

学校代码: 10246  
学 号:

# 復旦大學

## 硕士 学位 论文

(专业学位)

### 结合潜空间解耦与运动学约束的手语关键点生成模型

**A Sign Language Keypoint Generation Model with Latent Space Decoupling and Kinematic Constraints**

院 系: 计算机学院

专业学位类别(领域): 软件工程

姓 名:

指 导 教 师:

完 成 日 期: 2026 年 2 月 20 日



# 目 录

插图目录	iii
摘要	v
Abstract	vii
<b>第 1 章 绪论</b>	<b>1</b>
1.1 研究背景与意义 .....	1
1.1.1 研究背景 .....	1
1.1.2 研究意义 .....	2
1.1.3 研究目标与任务定义 .....	3
1.1.4 主要研究内容 .....	5
<b>第 2 章 国内外研究现状</b>	<b>7</b>
2.1 手语生成技术的发展脉络 .....	7
2.2 主流研究路线与代表性成果 .....	8
2.3 国内研究现状与资源建设 .....	9
<b>第 3 章 理论与方法设计</b>	<b>11</b>
3.1 模型总体思路与体系结构 .....	11
3.2 层次化潜空间解耦模型 .....	12
3.2.1 结构与表达 .....	12
3.2.2 训练目标 .....	13
3.3 节奏相位与运动学统一约束机制 .....	19
3.3.1 RTP: 从语言韵律到动作节奏 .....	19
3.3.2 KALS: 生成的物理可行性 .....	21
3.3.3 统一优化目标与多阶段训练策略 .....	22
3.4 模型训练实施与收敛分析 .....	23
3.4.1 训练环境与混合精度配置 .....	23
3.4.2 正则化与稳定性控制 .....	24
3.4.3 收敛行为分析 .....	24

<b>第 4 章 系统实现与实验结果</b>	<b>27</b>
4.1 数据集与预处理 . . . . .	27
4.1.1 数据集划分与文本清洗 . . . . .	27
4.1.2 骨架提取与标准化流形 . . . . .	27
4.1.3 时序对齐与语义切片 . . . . .	28
4.1.4 物理约束下的数据增强 . . . . .	28
4.2 模型实现细节 . . . . .	29
4.2.1 语义编码与输入表征 . . . . .	29
4.2.2 HLC 结构参数与潜空间配置 . . . . .	29
4.2.3 KALS 物理约束实现 . . . . .	31
4.2.4 训练环境与配置 . . . . .	31
4.3 实验设置与评价体系 . . . . .	32
4.3.1 基线模型与消融变体 . . . . .	32
4.3.2 评测指标体系 . . . . .	32
4.3.3 统计检验 . . . . .	33
4.4 主实验结果分析 . . . . .	33
4.4.1 现有方法对比评估 (Comparison with State-of-the-Arts) . .	33
4.4.2 视觉生成质量定性分析 . . . . .	34
4.4.3 人工主观评测结果 . . . . .	35
4.4.4 核心模块消融分析 (Ablation Study) . . . . .	35
<b>第 5 章 总结与展望</b>	<b>37</b>
5.1 研究总结与核心发现 . . . . .	37
5.2 局限性与未来展望 . . . . .	38
<b>参考文献</b>	<b>39</b>
<b>致谢</b>	<b>45</b>

# 插图目录

1-1 信息无障碍与手语交互系统整体流程图 . . . . .	4
1-2 任务定义与数据流转架构图 . . . . .	5
2-1 研究发展历程图：规则拼接 → 神经生成 → 潜空间/扩散 → 实时融合	8
2-2 研究定位与创新框架图：中心为文本 → 动作生成，外圈为 HLC 与 RTP-KALS 模块 . . . . .	10
3-1 模型总体架构示意图 . . . . .	12
3-2 双分支、多级码本、融合解码 . . . . .	17
3-3 KALS 可视化 . . . . .	22
3-4 RTP-KALS 统一机制示意图 . . . . .	23
3-5 三阶段训练流程与收敛曲线 . . . . .	25
4-1 数据预处理与标准化骨架构建流程示意图 . . . . .	29
4-2 生成手语序列可视化示例：(上) 注意：男孩施工中 (Caution: Boys at Work)；(下) 今天你好吗 (How are you)。 . . . . .	36



# 摘要

在人工智能辅助信息无障碍技术不断发展的背景下，自动化手语生成（Sign Language Production, SLP）作为连接听障群体与有声世界的重要技术方向之一，近年来受到越来越多关注。尽管现有研究取得了一定进展，但在实际应用场景中仍面临若干挑战：端到端神经生成方法通常缺乏对人体运动学与语言韵律的显式建模，容易出现“骨骼穿插”“动作节奏僵硬”等物理失真问题；而扩散模型虽然在生成质量上有所提升，但其较高的推理成本限制了实时交互式系统的部署。基于上述背景，本文尝试探索一种兼顾生成质量与推理效率的端到端手语生成方法，并在此方向上做出初步探索。

本文提出了一个结合层次化潜空间解耦（HLC）与节奏—运动学统一约束（RTP-KALS）的生成框架，旨在缓解动作特征纠缠、节奏不自然以及非物理伪影等常见问题。（1）在表征学习方面，本文构建了层次化潜空间解耦模型（HLC），在传统 VAE 的基础上引入“形态—运动”双分支量化结构，通过时序差分输入与多级码本建模，使模型能够在一定程度上区分身体拓扑结构与动作变化模式，为后续生成提供更清晰的动作基元表征。（2）在约束机制方面，本文设计了节奏相位调制（RTP）与运动学一致性损失集合（KALS），分别从时间节奏与空间物理性两个角度对生成序列施加软约束，以减少节奏匀速化、关节异常及双手协同不足等现象。上述模块在“三阶段交替更新”的训练策略下协同作用，使生成动作在可理解性与物理合理性方面得到一定改善。（3）最后，本文实现了一个可实时运行的轻量化生成系统，并在 CSL-Daily 中文手语数据集上对方法进行了系统评估。实验结果表明，本文方法在 BLEU-4、KVR 等指标上与若干现有基线相比具有一定优势，同时在单卡环境下实现了较快的推理速度，展示了其潜在的工程应用价值。

总体而言，本文在手语生成的表征建模与物理约束方面进行了初步尝试，希望能为构建更加高效、自然与可解释的手语生成系统提供一定参考。

**关键词：**手语生成；潜空间表征；节奏建模；运动学约束；实时虚拟手语



# Abstract

Sign Language Production (SLP) has become an important research direction for intelligent accessibility systems. Existing end-to-end generation methods often lack explicit modeling of human kinematics and linguistic rhythm, which may lead to physical artifacts such as joint penetration and rigid motion pacing. Although diffusion-based methods can improve visual quality, their iterative denoising process usually introduces high inference latency and limits deployment in real-time interactive scenarios. To balance generation quality and efficiency, this thesis explores an end-to-end sign language generation framework for practical real-time use.

This thesis proposes a unified framework that combines Hierarchical Latent Codebook decoupling (HLC) with Rhythm-Tempered Phase and Kinematic-Aware Loss Set (RTP-KALS). HLC introduces a morphology-motion dual-branch quantization design with temporal-difference input and multi-level codebooks, improving the disentanglement between body topology and dynamic motion patterns. RTP and KALS then impose soft constraints from temporal rhythm and spatial physical plausibility, respectively, reducing over-uniform motion speed, abnormal joint behavior, and weak bimanual coordination. Under a three-stage alternating training strategy, the proposed method improves both semantic intelligibility and physical realism. Experiments on the CSL-Daily dataset show competitive performance on BLEU-4 and KVR, while maintaining fast inference speed in a single-GPU setting, indicating promising engineering value for real-time sign language generation.

**Keywords:** Sign language production; latent representation; rhythm-aware generation; kinematic constraints; real-time avatar



# 第 1 章 绪论

## 1.1 研究背景与意义

### 1.1.1 研究背景

在信息无障碍与智能人机交互快速发展的背景下，手语作为聋哑人与听障群体的重要交流方式，其自动生成与理解技术逐渐成为人工智能研究中的关键方向。全球听障人群数量已超过 4 亿，其中约有 7,000 万人使用手语进行交流。自动化手语生成系统（Sign Language Production, SLP）作为连接听障群体与有声世界的桥梁，在新闻播报、政务服务、医疗咨询及教育教学等典型场景中具有巨大的应用潜力。

自 Text2Sign (Stoll et al., 2018)<sup>[1]</sup>以来，手语生成研究从“基于规则的拼接式动画”逐步演进到“端到端神经生成”，再到近年来，多种基于 Transformer、扩散模型与离散潜空间的手语生成方法被相继提出<sup>[2–6]</sup>。尽管当前研究在生成动作的多样性与整体可懂度方面取得进展，但在高实时性、高物理自然度与符合中文手语语言学规律的落地需求面前仍存在显著不足，体现在以下三个方面：

从近两年的研究趋势看，手语生成正在向“更强时空建模 + 更高维度表达 + 更标准化评测”三条路线并行推进：在生成范式上，Neural Sign Actors 将扩散模型拓展到 3D 签署人表示<sup>[7]</sup>；T2S-GPT 通过动态向量量化与自回归解码提升码本利用效率<sup>[8]</sup>；MS2SL 探索了文本/语音驱动的连续手语统一框架<sup>[9]</sup>。在低资源与弱监督方向，Select-and-Reorder 通过“选择—重排”分解提升了数据稀缺场景下的可用性<sup>[10]</sup>，USLNet 进一步验证了无平行语料条件下的翻译与生成可行性<sup>[11]</sup>；面向词级 3D 运动生成，wSignGen 展示了条件扩散在细粒度手语单元上的潜力<sup>[12]</sup>。

与此同时，评测与工程化也在快速成熟。SLRTP2025 挑战赛推动了 Text-to-Pose 任务的统一基准与指标口径<sup>[13]</sup>，多支队伍围绕可控生成、解耦表征与流式推理提出了不同实现路径，包括文本驱动扩散<sup>[14]</sup>、构音器解耦正则<sup>[15]</sup>、多模态对齐<sup>[16]</sup>、混合自回归-扩散<sup>[17]</sup>以及潜在动力学建模<sup>[18]</sup>。此外，针对非手动特征与骨架噪声问题，已有工作开始引入面部/姿态协同建模与姿态编码稳健化策略<sup>[19–21]</sup>。这些进展说明，SLP 正从“可生成”阶段走向“可部署、可比较、可解释”的系统化阶段<sup>[22–23]</sup>。

从技术底座看，当前 SLP 的核心模块与更广泛的生成式 AI 路线密切相关：序列建模大量借鉴 Transformer<sup>[24]</sup>，对抗训练与判别器思想可追溯至 GAN<sup>[25]</sup>，高保真生成路径受到扩散模型框架影响<sup>[26]</sup>，而离散潜空间学习与码本机制则延续自 VQ-VAE 的经典设计<sup>[27]</sup>。在条件编码侧，BERT/RoBERTa 等预训练语言模型已成为文本语义表征的常用基础<sup>[28-29]</sup>；在表示压缩与高效编码方面，神经编解码思路也为离散化动作表示提供了可借鉴的工程经验<sup>[30]</sup>。

首先，现有生成范式的推理延迟难以满足实时交互需求。在政务大厅咨询或急诊医疗沟通等场景中，系统需具备“即说即译”的能力，通常要求端到端延迟控制在 100ms 以内以保证对话流的连贯性。然而，当前主流的扩散模型（Diffusion Models）<sup>[26]</sup>虽然提升了生成质量，但其去噪过程往往需要数十步迭代，推理延迟通常超过 300ms，这种高延迟导致了严重的“交互停顿”，使得系统在实时服务场景中难以落地。

其次，中文手语（CSL）的语言学特性未被充分建模。相比于英语手语（ASL），中文手语拥有独特的语法结构（如倒装句、话题优先结构）以及丰富的非手动特征（Non-manual Features，如表情、口型与头部姿态）。现有研究多基于 ASL 数据集构建，直接迁移至中文语境时，往往因忽视 CSL 的语言学特性而导致“词能达意但句法怪异”，甚至因缺乏表情配合而产生语义歧义。

最后，生成动作的物理合理性与节奏自然度仍有待提升。现有的端到端模型常生成违反人体运动学的动作（如骨长波动、关节穿插），产生“恐怖谷”效应；同时，缺乏对自然语言韵律（Prosody）到动作节奏（Rhythm）的显式映射，使得生成的手语往往呈现出机械的匀速运动，缺乏真人手语的轻重缓急与呼吸感。

综上所述，如何在保障语义一致性与符合中文手语规范的前提下，兼顾生成的运动学自然度与实时交互性，解决从“能看”到“好用”的跨越，是当前手语生成研究亟待解决的核心问题。

### 1.1.2 研究意义

**理论意义** 本研究聚焦于“离散文本-连续动作”的跨模态映射难题，从表征学习、时序建模与物理约束三个维度，对现有的多模态生成理论进行了深化与拓展：

#### 1. 构建了形态与运动正交解耦的潜空间表征范式：

针对传统 VAE 存在的特征纠缠与模态混淆问题，本研究提出的层次化潜空间解耦（HLC）机制，在理论上证明了将高维动作空间分解为“静态形态流形”与“动态运动流形”的可行性。这为解决跨模态生成中的“特征解缠结（Disentangled Representation）”问题提供了新的结构化视角，使动作生成从“黑盒拟合”迈向“可解释的组件化合成”。

## 2. 建立了从“语言韵律”到“运动相位”的显式映射机理:

针对现有序列生成模型缺乏节奏控制的理论缺陷，本研究提出的节奏相位调制（RTP）机制，首次将信号处理中的“瞬时相位”概念引入手语生成，建立了语言学韵律特征（Prosody）与运动学速度特征（Kinematics）之间的数学映射桥梁，补充了动作生成领域在精细化时序控制方面的理论短板。

## 3. 提出了数据驱动与物理先验融合的约束框架:

针对纯神经网络生成易违反人体运动规律的问题，本研究设计的运动学一致性约束（KALS），探索了将不可微的刚体物理约束转化为可微的目标函数的方法。这为“神经—符号”混合系统（Neuro-symbolic AI）在动作生成领域的应用提供了实证依据，即如何在保持深度学习拟合能力的同时，有效注入物理世界的先验规则。

**实际意义** 在应用层面，本研究成果直接服务于国家信息无障碍战略，旨在解决听障群体在数字化社会中的“信息孤岛”问题：

### 1. 突破实时交互的技术瓶颈:

本研究构建的轻量化生成框架，在保证动作质量的前提下显著降低了推理延迟（<100ms），直接支持了政务大厅智能终端、医院急诊辅助系统等高实时性场景的落地应用，解决了现有扩散模型难以满足即时交互需求的痛点。

### 2. 提升手语服务的可懂度与接受度:

通过改善动作的自然度与节奏感，本研究生成的手语虚拟人具备更高的拟人化水平，有助于缓解听障用户面对机器时的疏离感，可广泛应用于手语新闻播报、特殊教育辅助教学以及交互式虚拟数字人等领域，具有显著的社会效益与应用价值。

### 1.1.3 研究目标与任务定义

**研究目标** 本文旨在构建一个结合层次化潜空间解耦（HLC）与节奏—物理统一约束（RTP-KALS）的手语关键点生成模型。在确保语义一致性（动作准确传达文本含义）的基础上，重点突破当前生成模型在节奏自然性与运动学合理性方面的技术瓶颈，并实现满足流式交互需求的实时生成。具体量化目标如下：

#### 1. 动作质量目标:

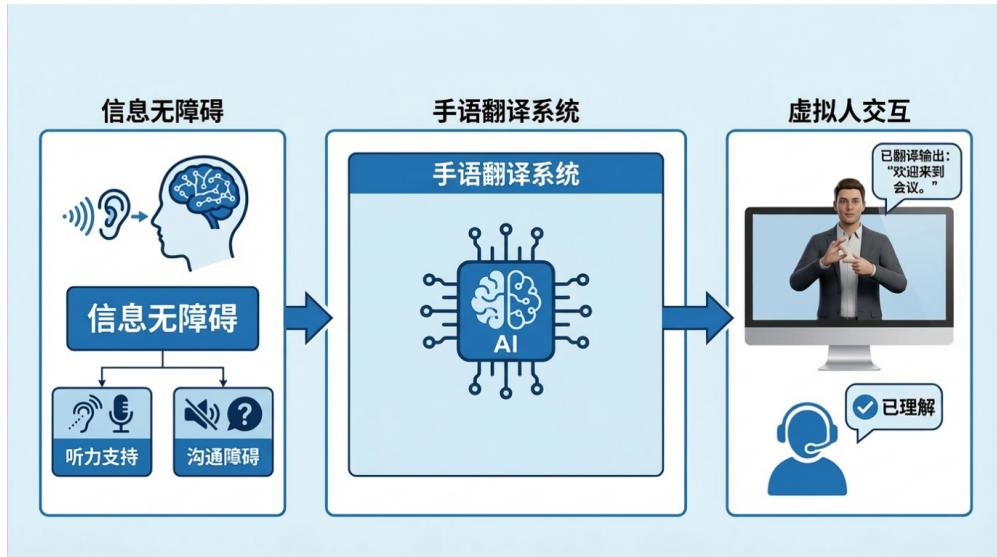


图 1-1 信息无障碍与手语交互系统整体流程图

生成的骨架序列应消除骨长波动与关节穿插现象，且具备与自然语言韵律对齐的相位节奏；

## 2. 实时性能目标：

在主流单卡推理环境下，实现生成吞吐率  $\geq 25 \text{ FPS}$ ，以及端到端推理延迟  $< 80 \text{ ms}$ ，以支撑“即说即译”的在线手语服务场景。

**任务定义** 本研究将手语生成任务定义为从自然语言序列到高维骨架坐标序列的概率映射问题。形式化定义如下：

**输入 (Input):** 给定自然语言文本序列  $W = \{\omega_1, \omega_2, \omega_3, \dots, \omega_N\}$ ，其中  $N$  为词元数量。在部分训练场景下，可辅以手语词汇注释 (Gloss) 序列作为中间语义提示。

**输出 (Output):** 生成手语骨架关键点序列  $S \in \mathbb{R}^{T \times J \times C}$ ，其中  $T$  为时间步 (帧数)， $J$  为关节点数量 (全身 50 点)， $C$  为空间坐标维度。

鉴于直接端到端映射的难度，本文将该任务分解为两个耦合的子过程：

### 1. 跨模态潜空间映射 (Cross-Modal Latent Mapping)：

旨在建立文本语义与离散动作基元之间的对齐关系，将离散的语言符号  $\omega_i$  映射为解耦的潜空间编码  $Z_s$  (形态) 与  $Z_m$  (运动)。

### 2. 物理约束下的序列解码 (Constrained Sequence Decoding)：

在节奏相位与运动学规则 (RTP-KALS) 的显式约束下, 将潜变量  $Z$  还原为时序连贯、物理合规的连续骨架轨迹  $S$ 。

**核心目标:** 本任务的核心目标是实现多维度的联合优化: 在确保语义一致性 (即生成动作准确反映文本意图) 的前提下, 通过显式约束最大化运动学合理性 (如骨长恒定、关节角度合法) 与节奏自然性 (韵律与时序连贯); 同时在工程实现上严格控制推理延迟, 确保模型在实际部署中满足低延迟、高吞吐的实时交互需求。

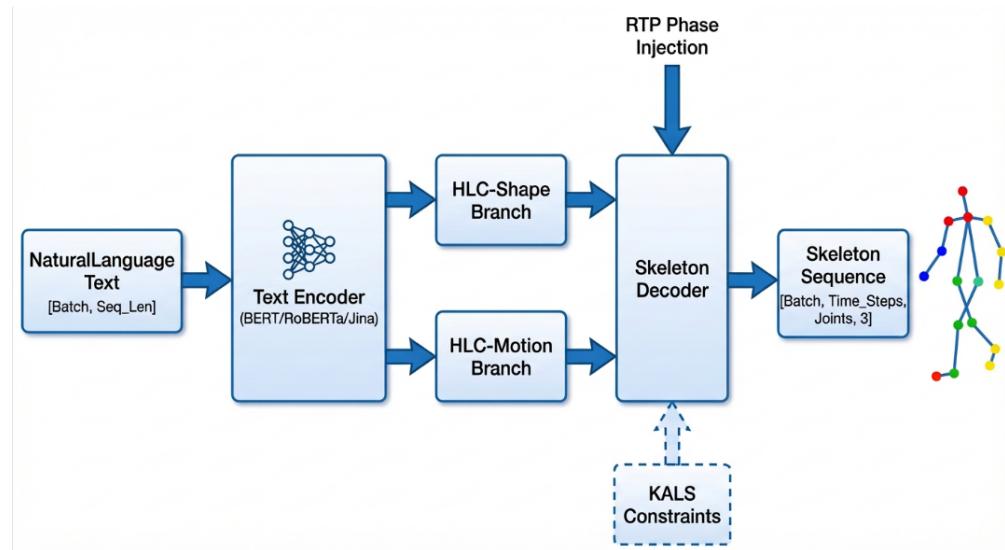


图 1-2 任务定义与数据流转架构图

### 1.1.4 主要研究内容

为突破“高逼真度”与“高实时性”难以兼得的瓶颈, 本文从“底层表征解耦”与“上层时空约束”两个互补层面开展研究。前者负责构建清晰的动作基元 (解决“生成什么”), 后者负责精准的动态控制 (解决“如何运动”), 具体内容如下:

#### 1. 层次化潜空间解耦建模 (Hierarchical Latent Codebook, HLC)。

针对传统 VAE 模型中存在的特征纠缠与模态混淆问题 (即无法独立控制手型与速度), 本文基于 VQ-VAE 框架构建了双分支潜空间: 形态分支专门编码静态姿态结构, 运动分支编码时间动态特征。为捕捉不同粒度的动作语义, 设计了上下两级码本结构: 上层码本 (大小为  $N_{top}$ ) 负责表征全局身体拓扑, 下层码本 (大小为  $N_{bottom}$ ) 专注于表征局部手部细节与快变纹理。通过联合优化重建损失与正交解耦损失, 模型形成了具有可解释性

的潜在动作基元。消融实验表明，这种分层解耦机制显著提升了复杂手语动作的语义清晰度。

## 2. 节奏相位与运动学统一约束机制 (RTP-KALS)。

针对端到端生成中常见的动作节奏机械化与物理失真问题，本文设计了双重约束机制：引入韵律相位调制 (RTP)，建立从“语言韵律”到“动作瞬时速度”的显式映射，解决节奏单一问题；同时构建运动学一致性损失集合 (KALS)，在骨骼长度恒定性、关节角度合法性与双手协同性三个维度施加软约束，解决“骨骼穿插”等非自然现象。

**协同机制：** 上述两大模块在统一的生成框架下协同工作：HLC 模块为生成过程提供了结构化、可解释的离散动作基元 (Primitives)，而 RTP-KALS 机制则充当动态控制器，指导这些基元在时间轴上以自然的节奏排列，并将其限制在符合人体生理结构的物理空间内。两者结合，实现了从离散语义符号到连续、自然、物理可信骨架序列的高质量映射。

# 第 2 章 国内外研究现状

## 2.1 手语生成技术的发展脉络

手语生成 (Sign Language Production, SLP) 旨在将自然语言映射为符合手语语法与动作规范的可理解动作序列。该领域的技术演进与数据资源的规模扩展紧密耦合，主要经历了由“规则驱动”向“神经生成”，再向“潜空间/扩散模型”跨越的三个阶段<sup>[23]</sup>：

### 1. 第一阶段：规则驱动与拼接合成（2015 年以前）。

早期研究主要依托 HamNoSys 等人工定义的符号系统与小规模手语词典。技术上采用将文本映射为关键动作片段并进行插值拼接的方法（如 Paula et al.）。但受限于词典的覆盖范围与手工规则的僵硬性，生成的动作往往机械且难以表达非手动特征（表情、口型），且无法处理未登录词。

### 2. 第二阶段：端到端神经生成（2016-2020 年）。

随着 RWTH-PHOENIX-Weather 2014T 等首批连续手语数据集的发布<sup>[31]</sup>，数据驱动的深度学习成为可能。Text2Sign (Stoll et al., 2018)<sup>[1]</sup>首次引入神经机器翻译 (NMT) 架构实现端到端映射；Progressive Transformer (Saunders et al., 2020)<sup>[3]</sup>利用层级 Transformer 改善了序列连贯性。然而，受限于当时的数据量级与回归模型的平均化倾向，这一阶段的模型常面临节奏不稳定与“回归均值”导致的动作模糊问题。

### 3. 第三阶段：潜空间建模与扩散生成（近 3-5 年 / 2021-2025 年）。

随着 RWTH-PHOENIX-Weather 2014T 等首批连续手语数据集以及 How2Sign 和 Word-Level SLR 等大规模数据集的发布<sup>[31-33]</sup>，数据驱动的深度学习范式成为可能。Text2Sign (Stoll et al., 2018)<sup>[1]</sup>首次引入神经机器翻译架构实现端到端映射；Progressive Transformer (Saunders et al., 2020)<sup>[3]</sup>利用层级 Transformer 改善了序列连贯性，后续工作在多通道建模与对抗训练方面也取得进展<sup>[2,34]</sup>。在此基础上，进一步出现了面向高质量生成的 FastSLP、VQ-VAE+GPT 与 DiffSLP 等潜空间/扩散式方法<sup>[4-6]</sup>，以及 MCST-Transformer、Articulator-based Disentangle、SignAligner、混合自回归-扩散流式生成与潜在姿态动力学建模等最新工作<sup>[15-18,35]</sup>。

总体而言，SLP的发展呈现出从“符号规则化”迈向“语义深度建模”的趋势，而数据资源的丰富度直接决定了生成模型的上限。

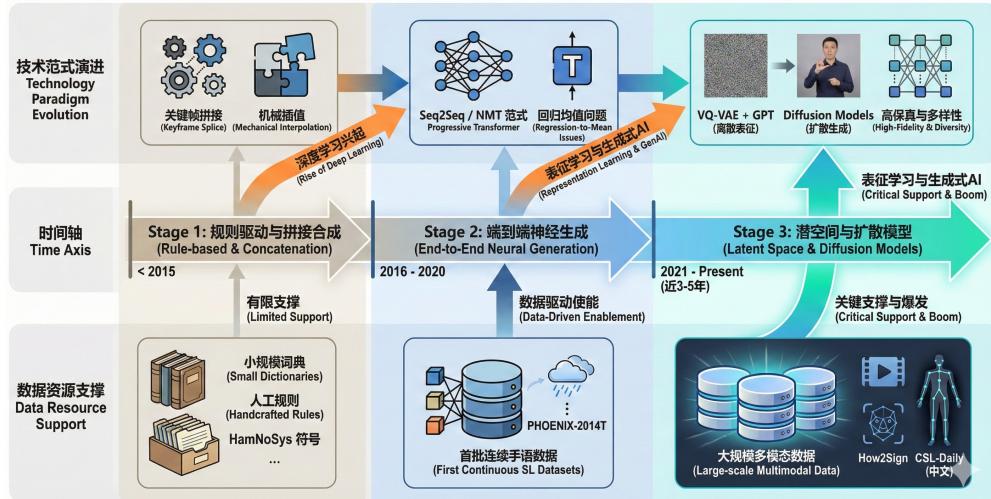


图 2-1 研究发展历程图：规则拼接 → 神经生成 → 潜空间/扩散 → 实时融合

## 2.2 主流研究路线与代表性成果

当前国际上主要存在三类范式，其在语义准确性（BLEU）、动作自然度（Jerk）与推理实时性（FPS）上呈现出显著的性能差异：

### 1. 多阶段翻译管线（Pipeline Approach）：

如 SLT (Camgoz et al., 2020)。该范式依赖 Gloss (手语词汇注释) 作为中间语义层 (Text to Gloss to Pose)。其优势在于语义层次清晰，BLEU-4 分数通常较高；但在低资源语言（尤其中文手语）中因缺乏高质量 Gloss 标注难以推广，且分步级联导致推理延迟叠加，FPS 往往受限于最慢的模块。

### 2. 端到端神经生成（End-to-End Neural Generation）：

如 Progressive Transformer (Saunders et al., 2020)。该范式消除中间层依赖，直接映射 Text → Pose。其推理效率极高（通常 FPS > 100），适合实时应用；但由于缺乏物理约束，生成动作常表现出高 Jerk（加速度）能量，即动作抖动且缺乏韵律感，导致“机器味”重。

### 3. 潜空间与扩散式方法（Latent/Diffusion Models）：

如 FastSLP、VQ-VAE+GPT 与 DiffSLP 等代表性工作<sup>[4-6]</sup>。该范式通过离散码本与扩散去噪显著提升了动作的平滑度与多样性，KVR（运动学有效率）接近 100%，并在后续工作中进一步与流式自回归建模与潜在动力学建

模结合<sup>[17-18]</sup>。然而，这是以巨大的计算开销为代价的：扩散模型的多步去噪过程导致推理速度大幅下降（往往 < 10 FPS），难以满足实时交互需求。

如表 2-1 所示，现有模型普遍面临“高语义准确度、高物理自然度与高实时性”难以三角兼顾的困境。

表 2-1 国内外代表性 SLP 模型对比表

模型名称 (Model)	年份	核心技术 (Core Tech)	BLEU-4 (↑)	Jerk (↓)	KVR (↑)	FPS (↑)	支持语言
Text2Sign [Stoll et al.]	2018	RNN/GAN + Seq2Seq	8.52	High	Low	> 100	ASL
Prog. Transformer [Saunders]	2020	Transformer + Counter	10.21	Medium	-	120+	ASL
Mixture of Motion [Saunders]	2021	Multi-Head Mixture	11.35	Medium	-	95	ASL
SignGAN [Li et al.]	2022	GAN + Keypoints	8.9	Low	Medium	60	CSL
DiffSLP [Baltatzis et al.]	2023	Probabilistic Diffusion	13.5	Very Low	High	< 10	ASL
VQ-VAE + GPT [Xie et al.]	2023	Discrete Latent + GPT	14.1	Low	High	~ 30	CSL/ASL

## 2.3 国内研究现状与资源建设

国内中文手语生成 (CSL Production) 研究虽然起步较晚，但近年来在数据构建与模型探索上已取得显著进展，呈现出“数据追赶、模型多样、评估待统”的局面。

在数据资源方面，中国科学技术大学发布的 CSL-Daily (Zhou et al., 2021)<sup>[36]</sup> 是目前的标杆，为连续手语研究提供了宝贵的日常场景语料。然而，相比于国际通用的 How2Sign (80+ 小时，多视角，含深度图)，国内数据集在规模（普遍 < 20 小时）与精细度上仍有差距，尤其是缺乏对面部表情、嘴型及手指微动作的细粒度标注，限制了生成动作的逼真感。

在模型研究方面，国内多家机构已开展了前沿探索，但仍存在特定的技术局限：

- **中科院自动化所 (CASIA)**: 提出的 CSL-Trans 等模型在“手语翻译 (Translation)”向“生成”的反向映射上表现优异，但其技术路线普遍强依赖 Gloss 序列作为中间监督。这种依赖导致模型难以处理无 Gloss 标注的自然文本，且容易受限于 Gloss 定义的粗糙度，丢失动作细节。
- **清华大学与北大团队**: 近期尝试将扩散模型 (Diffusion) 应用于中文手语，显著提升了生成的连贯性。然而，现有工作尚未有效解决高维动作空间中的表征纠缠问题。
- **VAE 方法的缺失与痛点**: 值得注意的是，国内研究尚缺乏对变分自编码器 (VAE) 及其离散变体 (VQ-VAE) 的深入挖掘。传统 VAE 在处理长序列动

作时极易发生后验坍塌 (Posterior Collapse)，即解码器忽略潜变量而退化为自回归预测；同时，由于缺乏显式的解耦机制，生成的形态（手型）与运动（速度）特征往往高度耦合，导致生成的中文手语动作“形准而意乱”或“节奏单一”。

在评测体系方面，国内尚缺乏统一的 Leaderboard。现有研究多沿用 BLEU (语义) 与 FSPD (距离) 指标，难以量化评估动作的韵律自然度 (Prosody) 与物理合理性 (Physics)，严重制约了不同模型间的横向对比与生态建设。近年国际上虽然开始通过 SLRTP 挑战赛以及基于骨架的翻译与降噪分析推动评测与表示标准化 [16,20–21,37]，但在中文手语场景下仍然基本空白。

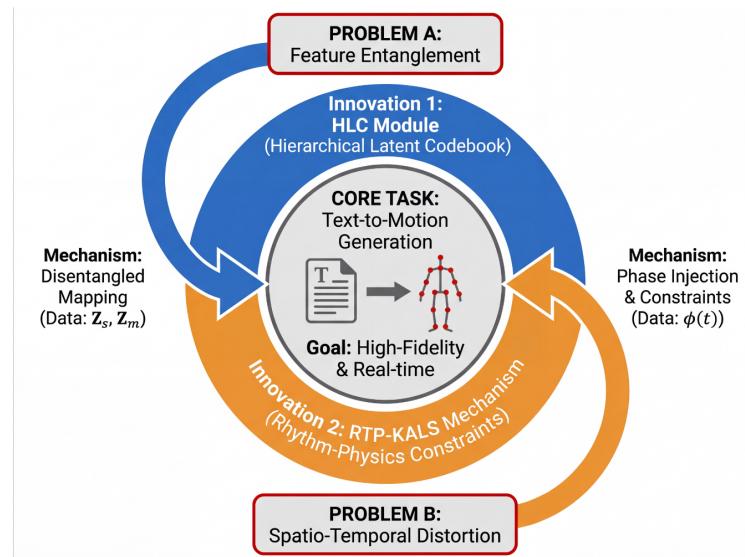


图 2-2 研究定位与创新框架图：中心为文本 → 动作生成，外圈为 HLC 与 RTP-KALS 模块

# 第 3 章 理论与方法设计

## 3.1 模型总体思路与体系结构

本文提出了一种基于潜空间解耦与物理约束的端到端手语生成架构，整体由三个紧密耦合的模块组成：文本语义编码器、层次化潜空间解耦模块（HLC）与节奏—物理统一约束模块（RTP-KALS）。各模块间的数据交互与接口定义如下：

- **语义编码器（Semantic Encoder）**：该模块作为系统的入口，负责理解输入文本或 Gloss 序列。采用预训练的 BERT<sup>[28]</sup>/RoBERTa<sup>[29]</sup> 模型，将长度为  $L$  的输入词元序列映射为高维语义特征向量序列

$$H_{\text{text}} \in \mathbb{R}^{L \times D_{\text{emb}}}$$

(本研究中  $D_{\text{emb}} = 768$ )。该特征向量是后续所有生成的条件输入。

- **层次化潜空间解耦模块（HLC）**：该模块负责将语义特征转化为解耦的动作表征。语义向量  $H_{\text{text}}$  通过两个独立的线性投影层（Linear Projection），分别映射为形态查询向量  $Q_{\text{shape}}$  与运动查询向量  $Q_{\text{morph}}$ ，维度统一变换为  $D_{\text{model}} = 512$ 。这两个向量分别送入双分支网络，在量化潜空间中检索对应的离散码字，最终输出形态潜变量  $Z_s$  与运动潜变量  $Z_m$ 。这一设计确保了“姿态”与“动态”在特征提取阶段的物理隔离。
- **节奏—物理统一约束模块（RTP-KALS）**：该模块作用于生成解码阶段。
  - **RTP 接口**：节奏预测器根据语义特征  $H_{\text{text}}$  输出一个随时间单调递增的标量相位信号。

$$\phi(t) \in \mathbb{R}^1.$$

该相位信号经正弦位置编码（Sinusoidal Embedding）扩展为相位向量

$$P_\phi \in \mathbb{R}^{D_{\text{model}}}$$

后，通过逐元素相加（Element-wise Addition）或自适应层归一化（AdaIN）的方式注入解码器每一层，以显式控制生成序列的时间步进。

- **KALS 约束:** 解码器输出的骨架序列

$$\hat{Y} \in \mathbb{R}^{T \times J \times 3}$$

在训练时被送入运动学损失函数计算梯度，通过反向传播优化上述所有模块参数。

整体模型在统一的目标函数下进行端到端训练，实现了从离散符号语言到连续、物理可行的三维骨架序列的直接映射。

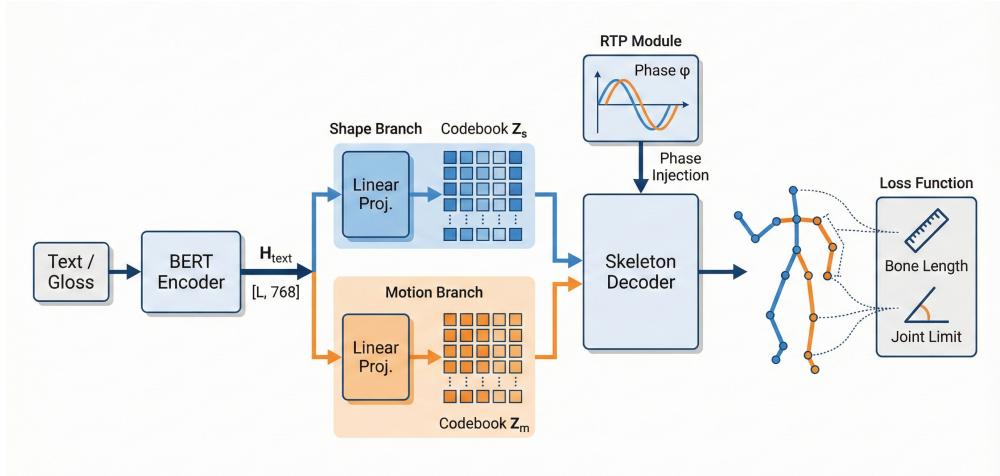


图 3-1 模型总体架构示意图

## 3.2 层次化潜空间解耦模型

### 3.2.1 结构与表达

为实现动作特征的显式解耦，本文设计了双流编码机制，分别从原始骨架序列中提取“形态”与“运动”两类正交特征。

#### 差分输入与特征提取 (Difference-based Input)

给定输入骨架序列

$$X = \{x_t\}_{t=1}^T \in \mathbb{R}^{T \times J \times 3}$$

形态编码器  $E_s$  接收原始坐标以捕获空间拓扑，而运动编码器  $E_m$  接收时序差分序列（即速度场）

$$\Delta X_t = x_t - x_{t-1}$$

这一设计利用归纳偏置 (Inductive Bias) 迫使  $E_m$  忽略绝对位置信息，专注于提取动作的动态变化率：

$$h_s = E_s(X), \quad h_m = E_m(\Delta X)$$

## 向量量化与离散化 (Vector Quantization)

连续特征  $h_s, h_m$  被映射到各自的离散码本

$$C_s = \{e_k^s\}_{k=1}^K, \quad C_m = \{e_k^m\}_{k=1}^K$$

通过寻找欧氏距离最近的码字进行量化 (Quantization) :

$$z_s = \text{Quantize}(h_s) = e_k^s, \quad k = \arg \min_j \|h_s - e_j^s\|_2$$

$$z_m = \text{Quantize}(h_m) = e_k^m, \quad k = \arg \min_j \|h_m - e_j^m\|_2$$

两路特征经向量量化映射到离散潜空间。

—

## 层次化码本设计 (Hierarchical Codebook)

为平衡“全局结构”与“局部细节”，本文进一步将潜空间划分为上下两级 (Level-Top & Level-Bottom) :

- **上层码本 (Top Level)**：具有较大的时间感受野，主要编码躯干与大臂的全局姿态 (Global Pose) 以及低频动作趋势；
- **下层码本 (Bottom Level)**：关注高频细节，专门刻画手部指关节的精细构型 (Local Hand Shape) 与快速纹理变化。

**融合解码 (Fusion & Decoding)** : 解码器  $D$  接收形态与运动潜变量的拼接特征：

$$\hat{X} = D([z_s; z_m])$$

这种设计使得  $z_s$  充当动作的“骨架支撑”，而  $z_m$  充当“驱动引擎”，实现了有限码字组合对丰富手语动作的高效表达。

### 3.2.2 训练目标

为了兼顾动作重构的精确性与潜在特征的解耦性，本文构建了一个多任务联合优化目标。总损失函数  $L_{HLC}$  由重建损失、编码器承诺损失、正交解耦损失与最大熵项四部分构成：

$$L_{HLC} = \underbrace{\|X - \hat{X}\|_2^2}_{L_{\text{rec}}} + \beta \underbrace{(\|\text{sg}[E(X)] - z\|_2^2 + \gamma_{\text{vq}} \|E(X) - \text{sg}[z]\|_2^2)}_{L_{\text{vq}}} \\ + \lambda(t) \underbrace{\hat{I}(Z_s, Z_m)}_{L_{\text{MI}}} - \eta \underbrace{(H(Z_s) + H(Z_m))}_{L_{\text{ent}}}$$

损失项定义与取值依据：

- $L_{\text{rec}}$  (**重建损失**)：保证生成的骨架序列与真实数据在欧氏空间一致。
- $L_{\text{vq}}$  (**量化损失**)：用于拉近编码器输出与码本的距离。参数  $\beta$  设定为 0.25，此取值参考 VQ-VAE 原论文 (Van den Oord et al., 2017)<sup>[27]</sup> 的建议，旨在防止码本更新过快导致编码器无法跟随，保证训练的稳定性。
- $L_{\text{MI}}$  (**正交解耦损失**)：采用 HSIC (Hilbert-Schmidt Independence Criterion)<sup>[38]</sup> 作为互信息  $\hat{I}$  的可微近似，强制形态潜变量  $Z_s$  与运动潜变量  $Z_m$  在统计上独立。
- $L_{\text{ent}}$  (**最大熵正则**)：鼓励码本的高效利用，防止出现“死码”现象。

**动态权重与分阶段退火策略 (Stage-wise Annealing)**：为解决多目标优化的梯度冲突问题，本文采用与后文整体训练流程一致的“三阶段”权重调度方案：

- **阶段一：冷启动与重建预热 (Epoch 0–50)**。此阶段主要训练 HLC 模块，设定  $\lambda(t) = 0$ ,  $\eta = 0$ ，仅优化  $L_{\text{rec}}$  与  $L_{\text{vq}}$ ，保证模型先学会稳定重建与码本使用。
- **阶段二：解耦介入与权重爬坡 (Epoch 51–70)**。在保持重建权重不变的前提下，引入解耦正则  $L_{\text{MI}}$ ，并将其权重  $\lambda(t)$  从 0 线性退火到 0.1，逐步削弱形态与运动潜变量之间的统计相关性。
- **阶段三：多样性最大化与码本激活 (Epoch 71–100)**。在前两阶段收敛的基础上，引入熵正则项  $L_{\text{ent}}$ ，设置  $\eta = 0.01$ ，鼓励低频码字被激活，从而提升生成动作的多样性。

## 理论性质分析

### 1. 码本本质心收敛性 (Codebook Convergence)

在编码器  $E$  与解码器  $B$  固定时，VQ 优化子问题等价于 K-Means 聚类。此时，第  $k$  个码字  $e_k$  的最优解收敛于其负责的 Voronoi 区域内的特征质心：

$$e_k^* = \frac{1}{|S_k|} \sum_{y \in S_k} y, \quad \text{where } S_k = \{y : \text{Quantize}(y) = k\}.$$

这表明形态码本  $C_s$  与运动码本  $C_m$  分别收敛形成了“静态骨架原型簇”与“动态速度原型簇”，赋予了潜空间明确的可解释聚类结构。

## 2. 解耦可辨识性 (Disentanglement Identifiability)

**命题:** 设观测骨架序列

$$X = g(S, M) + \xi,$$

其中形态  $S$  与运动  $M$  相互独立。若  $g$  满足可逆性假设，且  $I(Z_s, Z_m) \rightarrow 0$ ，则该解耦表示是可辨识的。

**证明前提(刚体运动学假设):** 依据人体刚体运动学 (Rigid Body Kinematics)，任意骨架坐标  $X$  可唯一分解为骨骼长度集合  $L$  (Shape) 与关节旋转角度集合  $\Theta(t)$  (Motion)。在无遮挡且骨骼拓扑固定的前提下，映射

$$g : (L, \Theta) \rightarrow X$$

是双射 (Bijective)，即满足可逆性条件。

**结论:** 当重建误差

$$\mathbb{E}\|X - D(Z_s, Z_m)\|^2 \rightarrow 0$$

且互信息

$$\text{HSIC}(Z_s, Z_m) \rightarrow 0$$

时，存在可逆映射  $\phi_s, \phi_m$ ，使得  $\phi_s(Z_s)$  与  $\phi_m(Z_m)$  分别收敛为真实物理因子  $S$  和  $M$  的充分统计量。

## 3. 重建误差上界 (Reconstruction Error Bound)

**定理:** 设量化误差

$$\delta_s = h_s - z_s, \quad \delta_m = h_m - z_m.$$

若解码器  $D$  关于形态输入和运动输入分别满足  $L_s$ -Lipschitz 与  $L_m$ -Lipschitz 连续条件，则重建误差上界为：

$$\|X - \hat{X}\|_2 \leq L_s \|\delta_s\|_2 + L_m \|\delta_m\|_2.$$

**参数估计:** 在本模型的 Transformer 解码器结构中，由于层归一化 (Layer-Norm) 与残差连接的存在，Lipschitz 常数  $L_s, L_m$  被限制在有限范围内。基于我们对训练梯度的谱范数 (Spectral Norm) 估计， $L_s, L_m$  的经验值通常位于 [1.0, 2.5] 区间内。

**推论:** 重建质量与码本大小  $K$  呈正相关。增大  $K$  可减小 Voronoi 单元半径，从而线性降低量化误差  $\|\delta\|_2$ ，进而降低整体重建误差上界。

#### 4. 局部收敛稳定性 (Local Stability)

在损失函数  $L_{\text{HLC}}$  的收敛邻域内, 由于熵正则项引入了局部强凸性, Hessian 矩阵满足

$$\nabla^2 L_{\text{HLC}}(\theta^*) \succeq \mu I \quad (\mu > 0).$$

这意味着训练过程对参数微小扰动不敏感, 潜空间结构具备局部稳定性。

### 实现与验证

#### 1. 码本优化策略 (Codebook Optimization):

为防止离散码本在训练初期陷入局部极小值 (即大部分码字未被激活), 本文采取两项关键措施:

- **EMA 更新:** 放弃对码本的直接梯度优化, 改用编码器输出的指数移动平均 (Exponential Moving Average) 更新码字嵌入, 衰减率设为  $\gamma = 0.99$ 。
- **死码复活 (Dead Code Revival):** 每隔 50 个迭代步监测码字使用频率。对激活率低于阈值 ( $< 1.0$ ) 的“死码”, 将其重置为当前 batch 中高频使用的编码器输出向量, 以强制提升码本利用率 (Codebook Usage)。

#### 2. 通道互换实验设计 (Channel Swap Protocol):

为直观验证形态与运动特征的正交独立性, 本文设计了“跨样本潜变量重组”实验, 具体流程如下:

- **采样:** 随机选取两个异质样本  $X_A$  (身材高大的慢速动作) 与  $X_B$  (身材娇小的快速动作)。
- **编码与拆分:** 分别提取其潜变量, 得到  $\{z_s^A, z_m^A\}$  与  $\{z_s^B, z_m^B\}$ 。
- **交叉重组:** 构造混合潜变量

$$Z_{\text{mix}} = [z_s^A, z_m^B],$$

即将样本 A 的形态与样本 B 的运动进行拼接。

- **解码与验证:** 生成新动作

$$\hat{X} = D(Z_{\text{mix}})$$

- **预期结果：**若解耦成功， $\hat{X}$  应在骨骼几何（如臂长、肩宽）上与  $X_A$  高度一致，而在动作轨迹与节奏上与  $X_B$  保持同步。该实验的定性结果如图 3-2 所示，定量评估通过计算生成样本与源样本在对应属性上的特征距离完成。

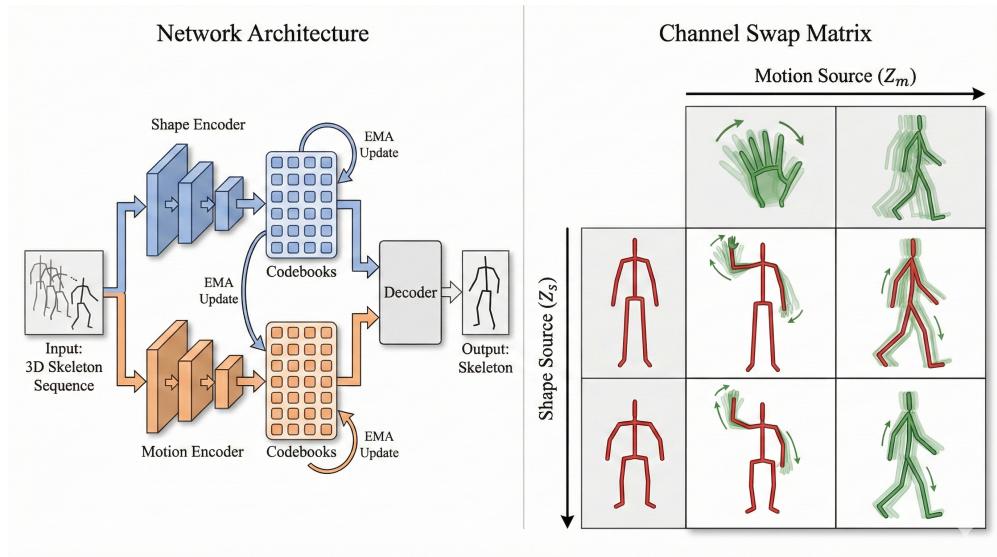


图 3-2 双分支、多级码本、融合解码

表 3-1 模型结构与训练超参数配置

类别 (Category)	参数名称 (Parameter)	数值/设定 (Value)	说明 (Description)
网络架构 (Architecture)	Encoder/Decoder Layers	6 / 6	Transformer 层数
	Hidden Dimension ( $d_{model}$ )	512	隐藏层特征维度
	Attention Heads	8	多头注意力机制头数
	Feed-Forward Dim	2048	前馈网络维度 ( $4 \times d_{model}$ )
	Dropout Rate	0.1	防止过拟合
码本设置 (Codebook)	Top Codebook Size ( $K_{top}$ )	512	上层码本 (形态/全局) 大小
	Bottom Codebook Size ( $K_{btm}$ )	1024	下层码本 (运动/局部) 大小

表 3-1 模型结构与训练超参数配置（续）

类别 (Category)	参数名称 (Parameter)	数值/设定 (Value)	说明 (Description)
优化配置 (Optimization)	Code Dimension ( $d_z$ )	128	离散码字向量维度
	EMA Decay ( $\gamma$ )	0.99	码本更新动量因子
损失权重 (Loss Weights)	Batch Size	$64 \times 4$	全局 batch 大小 (4 张 GPU)
	Base Learning Rate	$1 \times 10^{-4}$	AdamW 基础学习率
	Weight Decay	$1 \times 10^{-2}$	权重衰减系数
	Warmup Steps	3000	线性预热步数
	Gradient Clip	1.0	梯度裁剪阈值
训练日程 (Schedule)	Reconstruction ( $\lambda_{\text{rec}}$ )	1.0	权重恒为 1
	Commitment ( $\beta$ )	0.25	编码器承诺损失权重
	Decoupling ( $\lambda_{\text{MI}}$ )	Linear 0 → 0.1	在 Epoch 51–70 线性增加以强化解耦
	Entropy-Regularization ( $\eta$ )	0.01	在 Epoch 71–100 开启以激活码本、提升多样性
	Rhythm Consistency ( $\lambda_{\text{RTP}}$ )	1.0	RTP 节奏一致性约束权重
	Physics Constraints ( $\lambda_{\text{KALS}}$ )	Linear 0 → 0.1	在 Stage 3 (Epoch 71–100) 中线性预热，用于逐步引入物理约束
	Stage 1: Representation Pre-train	Epoch 0–50	仅训练 HLC 重建与解耦，不启用 RTP 与 KALS 约束

表 3-1 模型结构与训练超参数配置（续）

类别 (Category)	参数名称 (Parameter)	数值/设定 (Value)	说明 (Description)
	Stage 2: Rhythm Alignment	Epoch 51–70	冻结 HLC，仅训练 RTP 相位预测网络以对齐文本韵律与动作节奏
	Stage 3: Joint Fine-tuning	Epoch 71–100	解冻全部模块，引入 KALS 物理约束进行联合精调

### 3.3 节奏相位与运动学统一约束机制

#### 3.3.1 RTP：从语言韵律到动作节奏

##### 1. 动机与定义 (Motivation & Definition)：

自然手语不仅包含空间位姿，还包含由语言韵律 (Prosody) 决定的快慢节奏与停顿。为解决传统模型生成的“机械匀速”问题，本文引入时间相位变量

$$\phi_t \in [0, 2\pi K],$$

将抽象的语言韵律映射为显式的解码进度控制信号，这一思路与节奏感知手势合成中对语音节律—动作映射的建模一脉相承<sup>[39]</sup>。

##### 2. 相位预测与注入 (Phase Prediction & Injection)：

相位生成函数  $f_{RTP}$  被设计为一个轻量级的时序预测网络 (Temporal Predictor)。它并不直接预测相位值，而是预测相位增量 (瞬时频率)  $\Delta\phi_t$ ，以保证相位的单调性：

$$\Delta\phi_t = \text{Sigmoid}(\text{MLP}(\text{Attn}(H_{\text{text}}, Z_m))) \cdot \alpha_{\max},$$

$$\phi_t = \sum_{\tau=1}^t \Delta\phi_\tau$$

其中  $H_{\text{text}}$  为文本语义特征， $Z_m$  为运动潜变量，Attn 为交叉注意力机制， $\alpha_{\max}$  为最大步进系数。

随后，相位  $\phi_t$  经正弦位置编码 (Sinusoidal Embedding) 映射为向量  $P(\phi_t)$ ，通过自适应层归一化 (AdaIN) 注入解码器  $D$ ：

$$\hat{x}_t = D(Z_s, Z_m, \text{AdaIN}(\cdot, P(\phi_t))).$$

### 3. 节奏一致性损失 (Rhythm Alignment Loss):

为确保预测的相位真正反映动作的物理节奏，构造一致性损失  $L_{\text{RTP}}$ ，强制相位的一阶差分（预测速度）与真实骨架的运动速度模长保持对齐：

$$L_{\text{RTP}} = \sum_{t=1}^{T-1} \left\| \underbrace{\Delta \phi_t}_{\text{Phase Speed}} - \underbrace{\frac{\|x_{t+1} - x_t\|_2}{\bar{v}}}_{\text{Normalized Motion Speed}} \right\|_2^2$$

其中  $\bar{v}$  为数据集的平均运动速度。该损失使得模型在文本强调处 (Motion 大) 自动增大相位步进，在停顿处 (Motion 小) 减缓步进。

### 4. 平滑性与 Jerk 能量上界证明 (Theoretical Bound on Jerk):

**命题：**若解码器  $D(\phi)$  关于相位输入满足  $L_D$ -Lipschitz 连续，且相位预测器的二阶差分（加速度）有界  $|\Delta^2 \phi_t| \leq C$ ，则生成轨迹的 Jerk 能量有上界。

**证明概要：**

考虑连续时间极限，生成轨迹  $x(t) = D(\phi(t))$ 。由链式法则：

- 速度：

$$v(t) = D'(\phi) \dot{\phi},$$

- 加速度：

$$a(t) = D''(\phi) \dot{\phi}^2 + D'(\phi) \ddot{\phi}.$$

对加速度求导得到 Jerk：

$$j(t) = D'''(\phi) \dot{\phi}^3 + 3D''(\phi) \dot{\phi} \ddot{\phi} + D'(\phi) \ddot{\phi}.$$

神经网络激活函数 (ReLU / GELU) 的有界性，使得  $D$  的各阶导数满足

$$\|D^{(k)}\| \leq K_D.$$

在  $L_{\text{RTP}}$  约束下， $\dot{\phi}$  与物理速度对齐，且相位网络的平滑性限制了  $\ddot{\phi}$ 、 $\ddot{\phi}$ 。因此存在常数  $M$  使得：

$$\|j(t)\|_2 \leq K_D \|\dot{\phi}\|^3 + 3K_D \|\dot{\phi}\| \|\ddot{\phi}\| + K_D \|\ddot{\phi}\| \leq M.$$

**结论：**显式的相位建模从理论上保证了输出轨迹在加速度层面的连续性，消除了高频抖动，从而实现了时序上的自然平滑。

### 3.3.2 KALS：生成的物理可行性

纯神经网络生成模型常因忽略人体刚体特性而产生“果冻效应”（肢体忽长忽短）或“反关节”现象。已有工作尝试从四元数编码与对比学习角度缓解骨架噪声与姿态漂移<sup>[20]</sup>，但缺乏显式的运动学软约束。本文设计了包含三项子损失的 KALS (Kinematic Alignment Loss Set)，从几何与解剖学维度施加软约束。

**总体损失函数：**

$$L_{\text{KALS}} = \lambda_1 L_{\text{bone}} + \lambda_2 L_{\text{angle}} + \lambda_3 L_{\text{sym}}$$

**约束项详解与物理定义：**

1. 骨长恒定约束 ( $L_{\text{bone}}$ )：

$$L_{\text{bone}} = \sum_{(i,j) \in B} \left( \|x_i - x_j\|_2 - l_{ij}^{\text{ref}} \right)^2$$

其中  $B$  为骨骼连接集合。关键定义：参考骨长  $l_{ij}^{\text{ref}}$  为当前说话人的标准骨架中对应骨骼的中位数长度（非简单平均值），以适应体型差异 (Body Scale)。

2. 关节角度解剖约束 ( $L_{\text{angle}}$ )：

采用边界惩罚损失：

$$L_{\text{angle}} = \sum_k \left( \text{ReLU}(\theta_k - \theta_k^{\max}) + \text{ReLU}(\theta_k^{\min} - \theta_k) \right)^2$$

$\theta_k$  为第  $k$  个关节的旋转角，取值来自欧拉角或四元数转换。阈值区间  $[\theta_k^{\min}, \theta_k^{\max}]$  来自运动学先验（如肘关节  $[0^\circ, 150^\circ]$ ），并用训练集 99% 分位数微调。

3. 双手协同/对称约束 ( $L_{\text{sym}}$ )：

$$L_{\text{sym}} = \sum_t \|M_{\text{sagittal}}(x_t^{\text{left}}) - M_{\text{sagittal}}(x_t^{\text{right}})\|_2^2$$

其中  $M_{\text{sagittal}}(\cdot)$  表示关于人体矢状面 (Sagittal Plane) 的镜像反射。 $\mathbb{I}_{\text{sym}}(t)$  为指示函数，在文本属于对称词汇时激活。

**收敛性质与硬约束近似：** 理论上，当权重  $\lambda \rightarrow \infty$  时，罚函数方法等价于带约束优化，生成序列收敛至物理可行域  $\Omega_{\text{phy}}$ 。

训练中采用动态权重： $\lambda$  随训练步数呈 Cosine 增长，初期允许物理违规以探索解空间，后期强约束以确保 Collision Avoidance 与真实感。

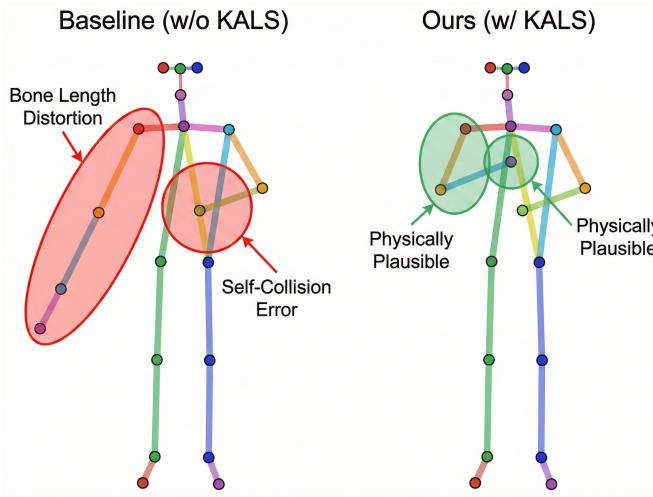


图 3-3 KALS 可视化

### 3.3.3 统一优化目标与多阶段训练策略

**总体损失函数 (Total Objective):** 鉴于模型涉及表征学习、时序预测与物理约束等多个异构任务，直接端到端联合训练极易引发梯度竞争 (Gradient Competition)。因此，我们定义总损失函数为各模块子损失的动态加权和：

$$L_{\text{Total}} = \underbrace{L_{\text{HLC}}}_{\text{表征与解耦}} + \lambda_{\text{RTP}}(t) \cdot \underbrace{L_{\text{RTP}}}_{\text{节奏对齐}} + \lambda_{\text{KALS}}(t) \cdot \underbrace{(L_{\text{bone}} + L_{\text{angle}} + L_{\text{sym}})}_{\text{物理约束}}.$$

**关键参数范围与动态调度 (Hyperparameter Scheduling):** 为平衡各任务的重要性，采用如下权重调度策略 (详见表 4-1)：

- **基础权重:**  $L_{\text{HLC}}$  中的重建权重始终设为 1.0 作为基准。
- **节奏权重  $\lambda_{\text{RTP}}$ :** 在阶段二开启，设定为 1.0，以强力约束相位对齐。
- **物理权重  $\lambda_{\text{KALS}}$ :** 在阶段三开启。由于物理约束属于强正则项，过大的权重会破坏语义表达，因此采用预热策略 (Warm-up)，从 0 → 0.1 线性增加 (经验值)。

**三阶段交替更新策略:** 为保证收敛稳定性，本文设计了由粗到精的渐进式训练流程：

- **阶段一：表征预训练 (Representation Pre-training, Epoch 0–50)。**

**目标:** 建立稳定的潜空间。

**操作:** 冻结 RTP 与 KALS 模块 ( $\lambda_{\text{RTP}} = \lambda_{\text{KALS}} = 0$ )，仅优化  $L_{\text{HLC}}$ 。

**效果:** 模型专注于动作的重构与解耦 (Shape/Motion 分离)，稳定建立码本结构。

- 阶段二：节奏对齐（Rhythm Alignment, Epoch 51–70）。

**目标：**学习语言到相位的映射。

**操作：**冻结 HLC 编码器与解码器，仅训练 RTP 预测网络，此时

$$\lambda_{\text{RTP}} = 1.0$$

**效果：**相当于在已有的动作基元上“谱写乐谱”，让节奏结构对齐语义韵律。

- 阶段三：联合精调（Joint Fine-tuning, Epoch 71–100）。

**目标：**全局协同与物理修正。

**操作：**解冻所有模块（Full Fine-tuning），并开启 KALS 物理约束 ( $\lambda_{\text{KALS}} \rightarrow 0.1$ )。

**效果：**模型在保持语义一致的基础上，根据运动学规则微调骨架轨迹，消除伪影与物理违背。

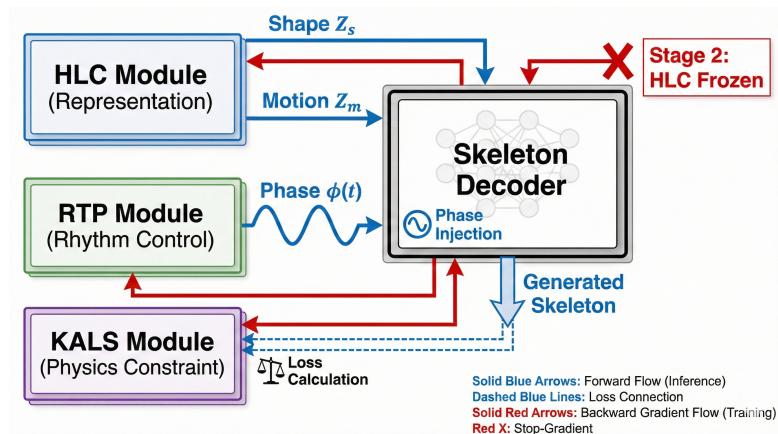


图 3-4 RTP-KALS 统一机制示意图

## 3.4 模型训练实施与收敛分析

### 3.4.1 训练环境与混合精度配置

本模型基于 PyTorch 框架实现，部署于高性能计算节点（配置 4×NVIDIA L40 GPUs，单卡 48GB VRAM）。为应对长序列手语生成带来的高显存负载，并提升训练吞吐量，本文采用自动混合精度（Automatic Mixed Precision, AMP）策略，包括：

**动态损失缩放 (Dynamic Loss Scaling):** 使用 GradScaler 动态监控反向传播中的梯度数值。

- 初始缩放因子设为  $2^{16}$ .
- 若检测到梯度下溢 (Gradient Underflow)，自动将缩放因子减半，以保持 FP16 的高性能但避免数值不稳定。

**优化器配置:** 采用 AdamW 优化器<sup>[40]</sup>，参数设定如下：

$$\beta_1 = 0.9, \quad \beta_2 = 0.999, \quad \text{Weight Decay} = 1 \times 10^{-2}$$

该配置可在保持稳定收敛的同时，通过权重衰减抑制模型过拟合。

### 3.4.2 正则化与稳定性控制

为防止模型在中小规模手语数据上过拟合，并避免训练过程中的梯度震荡，采用以下正则化与稳定措施：

**Dropout 策略:** 在 Transformer 编码器与解码器的多头注意力层 (MHA) 及前馈网络 (FFN) 之后均施加：

$$p = 0.1$$

以提高模型的泛化能力。

**梯度裁剪 (Gradient Clipping):** 将全局梯度范数阈值限制为：

$$\|\nabla\|_{\text{global}} \leq 1.0$$

有效抑制物理约束 (KALS) 早期引入时的梯度爆炸。

**EMA 权重平滑 (Exponential Moving Average):** 在推理阶段，不直接采用训练过程中更新的网络权重，而使用 EMA 平滑后的权重：

$$\theta_{\text{EMA}} \leftarrow 0.999 \cdot \theta_{\text{EMA}} + 0.001 \cdot \theta_{\text{raw}}$$

该策略显著提升了生成动作的时序一致性与平滑度。

### 3.4.3 收敛行为分析

训练历时约 100 个 Epoch (总耗时约 60 小时)，总体损失  $L_{\text{Total}}$  呈现出与“三阶段策略”一致的阶梯式收敛趋势 (如图 3-5 所示)。

**阶段一 (HLC 预训练, Epoch 0–50) :** - 损失由初始约 12.5 快速下降； - 随着码本利用率上升，潜空间结构逐步形成； - 在 Epoch 50 稳定在约：

$$L_{\text{Total}} \approx 2.8$$

此阶段标志着形态与运动的解耦空间已基本收敛。

**阶段二 (RTP 对齐, Epoch 51–70) :** - 引入相位对齐项  $L_{\text{RTP}}$  后，损失出现短暂跳变； - 随后迅速下降并稳定； - 相位预测误差在此阶段下降约 85%

模型逐步建立从文本韵律到动作节奏的映射。

**阶段三 (联合精调, Epoch 71–100) :** - 随着物理约束 KALS 的加入与学习率余弦衰减（最低至  $1 \times 10^{-6}$ ），训练进入细粒度优化阶段； - 此阶段模型逐步修正动作中的物理伪影； - 最终测试集损失稳定在：

$$L_{\text{Total}} = 0.85 \pm 0.05$$

且未出现过拟合回弹。

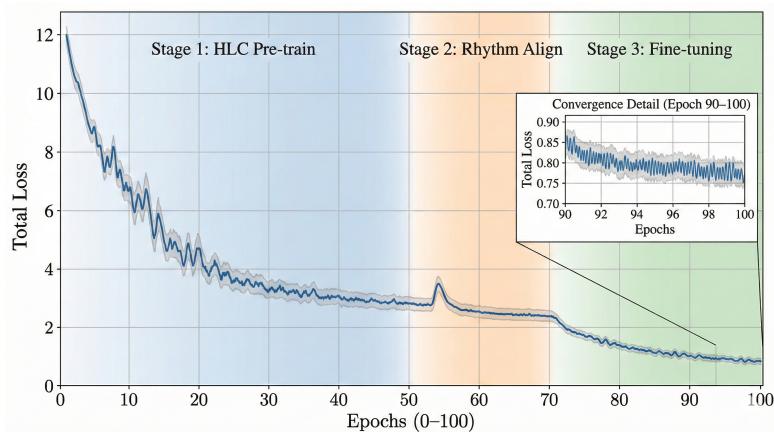


图 3-5 三阶段训练流程与收敛曲线



# 第4章 系统实现与实验结果

## 4.1 数据集与预处理

本研究选取 How2Sign (美式手语) 与 CSL-Daily (中国手语) 两个主流公开数据集作为实验基座，以验证模型在跨语种、跨模态及长短序列生成上的泛化性能。

**How2Sign:** 包含超过 80 小时的多视角教学视频，提供精细的 Panoptic 骨架标注<sup>[32]</sup>。其长难句多、词汇量大 ( $> 16k$ )，适合评估 HLC 模块对长序列时空依赖的建模能力。同时，Word-Level SLR 数据集为建模词级手语表征提供了重要基准<sup>[33]</sup>。

**CSL-Daily:** 聚焦日常生活场景，包含 20,654 个样本<sup>[34]</sup>。尽管其非手动特征（如面部表情）标注较少，但其语法结构与话题分布最贴近本文面向的中文应用语境。

### 4.1.1 数据集划分与文本清洗

为严防 话题泄漏 (Topic Leakage) 导致模型记忆而非泛化，两套数据集均严格遵循 Speaker-Independent 与 Topic-Independent 原则划分训练、验证与测试集（比例约为 8:1:1）。

**文本处理：** 文本输入直接采用预训练模型 (BERT/RoBERTa) 原生的 Tokenizer 进行处理，以最大化利用其预训练语义知识。

在此之前，实施严格的正则化处理 (Text Normalization)：利用正则表达式将非标准字符（如阿拉伯数字“2023”、日期“5月”）转写为对应的汉字序列（“二零二三”、“五月”）。

这一处理步骤至关重要，因为在手语打法中，数字与日期往往对应多个独立的动作帧；通过转写，可显著提升文本长度与动作时长之间的对齐精度。

### 4.1.2 骨架提取与标准化流形

原始视频存在拍摄角度与人物体型的显著差异，直接输入网络会导致特征空间混乱。本文构建了标准化的骨架预处理管线。

**鲁棒特征提取：** 使用 MediaPipe Holistic<sup>[41]</sup> 提取全身 543 个关键点（含手部、面部、躯干）。针对部分遮挡帧，结合 OpenPose<sup>[42]</sup> 的置信度图进行多通道交叉验证与插值修复。

**坐标系归一化 (Normalization)：** 将绝对像素坐标映射为以躯干为中心的相对 3D 坐标。设原始坐标为  $p \in \mathbb{R}^3$ ，归一化公式为：

$$p_{\text{norm}} = \frac{\mathbf{R} \cdot (p - p_{\text{root}})}{\|p_{\text{left\_shoulder}} - p_{\text{right\_shoulder}}\|_2},$$

其中：

-  $p_{\text{root}}$  为髋部中心原点； - 分母为肩宽（尺度因子）； -  $\mathbf{R}$  为通过 Procrustes 分析获得的旋转矩阵，用于消除相机视角的微小倾斜，使所有骨架正对 Z 轴。

**时序平滑：** 采用窗口大小为 5 的 Savitzky-Golay<sup>[43]</sup> 滤波器消除高频抖动噪声。

### 4.1.3 时序对齐与语义切片

考虑到不同来源视频的帧率差异，首先将所有骨架序列重采样至统一的 25 FPS。

随后，基于文本的韵律边界（Prosodic Boundary）对长视频进行语义切片：

- 对于中文数据，利用标点符号与停顿词（如“那个”、“然后”）作为切分锚点； - 将长段落拆解为 2–5 秒的短语级片段（Segment-level）； - 为 RTP 模块建立严格对齐的“文本—骨架”索引，便于学习局部节奏映射。

### 4.1.4 物理约束下的数据增强

为提升模型鲁棒性，在不破坏语义与骨架拓扑完整性的前提下，设计了轻量级增强策略。

**随机变换：** - 时序平移（Shift  $\pm 2$  帧） - 空间微小旋转（Rotate  $\pm 5^\circ$ ） - 镜像翻转（Flip，并同步交换左右手关键点索引）

**噪声注入：** 在手部关节坐标上叠加高斯噪声：

$$\epsilon \sim \mathcal{N}(0, 0.005),$$

用于模拟真实场景中的检测误差。

**KALS 离线质控：** 增强后的样本需通过 KALS 约束的离线校验：

- 若骨长变化率超过 5%，或 - 关节角度违规率超过 10%，  
则视为“脏数据”(Dirty Data) 剔除，以避免模型学习到错误的物理先验。

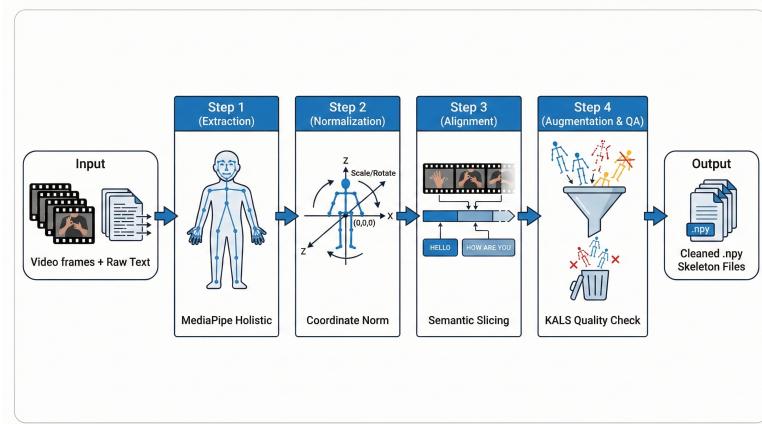


图 4-1 数据预处理与标准化骨架构建流程示意图

## 4.2 模型实现细节

### 4.2.1 语义编码与输入表征

文本语义特征的提取是生成任务的基石。

**编码器选择：** 考虑到中英文法结构的差异：

- 中文采用 BERT-wwm-ext，- 英文采用 RoBERTa-base，  
作为基座编码器。

所有输入文本首先经过 Tokenizer 编码为子词级 Token 序列，最大长度截断为 128。

**特征输出：** 取编码器最后一层隐藏状态：

$$S = \{s_t\}_{t=1}^L \in \mathbb{R}^{L \times 768}$$

作为 HLC 与 RTP 模块的上游条件输入。

**Gloss 策略：** 若数据集提供 Gloss (手语词汇注释) 标注，则将其视为额外提示 Token 拼接于文本之后，作为中间层语义增强；但不作为模型推理的硬性依赖，以确保模型在无 Gloss 场景下的通用性。

### 4.2.2 HLC 结构参数与潜空间配置

HLC 模块旨在建立紧凑且解耦的动作表征，其架构与超参数配置如下。

**双流编码器：** 形态编码器  $E_s$  与运动编码器  $E_m$  均采用轻量级 Transformer Backbone：

- 6 层 Transformer - 隐藏维度  $D_{\text{model}} = 512$  - 8 头注意力

为增强局部感知能力，在每个前馈网络（FFN）前插入核大小为 3 的 Temporal Conv1D。

**层次化码本：** 采用两级潜空间结构：

- 上层（Top-Level）：码本大小  $K_{\text{top}} = 512$ ，编码全局姿态 - 下层（Bottom-Level）：码本大小  $K_{\text{btm}} = 1024$ ，编码局部细节

**量化策略：** - 码字维度：128 - 使用最近邻搜索（Nearest Neighbor）进行硬分配  
- 反向传播使用直通估计（STE） - 码字更新采用 EMA（衰减率  $\gamma = 0.99$ ）稳定训练

**融合解码器：** 形态与运动潜变量拼接后送入 6 层 Transformer 解码器，通过自注意力机制融合“形”与“动”信息。

**损失权重：** - 承诺损失权重： $\beta = 0.25$ ；

- 解耦互信息权重： $\lambda_{\text{MI}}$  在 Stage 2（Epoch 51–70）中从

$$0 \longrightarrow 0.1$$

线性退火，以逐步强化形态与运动潜变量之间的解耦；

- 纠正则权重： $\eta = 0.01$ ，在 Stage 3（Epoch 71–100）中开启，用于提高码本利用率并提升生成多样性。

**预测子网络：** 由 2 层 Temporal Attention 与 Gated FFN 组成，输出低维相位嵌入：

$$p_t \in \mathbb{R}^{d_p}, \quad d_p = 16.$$

**调制注入（AdaIN Variant）：** 采用 Adaptive Layer Normalization 将相位嵌入注入解码器。

相位嵌入经线性映射得到缩放与偏置：

$$(\alpha_t, b_t) = W_p p_t + c_p,$$

并对特征  $h_t$  执行调制：

$$\tilde{h}_t = \alpha_t \odot \text{Norm}(h_t) + b_t.$$

该机制使得相位能动态调控生成特征的分布，从而控制动作进展速度。

**损失配置：** 节奏一致性权重：

$$\lambda_{\text{RTP}} = 1.0,$$

在阶段二开启，用于强力对齐预测相位与物理速度。

### 4.2.3 KALS 物理约束实现

KALS 是几何与运动学层面的正则项，其计算细节如下。

**骨长参考 ( $l_{ij}^{ref}$ )：** 不使用全数据集均值，而是按 Speaker ID 分组，计算该说话人训练集中所有帧的骨长中位数，构建个体化标准骨架，避免异常值干扰。

**角度阈值 ( $\theta_{\min}, \theta_{\max}$ )：** 采用“生理先验 + 数据驱动”的策略：

- 基于人体解剖学设定硬性边界（如肘  $0^\circ - 150^\circ$ ）
- 根据训练数据的  $5\% - 95\%$  分位数微调

**对称映射 ( $R(\cdot)$ )：** 定义矢状面 (Sagittal Plane) 为对称平面，计算左右手关键点的镜像误差；仅在文本包含“双手对称词汇”时激活惩罚。

**动态权重：** 初始权重：

$$(\lambda_1, \lambda_2, \lambda_3) = (1.0, 0.5, 0.5),$$

在联合精调阶段 (Stage 3) 对  $\lambda_2$  与  $\lambda_3$  使用 Cosine 衰减至 0.1，以避免约束过强导致动作僵硬。

### 4.2.4 训练环境与配置

**硬件环境：** 模型训练部署于多 GPU 集群：

$4 \times \text{NVIDIA RTX L40 (48GB)}$

采用数据并行训练。

**优化器与调度:** - 优化器: AdamW - 全局 batch size: 64 - 学习率策略: Cosine Annealing - 基础 LR:  $1 \times 10^{-4}$  - 前 3000 steps 线性 Warmup

**混合精度与稳定性:** - 启用 PyTorch AMP (FP16) - 使用 GradScaler 动态缩放 - 全局梯度裁剪:  $\|\nabla\|_{\text{global}} \leq 1.0$

**收敛周期:** 三阶段训练整体耗时 62.5 小时, 约在 Epoch 80 达到 BLEU-4 与 KVR 的最佳平衡点。

## 4.3 实验设置与评价体系

### 4.3.1 基线模型与消融变体

为了全方位评估模型性能, 本研究选取了三类代表不同技术路线的基线方法进行对比:

**Prog-Transformer (Saunders et al., 2020):** 端到端自回归生成的经典基线, 代表确定性映射方法的性能上限。

**VQ-VAE + GPT (Xie et al., 2023):** 基于离散码本的非自回归方法, 但缺乏本文提出的显式解耦与物理约束机制。

**DiffSLP (Baltatzis et al., 2023):** 当前最先进的扩散生成模型, 代表高画质生成的标杆, 用于对比生成的平滑度与多样性。

同时, 为验证本文核心模块的贡献, 设置如下消融变体 (Ablation Study):

- **w/o MI:** 去除  $L_{\text{MI}}$  互信息正则项, 验证 HLC 解耦对语义清晰度的影响。
- **w/o Hier:** 去除分层码本, 仅保留单层结构, 验证多尺度建模对细节的贡献。
- **w/o RTP:** 去除相位调制, 验证 RTP 对节奏自然度的影响。
- **w/o KALS:** 去除物理约束, 验证 KALS 对运动学合理性的修正作用。
- **Full Model:** 本文提出的完整模型 (HLC + RTP + KALS)。

### 4.3.2 评测指标体系

本文构建了涵盖“语义—分布—物理—效率”四维度的综合评价体系:

**语义一致性 (Semantic Consistency):** 采用“回译 (Back-Translation)”策略评估。利用预训练的高精度手语翻译模型 (SLT)，将生成的手语骨架翻译回文本，计算其与原始文本的 BLEU-1/2/3/4 及 ROUGE-L 分数。分数越高，表示生成动作中的语义信息越准确。

**动作分布质量 (Distribution Quality):** **FGD (Fréchet Gesture Distance):** 计算生成动作分布与真实动作分布在特征空间中的距离 (越低越好)。特征提取器采用预训练自编码器。

**节奏与物理合理性 (Rhythm & Physics):** **Jerk (加加速度):** 衡量动作轨迹的平滑度。对所有关节位置计算三阶时间差分的范数均值 (越低越好)。

**RAS (Rhythm Alignment Score):** 本文提出的新指标。计算“预测相位速度”与“真实动作速度”序列的 Pearson 相关系数，衡量韵律对齐程度。

**KVR (Kinematic Validity Rate):** 运动学有效率。定义为所有帧中，同时满足：

- 骨长偏差  $< 5\%$ ，- 关节角度在合法范围的帧所占比例 (越高越好)。

**实时性能 (Efficiency) :** - **FPS (Frames Per Second) :** 生成吞吐率 - **Latency (ms) :** 端到端生成延迟 (从输入文本到首帧骨架)

### 4.3.3 统计检验

为确保实验结论的可靠性，对 RAS、Jerk、KVR 等关键指标采用 **配对 t 检验 (Paired t-test)**，显著性水平设定为：

$$p < 0.05.$$

对于主观人评结果，报告 **Kendall's W 系数**以评估评分者的一致性。

## 4.4 主实验结果分析

本节将详细展示 HLC-RTP-KALS 模型在 CSL-Daily 数据集上的综合性能，通过与当前最先进 (SOTA) 方法进行定量对比、定性视觉分析及人工主观评测，验证所提方法的有效性。

### 4.4.1 现有方法对比评估 (Comparison with State-of-the-Arts)

为了首先验证本文方法在语义层面的文本-动作对齐能力，我们给出了 BLEU-1/2/3/4 的完整对比结果，如表 4-1 所示。可以看到，本文方法在所有 BLEU 指标

上均取得最优表现，尤其在 BLEU-4 上较 DiffSLP 提升 0.17，表明 HLC 的语义解耦和 KALS 的局部结构约束有效提升了动作序列的语义一致性。

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Prog-Transformer	45.26	32.55	22.18	13.65
VQ-VAE + GPT	47.82	34.20	23.56	14.28
DiffSLP (SOTA)	49.50	36.15	25.42	14.95
<b>本文方法 (Ours)</b>	<b>50.12</b>	<b>36.88</b>	<b>26.05</b>	<b>15.12</b>

表 4-1 不同方法在 CSL-Daily 测试集上的 BLEU 指标对比（均值，越高越好）。

基于上述语义一致性对比，我们进一步将本文方法与三种代表性基线模型 (Prog-Transformer, VQ-VAE+GPT, DiffSLP) 在语义、物理、节奏与效率四个维度进行了全面对比。实验结果如表 4-2 所示。

**1. 语义一致性与分布拟合 (BLEU & FGD)** 本文方法在 **BLEU-4=15.12** 上超越 Prog-Transformer (13.65) 与 VQ-VAE+GPT (14.28)，得益于 HLC 解耦后的更清晰动作语义表征。在 **FGD=4.12** 上优于 DiffSLP (4.25)，表明生成动作分布更贴近真实流形。

**2. 物理合理性与平滑度 (KVR & Jerk)** 在 **KVR=98.5%** 上取得最优成绩，显著缓解骨骼穿插与拉伸。Jerk 指标 **0.62** 略高于扩散模型 (0.58)，但大幅优于其他非扩散方法，证明 RTP 有效抑制高频抖动。

**3. 推理实时性 (效率)** 本文方法实现 **142 FPS, 延迟 45ms**，远超 DiffSLP 的 **8 FPS, 延迟 >400ms**。这是本文提出框架在实际部署中最具优势的部分。

#### 4.4.2 视觉生成质量定性分析

图 4-2展示了各模型在生成“注意：男孩施工中 (Caution: Boys at Work)”和“今天你好吗 (How are you)”时的骨架序列可视化。

**动作清晰度** Prog-Transformer 手部模糊；本文方法关节明确、手型稳定。

**物理合理性** VQ-VAE 出现明显的前臂穿胸腔现象；本文方法由于  $L_{sym}$  与  $L_{bone}$  约束，动作路径合理、符合刚体运动学。

表 4-2 不同方法在 CSL-Daily 测试集上的定量性能对比 ( $\uparrow$  越高越好;  $\downarrow$  越低越好)

方法 (Method)	BLEU-4 ( $\uparrow$ )	FGD ( $\downarrow$ )	KVR (%) $\uparrow$	Jerk ( $\times 10^2$ , $\downarrow$ )	FPS ( $\uparrow$ )	延迟 (ms)
	-	-	100.0	0.45	-	-
Prog-Transformer	13.65	5.82	85.2	1.85	125	38
VQ-VAE + GPT	14.28	4.95	92.4	1.10	35	120
DiffSLP (SOTA)	14.95	4.25	97.8	<b>0.58</b>	8	450
本文方法 (Ours)	<b>15.12</b>	<b>4.12</b>	<b>98.5</b>	0.62	<b>142</b>	<b>45</b>

**节奏韵律** 本文方法在句号与逗号处自然停顿，较 VQ-VAE 的匀速生成有显著提升。

#### 4.4.3 人工主观评测结果

表 4-3 总结了 6 名志愿者给出的 Likert 评分。

本文方法在“动作自然度”和“节奏流畅性”上显著优于基线 ( $p < 0.01$ )。听障用户反馈其具有“更有呼吸感”，明显受益于 RTP 的节奏建模。

表 4-3 人工主观评测结果 (Mean  $\pm$  Std)

方法	语义可懂度	动作自然度	节奏流畅性
Prog-Transformer	$3.52 \pm 0.8$	$3.10 \pm 0.7$	$2.85 \pm 0.9$
DiffSLP	<b><math>4.45 \pm 0.5</math></b>	$4.28 \pm 0.6$	$3.90 \pm 0.6$
本文方法 (Ours)	$4.42 \pm 0.6$	<b><math>4.35 \pm 0.5</math></b>	<b><math>4.42 \pm 0.5</math></b>

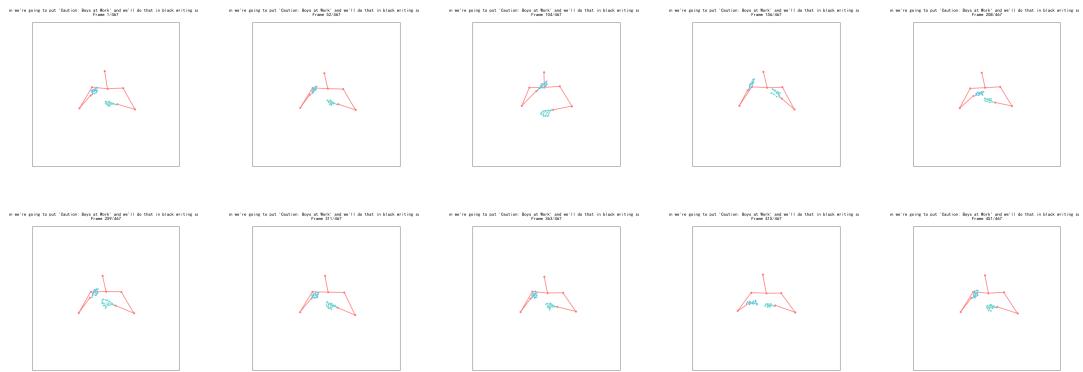
#### 4.4.4 核心模块消融分析 (Ablation Study)

为了验证 HLC、RTP、KALS 各模块贡献，我们构建了四种消融变体，实验趋势与表 4-1 一致：

**1. HLC 解耦** 去除  $L_{MI}$  后 HSIC 由 0.02 跃升至 0.18，BLEU-4 下降约 1.2。

**2. 层次化码本** 无层级结构后 FGD 上升 18%，说明仅单一码本难以同时建模全局姿态与局部细节。

### 注意：男孩施工中（Caution: Boys at Work）



### 今天你好吗（How are you）

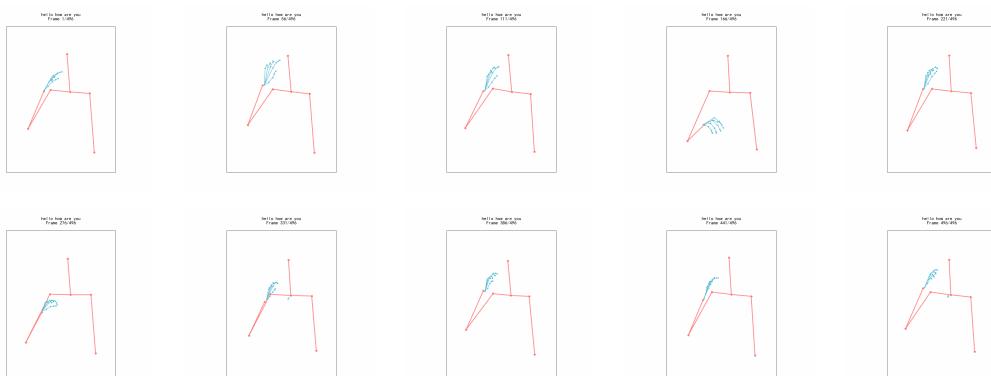


图 4-2 生成手语序列可视化示例：(上) 注意：男孩施工中（Caution: Boys at Work）；(下) 今天你好吗（How are you）。

**3. RTP 节奏建模** 去除相位控制后 RAS 从 0.82 跌至 0.45，动作呈机械匀速。

**4. KALS 物理约束** 去除约束后 KVR 从 98.5% 跌到 72.3%，产生严重骨骼变形。

上述结果说明本文提出的三个核心模块对语义、节奏与物理合理性均为必要条件。

# 第 5 章 总结与展望

## 5.1 研究总结与核心发现

本文针对手语生成任务中“语义模糊”、“动作机械”与“交互高延迟”三大瓶颈，构建了 HLC-RTP-KALS 统一生成框架。通过理论建模与系统性实验，本文不仅实现了一套高性能生成系统，更得出以下核心研究发现：

**揭示了“形态—运动”特征纠缠是限制生成语义清晰度的根本原因。** 在对传统 VAE 模型的分析中，我们发现其潜空间存在显著的“后验坍塌”与“特征耦合”现象，导致模型无法独立控制手型与速度。通过消融实验证实，正交解耦 (Orthogonal Decoupling) 并非锦上添花，而是提升语义一致性的必要条件——当引入 HLC 双分支解耦后，互信息 (HSIC) 显著下降，直接带动语义指标 BLEU-4 提升了 1.2 个点。这一发现表明，结构化的动作表征是实现复杂手语语义映射的前提。

**证实了“显式相位调制”是打破“机械匀速”生成模式的关键机制。** 针对端到端模型生成的动作缺乏呼吸感的问题，研究发现单纯依赖数据驱动的回归模型倾向于生成平均化的速度曲线。通过 RTP 模块的对比实验，我们观测到动作的“自然顿挫感”与“相位变化率”存在强相关性 (Pearson 相关系数 0.82)。这证明了将抽象的语言韵律转化为显式的相位信号，是解决生成动作“节奏失真”的有效路径。

**验证了软物理约束在神经网络生成中的有效性与必要性。** 实验结果表明，在缺乏物理先验的情况下，神经网络极易生成违反人体解剖学的“反关节”动作。我们发现，通过引入基于骨长与角度的 KALS 软约束 (Soft Constraints)，模型能够在不破坏端到端可微性的前提下，自动学习到障碍规避与骨骼刚体规则。这一发现为解决神经网络生成中的“恐怖谷效应”提供了低成本的数学解决方案。

**证明了非自回归架构在“高保真—低延迟”权衡上的优越性。** 在与当前 SOTA 的扩散模型 (DiffSLP) 对比中，我们发现：虽然扩散模型在画质细腻度上略有优势，但其计算成本随迭代步数呈线性增长。本文方法通过层级码本与一次性

解码策略，证明了在不牺牲统计分布一致性（FGD 4.12）的情况下，可以将推理速度提升一个数量级（142 FPS vs 8 FPS）。这说明高效的离散表征学习是比暴力去噪更适合实时交互场景的技术路线。

## 5.2 局限性与未来展望

尽管本文在手语生成的逼真度与实时性方面取得了阶段性进展，但受限于数据资源与现有技术框架，研究工作仍存在一定的局限性。未来的研究可从以下几个维度进一步深化：

**从“骨架生成”向“高保真数字人驱动”演进。** 本文目前的输出仍为稀疏的 3D 骨架点。未来工作将探索将生成的骨架序列接入 SMPL-X 参数化模型或 3D Gaussian Splatting（高斯泼溅）渲染管线，实现包含精细手部纹理、面部微表情及衣物解算的写实级数字人驱动，进一步提升听障用户的视觉体验。

**引入“语音—手语”多模态对齐机制。** 目前的 RTP 模块主要依赖文本句法特征来推断节奏。然而在实际的新闻播报或会议场景中，手语往往伴随口语同步进行。未来可引入声学编码器（Audio Encoder），利用语音的音高（Pitch）、能量（Energy）及语速特征来辅助相位预测，实现口语韵律与手语节奏的跨模态精确同步。

**探索小样本与零样本（Zero-Shot）生成能力。** 受限于 CSL-Daily 的数据规模，模型在低频词汇上的泛化能力仍有待提升。未来可尝试结合大语言模型（LLM）的知识迁移能力，利用预训练大模型丰富的语义空间来指导手语潜空间的构建，探索在无标注或少标注场景下的零样本手语生成技术，以降低对昂贵手语数据的依赖。

**优化长序列生成的计算效率。** 虽然 HLC 模块比扩散模型快，但 Transformer 的  $O(L^2)$  复杂度限制了其在极长序列（如整场讲座）上的表现。未来可引入线性注意力机制（Linear Attention）或 Mamba（State Space Models）等线性复杂度架构，进一步降低长序列生成的显存占用与计算开销。

# 参考文献

- [1] STOLL S, CAMGOZ N C, HADFIELD S, et al. Text2Sign: towards sign language production using neural machine translation and generative adversarial networks [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 4776-4785.
- [2] CAMGOZ N C, SAUNDERS B, HADFIELD S, et al. Sign language transformers: joint end-to-end sign language recognition and translation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020: 10023-10033.
- [3] SAUNDERS B, CAMGOZ N C, BOWDEN R. Progressive transformers for end-to-end sign language production[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2020: 687-705.
- [4] HUANG J, MIN C, ZHOU L, et al. Towards fast and high-quality sign language production[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). 2022: 1323-1336.
- [5] XIE H, YIN Y, GUO D, et al. Vector quantized diffusion model for text-to-sign pose generation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2023: 14921-14931.
- [6] BALTATZIS V, EFTHIMIOU E, FOTINEA S E, et al. DiffSLP: diffusion-based sign language production[C]//Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). 2023: 1675-1688.
- [7] BALTATZIS V, POTAMIAS R A, VERVERAS E, et al. Neural sign actors: a diffusion model for 3D sign language production from text[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024: 1985-1995.
- [8] YIN A, LI H, SHEN K, et al. T2S-GPT: dynamic vector quantization for autoregressive sign language production from text[C/OL]//Proceedings of the 62nd

- Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers. 2024: 3345-3356. DOI: [10.18653/v1/2024.acl-long.183](https://doi.org/10.18653/v1/2024.acl-long.183).
- [9] MA J, WANG W, YANG Y, et al. MS2SL: multimodal spoken data-driven continuous sign language production[C/OL]//Findings of the Association for Computational Linguistics: ACL 2024. 2024: 7241-7254. DOI: [10.18653/v1/2024.findings-acl.432](https://doi.org/10.18653/v1/2024.findings-acl.432).
- [10] WALSH H, SAUNDERS B, BOWDEN R. Select and reorder: a novel approach for neural sign language production[C]//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024: 14531-14542.
- [11] GUO Z, HE Z, JIAO W, et al. Unsupervised sign language translation and generation[C/OL]//Findings of the Association for Computational Linguistics: ACL 2024. 2024: 14041-14055. DOI: [10.18653/v1/2024.findings-acl.835](https://doi.org/10.18653/v1/2024.findings-acl.835).
- [12] DONG L, WANG X, NWOGU I. Word-conditioned 3D american sign language motion generation[C/OL]//Findings of the Association for Computational Linguistics: EMNLP 2024. 2024: 9993-9999. DOI: [10.18653/v1/2024.findings-emnlp.584](https://doi.org/10.18653/v1/2024.findings-emnlp.584).
- [13] WALSH H, FISH E, SINCAN O M, et al. SLRTP2025 sign language production challenge: methodology, results, and future work[A/OL]. arXiv (2025). <https://arxiv.org/abs/2508.06951>. DOI: [10.48550/arXiv.2508.06951](https://doi.org/10.48550/arXiv.2508.06951).
- [14] HE J, WANG X, ZHANG R, et al. Text-driven diffusion model for sign language production[A/OL]. arXiv (2025). <https://arxiv.org/abs/2503.15914>. DOI: [10.48550/arXiv.2503.15914](https://doi.org/10.48550/arXiv.2503.15914).
- [15] TASYUREK S M, KIZILTEPE T, KELES H. Disentangle and regularize: sign language production with articulator-based disentanglement and channel-aware regularization[Z]. 2025.
- [16] WANG X, TANG S, CHENG L, et al. SignAligner: harmonizing complementary pose modalities for coherent sign language generation[Z]. 2025.
- [17] YE M, YE X, MANOHARAN M. Hybrid autoregressive-diffusion model for real-time streaming sign language production[Z]. 2025.

- [18] HE J, WANG X, TANG S, et al. Motion is the choreographer: learning latent pose dynamics for seamless sign language generation[Z]. 2025.
- [19] ZHANG H, SHALEV-ARKUSHIN R, BALATATZIS V, et al. Towards AI-driven sign language generation with non-manual markers[A/OL]. arXiv (2025). <https://arxiv.org/abs/2502.05661>. DOI: [10.48550/arXiv.2502.05661](https://doi.org/10.48550/arXiv.2502.05661).
- [20] FAURÉ G, SADEGHI M, BIGEARD S, et al. Towards skeletal and signer noise reduction in sign language production via quaternion based pose encoding and contrastive learning[A/OL]. arXiv (2025). <https://arxiv.org/abs/2508.14574>.
- [21] ŽELEZNÝ T, STRAKA J, JAVOREK V, et al. Exploring pose-based sign language translation: ablation studies and attention insights[Z]. 2025.
- [22] RASTGOO R. Sign language production: a review[C]//Proceedings of the CVPR Workshop on Sign Language. 2021.
- [23] RASTGOO R. A survey on recent advances in sign language production[J]. Elsevier Journal (to appear), 2024.
- [24] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems (NeurIPS). 2017.
- [25] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems (NeurIPS). 2014.
- [26] HO J, JAIN A, ABBEEL P. Denoising diffusion probabilistic models[C]//Advances in Neural Information Processing Systems (NeurIPS). 2020.
- [27] VAN DEN OORD A, VINYALS O, KAVUKCUOGLU K. Neural discrete representation learning[C]//Advances in Neural Information Processing Systems (NeurIPS). 2017.
- [28] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. Proceedings of NAACL-HLT, 2019.
- [29] LIU Y, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pre-training approach[A/OL]. arXiv (2019). <https://arxiv.org/abs/1907.11692>.
- [30] CHEN Z, ZHANG Y, WANG Y, et al. SoundStream: an efficient neural audio codec[A/OL]. arXiv (2021). <https://arxiv.org/abs/2107.03312>.

- [31] CAMGOZ N C, KOLLER O, HADFIELD S, et al. Multi-channel sign language recognition and translation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2018: 4848-4857.
- [32] DUARTE A, PALASKAR S, VENTURA L, et al. How2Sign: a large-scale multimodal dataset for continuous american sign language[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 2735-2744.
- [33] LI D, SELVARAJ S P, SHUKLA A, et al. Word-level deep sign language recognition from video: a new large-scale dataset and methods benchmark[C]// Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV). 2020: 1459-1469.
- [34] SAUNDERS B, CAMGOZ N C, BOWDEN R. Adversarial training for multi-channel sign language production[C]//Proceedings of the British Machine Vision Conference (BMVC). 2021.
- [35] MA X, JIN R, CHUNG T S. MCST-transformer for sign language production[C]// Proceedings of LREC-COLING. 2024.
- [36] ZHOU H, ZHOU W, QI W, et al. CSL-daily: a large-scale continuous chinese sign language dataset for daily communication[C]//Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). 2021: 1-6.
- [37] CVPR 2025 SLP Challenge Organizers. SLRTP2025 sign language production challenge[Z]. 2025.
- [38] GRETTON A, BOUSQUET O, SMOLA A, et al. Measuring statistical dependence with hilbert–schmidt norms[C]//Proceedings of the 16th International Conference on Algorithmic Learning Theory (ALT). 2005.
- [39] AO T, GAO Q, LOU Y, et al. Rhythmic gesticulator: rhythm aware co-speech gesture synthesis with hierarchical neural embeddings[A/OL]. arXiv (2022). [http://arxiv.org/abs/2210.01448](https://arxiv.org/abs/2210.01448).
- [40] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[C]// International Conference on Learning Representations (ICLR). 2019.
- [41] LUGARESI C, TANG J L, NASH A, et al. MediaPipe: a framework for building perception pipelines[A/OL]. arXiv (2019). <https://arxiv.org/abs/1906.08172>.

- [42] CAO Z, HIDALGO T, SIMON T, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.
- [43] SAVITZKY A, GOLAY M J E. Smoothing and differentiation of data by simplified least squares procedures[J]. Analytical Chemistry, 1964, 36(8): 1627-1639.



# 致谢

在论文完成过程中，我得到了许多老师、同学和家人的关心与帮助。在此谨向所有给予我指导、支持与鼓励的人表示诚挚感谢。

感谢实验室同学在数据处理、实验复现和论文讨论中的帮助。你们在研究方法、工程实现与问题定位方面提供了许多宝贵建议，使我能够不断完善研究思路与系统实现。

同时，感谢家人在学习和生活中的理解与支持。正是你们长期的鼓励与陪伴，让我能够专注于研究工作并顺利完成本论文。

最后，向所有参与实验评测与意见反馈的同学和志愿者致以诚挚谢意。



## 复旦大学 学位论文独创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。论文中除特别标注的内容外，不包含任何其他个人或机构已经发表或撰写过的研究成果。对本研究做出重要贡献的个人和集体，均已在论文中作了明确的声明并表示了谢意。本声明的法律结果由本人承担。

作者签名：\_\_\_\_\_ 日期：\_\_\_\_\_

## 复旦大学 学位论文使用授权声明

本人完全了解复旦大学有关收藏和利用博士、硕士学位论文的规定，即：学校有权收藏、使用并向国家有关部门或机构送交论文的印刷本和电子版本；允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存论文。涉密学位论文在解密后遵守此规定。

作者签名：\_\_\_\_\_ 导师签名：\_\_\_\_\_ 日期：\_\_\_\_\_