CCL24-Eval任务10系统报告：结合LLM与3D动画技术的手语数字人系统

杨阳[1,2]，张颖[2]，黄锴宇[2]，徐金安[2,*]
1. 果不其然无障碍科技（苏州）有限公司
2. 北京交通大学
{y.yang, novzying, kyhuang, jaxu}@bjtu.edu.cn

摘要

手语翻译（Sign Language Translation, SLT）系统作为一种重要的辅助技术，为听障人士提供了与他人沟通的有效途径。然而，传统手语翻译系统在准确性、流畅性差等方面存在问题。本文提出了一种结合大语言模型（Large Language Model, LLM）和3D动画技术（3D Animation Technology）的手语翻译系统，旨在克服这些局限，提高翻译的准确性和流畅性。本文详细介绍了系统的设计与实现过程，包括提示词设计、数据处理方法以及手语数字人翻译系统的实现。实验结果表明，采用LLM方法在手语翻译中能够生成较为自然和准确的结果。在标准评估和人工评估的两种评估方法下，本系统在大多数情况下能够较好地完成手语翻译任务，性能优于传统方法。本文的研究为进一步改进手语翻译系统提供了有益的参考和启示。

关键词： 手语数字人；手语翻译；大语言模型

System Report for CCL24-Eval Task 10: A Sign Language Avatar System Integrating LLM and 3D Animation Technology

Yang Yang[1,2], Ying Zhang[2], Kaiyu Huang[2], Jinan Xu[2,*]
1. GoBetterStudio
2. Beijing Jiaotong University
{y.yang, novzying, kyhuang, jaxu}@bjtu.edu.cn

Abstract

Sign Language Translation (SLT) systems serve as a crucial assistive technology, providing an effective means of communication for individuals with hearing impairments. However, traditional SLT systems face challenges in terms of accuracy and fluency. This paper proposes a novel SLT system that combines Large Language Models (LLM) and 3D Animation Technology to address these limitations and enhance translation accuracy and fluency. The paper provides a detailed account of the system's design and implementation process, including prompt design, data processing methods, and the implementation of the sign language digital human translation system. Experimental results demonstrate that the LLM-based approach can generate more natural and accurate translations in SLT. Under both standard and human evaluations, this system performs the SLT tasks better than traditional methods in most cases. This research offers valuable insights and references for further improving SLT systems.

Keywords: Sign Language Avatar , Sign Language Translation , Large Language Model

*通讯作者/Corresponding author

第二十三届中国计算语言学大会论文集，第393页-第404页，太原，中国，2024年7月25日至28日。
卷3：评测报告

1 引言

手语是一种复杂的视觉语言，主要用于听障人群的交流。它不仅包括手部动作，还涉及面部表情和身体姿态等多方面的表达。这些特性使得手语的翻译和生成面临巨大挑战。近年来，随着人工智能（Artificial Intelligence, AI）和自然语言处理（Natural Language Processing, NLP）技术的快速发展，开发高效的手语翻译系统成为可能。手语数字人（Sign Language Avatars）可以模拟手语动作，为聋人提供实时或非实时的翻译服务，不仅能极大地提升听障人士的交流能力，还能促进社会的包容性和无障碍环境的建设。

手语数字人翻译属于手语生成（Sign Language Production, SLP）领域，即以口语作为源语言，生成相应的手语表达(Yin et al., 2021)。现有的手语生成方法主要分为以下几类，各自存在明显的局限性:

- 拼接手语图片：通过从现有语料中标记图片并进行拼接来生成翻译结果。其局限性在于，准确性受限于语料的质量和覆盖范围，难以应对复杂的手语表达。
- 姿态估计技术(Zuo et al., 2024; Forte et al., 2023)：这一关键方法通过分析输入文本来解码并确定手部姿态信息，如手部的位置、形状和运动轨迹。然而，这种方法需要大量高质量的手语视频数据来训练姿态估计模型，数据的质量和多样性直接影响生成效果。
- 生成式模型（如扩散模型）(Baltatzis et al., 2024; Saunders et al., 2020a)：这些模型通过不断预测下一帧图像，以生成连续的手语动作。尽管能够生成高质量的连续图像序列，但其计算复杂度高，难以实现实时翻译。

在实际应用中，选择合适的方法并在现有方法基础上进一步提高手语生成的准确性和自然性，是未来研究的重要方向。

本研究提出了一种创新性手语翻译系统，该系统融合LLM的文本处理能力与3D动画技术的表现力，具备深度语义理解，能够深入分析复杂的语言结构和上下文信息，生成更为准确和自然的手语翻译。此外，系统设计了多样化的上下文学习样本，确保翻译结果在语义和语境上的双重精准，满足多样化的交流场景需求。3D动画技术的应用，进一步丰富了手语的视觉呈现，创造出既流畅又逼真的手语动作，这些动作不仅在视觉上吸引人，更在表达上贴近自然手语的非言语特征，如节奏、力度和表情。系统具备实时交互与反馈的能力，确保用户输入能够得到即时响应，有效提升了沟通的效率和用户体验。多模态融合技术的应用，整合了手部动作、面部表情和身体语言，实现了全面的交流表达，使手语翻译更加生动和直观。系统的可扩展性与兼容性设计，也使其能够灵活地集成到不同的平台和设备中，为用户提供了便捷的接入和使用体验。

综上所述，本文研究在手语翻译技术领域提供了一种全新的视角和解决方案，不仅推动了技术的发展，也为听障社群带来了更加丰富和便捷的沟通方式，有助于构建一个更加包容和无障碍的交流环境。

2 数据处理

数据预处理是开发手语数字人翻译系统的关键步骤之一。高质量的数据能显著提升模型性能，进而提高翻译准确性。本文在数据预处理过程中主要包括数据准备和数据格式化两个部分，这些步骤确保了模型能够有效地学习和翻译手语。

2.1 数据集

本研究严格遵循评测标准，所采用的手语语料库均来源于经过评测任务[1]授权的数据库，如表 1所示，包括“XMU_CSL”、“BUU_CSL”以及“ZZSZY_CSL”。这些语料库覆盖了日常生活中的各类典型情境，如医疗沟通、客户服务、交通指示和购物交流等多个关键领域。通过这种跨领域的数据使用策略，本文确保了数据的全面性和多样性，为手语生成模型的训练与评估提供了坚实基础。

[1] https://github.com/ann-yuan/QESLAT-2024

Dataset	Lang.	Sentences	版权所有者
XMU_CSL	CSL	500	厦门大学
BUU_CSL	CSL	500	北京联合大学
ZZSZY_CSL	CSL	74	株洲手之声信息科技有限公司

Table 1: 本研究中使用的手语语料库概览，包括语料库名称、使用的语言（CSL代表中国手语）、句子数量以及版权所有者。表格展示了三个主要的语料库，共有1074个句子。

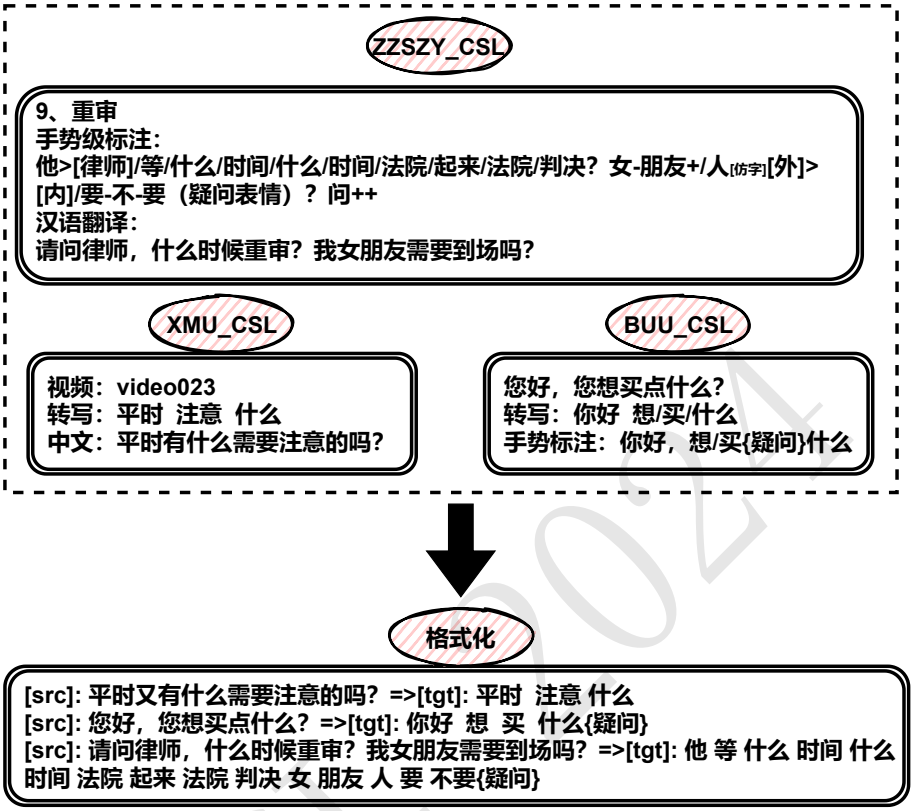

Figure 1: 数据格式化流程图，展示了如何将不同标注风格的语料库内容转换为统一格式。

如图 1所示，这些语料库不仅包含手语记录，还包含丰富的标注信息。每个条目都由中文源文本及其对应的Gloss组成。Gloss作为一种标准化的手语文本表现形式，为每个具体的手语动作提供了精确描述。这种一对一的映射关系极大地方便了从文本到手势的直接转换，提升了模型训练的效率和准确性。例如，中文句子“平时有什么需要注意的吗？”在Gloss中可能被分解为“平时”、“注意”和“什么”，这种分解不仅体现了手语的构词特点，也为理解和生成手语提供了结构化的视角。

如Muller(2023)所述，Gloss存在以下缺点：

- 信息丢失：Gloss并非手语的完整表示，缺乏手语的非手动通道（如面部表情和身体语言）和三维空间使用等语言线索。图 1中，“XMU_CSL”是词级标注，只标注出了手语动作。为了减轻信息丢失的问题，“BUU_CSL”加入了表情信息标注，非手部动作标记等。“ZZSZY_CSL”则更进一步，使用了更加详细的手势级标注，同样包括表情信息和非手部动作标记等。

- 不一致性：不同语料库中的Gloss转写标准差异很大，导致不同语料库或跨语言的Gloss不可比。如图 1所示，“XMU_CSL”、“BUU_CSL”以及“ZZSZY_CSL”存在不同的标注风格。

- 资源密集：Gloss的转写过程需要由专家语言学家完成，非常劳动密集，如果简单应用一些数据增强方法，容易导致翻译错误。
- 非实际应用：Gloss是语言学工具，并非聋人社区中确立的书写系统，手语用户通常在日常生活中不阅读或书写Gloss，所以很难从其他来源获取Gloss。

尽管存在这些缺点，Gloss也有许多优势，这些优势使其在手语翻译系统中仍然具有重要作用：

- 文本兼容性：Gloss作为手语的语义标签，在英语中通常由口语单词的大写基本形式组成，汉语中则由字词组成，能够无缝融入现有的机器翻译(Machine Translation)流程。
- 简化处理：由于Gloss是文本形式的，现有的机器翻译方法只需最少修改就能应用。
- 易于理解：对于机器翻译研究者来说，Gloss提供了一种相对容易理解手语的方式，因为它们以文本形式呈现。

利用这些高质量的手语语料库，本文旨在深入探索并推动手语翻译技术的发展，为听障群体提供更加准确、自然且高效的交流辅助工具。这些资源的多样性和丰富性，为本研究的深度分析和模型优化提供了坚实的数据支持。

2.2 数据格式化

数据格式化是数据处理流程中的关键步骤，涉及将原始数据转换成适合LLM进行上下文学习的形式。这一过程对于确保模型能够有效学习Gloss的翻译规则至关重要。

首先，将每条数据中的中文源文本与其对应的Gloss进行配对，形成源文本和目标文本对。这些文本对构成了模型训练的基本单元，使得模型能够在上下文中学习从中文到Gloss的映射。然后，本文利用生成的源文本和目标文本对，构建模型的上下文学习样本。

具体细节如图 1所示。为了确保所有数据的格式统一，便于模型批量处理，本文将不同来源的语料处理为相同格式的数据。在源文本和目标文本中添加特殊标记，以帮助模型识别源目语言的边界。例如，在源文本前添加“[src]”标记，在目标文本前添加“[tgt]”标记。在处理表情时，本文将表情块置于手语块之后，手语和表情视为一个同步块，同步块表示手语和表情在客户端会同时启动。

通过上述步骤的数据预处理，确保了数据的高质量和一致性，为后续模型的训练和评估提供了坚实基础。数据预处理不仅提升了模型性能，也为手语翻译系统的实际应用奠定了基础。本文将在未来的工作中继续扩展数据集，探索更加复杂的手语表达形式，并进一步优化数据预处理流程，以应对更复杂的手语翻译任务。

3 方法

为有效解决手语识别中的歧义性和边界模糊性问题，本文提出了一种结合LLM和3D动画技术的手语翻译系统，为听障者提供更加准确和便捷的沟通工具。系统整体架构如图 2所示，主要包括推理阶段和可视化阶段。在推理阶段，系统首先读取用户的中文输入，并结合知识库语料与LLM将中文翻译为Gloss。随后，手语数字人模块生成手语的视觉表示。系统的核心在于利用LLM结合上下文信息动态生成准确的Gloss。在可视化阶段，系统以多模态方式展示结果，使用户能够直观理解和验证手语转换结果。这一流程不仅提高了手语翻译的准确性和效率，而且通过生动的视觉表现增强了交流的自然性和直观性。

3.1 Gloss推理阶段

在推理阶段，系统主要集成了知识库语料与LLM。LLM的翻译任务提示词设计如表 2所示，包含任务描述、示例和翻译任务三部分(Agrawal et al., 2022; Vilar et al., 2023; Zhang et al., 2023)。任务描述提供了手语翻译任务的背景信息，解释了[src]和[tgt]的含义，即从中文原文翻译到手语的文本表示形式——Gloss，以及Gloss在机器学习模型中的重要性。示例部分展示了人类专家编撰的翻译例子，帮助模型理解如何进行翻译，“example”是占位符，代表上下文学习样本。翻译任务部分明确说明需要翻译的具体中文句子，即占位符“chinese”，并提示模型生成对应的Gloss。这种设计旨在提高模型对手语翻译准确性和自然性的理解，进而提升翻译质量。

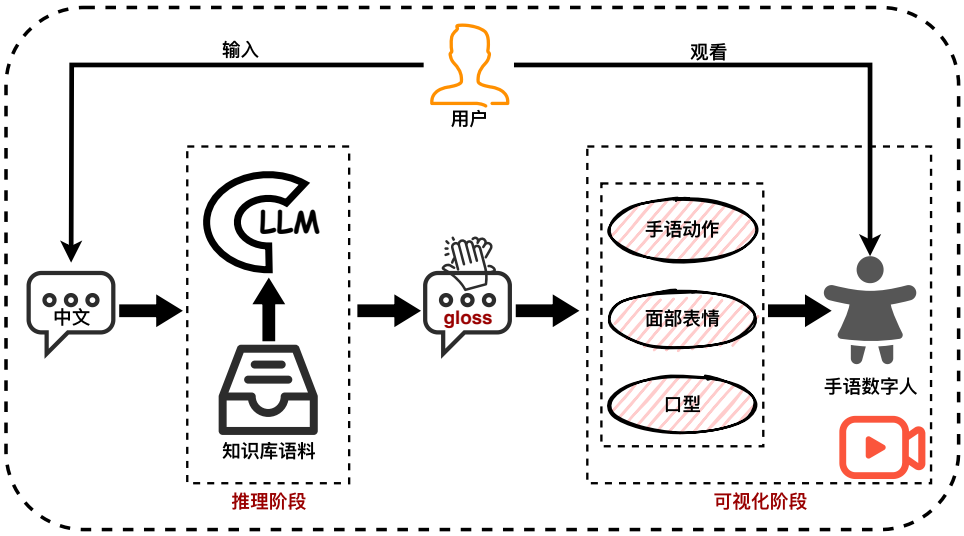

Figure 2: 手语翻译系统的架构示意图，包括两个核心阶段：推理阶段和可视化阶段。

分类	提示词
任务描述	你是一个手语翻译专业助手。需要将[src]翻译到[tgt]，[src]中是中文，[tgt]中是对应的gloss，gloss是手语的文本表示， 对于机器学习模型来说，gloss是非常友好的一种表示。
示例	下面给出了人类专家编撰的例子，你需要仔细学习，然后给出最后一行翻译的gloss。{example}
翻译任务	[src]: {chinese} ⇒ textnormal{[tgt]}:

Table 2: 用于LLM翻译任务的提示词设计。

如表 3所示，本文充分利用LLM出色的文本处理能力，将处理后的数据输入到模型中，使其能够深刻理解知识库语料并生成最终的Gloss。除此之外，本文对比了包括文心4.0、通义Plus和星火Pro在内的多个模型，旨在评估它们在翻译准确性和自然性方面的表现。通过这种比较，可以洞察不同模型在理解和转换自然语言到手语Gloss序列的能力上的差异。整个过程涉及复杂的语义转换和上下文理解，不仅提升了翻译的准确性，还增强了翻译的自然性，使手语翻译系统在实际应用中更加实用和可靠。

3.2 可视化阶段

在可视化阶段，将LLM输出的Gloss视为一系列具体的指令，指导数字人的行为表现。客户端采用了一个创新的3通道动画状态机设计，包括手语通道、表情通道和口型通道，以实现动作、表情和口型的同步播放。

3D手语动作库的构建 本文在构建3D手语动作库的过程中，严格遵循了国家标准手语词典的规范，精心制作了所有国家标准手语词典中的手语动作文件，总计超过8000个GLTF格式的动作文件。为了确保翻译的准确性和一致性，本文将每个动作与相应的Gloss进行了一一对应。此外，本文对每个动作进行了详细的标注，包括手势起始位置、结束位置、动作时长以及手势轨迹等信息，以便于在实际应用中能够精确调用和展示。

为了进一步提高手语动作库的实用性和覆盖面，本文还扩展了国家标准手语词典的内容，增加了常用的专业术语和新兴词汇的手语动作。同时，针对不同手语使用者的习惯差异，本文提供了多种手语动作的变体，以满足不同用户的需求。

手语表现 在手语表现阶段，系统首先对Gloss进行解析，将其拆分为手语动作、表情和口

	Source	Reference	LLM	Inference
Train	两斤多点，行不行?	两斤 多 一点 行 不行	文心4.0 通义Plus 星火Pro	两斤 多 一点 行 不行 两斤 多 一点 行 不行 两斤 多 一点 行 不行
	是的，我需要手语翻译。	是 手翻 我 需要	文心4.0 通义Plus 星火Pro	是 手翻 我 需要 是 手翻 我 需要 是 手翻 我 需要
Test	您的外卖到了，请尽快领取，谢谢。	外卖 到 拿 快 谢谢	文心4.0 通义Plus 星火Pro	外卖 你 到 快 取 谢谢 您的外卖 到 请 快领 谢谢 外卖 到 你 快 领取 谢谢
	具体需要做哪些检查呢?	具体 检查 哪些	文心4.0 通义Plus 星火Pro	具体 检查 哪些 具体 检查 哪些 需要 检查 具体 哪些 需要
	如果发现有诈骗情况，要及时报警	如果 发现 骗子 有 打电话 报警 快	文心4.0 通义Plus 星火Pro	如果 发现 诈骗 有 报警 及时 如果 发现 诈骗 情况 要 报警 发现 诈骗 情况 及时报警
	请问您需要将行李托运至何处?	问问 你 行李 托运 到 哪里	文心4.0 通义Plus 星火Pro	你 行李 托运 哪里 请问 问 你 行李 托运 哪里 行李 托运至何处 你 需要

Table 3: 不同LLM在处理手语翻译任务时的推理结果对比。其中Source列展示原始的中文句子，Reference列提供这些句子的标准标注结果，Inference列展示各个LLM生成的推理结果。

型，分别对应动画状态机的三个通道。手语动作的过渡采用了动画混合技术，该技术通过插值计算在不同动作之间生成平滑过渡，使得手语动画更加自然流畅。具体而言，动画混合技术能够在动作切换时计算中间帧，从而避免生硬的动作切换，提升视觉连贯性。

表情部分使用预先制作的常见情绪表情，这些表情存储在YAML格式的配置文件中，通过morph通道控制数字模型的面部表情变化，确保面部表情的自然性和准确性。对于口型部分，本文采用ARKit BlendShape技术，口型数据来源于Peng(2023)的工作结果。该模型通过对输入句子的情感和TTS语音特征进行分析，生成对应的口型变化数据，尽可能保证口型动作与手语内容的高度一致和同步。

可视化阶段与用户交互环节共同构成一个闭环系统，旨在提供高质量的多模态交流体验。手语翻译系统通过多模态方式展示结果，使用户能够直观地理解和验证手语的转换结果。整个系统的设计和实现，不仅提升了手语翻译的准确性和自然性，还增强了用户的互动体验，为手语使用者提供了一个更加便捷和高效的交流工具。

4 结果

4.1 实验评估

数据集划分与评估方法 为了评估不同LLM在手语翻译中的表现，本文将数据集划分为训练集和测试集。考虑到商业LLM的token限制，本文随机选择了语料库中100条数据作为训练集（即上下文学习样本），其余数据作为测试集。这种划分方式确保了在有限的资源下能够充分训练模型并评估其泛化能力。本文选择了三种不同的知名商业大模型进行评估，通义千问(Bai et al., 2023)，百度文心(Sun et al., 2021)，讯飞星火[2]，并从训练集和测试集中各选择了10句，通过LLM翻译成Gloss，对结果进行评估。评估指标包括BLEU-1（B1）、BLEU-2（B2）、BLEU-3（B3）、BLEU-4（B4）和ROUGE评分，这些指标能够综合反映模型在翻译质量上的表现。

[2] https://gitee.com/iflytekopensource/iFlytekSpark-13B

自动评测结果 表 4展示了不同LLM在训练和测试阶段的表现。评估采用了BLEU-1至BLEU-4和ROUGE指标，这些指标衡量了模型生成文本与参考文本之间的相似度。表格中列出了中国部分知名商业大模型的表现。从总体上看，各模型的表现均较好，显示了本文系统的通用性。然而，在测试阶段，文心4.0的表现最为优异，特别是在生成测试集的Gloss序列方面具有显著优势。这表明文心4.0在实际应用中具有更高的可靠性和准确性。评估训练集样本的目的是验证LLM在上下文学习中的有效性，而测试集样本的评估则用于测试LLM在实际应用中的泛化能力。文心4.0在所有评测指标上均优于其他模型，特别是在生成测试集序列Gloss时表现尤为突出。这可能是因为文心4.0在语义理解和上下文处理上有更强的能力。

LLM	Train					Test				
	B1	B2	B3	B4	ROUGE	B1	B2	B3	B4	ROUGE
星火3.5	75.94	75.32	74.53	71.41	78.33	32.98	19.61	7.70	4.18	33.56
星火Pro	82.88	81.45	80.57	76.77	83.71	35.02	25.94	21.29	14.42	37.74
通义Plus	79.31	73.84	70.34	67.03	82.55	30.99	14.25	5.70	3.07	34.73
通义Turbo	87.93	82.33	78.58	74.95	87.31	40.79	15.72	5.93	3.08	37.23
文心4.0	**96.30**	**94.73**	**93.59**	**91.22**	**96.67**	**66.67**	**46.06**	**31.85**	**18.79**	**63.20**
文心3.5	83.64	74.67	69.67	63.35	83.22	56.56	30.47	20.62	13.40	47.67

Table 4: 不同大型语言模型（LLM）在训练集和测试集上的自动评测结果。

组别	准确性	自然性	可读性	文化适应性
专家组	2.1	2.25	2.21	2.07
采集组	2.14	2.07	1.96	1.11
普通组	2.93	2.89	2.89	2.57

Table 5: 人类评审对手语数字人翻译质量的评分结果，分别从准确性、自然性、可读性和文化适应性四个维度进行评估。

人类评测结果 表 5展示了手语数字人翻译质量的人工评测结果，采用5分制，包括四个主要指标：手语语法准确性、自然性、可读性以及文化适应性，各项指标的得分反映了数字人在手语表达上的综合表现。评测重点关注手语数字人在准确表达手语的能力，强调自然性和可读性，同时兼顾文化适应性。从结果可以看出，普通组在所有指标上得分最高，表明该系统在实际应用环境中的表现较好。专家组和采集组的评分略低，可能是因为他们对手语有更高的专业要求。这些结果表明本文的系统在实际应用中具有较高的用户接受度，但仍有提升空间，对于理解手语翻译系统在实际应用中的用户接受度和改进方向具有重要意义。

4.2 结果分析

训练集推理的完美表现 所有模型在训练集上的推理表现都非常优秀，这说明它们在训练过程中成功地记住了训练数据的语境和模式。具体来说，训练集中的每个句子在不同模型生成的推理结果中都能高度一致，几乎没有误差。这表明模型在面对已知数据时，能够很好地应用其学习到的知识进行准确的推理。

测试集推理结果的差异 在测试集上，不同模型推理结果出现了显著的差异。具体可分为以下3类：

- Gloss不一致：例如，在句子“您的外卖到了，请尽快领取，谢谢。”的推理中，不同模型使用的Gloss略有不同。文心4.0的结果是“外卖 你 到 快 取 谢谢”，而通义Plus的结果是“您的外卖 到 请 快领 谢谢”。尽管从中文角度看，这些Gloss在语义上相近，但在3D数字人手语生成时，这种不一致可能导致手语翻译的不准确。

- Gloss顺序不一致：在句子“请问您需要将行李托运至何处？”的推理中，文心4.0的结果是“你 行李 托运 哪里 请问”，而通义Plus的结果是“问 你 行李 托运 哪里”。Gloss顺序的不同会影响手语翻译的准确性，因为手语的语序与口语可能不同，需要特别的注意。

- Gloss缺失或多增：例如在句子“请问您需要将行李托运至何处？”的推理中，文心4.0的结果是“你 行李 托运 哪里 请问”，而星火Pro的结果是“行李 托运至何处 你需要”。这里可以看到星火Pro的推理结果中多了一些Gloss，而文心4.0则缺失了一些Gloss。Gloss的缺失或多增会影响到句子的完整性和准确性。

评测任务的综合得分分析 在表 6中展示了不同参赛队伍在专家组、采集组和普通组的得分情况。表格中列出了四个组别的评分结果：A组，B组，D组以及本研究提出的方法（标记为“Ours”）。每组得分反映了各队伍的数字人在表达手语语义时的有效性。在总共14支参赛队伍中，包括本研究在内的4支队伍能够有效表达手语语义。根据专业评审的打分，本研究的方法在所有参赛者中排名第二，显示出较高的翻译准确性和用户接受度，尤其在专家组和普通组中得分较高。这表明本方法在处理实际应用中的手语翻译任务时具有较好的通用性和准确性。然而，与最佳表现的模型相比，仍存在一定差距，特别是在采集组中的表现需要进一步提升。

可能的改进方向 为了进一步提升LLM在gloss生成上的一致性和准确性，以及3D手语数字人的表现力，以下是一些潜在的改进方向：

- 同步参考标准化手语词库：模型在推理时需同步参考一个权威的手语词库，确保生成的Gloss与标准化手语一致。这有助于避免因语义相近而选择错误的Gloss，从而提高翻译的准确性。

- 优化Gloss顺序：通过扩充上下文学习样本库，并增加不同语境和复杂句子的训练样本，使模型能够学习到正确的Gloss顺序。此外，制定基于手语专家的建议和手语语法知识的翻译规则，以指导模型生成符合手语习惯的Gloss序列。

- 减少Gloss的缺失或多增问题：加强对句子结构的理解，通过强化语法校验机制，确保生成的句子结构完整且语义准确。同时，在生成过程中引入纠错机制，以及时发现并修正Gloss的缺失或多增问题(Yano and Utsumi, 2021)。

- 加强手语韵律的建模：在Gloss推理阶段，引入关于手语韵律的信息，如手势的节奏和力度(Inan et al., 2022)，以使数字人的手语表现更贴近真实的手语交流。

- 增强表情和口型的同步：通过增加表情和口型数据的训练量，并应用先进的同步算法，进一步优化表情和口型的同步技术，确保其能够准确反映手语的情感和意图。

- 语境理解的增强：利用向量库检索技术，提供与输入语句语义相近的上下文学习样本，以增强模型对语境的理解能力。这种技术可以帮助模型更好地捕捉到语句的深层含义和上下文联系，从而提高手语Gloss的生成能力。

总体来看，文心4.0在自动评测和人工评测中均表现出色，特别是在实际应用中的表现上显示出显著优势。这表明，结合LLM和3D动画技术的手语翻译系统在提高手语翻译准确性和用户体验方面具有巨大潜力。

组别	A	B	Ours	D
专家组	3.250	2.098	2.321	1.705
采集组	3.009	1.580	1.830	1.143
普通组	3.777	2.375	2.839	2.313

Table 6: 不同参赛队伍在手语翻译评测任务中的综合得分。

第二十三届中国计算语言学大会论文集，第393页-第404页，太原，中国，2024年7月25日至28日。
卷3：评测报告

5 相关工作介绍

手语数字人翻译属于手语生成任务，已有大量研究致力于开发高效、准确的翻译系统，以便更好地服务于听障人士。这些研究主要集中在手语翻译与生成两个方向，旨在通过技术手段实现自然语言到手语间的翻译。

5.1 手语翻译

在手语翻译方面，研究者们提出了多种方法来实现从自然语言到手语的翻译。早期的研究大多依赖于将自然语言转换为Gloss，然后再生成手语动作。例如，Jin(2022)、Zhu(2023)和Kan(2022)采用了这种两步法，通过Gloss作为中间层来实现翻译。这种方法的优点是能够利用大量的Gloss数据进行训练，从而提高翻译的准确性。然而，它也存在一些局限性，例如Gloss数据的获取和标注成本较高，并且Gloss本身不能完全捕捉到自然语言中的复杂语义信息。

为了克服这些局限，近年来有研究者尝试直接从自然语言生成手语，而不依赖于中间的Gloss表示。例如，Lin(2023)提出了一种基于端到端的翻译模型，通过直接将自然语言映射到手语动作序列来提高翻译的流畅性和自然度。类似地，Wong(2024)和Yin(2023)也提出了不同的无Gloss翻译方法，这些方法在处理复杂句子结构和捕捉上下文信息方面表现出色。

随着自然语言处理（NLP）领域的快速发展，手语翻译模型也在不断演进。早期的模型主要基于传统的循环神经网络（Recurrent Neural Network, RNN）和长短期记忆网络（Long Short-Term Memory, LSTM），如Guo(2018)提出的层次化LSTM模型，用于捕捉手语中的时间依赖关系。随着注意力机制和Transformer模型的引入，翻译效果得到了显著提升。例如，Cihan Camgoz(2020)提出的基于Transformer 的手语翻译模型，通过全局注意力机制能够更好地捕捉长距离依赖关系，提高了翻译的准确性和流畅性。

近年来，预训练语言模型（如BERT 和GPT）的成功应用，进一步推动了手语翻译技术的发展。Zhao(2023)利用BERT进行手语翻译，通过预训练模型的强大语义理解能力，显著提升了翻译性能。同时，大语言模型（LLMs）如GPT也开始应用于手语翻译领域。Gong(2024)和Wong(2024)探讨了将LLM应用于手语翻译的可能性，这些模型能够处理更加复杂和多样化的语言输入，使翻译结果更加自然和连贯。

手语翻译技术在不断发展，从早期的基于Gloss的两步法到如今的端到端翻译，从传统的RNN和LSTM模型到现代的Transformer和大语言模型，这些技术的进步极大地提高了手语翻译的准确性和自然性。未来的研究将继续探索如何利用更先进的模型和方法，进一步提高翻译质量，满足实际应用需求。

5.2 手语生成

近年来，手语生成和手语数字人技术逐渐受到广泛关注。手语数字人是通过3D建模和动画技术，实现手语动作逼真展示的虚拟人形形象，这一技术在辅助听障人士的交流中具有重要作用。例如，Lacerda(2023)开发了一种基于Unity 3D引擎的手语数字人系统，该系统能够实时模拟复杂的手语动作，实现了高度的准确性和自然度。

在手语生成领域，姿态估计和关键点推理技术发挥了关键作用。Zuo(2024)、Forte(2023)和Yu(2024)提出了基于姿态估计的方法，通过捕捉人体关键点，并将这些关键点渲染成数字人，达到了逼真的手语展示效果。尤其是Forte(2023)，他们的研究展示了如何通过优化关键点推理算法，提高手语动作的流畅性和自然度，使手语数字人在实际应用中更加实用和可靠。

此外，扩散模型在手语生成中的应用也取得了显著进展。Baltatzis(2024)提出了基于扩散模型的手语生成方法，通过端到端的训练实现了从自然语言到手语的直接翻译。该方法有效地结合了文本信息和视觉信息，使生成的手语更加连贯和自然，克服了传统方法中由于信息割裂导致的动作生硬问题。

另一类重要的方法是基于生成对抗网络（Generative Adversarial Networks, GAN）的手语生成系统。Saunders(2020a)提出了一种利用GAN进行手语生成的系统，通过对抗训练，使生成的手语视频更加逼真和自然。该系统不仅提高了手语动作的视觉效果，还增强了模型对不同手语表达的适应性和鲁棒性。

手语数字人技术的进展为手语翻译系统的发展提供了新的可能性。例如，Lakhfif(2020)研究了手语数字人在教育服务中的应用，证明了其在增强用户体验和提高交流效率方面的潜力。通过手语数字人，听障人士可以更直观地理解教育内容，极大地改善了学习效果。

本研究在现有工作的基础上，提出了一种创新性的手语翻译技术，融合了LLM的文本处理能力和3D动画技术的表现力。与以往工作相比，本系统不仅在翻译准确性上有所提升，更在手语动画的流畅性和表现力上实现了质的飞跃。通过深度语义理解，本系统能够深入分析复杂的语言结构和上下文信息，生成更准确和自然的手语翻译。此外，本系统设计的多样化上下文学习样本，确保了翻译结果在语义和语境上的双重精准，满足了多样化的交流场景需求。此外，本研究还特别关注了手语的视觉呈现，通过3D动画技术创造出既流畅又逼真的手语动作，这些动作在视觉上吸引人，并在表达上贴近自然手语的非言语特征，如节奏、力度和表情。系统的实时交互与反馈能力，以及多模态融合技术的应用，使得手语翻译更加生动和直观，极大地提升了用户体验。

未来的研究将集中在优化手语数字人的动作自然度和系统的响应速度，以满足实际应用的需求。这包括更精细的3D建模技术、更高效的动画生成算法，以及更智能的自然语言处理模型。随着技术的不断进步，手语数字人将在更多领域中发挥重要作用，如教育、医疗和公共服务，为听障人士提供更好的支持和服务。

6 结论

本研究通过使用大规模语言模型和3D动画技术，开发了一个高效的手语数字人翻译系统。实验结果表明，系统在准确性和实用性方面表现出色，特别是文心4.0在自动评测和人工评测中均表现出显著优势，显示了其在手语翻译中的高可靠性和准确性。本研究开发的手语翻译系统在提高手语翻译准确性和用户体验方面具有巨大潜力，为听障者提供了更加准确和便捷的沟通工具。未来的研究和优化工作将进一步推动手语翻译技术的发展和应用。

致谢

本研究受国家自然科学基金面上项目(No. 62376019, 61976015, 61976016, 61876198, 61370130)资助。作者们还对匿名评审专家给予的宝贵建议表示衷心的感谢。

参考文献

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context Examples Selection for Machine Translation, December. arXiv:2212.02437 [cs].

Rotem Shalev Arkushin, Amit Moryossef, and Ohad Fried. 2023. Ham2Pose: Animating Sign Language Notation into Pose Sequences. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21046–21056, Vancouver, BC, Canada, June. IEEE.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report, September. arXiv:2309.16609 [cs].

Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. 2024. Neural Sign Actors: A diffusion model for 3D sign language production from text, April. arXiv:2312.02702 [cs].

Jiaxin Cheng, Soumyaroop Nandi, Prem Natarajan, and Wael Abd-Almageed. 2021. SIGN: Spatial-information Incorporated Generative Network for Generalized Zero-shot Semantic Segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9536–9546, Montreal, QC, Canada, October. IEEE.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign Language Transformers: Joint End-to-End Sign Language Recognition and Translation. In *2020 IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 10020–10030, Seattle, WA, USA, June. IEEE.

Maria-Paola Forte, Peter Kulits, Chun-Hao Huang, Vasileios Choutas, Dimitrios Tzionas, Katherine J. Kuchenbecker, and Michael J. Black. 2023. Reconstructing Signing Avatars From Video Using Linguistic Priors, April. arXiv:2304.10482 [cs].

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. LLMs are Good Sign Language Translators, April. arXiv:2404.00925 [cs].

Dan Guo, Wengang Zhou, Houqiang Li, and Meng Wang. 2018. Hierarchical LSTM for Sign Language Translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April.

Mert Inan, Yang Zhong, Sabit Hassan, Lorna Quandt, and Malihe Alikhani. 2022. Modeling Intensification for Sign Language Generation: A Computational Approach. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2897–2911, Dublin, Ireland. Association for Computational Linguistics.

Tao Jin, Zhou Zhao, Meng Zhang, and Xingshan Zeng. 2022. Prior Knowledge and Memory Enriched Transformer for Sign Language Translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3766–3775, Dublin, Ireland. Association for Computational Linguistics.

Navroz Kaur Kahlon and Williamjeet Singh. 2023. Machine translation from text to sign language: a systematic review. *Universal Access in the Information Society*, 22(1):1–35, March.

Jichao Kan, Kun Hu, Markus Hagenbuchner, Ah Chung Tsoi, Mohammed Bennamoun, and Zhiyong Wang. 2022. Sign Language Translation with Hierarchical Spatio-Temporal Graph Neural Network. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2131–2140, Waikoloa, HI, USA, January. IEEE.

Jung-Ho Kim, Eui Jun Hwang, Sukmin Cho, Du Hui Lee, and Jong Park. 2022. Sign Language Production With Avatar Layering: A Critical Use Case over Rare Words. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1519–1528, Marseille, France, June. European Language Resources Association.

Inês Lacerda, Hugo Nicolau, and Luisa Coheur. 2023. Enhancing Portuguese Sign Language Animation with Dynamic Timing and Mouthing, July. arXiv:2307.06124 [cs].

Abdelaziz Lakhfif. 2020. Design and Implementation of a Virtual 3D Educational Environment to improve Deaf Education, May. arXiv:2006.00114 [cs].

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. Gloss-Free End-to-End Sign Language Translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12904–12916, Toronto, Canada. Association for Computational Linguistics.

Amit Moryossef. 2023. sign.mt: Real-Time Multilingual Sign Language Translation Application, October. arXiv:2310.05064 [cs].

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.

Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. 2023. EmoTalk: Speech-Driven Emotional Disentanglement for 3D Face Animation, August. arXiv:2303.11089 [cs, eess].

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020a. Everybody Sign Now: Translating Spoken Language to Photo Realistic Sign Language Video, November. arXiv:2011.09846 [cs].

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020b. Progressive Transformers for End-to-End Sign Language Production. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, volume 12356, pages 687–705. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021. Mixed SIGNals: Sign Language Production via a Mixture of Motion Primitives. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1899–1909, Montreal, QC, Canada, October. IEEE.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation, July. arXiv:2107.02137 [cs].

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for Translation: Assessing Strategies and Performance, June. arXiv:2211.09102 [cs].

Liming Wang, Junrui Ni, Heting Gao, Jialu Li, Kai Chieh Chang, Xulin Fan, Junkai Wu, Mark Hasegawa-Johnson, and Chang Yoo. 2023. Listen, Decipher and Sign: Toward Unsupervised Speech-to-Sign Language Recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6785–6800, Toronto, Canada. Association for Computational Linguistics.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2GPT: Leveraging Large Language Models for Gloss-Free Sign Language Translation, May. arXiv:2405.04164 [cs].

Pan Xie, Taiying Peng, Yao Du, and Qipeng Zhang. 2024. Sign Language Production with Latent Motion Transformer. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3012–3022, Waikoloa, HI, USA, January. IEEE.

Ken Yano and Akira Utsumi. 2021. Pipeline Signed Japanese Translation Focusing on a Post-positional Particle Complement and Conjugation in a Low-resource Setting. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2021–2032, Online. Association for Computational Linguistics.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including Signed Languages in Natural Language Processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, and Zhou Zhao. 2023. Gloss Attention for Gloss-free Sign Language Translation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2551–2562, Vancouver, BC, Canada, June. IEEE.

Zhengdi Yu, Shaoli Huang, Yongkang Cheng, and Tolga Birdal. 2024. SignAvatars: A Large-scale 3D Sign Language Holistic Motion Dataset and Benchmark, April. arXiv:2310.20436 [cs].

Jan Zelinka and Jakub Kanis. 2020. Neural Sign Language Synthesis: Words Are Our Glosses. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3384–3392, Snowmass Village, CO, USA, March. IEEE.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting Large Language Model for Machine Translation: A Case Study, January. arXiv:2301.07069 [cs].

Weichao Zhao, Hezhen Hu, Wengang Zhou, Jiaxin Shi, and Houqiang Li. 2023. BEST: BERT Pre-training for Sign Language Recognition with Coupling Tokenization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3):3597–3605, June.

Dele Zhu, Vera Czehmann, and Eleftherios Avramidis. 2023. Neural Machine Translation Methods for Translating Text to Sign Language Glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12523–12541, Toronto, Canada. Association for Computational Linguistics.

Ronglai Zuo, Fangyun Wei, Zenggui Chen, Brian Mak, Jiaolong Yang, and Xin Tong. 2024. A Simple Baseline for Spoken Language to Sign Language Translation with 3D Avatars, January. arXiv:2401.04730 [cs].