GLOS：手语生成与时间对齐的词汇级条件

Taeryung Lee[1,3] **Hyeongjin Nam**[1,2,3] **Gyeongsik Moon**[4] **Kyoung Mu Lee**[1,2]

[1] IPAI & ASRI, [2] Dept. of ECE, Seoul National University, Korea
[3] KRAFTON [4] Dept. of CSE, Korea University, Korea

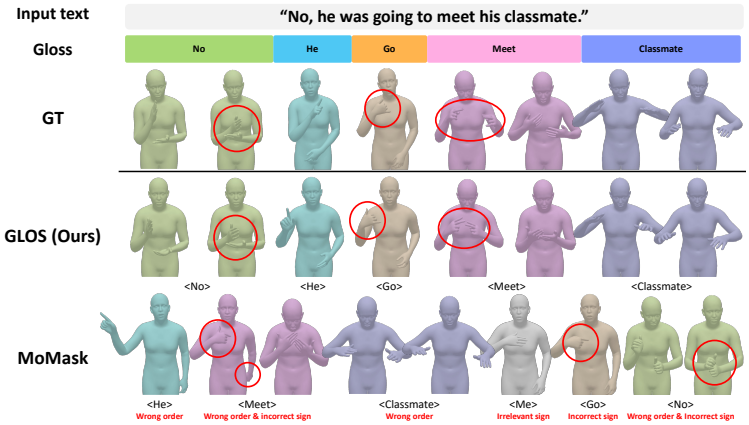


Figure 1: Qualitative comparison between our proposed GLOS and prior method (MoMask [12]). We represent the glosses with colors, and the sign segments corresponding to each gloss are painted in the same color. Human-interpreted glosses of generated signs are denoted in brackets < > below the signs. While the signs generated by the prior method exhibit incorrect lexical order and include inaccurate signs, our GLOS produces signs in the correct order with accurate movements.

Abstract

手语生成（SLG），或文本到手语的生成，弥合了手语使用者和非使用者之间的差距。尽管 SLG 在近期取得了进展，现有的方法在词汇顺序的正确性和语义准确性方面仍然常常存在不足。这主要是由于句子级条件的存在，即将输入文本的整个句子编码为一个单一的特征向量，作为 SLG 的条件。这种方法未能捕捉手语的时间结构，缺乏词级语义的细粒度，常常导致手语序列混乱和动作含糊。为了克服这些局限性，我们提出了 GLOS，一种具有时间对齐词汇级条件的手语生成框架。首先，我们采用词汇级条件，定义为与动作序列时间对齐的词汇嵌入序列。这使得模型能够在每个时间步访问手语的时间结构和词级语义。因此，这允许对手语进行细粒度的控制，更好地维护词汇顺序。其次，我们引入了一个条件融合模块，时间对齐条件（TAC），以有效传递词汇级条件提供的词级语义和时间结构到对应的动作时间点。我们的方法，包括词汇级条件和 TAC，生成的手语具有正确的词汇顺序和高语义准确性，在 CSL-Daily 和 Phoenix-2014T 上优于以往的方法。

Preprint. Under review.

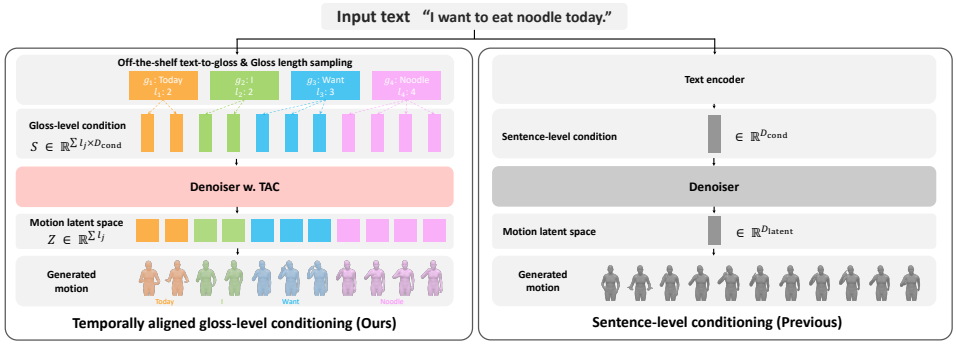


Figure 2: 概念比较。我们将我们的时间对齐词汇水平条件与传统的句子水平条件进行比较。我们的方法利用词汇水平条件与时间对齐条件（TAC）相结合，以确保词汇条件与相关动作片段之间的时间对齐，从而以正确的词汇顺序和高度语义准确性生成手势。

1 引言

手语是全球数百万名聋人及听力障碍人士交流的重要手段。然而，手语使用者与非手语使用者之间的沟通依然有限，依赖于专业翻译人员的帮助。作为降低这种沟通障碍的关键步骤，文本到手语生成（手语生成，SLG [42, 3, 43, 37]）近年来开始受到越来越多的研究关注。SLG 在使媒体和公共服务中的自动字幕和虚拟助手成为可能方面发挥着重要作用。手语通过一系列手势来传达意义，每个手势对应一个称为“手语词素”的类词语言单位。每个词素对齐的手势片段，被称为“词汇手语”，必须遵循正确的词汇顺序 [25, 8]，以确保所传达的意义准确无误。如图 1 所示，以往的方法常常存在词汇顺序不正确和语义准确性低的问题。一个关键原因在于使用了句子级条件，其中整个输入句子被编码为单一特征向量，而未考虑手语的时间顺序结构（见图 2，右）。这一策略引入了两个主要限制。首先，句子级条件无法捕捉手语的时间结构。由于缺乏明确的时间结构，模型常常产生顺序错误的手语。其次，它缺乏在每个时间步传达细粒度上下文的能力。这导致未能清晰对应个别词素的模糊手势。综上所述，这些限制阻碍了在生成准确手语时保持词汇顺序和语义准确性的能力。

为了解决上述两个限制，我们提出了 GLOS，一种具有时间对齐手语词汇级别条件的手语生成方法。如图 2（左）所示，我们的框架由两个部分组成：（1）词汇级别条件和（2）时间对齐条件（TAC）。首先，我们使用词汇级别条件，将其定义为与动作序列时间对齐的词汇嵌入的二维序列（时间和通道维度）。词汇级别条件捕捉了手语的顺序结构以及每个时间步骤所需的细粒度语义，从而克服了句子级别条件的限制。值得注意的是，由于语言顺序的差异，语音语言的标记化序列未能反映手语的时间结构。相反，我们使用现成的文本到词汇模型从输入文本中提取词汇，[7] 并在时间上将词汇嵌入与运动序列对齐。其次，我们引入了 TAC，一个条件融合模块，该模块在与每个词汇及其周围时间步相对应的局部时间范围内，将词汇级条件集成到运动潜在空间中。SLG 旨在从词汇序列生成一个结构化的运动帧序列，因此属于一般的序列到序列任务类别。虽然基于 Transformer 的交叉注意力 [34] 在机器翻译等序列到序列任务中被广泛使用，但我们发现其在手语生成中效果不佳。在交叉注意力中，每个运动时间步会全局关注所有词汇标记，导致时间上信息混杂过多。与之相反，我们的 TAC 能够在对应的时间步和时间局部环境中有效地将词汇级条件提供的词级语义传递到运动潜在空间中。我们利用 Adaptive Layer Normalization (AdaLN) [14] 进行时间对齐感知的调制，并使用一维卷积结合局部环境。这种对齐感知设计使得模型能够更好地捕捉手语的连续结构，并生成更准确且时间一致的手语序列。

大量实验表明，我们的方法在 CSL-Daily [41] 和 Phoenix-2014T [4] 上显著优于现有的 SLG 方法。我们的主要贡献总结如下。

• 我们提出了 GLOS，一种利用时间对齐的词汇级条件的手语生成框架。

我们引入了词义级别的条件，以在每个时间步提供细粒度的语义信息，从而解决句子级条件的局限性。

• 我们设计了时间对齐条件（TAC），以保持词语条件与动作时间步之间的时间对齐，从而实现准确且时间一致的手语生成。

2 相关工作

手语生成 手语生成（SLG） [3, 30, 29, 15, 43, 42, 38, 37] 旨在从文本输入中合成准确的手势序列，使与听力障碍人士的机器媒介交流成为可能。NSA [3] 构建了一个 3D 手语数据集并利用它来训练扩散模型。SOKE [42] 采用预训练的语言模型来生成手语，采用多头解码策略以有效融合不同的身体部位。T2S-GPT [37] 提出了一个基于 GPT 的自回归生成框架，该框架基于 VAE 编码的离散潜在空间。Spoken2Sign [43] 提出了一个实用的管道，能够从输入文本中提取词汇，对应词汇检索 3D 手势，并将它们连接以生成手语。

尽管在 SLG 领域取得了显著进展，但对文本输入的有效条件策略关注较少。现有大多数方法将输入文本编码为句子级条件，无法捕捉手语的时间结构，缺乏传达细粒度上下文的细致性。Spoken2Sign [43] 部分解决了这个问题，通过结合词汇级时间对齐，但它是通过独立检索每个词汇的预定义手势并进行插值实现的。这常导致不真实的过渡，并忽视了相邻词汇间的时间依赖关系。相比之下，我们提出的 GLOS 在端到端生成框架中对词汇级条件建模，确保时间对齐和局部上下文的整合。与基于检索的方法如 Spoken2Sign 不同，我们的方法通过基于扩散的解码 [10] 学习顺畅的过渡并捕捉相邻词汇间的依赖关系。

人类运动生成 人体动作生成目标是从文本描述中合成逼真且时间上连贯的身体运动。ACTOR 使用基于 Transformer 的 VAE 学习动作条件的运动嵌入，以生成多样化的人体运动。MDM 在运动空间中应用去噪扩散模型，以从自然语言描述中生成高质量的人体运动。MotionGPT 将人体运动视为离散的标记序列，并在字幕生成和语言理解任务中预训练一个统一的 transformer。MoMask 引入了一个具有掩码建模的分层 VQ 框架，以增强细粒度的运动生成。虽然现有的人体运动生成方法可以被改编用于手语生成，但简单的改编会导致与词汇序列的时间不对齐。我们提出的 TAC 设计增强了词汇和生成运动之间的时间对齐，以解决这一问题，从而实现更准确和富有表现力的手语合成。进一步的讨论见第 5.1 节。

3 初步

3.1 VQ-VAE

VQ-VAE [33, 39] 包含一个编码器 E ，一个解码器 G ，以及一个码本 $\mathcal{Z} = \{\boldsymbol{z}_k \in \mathbb{R}^d\}_{k=1}^K$ ，其中 K 是码本条目的数量，d 是每个码向量的维度。给定一个输入动作 $\boldsymbol{X} \in \mathbb{R}^{L \times D}$ ，其中 L 和 D 分别表示序列长度和特征维度，编码器将其映射为一个潜在表示 $\boldsymbol{Z}^e = E(\boldsymbol{X}) \in \mathbb{R}^{L' \times d}$ 。潜在序列的长度定义为 $L' = \lfloor L/r \rfloor$ ，其中 r 表示编码器 E 的时间缩小率。我们将编码器输出和量化后的潜在表示分别记为 $\boldsymbol{Z}^e$ 和 $\boldsymbol{Z}^q$ 。每个在时间步 i 的潜在向量 $\boldsymbol{Z}_i^e$ 被量化到最近的码本向量 $\boldsymbol{z}_k$ ：

$$\boldsymbol{Z}_i^q = \boldsymbol{z}_k, \quad \text{where } k = \arg\min_{k'} \|\boldsymbol{Z}_i^e - \boldsymbol{z}_{k'}\|_2. \tag{1}$$

解码器从量化的潜在码中重建动作。VQ-VAE 被训练以最小化以下损失函数：

$$L = \|\boldsymbol{X} - \hat{\boldsymbol{X}}\|_2^2 + \|\text{sg}(\boldsymbol{Z}^e) - \boldsymbol{Z}^q\|_2^2 + \beta\|\boldsymbol{Z}^e - \text{sg}(\boldsymbol{Z}^q)\|_2^2, \tag{2}$$

其中 $\text{sg}(\cdot)$ 是停止梯度操作，β 是一个损失权重。

3.2 VQ 扩散

VQ-Diffusion [10] 作用于离散的 VQ token，这是从一个预训练过的 VQ-VAE 中获得的有限集合的码本索引。与连续扩散模型 [13] 不同的是，这种方法在 VQ token 上定义了一个离散的马尔可夫扩散过程。

前向过程。令 $\boldsymbol{k}_0 \in \{1, \ldots, K\}^{L'}$ 表示来自码本的原始 VQ 令牌序列。前向扩散过程通过随机替换令牌，在 T 个时间步内逐渐破坏 $\boldsymbol{k}_0$ ，由一个转换矩阵 Q_t 定义：

$$q(\boldsymbol{k}_t \mid \boldsymbol{k}_{t-1}) = \boldsymbol{v}^\top(\boldsymbol{k}_t) Q_t \boldsymbol{v}(\boldsymbol{k}_{t-1}), \tag{3}$$

其中 $\boldsymbol{v}(\cdot)$ 表示 VQ 令牌的独热表示，$Q_t \in \mathbb{R}^{K \times K}$ 是一个行和为 1 的随机矩阵。转换矩阵是使用超参数 α_t 、β_t 和 γ_t 构建的，它们分别控制令牌保留、用随机令牌替换以及掩码的概率。在最后一个时间步 T ，所有令牌都被完全掩盖，并且掩盖的序列作为反向去噪过程的起始点，类似于在连续 DDPM 框架中从标准高斯分布进行采样。

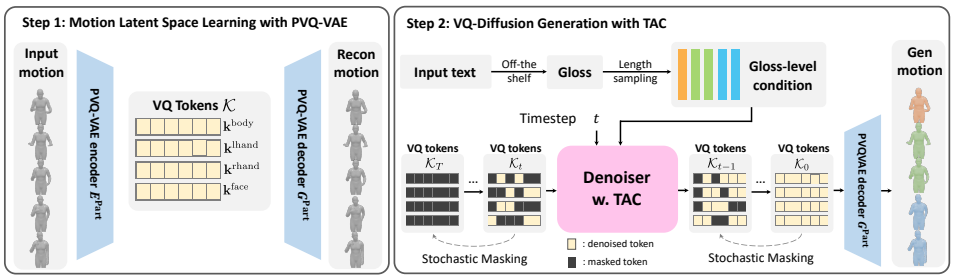


Figure 3: 概述。我们首先使用 PVQ-VAE 构建动作潜空间（左侧）。在生成阶段（右侧），我们首先用现成的文本到短语模型 [7] 和长度采样将输入文本编码为短语级条件。我们提出的具有 TAC 的去噪器生成动作潜空间中的符号，并将其转发到 PVQ-VAE 解码器以生成手语。

反向过程。反向过程的目标是从带噪的 $\boldsymbol{k}_t$ 中恢复 $\boldsymbol{k}_0$ 。这是通过估计逆转变来实现的：

$$p_\theta(\boldsymbol{k}_{t-1} \mid \boldsymbol{k}_t, \boldsymbol{y}) = \sum_{\tilde{\boldsymbol{k}}_0} q(\boldsymbol{k}_{t-1} \mid \boldsymbol{k}_t, \tilde{\boldsymbol{k}}_0)\, p_\theta(\tilde{\boldsymbol{k}}_0 \mid \boldsymbol{k}_t, \boldsymbol{y}), \tag{4}$$

，其中 $\tilde{\boldsymbol{k}}_0$ 是预测的去噪 VQ 符号序列，$\boldsymbol{y}$ 是条件输入（例如，文本提示）。去噪器模型可以为特定目的而设计，例如手语生成，预测初始状态 $p_\theta(\tilde{\boldsymbol{k}}_0 \mid \boldsymbol{k}_t, \boldsymbol{y})$ 。去噪器经过训练以用变分下界 [10] 估算后验转变分布 $q(\boldsymbol{k}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{k}_0)$ 。

4 GLOS

在本节中，我们描述我们的 GLOS，其中包含如图 3 所示的术语级条件和时间对齐条件（TAC）。我们的方法采用了一种基于 VQ-Diffusion [10] 的体部位感知生成模型。第 4.1 节详细介绍了使用 PVQ-VAE 学习体部位感知运动潜在空间的方法。第 **??** 节介绍了术语级条件的编码。第 **??** 节描述了 VQ-diffusion 生成的细节，包括具有 TAC 的去噪器的架构。

4.1 通过 PVQ-VAE 学习运动潜在空间

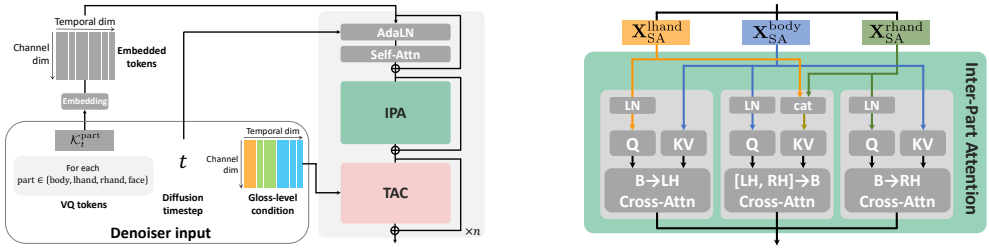
图 3 左边展示了我们使用提出的 PVQ-VAE 的运动潜在空间学习。我们使用 PVQ-VAE 学习运动潜在空间以增强手部和面部的表现力，这是受到之前研究工作的启发 [42, 5, 23] 。根据第 3.1 节中的符号，每个输入运动序列 $\boldsymbol{X}$ 被分解成 $\boldsymbol{X}^{\text{body}}$ 、$\boldsymbol{X}^{\text{lhand}}$, $\boldsymbol{X}^{\text{rhand}}$ 和 $\boldsymbol{X}^{\text{face}}$ 。我们用 $\boldsymbol{X}^{\text{part}}$ 表示每个部分的输入运动序列，其中 part $\in \{\text{body}, \text{lhand}, \text{rhand}, \text{face}\}$ 。

每个部分 $\boldsymbol{X}^{\text{part}}$ 的输入运动序列由一个专用的 1D 卷积编码器 E^{part} 和解码器 G^{part} 处理，具有每个部分对应的码书 $\mathcal{Z}^{\text{part}} = \{\boldsymbol{z}_k^{\text{part}} \in \mathbb{R}^{\text{d}}\}_{k=1}^{K}$ 。每个输入 $\boldsymbol{X}^{\text{part}}$ 被编码为 $\boldsymbol{Z}^{e,\text{part}} = E^{\text{part}}(\boldsymbol{X}^{\text{part}})$ ，然后使用码书量化为 $\boldsymbol{Z}^{q,\text{part}}$ 。量化特征解码为 $\hat{\boldsymbol{X}}^{\text{part}} = G^{\text{part}}(\boldsymbol{Z}^{q,\text{part}})$ 。这里，$\boldsymbol{Z}^{e,\text{part}}$ 和 $\boldsymbol{Z}^{q,\text{part}}$ 分别表示每个部分的编码器输出和量化潜在。训练目标是最小化所有部分的损失 $L = L^{\text{body}} + L^{\text{lhand}} + L^{\text{rhand}} + L^{\text{face}}$ ，每个项在方程 2 中定义。

如图 3 （右上）所示，我们首先使用现成的文本到词语转换方法 [7] 将输入文本转换为词语序列 $g = \{g_j\}_j$ 。我们从训练集中为每个词语 g_j 采样平均长度 l_j ，以构建长度序列 $l = \{l_j\}_j$ ，其中 j 表示词语索引。对于训练集中未包括的词语，我们使用所有词语的平均长度。

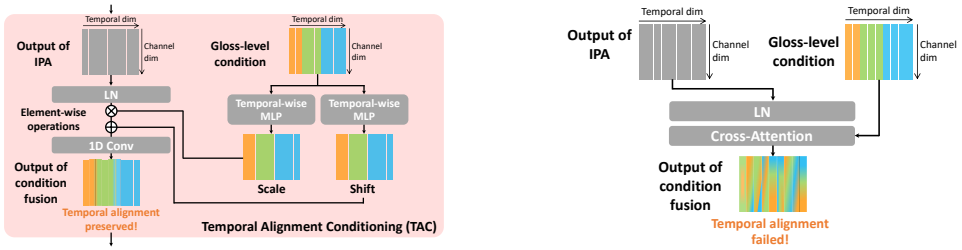
我们使用预训练的 mBART 提取每个 g_j 的光泽 $c_j \in \mathbb{R}^{D_{\text{cond}}}$ 的文本嵌入，其中 D_{cond} 表示条件向量的通道维度。然后，如图 2 左侧第二行所示，我们将每个光泽特征 c_j 重复 l_j 次并按顺序堆叠它们，以获得光泽级条件 $S \in \mathbb{R}^{L \times D_{\text{cond}}}$ 。请注意，$L = \sum_j l_j$ 表示所需的总运动长度。生成的光泽级条件在每个时间步提供了手语的顺序结构和词级语义，能够实现准确且时间顺序的手语生成。在本节中，我们描述了使用 TAC 的 VQ-Diffusion 生成过程。首先，我们描述去噪器的架构。然后，我们提出 TAC 以促进条件融合期间的时间对齐。我们还介绍了跨部件注意（IPA），以实现关节动作者之间的协调运动，这在单独使用 PVQ-VAE 编码身体部位时至关重要。

图 4a 可视化了所提出的去噪器的架构。除 IPA 模块内有交互外，每个身体部分——主体、左手、右手和脸部——是独立进行处理的。具体来说，对于每个部分，我们首先将 VQ tokens $k^{\text{part}} \in \mathbb{R}^{L}$ 表示为代码书索引的向量。VQ tokens 被形成嵌入 tokens $\boldsymbol{X}^{\text{part}}_{\text{embed}} \in \mathbb{R}^{L \times D_{\text{feat}}}$ 。D_{feat} 表示特征的通道维度，并在去噪器的正向传递过程中保持不变。然后，我们通过自适应层归



(a) 去噪器 (b) 跨部件注意力 (IPA)。

Figure 4: 模型架构。我们展示了所提出的去噪器和跨部位注意力（IPA）模块的详细架构。(a) 在我们的模型中，除了在 IPA 模块内，各个身体部位——身体、左手、右手和面部——被独立编码，没有跨部位的交互。(b) IPA 模块中的不同身体部位之间的交叉注意力促进了关节动作的更协调，使生成的手语质量更高。



(a) 时间对齐条件（TAC，我们的方法）。 (b) 交叉注意力。

Figure 5: 条件融合方法的比较。我们可视化了两种条件融合方法：(a) 我们的 TAC 由 AdaLN 和一维卷积组成，以保持与局部时间上下文的时间对齐。(b) 交叉注意力方法，该方法未能保持时间对齐约束。

一化（AdaLN）[14] 将扩散时间步 t 整合到嵌入 tokens $\boldsymbol{X}_{\text{embed}}^{\text{part}}$ 中，得到 $\boldsymbol{X}_{\text{time}}^{\text{part}}$。这会通过一个自注意力（SA）层，随后是残差连接，产生 $\boldsymbol{X}_{\text{SA}}^{\text{part}} = \text{SA}(\boldsymbol{X}_{\text{time}}^{\text{part}}) + \boldsymbol{X}_{\text{embed}}^{\text{part}}$。接着，$\boldsymbol{X}_{\text{SA}}^{\text{part}}$ 通过具有残差连接的部分间注意力（IPA）模块处理，生成 $\boldsymbol{X}_{\text{IPA}}^{\text{part}} = \text{IPA}(\boldsymbol{X}_{\text{SA}}^{\text{part}}) + \boldsymbol{X}_{\text{SA}}^{\text{part}}$。最后，特征与 TAC 中的光泽级条件 S 融合，产生输出 $\boldsymbol{X}_{\text{TAC}}^{\text{part}}$。

图

部件间注意力 (IPA)。 4b 展示了我们提出的 IPA 模块的架构。该模块旨在通过跨部位特征集成，使不同发音器之间实现协调运动，从而增强去噪过程。IPA 针对身体、左手和右手的特征进行操作。首先对每一个这些特征使用 LayerNorm [2] 进行标准化，然后在与其他身体部位的交叉注意中作为查询特征。具体来说，身体特征会关注手部特征的连接，而每只手会关注身体特征。模型通过这种交叉注意机制捕捉身体部位之间的依赖关系，生成精细的身体和手部特征。我们没有将 IPA 应用于面部分支，因为该方法并没有改善我们的实验。

时间对齐条件 (TAC)。 图 5a 显示了所提出的用于条件融合的 TAC，它由 AdaLN 和一个时间 1D 卷积组成。TAC 能够有效地将每个词汇级别的语义通过局部时间范围传递到运动潜在空间，映射到每个词汇及其周围时间步的对应时间步。这使模型能够捕捉手语的顺序结构，确保精确的对齐和语义准确性。

给定一个词汇层次条件 $S \in \mathbb{R}^{L \times D_{\text{feat}}}$，两个时序 MLP 独立地产生缩放 $u \in \mathbb{R}^{D_{\text{feat}}}$ 和移动 $v \in \mathbb{R}^{D_{\text{feat}}}$。这些向量用于通过 AdaLN 调节输入特征，随后进行一次一维卷积。整体操作定义为 $\boldsymbol{X}_{\text{TAC}}^{\text{part}} = \text{Conv1D}\left(\text{LN}(\boldsymbol{X}_{\text{IPA}}^{\text{part}}) \otimes (1+u) \oplus v\right)$，其中 $\otimes$ 和 $\oplus$ 分别表示逐元素相乘和相加。通过这一设计，AdaLN 确保每个词汇条件与其对应的时间步对齐，同时一维卷积使得模型能够考虑局部上下文，以实现稳定且连贯的生成。消融研究证明，我们的对齐感知融合策略优于交叉注意力，允许每个运动时间步关注所有词汇标记（图 5b）。

Table 1: 光泽水平条件和 TAC 的有效性。我们通过比较不同特征类型和条件方法的组合来验证光泽水平条件和 TAC 的贡献。我们的方法出现在最后一行。

Features	Temporal dim	Condition fusion	DTW-JPE ↓			Back translation		
			Body	Hand	All	WER ↓	BLEU-4 ↑	ROUGE ↑
Text	✗	AdaLN + 1D Conv (TAC)	0.242	0.579	0.411	93.53	1.89	14.70
Gloss	✗	AdaLN + 1D Conv (TAC)	0.275	0.722	0.447	98.24	0.45	10.99
Gloss	✓	Cross Attention	0.247	0.584	0.416	96.66	0.57	11.65
Gloss	✓	AdaLN + FC	0.212	0.492	0.364	77.87	7.47	25.78
Gloss	✓	AdaLN + 1D Conv (TAC, Ours)	0.201	0.443	0.347	62.24	13.27	35.83

Table 2: IPA 的有效性。我们通过测试多种注意力流的组合来验证 IPA 的贡献：身体 → 手（B → H）、手 → 身体（H → B）和身体 → 面部（B → F）。我们的方法出现在最后一行。

Attention			DTW-JPE ↓			Back translation		
B → H	H → B	B → F	Body	Hand	All	WER ↓	BLEU-4 ↑	ROUGE ↑
✗	✗	✗	0.207	0.451	0.356	70.11	10.33	31.99
✓	✗	✗	0.203	0.451	0.352	63.06	11.77	34.71
✗	✓	✗	0.206	0.464	0.353	69.37	10.67	32.11
✓	✗	✓	0.202	0.452	0.350	62.55	12.30	35.00
✗	✓	✓	0.202	0.453	0.348	63.91	12.48	35.31
✓	✓	✓	0.203	0.452	0.347	62.67	12.79	35.48
✓	✓	✗	0.201	0.443	0.347	62.24	13.27	35.83

5 实验

我们使用了两个大规模的手语数据集：CSL-Daily [41] 和 PHOENIX-2014T [4]，分别包含约 20K 和 8K 个样本，涵盖中文和德语手语。我们的实验在最大手语长度为 180 帧的情况下进行评估 [32, 39]。

根据之前的研究 [42, 3, 43]，我们使用多种针对运动准确性和语言忠实度的指标来评估生成手语的质量。对于运动级别的评估，我们使用结合关节位置误差的动态时间规整（DTW-JPE）[24]，该方法测量生成和参考运动的对齐关节轨迹之间的距离。为了评估语言忠实度，我们采用回译方法，使用一个连续手语识别（CSLR）模型 [7]，该模型将生成的手语映射回词汇或文本。由于 CSLR 模型 [7] 因资源有限无法在我们的系统中重现，我们重新实现了一个兼容的特征编码器。我们报告了手语到词汇评估的词错误率（WER）[21]，反映预测和参考词汇之间的编辑距离。对于手语到文本的评估，我们使用 BLEU-4 [26] 和 ROUGE [20]，分别计算 n-gram 精度和召回率。这两种指标对词序敏感，使其适用于评价生成手语的词汇顺序。

由于之前方法的评估系统不是公开可用的 [3, 42] 或者由于计算资源的限制而无法重现 [43, 7]，我们无法使用之前报告的定量结果与我们的结果进行比较。相反，我们在 CSL-Daily 和 PHOENIX-2014T 上重新运行了实验，使用官方实现（如果可用）[16, 32, 12]，或者在代码未公开发布时自己重现基线 [37, 3, 42]。

在本节中，我们展示了消融研究以验证我们的贡献。我们使用测试集样本的长度进行消融研究，以用于对齐释义条件、TAC 和 IPA，确保实验环境的可控性。为了进行稳健性分析，我们改为从训练集采样，以全面评估稳健性。

词汇对齐条件和时间对齐条件 (TAC)。 表格 1 清楚地展示了我们方法的有效性，该方法结合了词素级条件和 TAC。我们首先观察到没有时间结构的句子级条件会导致性能不佳（第一行）。类似地，1D 词素嵌入编码与句子级条件类似（如图 2 右上角所示）也导致低性能（第二行）。即使使用词素级条件，应用跨注意力而不保持时间对齐（第三行）也未能改善性能，这突显了时间对齐的重要性。这是因为，与我们的 TAC 模块不同的是，TAC 模块在运动空间中将词素条件集成到相应的局部时间范围内，而跨注意力允许每个运动时间步关注所有词素标记。这种全局整合扰乱了词素和运动之间的时间对齐，导致时间上的信息过度混合。最后，将 AdaLN 与全连接层（第四行）与我们的 TAC 块进行比较，证实了通过 1D 卷积引入局部上下文可以带来显著的性能提升。全连接层在每个时间步独立应用，没有利用相邻时间步的局部上下文。这些结果表明，性能提升源于通过词素级条件提供手语的时间结构和细粒度语义，并通过 TAC 保持时间对齐。

部件间注意力 (IPA)。 表 2 表明，IPA 显著提高了手语生成的性能。我们将体到手（B → H）对应于图 4b 的左侧和右侧分支，而手到体（H → B）的注意力对应于中央的分支。体到脸（B → F）的注意力在图中未展示，因为它未包含在我们最终的 IPA 设计中。单独应用 B → H 和 H → B 的注意力（第二行和第三行）在没有 IPA 的基线（第一行）上带来了性能提升。添加 B → F 的注意力则提供了额外的改进（第四行和第五行）。然而，当 B → H 和 H → B 的注意力都已经激活时（最后一行），添加 B → F（第六行）略微降低了性能。我们将

Table 3: 我们系统的鲁棒性。我们展示了我们方法在文本到手语模型选择和输入长度变化方面的鲁棒性。我们的方法出现在最后一行。

(a) 文本到手语模型的鲁棒性。

Text-to-gloss method	DTW-JPE ↓			Back translation		
	Body	Hand	All	WER ↓	B-4 ↑	ROUGE ↑
mT5 finetuned	0.238	0.536	0.422	73.84	16.74	41.76
M2M finetuned	0.231	0.526	0.401	70.19	16.15	41.60
mBART finetuned	0.229	0.523	0.395	70.17	17.21	41.42
Off-the-shelf [7]	0.214	0.467	0.377	55.15	15.70	40.43

(b) 对输入长度变化的鲁棒性。

Length variation	DTW-JPE ↓			Back translation		
	Body	Hand	All	WER ↓	B-4 ↑	ROUGE ↑
length + 8	0.229	0.479	0.420	58.12	14.13	38.08
length + 4	0.222	0.477	0.401	55.47	16.40	40.89
length - 4	0.208	0.468	0.365	56.88	14.91	38.62
length + 0	0.214	0.467	0.377	55.15	15.70	40.43

Table 4: 与现有技术方法的比较。我们报告了 CSL-Daily 和 Phoenix-2014T 数据集中与之前现有技术方法的定性比较。无论是从训练集还是测试集中采样的输入长度，我们的 GLOS 都在这两个数据集上以较大幅度超过之前的方法。

Method	CSL-Daily						Phoenix-2014T					
	DTW-JPE ↓			Back translation			DTW-JPE ↓			Back translation		
	Body	Hand	All	WER ↓	BLEU-4 ↑	ROUGE ↑	Body	Hand	All	WER ↓	BLEU-4 ↑	ROUGE ↑
T2S-GPT [37]	0.278	0.644	0.453	93.01	2.05	16.57	0.593	1.037	0.825	98.40	2.35	8.78
MotionGPT [16]	0.653	0.901	0.797	95.04	2.08	15.63	0.907	1.145	1.024	94.08	5.54	15.51
MDM [32]	0.318	0.748	0.529	95.54	3.01	17.29	0.427	1.061	0.748	98.27	5.44	14.90
MoMask [12]	0.314	0.641	0.492	91.36	3.57	19.77	0.355	0.675	0.538	91.33	6.92	21.77
NSA [3]	0.314	0.740	0.523	95.40	1.72	15.18	0.309	0.680	0.541	94.18	6.24	18.73
SOKE [42]	0.272	0.649	0.447	95.68	1.95	15.62	0.258	0.589	0.441	97.84	4.23	13.16
Ours (Test len)	0.201	0.443	0.347	62.24	13.27	35.83	0.222	0.520	0.381	87.93	8.54	24.38
Ours (Train len)	0.214	0.467	0.377	55.15	15.70	40.43	0.228	0.524	0.400	84.44	8.28	22.95

此归因于训练集中面部表情的多样性有限，使得面部特征信息量减少，并可能引入冗余或噪声信号，超出了已经丰富的身体与手部的互动信号。基于这些发现，我们采用身体和手之间的交叉注意力作为我们的最终 IPA 设计，如图 4b 所示。

针对文本到符号以及长度采样的鲁棒性。 表 3 展示了我们的方法在文本到手势模型和手势长度变化下的鲁棒性。左侧模块基于微调的 mT5 [36]、M2M [9] 和 mBART [22] 方法报告了定量结果。尽管在 DTW-JPE 和 WER 上有轻微的性能下降，我们的模型始终实现了较强的反向翻译准确性，表明在不同的文本到手势方法面前具有鲁棒性。右侧模块评估了在不同手势长度下的性能。此处，输入长度在原始长度的 -4 至 $+8$ 范围内进行了扰动。考虑到 CSL-Daily 中平均词汇手势长度为 16，这相当于约 -25% 至 $+50\%$ 的变化。我们的模型在此范围内保持稳定性能，展现出强大的时间鲁棒性。

5.1 与当前先进方法的比较

表 4 清楚地表明，我们提出的 GLOS 在 CSL-Daily [41] 和 Phoenix-2014T [4] 数据集上明显优于以往的方法。我们在两种手语长度采样策略下报告结果：使用测试集样本长度（第七行）和从训练集中采样（最后一行）。在这两种情况下，GLOS 始终优于所有基线方法。

在 CSL-Daily（左侧模块）上，GLOS 比现有方法取得了显著的改进。虽然 MoMask [12] 在现有工作中报告了最佳的返译分数，SOKE [42] 获得了最佳的 DTW-JPE 分数，但我们的方法在这两个指标上均以明显优势超越了它们。同样地，在 Phoenix-2014T（右侧模块）上，GLOS 也达到了最新的研究水平。尽管 MoMask 和 SOKE 分别在返译和 DTW-JPE 方面表现出色，但无论采样策略如何，我们的方法始终优于它们的最佳结果。图 **??** 显示，与 MoMask [12] 和 NSA [3]（在以前的方法中表现最佳）相比，我们的 GLOS 生成正确词汇顺序的准确手语。每个词素都用不同的颜色表示，相应的词汇手语也用与词素相同的颜色表示。人工解释的词素在括号 < > 内表示。

6 结论

我们提出了 GLOS，一种具有时间对齐的词汇级别条件的手语生成框架。通过结合词汇级别的条件和 TAC，GLOS 解决了句子级别条件的关键限制：缺乏手语的时间结构和无法传达每个时间步细粒度上下文的细微差别。通过克服上述两个限制，我们的 GLOS 成功生成了准确的手语。我们提出的方法在 CSL-Daily 和 Phoenix-2014T 数据集上的表现大大优于之前的方法。

References

[1] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. TEACH: Temporal Action Compositions for 3D Humans. In 3DV , 2022.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. In arXiv , 2016.

[3] Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. Neural Sign Actors: A diffusion model for 3d sign language production from text. In CVPR , 2024.

[4] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In CVPR , 2018.

[5] Changan Chen, Juze Zhang, Shrinidhi Kowshika Lakshmikanth, Yusu Fang, Ruizhi Shao, Gordon Wetzstein, Li Fei-Fei, and Ehsan Adeli. The language of motion: Unifying verbal and non-verbal language of 3d human motion. In arXiv , 2024.

[6] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In CVPR , 2023.

[7] Yutong Chen, Ronglai Zuo, Fangyun Wei, Yu Wu, Shujie Liu, and Brian Mak. Two-stream network for sign language recognition and translation. In NeurIPS , 2022.

[8] Qi Cheng and Rachel I Mayberry. Acquiring a first language in adolescence: The case of basic word order in american sign language. Journal of child language , 46(2):214–240, 2019.

[9] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. In arXiv , 2020.

[10] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In CVPR , 2022.

[11] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. TM2T: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In ECCV , 2022.

[12] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative masked modeling of 3d human motions. In CVPR , 2024.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In NeurIPS , 2020.

[14] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In ICCV , 2017.

[15] Eui Jun Hwang, Jung Ho Kim, Suk Min Cho, and Jong C. Park. Non-autoregressive sign language production via knowledge distillation. In BMVC , 2021.

[16] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. MotionGPT: Human motion as a foreign language. In NeurIPS , 2024.

[17] Taeryung Lee, Gyeongsik Moon, and Kyoung Mu Lee. MultiAct: Long-term 3d human motion generation from multiple action labels. In AAAI , 2023.

[18] Taeryung Lee, Fabien Baradel, Thomas Lucas, Kyoung Mu Lee, and Gregory Rogez. T2LM: Long-term 3d human motion generation from multiple sentences. In CVPR Workshop on Human Motion Generation , 2024.

[19] Shuai Li, Sisi Zhuang, Wenfeng Song, Xinyu Zhang, Hejia Chen, and Aimin Hao. Sequential texts driven cohesive motions synthesis with natural transitions. In ICCV , 2023.

[20] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out , 2004.

[21] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-04) , pages 605–612, 2004.

[22] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics , 8:726–742, 2020.

[23] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. In arXiv , 2023.

[24] Meinard Müller. Information retrieval for music and motion , volume 2. Springer, 2007.

[25] Donna Jo Napoli and Rachel Sutton-Spence. Order of the major constituents in sign languages: Implications for all language. Frontiers in psychology , 5:376, 2014.

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In ACL , 2002.

[27] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In ICCV , 2021.

[28] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In ECCV , 2022.

[29] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Adversarial training for multi-channel sign language production. In BMVC , 2020.

[30] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Progressive transformers for end-to-end sign language production. In ECCV , 2020.

[31] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In ICLR , 2024.

[32] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In ICLR , 2023.

[33] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In NeurIPS , 2017.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NeurIPS , 2017.

[35] Bizhu Wu, Jinheng Xie, Keming Shen, Zhe Kong, Jianfeng Ren, Ruibin Bai, Rong Qu, and Linlin Shen. Mg-motionllm: A unified framework for motion comprehension and generation across multiple granularities. In CVPR , 2025.

[36] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In NAACL , 2021.

[37] Aoxiong Yin, Haoyuan Li, Kai Shen, Siliang Tang, and Yueting Zhuang. T2S-GPT: Dynamic vector quantization for autoregressive sign language production from text. In ACL , 2024.

[38] Zhengdi Yu, Shaoli Huang, Yongkang Cheng, and Tolga Birdal. SignAvatars: A large-scale 3d sign language holistic motion dataset and benchmark. In ECCV , 2024.

[39] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2M-GPT: Generating human motion from textual descriptions with discrete representations. In CVPR , 2023.

[40] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. MotionDiffuse: Text-driven human motion generation with diffusion model. IEEE TPAMI , 2024.

[41] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In CVPR , 2021.

[42] Ronglai Zuo, Rolandos Alexandros Potamias, Evangelos Ververas, Jiankang Deng, and Stefanos Zafeiriou. Signs as Tokens: A retrieval-enhanced multilingual sign language generator. In arXiv , 2024.

[43] Ronglai Zuo, Fangyun Wei, Zenggui Chen, Brian Mak, Jiaolong Yang, and Xin Tong. A simple baseline for spoken language to sign language translation with 3d avatars. In ECCV , 2024.

7

NeurIPS 论文清单