, pp. 276–286.
[117] I. Tenney, D. Das, and E. Pavlick, "Bert rediscovers the classical nlp pipeline," in *ACL*, 2019, pp. 4593–4601.
[118] J. H. Low, O. M. Sincan, and R. Bowden, "Sage: Segment-aware gloss-free encoding for token-efficient sign language translation," in *ICCVW*, 2025.
[119] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang, "A survey on mixture of experts in large language models," *TKDE*, 2025.
[120] S. McIntosh-Smith, S. Alam, and C. Woods, "Isambard-ai: a leadership-class supercomputer optimised specifically for artificial intelligence," in *Proceedings of the Cray User Group*. ACM, 2024, pp. 44–54.

Ronglai Zuo is currently a Research Associate at Imperial College London. He received his Ph.D. degree from the Hong Kong University of Science and Technology in 2024 and his B.Eng. degree from the Special Class for the Gifted Young, University of Science and Technology of China, in 2020. His research focuses on sign language processing, generative models, and multimodal learning. He has published several papers in top-tier conferences, including CVPR, ICCV, ECCV, and NeurIPS.

Rolandos Alexandros Potamias is an Assistant Professor at Imperial College London. He received his PhD from Imperial College London and his M.Eng. degree from National Technical University of Athens. His research is focused on 3D Computer Vision and Embodied AI with a particular focus on human and robot dexterity. He has published several papers in top-tier journals and conferences, including CVPR, ICCV, ECCV, NeurIPS, and ICML.

Qi Sun is an Imperial College London graduate with an MSc degree in Computing, specializing in Artificial Intelligence and Machine Learning. She received her BSc degree in Computer Science with First Class Honors from the University of Bristol in 2024. Her research focuses on machine learning, computer vision, natural language processing, and multimodal learning. She has published a full paper at the ACM SIGGRAPH European Conference on Visual Media Production 2024, which received the Best Paper award.

Evangelos Ververas is currently a Visiting Researcher at Imperial College London. He received his PhD from Imperial College London in 2023 supervised by Prof. Stefanos Zafeiriou. Previously, Evangelos received his diploma of Electrical and Computer Engineering from Aristotle University of Thessaloniki, in 2016. His research focuses on 3D human-centric vision and modelling. He has published papers in top-tier conferences and journals including CVPR, ICCV, ECCV, NeurIPS, IJCV and TPAMI.

Jiankang Deng (Member, IEEE) is currently an Assistant Professor with the Department of Computing, Imperial College London. His research explores multimodal foundation models and generative modeling of the physical world. He is one of the main contributors to the widely used open-source platform Insightface. He has over 24,000 citations for his research with an h-index of 53. He is an active area chair of prestigious computer vision and machine learning conferences (e.g., CVPR, ICCV, ECCV, ICML, NeurIPS, ICLR and AAAI). He is also an Associate Editor of IEEE Transactions on Image Processing, Transactions on Machine Learning Research, and Neural Networks. He is a Member of the IEEE.

Stefanos Zafeiriou (Member, IEEE) is a Professor in machine learning and computer vision with the Department of Computing, Imperial College London and a holder of the prestigious UKRI Turing AI World-Leading Research Fellowship. He has co-authored over 250 papers in top-tier machine learning and computer vision venues, including IEEE TPAMI and IJCV, as well as at leading conferences such as CVPR, ICCV, ECCV, NeurIPS, and ICML. His research focuses on machine learning models applied to computer vision and biosignal analysis. His work has garnered more than 50,000 citations, resulting in an h-index of 94. In recognition of his work, he has received Imperial College's President's Medal for Excellence in Research Supervision (2016) and the President's Medal for Excellence in Innovation and Entrepreneurship (2022). His students are frequent recipients of highly competitive fellowships, such as the Google Fellowship (x2), the Intel Fellowship, and the Qualcomm Fellowship (x4). He has served as an (Guest) Associate Editor for premier journals such as IEEE Transactions on Pattern Analysis and Machine Intelligence, International Journal of Computer Vision, and IEEE Transactions on Affective Computing. He has guest-edited more than 8 journal special issues and co-organized over 25 workshops and challenges at top conferences. He also served as the General Chair for BMVC 2017.