

基于多模态的手势识别与生成算法 及应用研究

(申请清华大学电子信息硕士专业学位论文)

培 养 单 位： 深圳国际研究生院

专 业 领 域： 互联网 + 创新设计

申 请 人： 房 丰 仪

指 导 教 师： 杨 文 明 副教授

二〇二五年三月

Research on Multimodal Gesture Recognition and Generation Algorithm and Application

Thesis submitted to
Tsinghua University
in partial fulfillment of the requirement
for the professional degree of
Master of Electronic and Information Engineering

by
Fang Fengyi
(Internet+Innovation Design)

Thesis Supervisor: Associate Professor Yang Wenming

March, 2025

学位论文指导小组、公开评阅人和答辩委员会名单

指导小组名单

李 XX	教授	清华大学
王 XX	副教授	清华大学
张 XX	助理教授	清华大学

公开评阅人名单

刘 XX	教授	清华大学
陈 XX	副教授	XXXX 大学
杨 XX	研究员	中国 XXXX 科学院 XXXXXXXX 研究所

答辩委员会名单

主席	赵 XX	教授	清华大学
委员	刘 XX	教授	清华大学
	杨 XX	研究员	中国 XXXX 科学院 XXXXXXX 研究所
	黄 XX	教授	XXXX 大学
	周 XX	副教授	XXXX 大学
秘书	吴 XX	助理研究员	清华大学

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；（3）按照上级教育主管部门督导、抽查等要求，报送相应的学位论文。

本人保证遵守上述规定。

作者签名： _____

导师签名： _____

日 期： _____

日 期： _____

摘 要

手势是人类交流信息和表达意图的重要方法^[1]。近年来,手势驱动的人机交互技术取得了显著的进展^[2-4],使得手势识别(HGR)与手势生成(HGG)成为研究的热点之一。手语是听障人士日常交流的重要媒介,其学习与普及在社会交往中具有重要意义。然而,标准手语的推广受限,手语学习难度较高,且优质教学资源匮乏。基于手势的人工智能技术有望为手语辅助教学问题提供新的解决方案。本文基于深度学习技术,研究构建了多模态手势识别与协同手势生成算法;基于此,围绕手语学习过程中的核心挑战,设计实现了一个交互式手语学习助手系统。主要创新贡献如下:

(1) 多模态手势识别算法创新:提出了一种可插拔的多策略解耦与语义集成网络(MDSI),通过“姿势-运动”与“时空-通道”特征解耦,有效降低 RGB-D 手势识别中的信息冗余,并结合语义滤波与标签平滑机制提升语义区分能力。实验结果表明,该方法在 IsoGD 和 THU-READ 数据集上分别超越现有最优方法 2.48% 和 4.33%,同时保持轻量化设计,其附加参数量仅占主干网络的 6.84%。

(2) 协同手势生成算法创新:提出了 CoordSpeaker,一种新颖的协同字幕赋能的同声手势生成方法,实现了手势运动生成中的节奏同步和语义对齐。通过首次引入手势描述生成模块,有效解决了手势数据缺乏文本标注的问题。结合统一的运动表示与可控潜在扩散模型,实现语义与节奏的精准协同控制。实验表明,该方法可生成高质量的(Jerk 0.179 \rightarrow)、语音同步(BC 0.057 \uparrow)、语义相关(MM-Dist: 6.814)的协同手势运动,显著领先同类方法。同时,将平均推理时间(AITS)降低至 0.842 秒,较现有方法加速 6 倍以上,极大提升了实际应用价值。

(3) 交互式手语学习系统构建:基于上述算法,设计并实现了一套集成实时识别、标准动作生成与交互反馈的手语学习系统。系统采用模块化架构,支持手语学习、练习评估和自由练习三种核心交互模式,并提供清晰的界面布局与友好的交互设计。在 RGB-D 相机上的实验验证表明,系统在 12 类数据上的识别准确率超过 99%,推理延迟小于 0.1 秒;用户研究表明,生成的手部动作自然度偏好较同类方法提升 4.65%,充分验证了所提算法的实用价值。

本文提出的手势识别与生成算法不仅显著提升了识别准确率与生成质量,同时也为解决手语教学资源匮乏问题、提升手语学习效率提供了新的技术路径。

关键词: 手势识别; 手势生成; 多模态; 手语学习; 交互式系统

Abstract

Gestures are a vital means of conveying information and expressing intentions in human communication^[1]. In recent years, gesture-based human-computer interaction technologies have achieved remarkable progress^[2-4], making gesture recognition (HGR) and gesture generation (HGG) prominent research topics. Sign language serves as a crucial medium for daily communication among the hearing-impaired population, and its learning and popularization hold significant social importance. However, the promotion of standardized sign language faces limitations due to high learning difficulty and scarcity of quality educational resources. Gesture-based artificial intelligence technology offers promising solutions for sign language assisted teaching. Based on deep learning techniques, this thesis develops multimodal gesture recognition and collaborative gesture generation algorithms. Building upon these, we design and implement an interactive sign language learning assistant system that addresses core challenges in the sign language learning process. The main innovative contributions are as follows:

(1) Innovation in multimodal gesture recognition: We propose a plug-and-play Multi-strategy Decoupling and Semantic Integration Network (MDSI), which effectively reduces information redundancy in RGB-D gesture recognition by introducing "pose-motion" and "spatiotemporal-channel" feature decoupling. Additionally, semantic filtering and label smoothing mechanisms enhance semantic distinction. Experimental results demonstrate that MDSI surpasses the state-of-the-art methods by 2.48% and 4.33% on the IsoGD and THU-READ datasets, respectively. Furthermore, the model maintains a lightweight design, with additional parameters accounting for only 6.84% of the backbone network.

(2) Innovation in collaborative gesture generation: We propose CoordSpeaker, a novel coordinated caption-empowered co-speech gesture generation approach, realizing both rhythmic synchronization and semantic alignment in speaker motion generation. By introducing a gesture caption generation module for the first time, we effectively address the absence of textual annotations in gesture datasets. By leveraging a unified motion representation and a controllable latent diffusion model, our approach achieves precise coordination of semantic consistency and rhythmic alignment. Experimental results indicate that the proposed method generates high-quality gestures that align well with speech

content and semantic intent, receiving significantly higher user preference than competing methods. Moreover, an optimized latent diffusion process reduces the average inference time (AITS) to 0.406 seconds, achieving a 6-fold speedup over existing approaches and substantially enhancing practical applicability.

(3) Development of an interactive sign language learning system: Based on the proposed algorithms, we design and implement an integrated system that supports real-time recognition, standard gesture generation, and interactive feedback. The system employs a modular architecture and offers three core interaction modes: sign language learning, practice evaluation, and free practice, with an intuitive interface layout and user-friendly interaction design. Experimental validation with an RGB-D camera shows that the system achieves a recognition accuracy exceeding 99% for 12 gesture classes, with an inference latency of less than 0.1 seconds. Additionally, the naturalness preference for generated gestures surpasses existing methods by 17.22%, demonstrating the practical value of the proposed algorithms.

The proposed gesture recognition and generation algorithms not only significantly improve recognition accuracy and generation quality but also provide a novel technical approach to addressing the shortage of sign language educational resources and enhancing learning efficiency.

Keywords: Gesture Recognition; Gesture Generation; Multimodal Learning; Sign Language Learning; Interactive System

目 录

摘 要.....	I
Abstract.....	II
目 录.....	IV
插图和附表清单.....	VII
第 1 章 绪论	1
1.1 研究背景.....	1
1.2 国内外研究现状综述.....	3
1.2.1 多模态动态手势识别.....	3
1.2.2 协同手势运动生成.....	8
1.2.3 基于手势的人机交互应用系统.....	10
1.3 研究内容	12
1.4 章节安排	13
第 2 章 相关技术	15
2.1 计算机视觉基础.....	15
2.1.1 卷积神经网络.....	15
2.1.2 3D 卷积神经网络	15
2.1.3 视觉变换器.....	15
2.2 生成模型基础.....	16
2.2.1 变分自编码器.....	16
2.2.2 扩散模型.....	17
第 3 章 基于多模态的动态手势识别算法研究	19
3.1 多策略手势特征解耦网络	19
3.1.1 姿态-运动解耦 (PMD)	20
3.1.2 空间-时间-通道解耦 (STCD)	21
3.1.3 双分支视频编码器.....	22
3.2 多模态手势语义集成网络	22
3.2.1 语义嵌入.....	22
3.2.2 语义滤波器 (SF).....	23
3.2.3 语义标签平滑 (SLS)	24
3.2.4 整体损失.....	24

3.3 实验结果与分析	24
3.3.1 实验设置	24
3.3.2 与最先进方法的比较	25
3.3.3 消融研究	28
3.4 本章小结	32
第 4 章 基于手势描述的协同手势生成算法研究	33
4.1 运动表示	33
4.1.1 统一运动表示	33
4.1.2 潜在表示	34
4.2 可控手势潜在扩散模型	34
4.2.1 潜在扩散过程	35
4.2.2 分层条件注入	35
4.2.3 无分类器指导	35
4.2.4 训练与推理	35
4.3 手势描述	36
4.4 多粒度描述控制	37
4.5 实验结果与分析	38
4.5.1 实验设置	38
4.5.2 协同手势生成	39
4.5.3 文本驱动的运动生成	40
4.5.4 手势描述生成	41
4.5.5 消融实验	42
4.5.6 更多可视化结果	44
4.6 本章小结	44
第 5 章 交互式手语学习助手设计与实现	47
5.1 系统总体设计	47
5.1.1 设计目标	47
5.1.2 系统架构	47
5.2 核心功能模块设计	48
5.2.1 手语识别模块	48
5.2.2 手语生成模块	48
5.2.3 交互反馈模块	49

5.3 人机交互设计	50
5.3.1 手语学习模式	50
5.3.2 用户界面设计	50
5.3.3 交互引导设计	50
5.4 系统实现与部署	50
5.4.1 开发环境与技术栈	50
5.4.2 关键技术实现	51
5.5 系统测试与评估	51
5.6 本章小结	53
第 6 章 总结与展望	54
6.1 工作总结	54
6.2 研究展望	55
参考文献	56
致 谢	65
声 明	66
个人简历、在学期间完成的相关学术成果	67
指导教师评语	68
答辩委员会决议书	69

插图和附表清单

图 1.1	不同的手势识别技术	1
图 1.2	RGB-D 手势识别的主要挑战可归因于两个因素：(i) 信息冗余 (IR) 存在于类内，尤其是背景、照明和视角。(ii) 信息缺失 (IA) 存在于类间，尤其是视觉上相似的手势。	2
图 1.3	基于视觉的手势分类法 ^[5]	4
图 1.4	基于手势识别的人机交互应用系统	11
图 1.5	增强现实环境下的实时翻译应用	12
图 2.1	2D 和 3D 卷积示意图 ^[54] 。2D 卷积在图像或视频帧上运算得到特征图，而 3D 卷积可在视频序列上同时提取时空特征。	15
图 2.2	Video Transformer 不同设计选择的可视化 ^[72] 。数据标记采用浅灰色（如果使用标记则采用黑色描边），而增强标记采用深灰色；白色标记是初始化的可学习标记；[CLS] 标记用“C”表示（增强后填充黑色）。从侧面流入 (T) 变压器的数据用于交叉注意。	16
图 3.1	我们的可插入式多策略语义集成解耦 (MDSI) 框架概述。MDSI 可以无缝集成到基本编码器 ϵ 中，从两个方面增强手势识别性能：i) 对于信息冗余，MDN (图 3.2) 通过 PMD 和 STCD 强调不同维度和尺度的特征信息 (图 3.3)。ii) 对于信息缺失，SIN 通过 SF 将自然语言建模与 SLS 一起集成到手势识别中 (图 3.4)。	19
图 3.2	多策略解耦网络 (MDN)。MDN 配置为双分支视频编码器。i) 全局分支 (\mathcal{G}) 将原始视频 \mathbf{V} 作为输入并对全局特征进行编码。ii) 解耦分支 (\mathcal{D}) 利用 PMD 模块同时捕获解耦的细粒度姿势 \mathbf{h} 和粗粒度手部运动 \mathbf{m} 。iii) STCD 模块插入编码器 $\epsilon_{\mathcal{G}}$ 和 $\epsilon_{\mathcal{D}}$ 的不同阶段，以执行与维度无关的解耦和注意。 ..	20
图 3.3	多策略解耦管道。(a) 姿势-运动解耦 (PMD)，(b) 空间-时间-通道解耦 (STCD)。	20
图 3.4	语义整合网络 (SIN) 结合了语义滤波器 (SF) 和语义标签平滑 (SLS)，以促进语义知识整合。	22
图 3.5	可视化所提方法的增强效果。(a) 基础模型的特征分布可视化。(b) 所提出的 MDSI 的特征分布可视化。(c) 基础模型的混淆矩阵。(d) 所提出的 MDSI 的混淆矩阵。	31

图 3.6	阐明所提出的 MDSI 在解决信息冗余 (IR) 和信息缺失 (IA) 挑战方面的增强功能。(a) IR 场景下的性能改进 (对应于图 1.2(a) 中的样本)。(b) IA 场景下的性能改进。(对应于图 1.2(b) 中的样本)	31
图 4.1	CoordSpeaker 支持 手势字幕和定制的 协同的说话者动作生成, 既能与字幕保持语义一致, 又能与音频保持节奏同步。例如, 在演讲场景中, 我们的方法允许说话者在讲话时自然地向前走并鞠躬, 无缝地做出结束手势。	33
图 4.2	协同手势生成模型概览。我们的条件潜在扩散模型 (第 4.2 节) 由两个关键组件组成: (1) 手势变分自动编码器, 可学习统一的低维潜在表示, 从而实现紧凑的跨数据集运动建模, 以及 (2) 分层控制的降噪器, 确保分层条件注入并在该学习到的潜在空间中有效运行。	34
图 4.3	手势描述生成框架。我们的手势描述生成框架包含两个主要组件: 运动分词器和运动感知语言模型。运动分词器将手势序列编码为离散的运动 token 序列, 运动感知语言模型则基于这些 token 和提示模板生成对应的手势描述。	36
图 4.4	协同手势生成的定性比较。红色框突出显示语义不一致, 黄色框表示不自然的动作, 绿色框表示协同良好的自然手势。	39
图 4.5	手势描述生成结果示例。生成的描述准确地描述了整体运动模式和细粒度的手势细节。	41
图 4.6	定性消融研究。结果是使用语音音频和单字幕 “ <i>A person is raising both hands up while talking</i> ” 生成的。	42
图 4.7	协同手势生成的更多视觉结果。	45
图 4.8	更多手势字幕结果。彩色框突出显示了手势和文本字幕之间的精确映射。	46
图 5.1	交互式手语学习助手系统架构	48
图 5.2	用户界面设计	49
图 5.3	交互引导设计	51
图 5.4	奥比中光 RGB-D 视觉传感器 Gemini 2 正视实物图。	51
图 5.5	商业 RGB-D 相机自采数据集样本示例	52
图 5.6	手语识别模块推理时间 (单位: ms)	52
表 3.1	与 IsoGD 数据集上最先进的方法的性能比较。最佳和第二佳方法通过 加粗 和 <u>下划线</u> 标注。	26

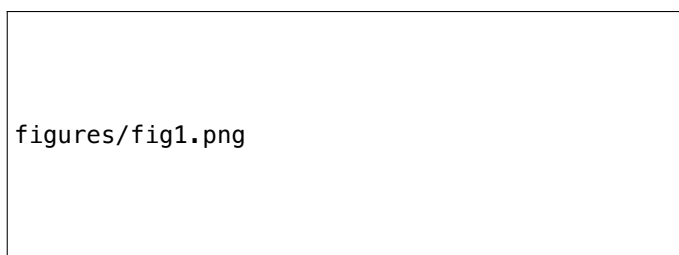
表 3.2	与 THU-READ 数据集上最先进的方法的性能比较。最佳和第二佳方法通过 加粗 和 <u>下划线</u> 标注。	27
表 3.3	MDN 各个子分支的性能比较。“Trans.”表示 Transformer。	28
表 3.4	THU-READ(CS4) 上 STCD 模块的消融研究。“Trans.”表示 Transformer。	29
表 3.5	SIN 组件的消融研究：语义过滤器 (SF) 和语义标签平滑 (SLS)。SF _{<i>n</i>} 表示使用 <i>n</i> 个语义过滤器混合一个视觉特征 (SF ₀ 表示视觉过滤器是随机初始化的，没有集成语义过滤器)。	29
表 3.6	每个子模块的可训练参数，其中 MDSI 的总参数被 加粗 。“Trans.”表示 Transformer。	30
表 4.1	手势描述框架中使用的提示模板示例。	37
表 4.2	与基线模型和消融研究进行比较的定量结果。‘→’表示越接近真实运动越好。每个指标均在 20 次运行的 95% 置信区间下报告。我们报告 BC ×10 ⁻¹ 和 Top-1 R-Precision。	40
表 4.3	与 HumanML3D ^[93] 测试集上的最新方法进行比较。我们按照 ^[97] 计算标准度量。‘→’表示越接近真实运动越好。每个度量均在 20 次运行的 95% 置信区间下报告。	41
表 4.4	多粒度字幕策略的消融研究。“Reg.”表示常规字幕策略，“Dyn.”表示动态字幕策略，“Hie.”表示分层字幕策略。每个指标均在 20 次运行的 95% 置信区间下报告。我们报告 BC ×10 ⁻¹ 和 Top-1 R-Precision。	43
表 5.1	手部动作生成结果的用户偏好胜率 (%)。结果表明我们生成的结果被认为更加真实和可控，在自然度和匹配度方面分别优于之前的工作 ^[31] 4.65% 和 1.87%。	53

第1章 绪论

1.1 研究背景

手势是人类传递信息、表达意图的重要方式之一，具有灵活性强、信息传递效率高等优点^[1]。随着计算机技术的快速发展，基于手势的人机交互技术已在多个领域取得了显著进展^[2-4]，这使得手势识别（HGR）与手势生成（HGG）成为研究的热点之一。全球约有 4.66 亿听力受损人群，手语作为听障人士之间最主要的沟通媒介，在其日常生活和社会交往中发挥着不可替代的作用。然而，由于标准手语的普及程度较低、手语学习难度较高、教学资源严重匮乏，听障人士与普通人与人之间往往存在较大的交流障碍。基于手势的深度学习技术的发展为解决手语教学问题提供了新的可能性。通过将手势识别（HGR）与手势生成（HGG）算法同人机交互技术相结合，有望解决手语辅助教学中的手语评估、动作生成等关键挑战，为手语学习者提供更加自然、便捷和高效的学习体验。

手势识别是指对人类手势进行跟踪、识别其含义，并将其转化为具有语义意义的指令的过程^[5]。图 1.1 展示了不同的手势识别技术。基于视觉的手势识别是指利用相机等视觉传感设备捕获手势的形状动作，并利用计算机视觉等技术对 2D/3D 手势进行识别，具有用户友好、设备易得的优势。深度神经网络^[6-7] 的进步显著增强了识别能力，通过改进的网络架构^[8-10]、多模态融合^[7,11-12] 和数据增强^[7,13]，在标准场景中实现了显著的准确性。尽管取得了这些成就，但识别 RGB-D 视频中的手势仍面临相当大的挑战，包括不同的光照、不同的背景、表演者的外观差异以及视觉上相似的手势。



(a) 基于触觉的手势识别: Cyber Glove II^[14]; (b) 基于计算机视觉的手势识别: SoftKinetic HD 相机^[15]。

图 1.1 不同的手势识别技术

我们观察到，RGB-D 手势识别的主要挑战可以归因于两个因素：(i) 信息冗余 (IR)。在纠缠的时空空间中，冗余信息很难处理^[6,16]。利用耦合建模结构的模型通

figures/IR.pdf

(a) 同一手势类别具有不同的背景、照明和视角。

figures/IA.pdf

(b) 不同的手势类别具有视觉上相似的代表。

图 1.2 RGB-D 手势识别的主要挑战可归因于两个因素: (i) **信息冗余 (IR)** 存在于类内, 尤其是背景、照明和视角。(ii) **信息缺失 (IA)** 存在于类间, 尤其是视觉上相似的手势。

常会在训练期间学习背景、照明和表演者外观等不相关的特征^[6], 这可能会导致误导性分类(图 1.2(a))。尽管现有方法在训练数据集上表现良好, 但在未见的场景中, 它们的准确性会显著下降。这种差异凸显了一个关键的泛化问题, 即模型在消除冗余信息方面遇到挑战, 阻碍了与任务相关的细微手势特征提取。(ii) **信息缺失 (IA)**。模型难以区分具有高视觉相似性的手势(图 1.2(b))。虽然现有方法建议加入额外的线索, 如姿势^[13,17]和光流^[11]来增强手势识别, 但这些方法的有效性仍然依赖于视觉, 并且受到运动模糊和透视变化等问题的严重影响, 特别是在区分视觉相似的手势时。鉴于这些挑战, 有必要开发一种有效的方法, 最大限度地减少不相关的信息冗余, 并解决整个手势识别过程中基本信息的缺失。

手势生成因其在人机交互领域的广泛应用而备受关注, 如虚拟现实、游戏和数字虚拟人等。为了增强手势生成的多样性和可控性, 研究人员探索了多种模态, 包括语音^[18-20]、文本转录^[21-23]、情感^[24-25]、风格^[18,26-27]以及说话者身份^[28]等。其中, 语音同步性和语义相关性是两个备受关注的方面。尽管深度神经网络^[20,23,29]的进步显著提高了生成质量和多样性, 但当前的方法主要关注自发性的伴随语音手势, 而忽视了文本驱动的非自发性手势^[30]。这通常导致虚拟人动作的模糊性^[31], 并限制了通过文本提示控制生成动作的灵活性, 阻碍了手势生成算法的实际应用。

一个关键挑战在于现有手势数据集缺乏直接的描述性文本标注：虽然语音数据自然且直接可用，但语义相关性通常仅通过语音转录推断，导致语义关联较弱。此外，手动标注手势语义成本过高，这阻碍了高质量生成和精细化的说话者语义控制。虽然一些方法尝试通过引入额外的动作数据集进行联合训练来缓解这一问题^[30]，但它们仍然存在手势数据的语义差距，只能在两个任务之间切换而非实现联合控制。另一种潜在的解决方案是通过动作-文本对齐预训练构建对齐的嵌入空间，以获取手势的隐式文本标签^[31]，但这会引入额外的训练和推理成本。此外，由于人体动作数据集和手势数据集之间存在显著的分布差异^[31]，联合嵌入空间的泛化能力仍有待验证。鉴于这些挑战，有必要开发一种有效的方法，既能解决手势数据标注缺失问题，又能实现多模态信号的协同生成控制，同时将成本降至最低。

基于上述背景，本文致力于提出一种新颖的多模态手势识别与手势生成算法，并基于此开发一个交互式手语学习助手系统。首先，通过引入一种多策略解耦与语义集成手势识别网络，实现手势识别的准确性提升；其次，通过提出一种描述驱动的协同手势生成网络，增强手势生成的描述性控制；最后，通过在手语学习系统中集成手势识别和生成算法，实现手语动作的智能评估和标准动作库的动态扩充生成。该系统能够有效缓解手语教学资源匮乏的问题，提升自主手语学习的效率和质量，还可以为听障人士的日常交流提供有力的技术支持，具有重要的社会价值和广阔的应用前景。

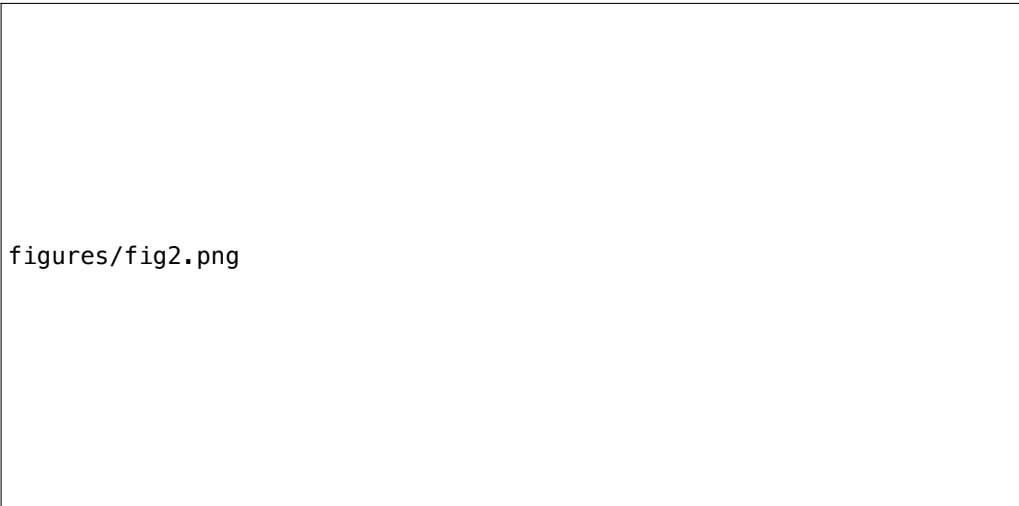
1.2 国内外研究现状综述

1.2.1 多模态动态手势识别

1.2.1.1 手势识别概述

手势是人类传递信息、表达意图的重要途径，文献^[5,32]将手势分类为两种类型：静态手势和动态手势。静态手势被定义为在任何时间内没有在空间中的方向和位置运动，如果上述时间持续时间有运动，则称为动态手势。动态手势又包括五种类型^[33]：标志，情感展示，调节器，适配器和解释器。图 1.3 显示了手势类别的分类学。由于动态手势中往往包含更丰富的语义表示，近年来更多的研究围绕动态手势展开。

手势识别是一个完整的过程，涉及跟踪手势、识别其含义并将其转化为有意义的指令^[5]。手势识别的主要方法可以分为基于触觉的方法和基于视觉的方法^[5,34]。基于触觉的方式是通过数据手套等设备检测手部动作、弯曲时的物理反应，收集相关数据并通过计算机或微控制器进行处理识别的方法。常用的接触式传感方法



figures/fig2.png

图 1.3 基于视觉的手势分类法^[5]

包括：数据手套、手部光学标记、加速度计、触摸屏、表面肌电等^[34]。尽管基于数据手套的方法具有可穿戴、灵活便捷等优势，但它们也有许多局限。例如：基于接触的设备并没有给用户太多的可接受性^[5]，并且不适合老年人；电线连接会影响设备的灵活性；长时间使用也受到设备寿命的限制；此外，一些传感器可能相当昂贵^[34]。视觉手势识别是通过相机等视觉传感器捕捉手势的形状和动作，并运用计算机视觉技术对 2D 和 3D 手势进行识别的过程。常用的视觉相机设备包括 2D 摄像头、Kinect、LeapMotion、Time of Flight (ToF) 相机等^[35]。尽管视觉手势识别容易受到遮挡、光照等问题的干扰，但由于其具有用户友好、设备易得的优势，因此视觉手势识别成为近年来研究的主流方向。

本研究将深入探讨基于视觉的动态手势识别，探讨多模态识别过程中的特征解耦与融合，通过引入先进的深度学习技术进一步提升识别的准确性和鲁棒性。

1.2.1.2 基于手工提取特征的手势识别方法

早期研究使用手工提取特征进行手势识别，包括检测、追踪和识别三个阶段^[5]。检测的目的是进行手的检测和相应图像区域的分割，常基于肤色^[36]、形状^[37]、3D 模型^[38]、运动^[39]、骨架^[40]等特征进行手部检测。追踪技术旨在获取手部在连续帧中的位置信息，构建动态轨迹特征。通过对这些包含手势基本信息的轨迹进行分析，可实现对手势类型的识别与理解。基于模板的方法 (Template based)^[41]与手部检测方法非常相似。此类方法在前一帧中检测到手的空间附近调用手检测器，从而在很大程度上限制图像搜索空间，但要求图像具有足够的采样帧率。基于最佳估计的方法 (Optimal estimation)^[42]采用卡尔曼滤波器^[43]提供的最优估计框架具有实时的性能并能为连续帧提供预测。基于粒子过滤 (Particle filtering) 的方法^[44]被

用来在密集的视觉混乱中跟踪手的位置和手指的配置，此类方法用一组粒子建模手的位置，对于复杂的模型需要更多的粒子。基于连续自适应均值偏移 (CamShift) 的方法^[45]参考了基于核的 mean shift 算法原理，通过迭代搜索找到帧序列中与其样本模式最相似的分布模式来跟踪目标，并能够在跟踪过程中自适应调整跟踪窗口的大小和目标的分布模式。CamShift 具有轻量、鲁棒、高效的优点，但在复杂场景中容易失败、且容易发生窗口漂移问题。识别阶段的目标是对手的位置、姿势提供最终的语义解释。静态手势识别可以使用模板匹配或基于机器学习的简单分类器^[35]，如：支持向量机 (SVM)^[46]、随机森林^[35]、K 近邻算法 (KNN)^[47]等。动态手势识别则需要对时间维度进行建模，隐马尔科夫模型 (HMM)^[48]，动态时间规整 (DTW)^[49]，时延神经网络^[36]等算法在动态手势识别系统中得到了广泛的应用。传统手势识别方法的关键在于如何合理地提取手工特征，具有计算成本低、速度快的优势。但由于过于依赖于技术人员经验与精巧的模型设计，算法的泛化性、鲁棒性往往较差。

1.2.1.3 基于深度学习的手势识别方法

伴随着计算能力的提升以及大规模多模态数据的积累，基于深度学习的手势识别技术实现了重大突破。与传统算法相比，深度学习不仅减少了对手工工程的需求，而且提升了算法的准确性、适用性与效率。因此，基于深度学习的手势识别方法已经成为该领域的主流研究方向，受到广泛关注。

卷积神经网络 (CNN) 已被证明是许多计算机视觉任务中的有效算法，2DCNN 方法被广泛用于处理静态手势图像的分类和识别任务。这些网络通过卷积层和池化层有效地捕获图像中的空间特征。针对视频中的动态手势动作，早期研究者试图在时域中扩展 2DCNN 的连接性以进行视频分类^[50]。Simonyan 等人^[51]提出了双流神经网络 (Two Stream CNN) 来分别学习空间和时间特征，其中空间流处理静态单帧图像，时间流则处理多帧光流。TSN^[52]在双流卷积神经网络的基础上进行稀疏时间采样，所提出的时间段网络可以对整个动作视频进行高效的学习。TRN^[53]进一步改进了 TSN 中的融合方式，并提出了时间维度上的多尺度特征融合，以提升算法的鲁棒性。2DCNN 结构简单，但存在时域信息建模的不足；而双流方法依赖于预先提取的密集光流进行运动表示，网络同样缺乏有效的时域特征提取模块。

受图像识别领域卷积神经网络突破的启发，3DCNN 被提出并广泛应用于视频理解任务。C3D 网络^[54]通过在时空维度上扩展二维卷积核，实现了对视频序列中时空特征的直接提取，该网络在多种视频分析任务中展现出优异的性能。此后，一系列基于 3DCNN 的神经网络被提出，并逐渐成为视频理解任务中的主流选择。Carreira 等人^[55]结合了双流神经网络与 3DCNN 的优势，引入了一种新的双流膨胀

3D ConvNet (I3D) 进行时空建模, 在 Kinetics 动作视频数据集上实现了最先进的水平。R2+1D^[56] 通过将 3D 卷积分解为 2D 空间卷积和 1D 时间卷积的方式, 实现了更高效的时空特征提取。S3D^[57] 考虑速度与准确度之间的平衡, 将 I3D 与时空分离卷积结合, 并引入一种特征门控机制以进一步提升网络的识别准确率。Zhang 等人提出了 Deformable 3DCNN^[58], 通过所设计的轻量级时空可变形卷积模块, 通过根据前序特征图学习额外的偏移量来增强 3D 卷积的时空采样位置, 以减轻手势识别中背景的干扰。3DCNN 能够有效提取视频中的局部时空特征, 但对长序列数据的建模能力较弱。此外, 当网络输入较长时, 往往需要更多层或更大的内核和步幅大小, 增加了网络计算成本的同时也容易出现过拟合现象。

循环神经网络(RNN)^[59]和长短期记忆网络(LSTM)^[60]是处理序列数据的常用模型。这一特点促使研究人员将 CNN 和 RNN/LSTM 的优势结合起来, 先后学习局部和全局时空特征。长期循环卷积网络 (LRCN)^[61]依次使用卷积神经网络 (CNN) 和长短期记忆 (LSTM) 网络学习空间和时间特征, 多提出的方法在动作识别、图像描述与视频描述多任务中取得了有竞争力的结果。Molchanov^[62]和 Cao^[63]等人分别将 3DCNN 与循环神经网络 (RNN) 和长短期记忆网络 (LSTM) 相结合以进行端到端的手势识别, 前者可以从多模态数据中同时检测和分类动态手势, 后者则额外设计了一个具有相邻时间片之间循环连接的时空转换模块, 可以在空间和时间维度上主动地将三维特征图转换为规范视图。

由于全连接长短期记忆网络 (FC-LSTM) 难以提取数据中的空间特征, Shi 等人^[64]提出了卷积长短期记忆网络 (ConvLSTM) 来处理连续图像以进行临近降水预报。研究表明, 相比 LSTM, ConvLSTM 能更有效地捕获时空特征, 并已在动作识别、手势识别等多个视觉任务中得到成功应用。Zhu 等人^[65]提出了一种基于 3D 卷积和 ConvLSTM 的孤立手势识别方法。该方法首先利用 3DCNN 提取短期时空特征, 随后通过 ConvLSTM 建模长期时空依赖关系, 并采用 SPP^[66]对特征进行归一化。该方法分别在 RGB 和深度模态上训练, 最终通过平均融合得到预测结果。在此基础上, Zhang 等人^[67]提出了 3DCNN+BiConvLSTM+2DCNN 网络用于孤立手势识别。具体地, 文章引入了双向卷积长短期记忆网络 (BiConvLSTM) 以替换网络中的 ConvLSTM, 并引入额外的 2DCNN 学习更高层次的时空特征。AttnConvLSTM^[68]与 ConvLSTMForGR^[8]进一步探索了 ConvLSTM 中空间卷积的冗余性以及注意力机制的影响。研究发现, 门控结构中的空间卷积对特征融合的作用有限, 同时在输入输出门中引入注意力机制也未能有效提升特征融合效果。在此基础上衍生出一种新的 GatedConvLSTM 变体, 其中卷积结构仅嵌入到 ConvLSTM 的输入到状态转换中。对于未经分割的手势视频, Zhu 等人^[9]进一步将类似网络

应用于连续手势识别，并提出了一种两阶段的连续手势识别方法。在分割阶段，使用时间扩张三维卷积神经网络将连续手势序列分割成孤立的手势实例；在识别阶段，则使用基于 3DCNN+ConvLSTM+2DCNN 的孤立手势识别网络进行识别分类。基于循环神经网络的动态手势识别方法能够有效进行长期时间建模。然而由于手势视频中不同阶段的帧往往具有不同的重要性^[69]，此类方法因为缺少时域上的注意机制，难以对不同重要性的视频帧进行有效建模。另外，这类网络的计算效率也受到限制，因为每个时间步的计算都需要等待前一步的结果。

Transformer 架构^[70]可以通过注意机制更好地捕获上下文信息。Vision Transformer (ViT)^[71]是一项开创性的工作，通过将视觉数据作为序列处理，而不是依赖于传统的卷积网络，引入了图像分类的新范式转变。受 ViT 的启发，Video Transformer (VT)^[72]通过引入额外的时间维度进行时空建模与视频理解。TimeSformer^[73]率先使用纯 Transformer 结构用于视频识别。ViViT^[74]探索了几种不同的时空注意力分解变体，验证了时空编码器分离方法的有效性。VTN^[75]同样采用时空编码器分离的方式，并进一步引入 Longformer 实现复杂度为 $O(n)$ 的长序列建模。该机制通过滑动窗口实现局部上下文的自注意力计算，并结合任务相关的全局注意力机制。Video Transformer (VT) 方法已经在行为识别等视频理解领域取得了显著的成功，然而目前却少有研究将 Video Transformer 方法应用于手势识别领域来解决动态手势的分类问题。这可能是由于 Video Transformer 在计算复杂性方面面临着挑战，通常需要大量的计算资源和大规模的数据集来实现最佳性能^[76]。

与以前利用耦合建模结构来处理纠结的跨模态特征的方法相比，本研究侧重于两个解耦维度中的多模态特征的相互作用，并且可以无缝集成到各种 RGB-D 方法中。

1.2.1.4 解耦手势特征学习

由于复合手势特征在纠缠空间^[6,16]中难以处理，一些研究人员尝试从各个角度解耦复杂手势特征学习，以提高识别性能。一些工作^[10,16]尝试将“时空”手势特征空间解耦。Zhou 等人^[6]提出了一种解耦时空表示学习网络 (DSN, DTN) 来学习特定于维度的表示。然而，明确地隔离整个网络的时空维度可能会破坏特征的时空连通性，从而损害模型捕获隐式关节时空信息的能力。其他一些研究侧重于将 3D 手势表示与 3D 骨架或点云数据解耦。Guo 等人^[77]强调，可以从两个角度看到手部骨架：显式关节云和隐式骨架拓扑。Liu 等人^[78]提出，各种手势类别在对不同尺度特征的依赖方面表现出多样性。Bigalke 等人^[79]提出了一种解耦的双流模型，用于从 3D 点云序列中独立学习局部姿势特征和全局运动特征。然而，这种方法受到高数据收集成本的限制，对实际应用提出了挑战。本研究试图同时将

RGB-D 手势与两个范例分离：i) 姿势-运动解耦 (PMD) 和 ii) 空间-时间-通道解耦 (STCD)。这种方法有效地将细微特征集成到两个子空间中，同时保留了联合特征，使其有别于其他方法。

1.2.1.5 多模态视觉-语言模型

随着多模态领域的蓬勃发展，大量视觉语言模型^[80-81]应运而生，并被应用于各种任务，例如，图像文本匹配和图像文本检索。CLIP^[80]是该领域的开创性工作之一，利用大规模图像文本对进行训练，共同优化文本编码器和视觉编码器，以实现图像和语义特征对齐。鉴于 CLIP 依赖于大规模预训练，一些研究人员旨在基于 CLIP 进行微调，以实现高效的少样本迁移和领域泛化^[82]。Zuo 等人^[13]避免使用密集文本编码器，而是选择使用 fastText 来提取词级嵌入，以利用注释（符号标签）中包含的隐性知识。手势本身具有自然的语义属性；然而，使用简单的词级嵌入不足以捕捉手势注释的复杂性。在这一努力中，我们首先将语义引入手势识别，利用预先训练的 CLIP 和适配器从手势注释线索中提取更丰富的语义，并促进语义视觉特征的充分交互，这与以前的视觉语言模型不同。

1.2.2 协同手势运动生成

1.2.2.1 手势生成概述

手势生成任务是从多模态数据中生成手势动作序列的复杂任务。为了实现更多样、可控的协同语音手势生成，多模态数据被辅助建模，包括语音音频^[18-20]、文本转录^[21-23]、情感^[24-25]、风格^[18,26-27]、说话人 ID^[28]。其中，语音同步性和语义相关性是两个关键方面。由于语音数据的自然可获得性，协同语音手势生成一直是主要的研究重点。然而，由于手势数据集缺乏直接的描述性文本标注，语义相关性通常只能通过语音转录来推断，这限制了语义关联的强度。Yang 等人^[30]提出了一种协同语音和文本驱动的方法，但仍然受到标注不足的限制，且无法提供基于两种信号的协同生成。Chen 等人^[31]利用提示-动作对齐预训练来生成隐式文本标签，但引入了额外的训练和推理成本。与现有方法不同，我们提出了一种基于描述的手势潜在扩散模型，在保持低标注和计算成本的同时，实现了语义相关性和节奏一致性。

为了解决这些限制，我们提出了一种联合描述和语音驱动的协同手势生成框架，在实现生成手势的语义相关性和节奏一致性的同时，保持低标注和计算成本。

1.2.2.2 基于深度学习的手势生成方法

传统方法依赖基于规则或统计模型来学习语音-手势映射，而深度学习方法利用神经网络来建模音频和手势之间的复杂关系^[83]。

早期方法主要采用 RNN^[84-85] 进行手势序列建模。Liu 等人^[84] 提出了一个大规模身体-表情-音频-文本数据集 BEAT，它具有 76 小时的高质量多模态数据。基于此他们构建了一个级联运动网络 (CaMN)，它由上述六种模态组成，这些模态以级联架构构建，用于手势合成。Yoon 等人^[85] 引入了一种基于多模态上下文的手势生成方法，通过融合语音、文本和说话者特征来生成自然的手势动作。但基于 RNN 的方法在捕获长期依赖性和计算效率方面存在诸多局限。随后，Transformer^[21-22] 被引入以通过自注意力机制更好地建模时序关系。Zhi 等人^[21] 引入了 LivelySpeaker，这是一个实现语义感知协同语音手势生成的框架，并提供了多个控制句柄。该方法任务分为两个阶段：基于脚本的手势生成和音频引导的节奏优化。Bodyformer^[22] 提出了一个变分转换器来有效地模拟手势的概率分布，这可以在推理过程中产生不同的手势，并引入了一个模式位置嵌入层和一个模态内预训练方案来捕捉不同说话模式下的不同运动速度并缓解数据的稀缺性，以从有限的中学习语音和 3D 手势之间的复杂映射。然而，Transformer 首先于其相对于序列长度的二次方复杂度，带来了显著的计算挑战。近期，MambaTalk^[20] 首次将选择性状态空间模型 (SSM) 引入手势合成领域，通过线性时间计算实现了高效的序列建模。该方法实现了具有离散运动先验的两阶段建模策略，以提高手势的质量。利用基础 Mamba 块，通过多模态集成增强手势多样性和节奏。

为了提升生成质量和可控性，研究者开始探索基于潜变量的生成方法。一些方法采用 VQ-VAE 架构对手势动作进行离散编码，实现更高质量的手势生成。EMAGE^[23] 利用四个合成 VQ-VAE 与掩码身体手势先验来提高结果的保真度和多样性，该研究同时引入了 BEAT2 (BEAT-SMPLX-FLAME)，这是一个新的网格级整体共语音数据集，优化了头部、颈部和手指运动的建模，提供了社区标准化的高质量 3D 动作捕捉数据。QPGesture^[86] 提出了一个手势 VQ-VAE 模块来学习 codebook 来总结有意义的手势单元。每个代码代表一个独特的手势，有效缓解了随机抖动问题。TalkSHOW^[87] 提出了一个新颖的语音到动作生成框架，其中分别对面部、身体和手进行建模，结合 VQ-VAE 和跨条件自回归模型 (Gated PixelCNN) 产生连贯且一致逼真的运动。

由于其强大的生成能力，扩散模型^[18,29,88-89] 近期被广泛应用于运动和手势生成，显著增强了生成输出的多样性。SIGGesture^[29] 通过大规模预训练扩散模型的语义注入进行泛化协同语音手势合成。该研究收集了一个名为 LSMoG 的大规模手

势数据集进行预训练,时间长达 400 小时。在运动生成领域, MotionDiffuse^[90] 代表了第一个基于文本的运动扩散模型,该模型提供了对身体部位的细粒度指令,并实现了随时间变化的文本提示的任意长度的运动合成。MDM^[89] 引入了一种基于原始运动数据的运动扩散模型,实现了高质量的生成和通用的条件化,共同构成了新的运动生成任务的良好基线。Yang 等人^[18] 使用使用扩散模型生成风格化音频驱动的协同语音手势,可以生成自然、语音匹配、风格一致的手势。然而,现有基于扩散的方法通常直接在动作空间进行建模采样,导致计算开销大且训练不稳定。相比之下,潜在扩散模型 (LDMs)^[91] 通过在低维潜空间进行操作,在保持强大生成能力的同时实现了计算效率。

与现有方法不同,我们提出了一种基于描述增强的手势潜在扩散模型,该模型在保持低标注和计算成本的同时,实现了语义相关性和节奏一致性的双重目标。

1.2.2.3 动作-文本转换

人类动作展现出类似于自然语言的语义耦合,常被视为一种肢体语言形式^[92]。以往关于人体运动的研究探索了各种与文本相关的任务,包括文本到动作生成^[89,93]、动作到文本描述^[92,94] 以及统一的动作-语言建模^[92,95-96]。近期的文本到动作工作 (MDM^[89]、MLD^[97]、MotionLCM^[98]) 利用预训练语言模型提取语义信息来控制动作生成。动作描述旨在使用自然语言描述人类动作,早期方法依赖统计模型和 RNN 来学习动作到语言的映射^[99-100]。近年来,双向动作-文本转换受到了更多关注。TM2T^[94] 通过标记化首次实现了动作与文本的双向生成,尽管仅限于单一统一框架。随着大语言模型的发展,统一的动作-语言模型 (如 MotionGPT^[92]、MotionChain^[96]、M3GPT^[101]) 通过融合语言数据和大规模动作模型而涌现。

尽管手势本身具有自然的语义属性,但目前尚未有研究探索手势理解和描述生成。为了解决这些局限性,本文首先探索了一种手势描述生成方法来填补手势数据中文本标注的空白,并开发了一种多粒度描述控制机制,实现对非自发手势的精确语义控制。

1.2.3 基于手势的人机交互应用系统

手势作为人与人之间交流的自然媒介形式,是人机交互最适合的形式之一^[34],基于手势的人机交互技术已经被广泛应用在各个领域。例如:

(1) 机器人控制。基于手势的人机交互,可被广泛用于机器人导航控制等领域^[102]。国内无人机制造商大疆公司所发布的晓 Spark 的无人机可以识别用户手势,完成上下起降、左右移动和拍照等基本操作,极大的提高了用户体验 (图 1.4(a))。

(2) 车载手势。Alba-Castro 等基于手势识别技术,构建了一个基于手势操作的

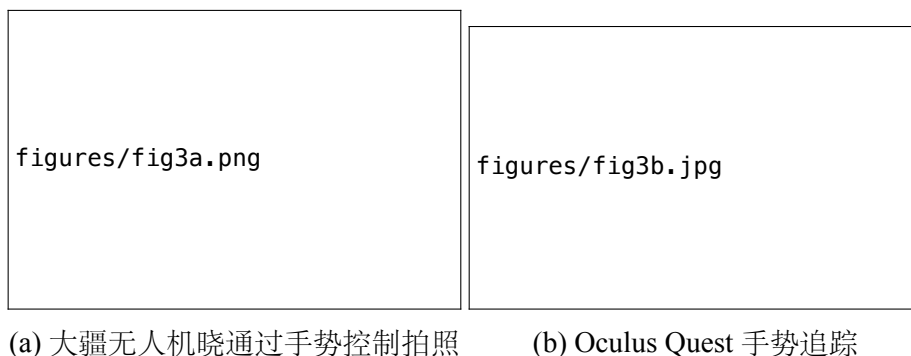


图 1.4 基于手势识别的人机交互应用系统

车载娱乐媒体设备控制系统^[103]。Manawadu 等基于人车接口，利用手势识别的技术实现了自动驾驶汽车的经纬度控制^[104]。

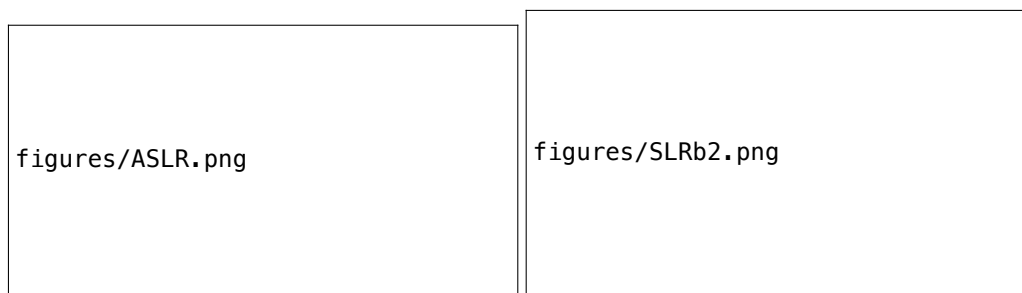
(3) 智能家居。手势识别技术在家庭自动化领域具有广泛应用，可实现对照明系统、通风设备、视听设备等智能家电的便捷操控。三星集团在 2012 年推出的智能电视机，能够通过捕捉用户的手势动作对电视实现远程控制。Desai 等人的研究表明，相关技术可以用来改善老年人的生活质量^[3]。

(4) 手语翻译。聋哑人由于生理限制，无法使用声音进行交流，手势和手语成为了他们最依赖的沟通形式。然而，手语翻译是在实时翻译领域容易被忽略的交流形式^[105]。近年来，手势识别在手语翻译系统中的应用研究得到了越来越多的关注^[2]。纽约大学的 Li 等人研发了一款名为“ASLR”的 AR 实时手语翻译应用原型^[105]。该应用将计算机视觉与 AR 技术相结合，能够捕捉摄像头前执行的特定手语手势，并以用户的母语提供实时翻译（如图 1.5(a) 所示）。然而，由于该原型设计用于手机，用户在执行手语的过程中需要反复拿起和放下手机，增加了使用的负担。Will 等人构建了一个 AR 眼镜上的实时语音翻译系统原型^[106]。该系统可以聆听语音，将其翻译成 37 种语言之一，并将生成的文本直接显示在用户眼镜上作为字幕。如图 1.5(b) 所示，用户可以在 AR 眼镜上享受实时的字幕翻译，同时自然地进行下棋等活动。这表明在 AR 眼镜上部署实时翻译应用是一种自然而灵活的方式。

(5) 虚拟现实交互。手势能够使用户与虚拟场景的交互更加自然方便，因此被广泛应用于增强现实、虚拟现实等环境中（图 1.4(b)）。Taylor 等人使用手势识别系统对采集到的手势进行实时匹配以弹奏虚拟钢琴^[107]。

(6) 游戏娱乐。手势识别技术在游戏领域（Kinect、Xbox）得到了广泛应用。例如，通过摄像头传感器捕捉玩家的手部和身体动作，实现与游戏的自然交互，提升游戏体验。

(7) 临床与健康。手势识别技术可以帮助医生在手术室内实现非接触式操控医



(a) ASLR: 使用手机进行 AR 实时手语翻译的应用原型 (b) AR 眼镜上的实时语音翻译应用

图 1.5 增强现实环境下的实时翻译应用

疗设备和影像系统^[4]。同时，该技术也被应用于辅助行动不便人士的康复训练和日常生活，例如通过手势来控制智能轮椅^[108]。

(8) 计算机交互。手势交互作为一种创新的人机界面形式，能够取代鼠标键盘等传统设备，为用户提供更加直观的界面操作体验^[109]。

本研究将聚焦动态手势识别与手势生成算法在手语教学中的应用。基于构建的交互式手语学习助手系统，实现手语学习、辅助练习、实时反馈等功能。这不仅能够有效缓解手语教育师资与教学资源短缺的难题，还能提升手语教学中的人机交互体验，具有广泛的社会价值与应用潜力。

1.3 研究内容

本文针对多模态动态手势识别算法与协同手势生成算法进行了研究，并基于此构建了一个交互式手语学习助手。具体研究内容如下：

多模态手势识别算法研究 针对动态手势识别任务中存在的“信息冗余”与“信息缺失”挑战，本研究提出了一种用于 RGB-D 手势识别的新型可插拔方法，称为多策略解耦和语义集成手势识别网络 (Multi-strategy Decoupling with Semantic Integration Network, MDSI)。首先，为了缓解 IR，我们引入了多策略解耦网络 (MDN)。该网络通过多种策略巧妙地解耦了纠缠的手势特征：a) 姿势-运动解耦 (PMD) 和 b) 空间-时间-通道解耦 (STCD)。PMD 将手势视频解耦为细粒度 姿势和粗粒度 运动，由冻结的预训练姿势估计器^[110] 支持。STCD 采用与维度无关的自注意力技术，从而有效地消除了多余的信息并增强了微妙但关键的特征。其次，为了解决 IA，我们将自然语言建模集成到手势识别中，开发了语义整合网络 (SIN)。SIN 由两个核心组件支撑：语义过滤器 (SF) 和 语义标签平滑 (SLS)。SF 采用跨模态混合过滤将语义信息集成到视觉建模中，从而提高了模型在潜在特征空间中辨别细微差别的

能力。SLS 利用标签的语义相似性在训练阶段深化模型的语义理解。此过程利用强大的预训练 CLIP 文本编码器来提取深度语义特征，编码器在训练期间冻结，不会产生额外的计算成本。据我们所知，我们是第一个将多模态语义信息引入该领域的，标志着一项有希望的探索。我们在基准 IsoGD 和 THU-READ 数据集上进行的大量实验结果表明，所提出的方法可以显著提高基于 RGB-D 的手势识别性能。

协同手势生成算法研究 针对手势数据的描述文本缺失和多模态协同控制困难的挑战，我们提出了一种新颖的手势描述与生成 (CoordSpeaker) 方法，通过手势到描述的转换填补手势语义标注的空白，并通过描述增强的伴随语音手势生成实现对手势生成的精确语义控制。首先，为应对协同多模态控制生成的挑战，我们开发了一个手势潜在扩散模型，该模型包含一个手势变分自编码器 (VAE) 用于学习统一的动作潜在表示，以及一个具有层次控制去噪器的潜在扩散模型，用于实现精细的多条件控制。其次，为缓解手势数据标注缺失的问题，我们引入了一个手势描述框架，该框架利用动作-语言模型以低成本为手势数据生成描述性文本，并通过多粒度描述控制机制实现精确的语义控制。据我们所知，这项工作是首次探索解决手势语义标注挑战的手势描述方法，从而实现了对手势生成的有效描述和语音控制，为手势-文本双向转换提供了新的视角。大量实验表明，我们的方法能够有效生成描述性手势描述，并实现语义连贯、节奏同步的手势生成。

交互式手语学习助手 进一步地，本文将深入探讨算法在人机交互中的应用前景。基于所提出的 MDSI 动态手势识别算法与 CoordSpeaker 协同手势生成算法构建一个交互式手语辅助学习系统，结合多模态信息融合技术和实时动作捕捉机制，实现精准、高效的手势识别响应，与流畅、自然的手势动作生成。这不仅有助于改善提升手语教学过程中的人机交互体验，缓解手语教育师资与教学资源短缺的难题^[11]，更可以充分发挥人工智能手势技术在人机交互环境中的潜力，具有广泛的社会意义与应用价值。

1.4 章节安排

本文共分六章，主要内容组织如下：

第一章为绪论，分析了研究背景及意义，介绍了国内外研究现状，并阐述了本文的研究内容和章节安排。

第二章为相关技术，介绍了本文所涉及的计算机视觉基础知识和生成模型基础理论，包括 3D 卷积神经网络、视觉变换器、变分自编码器和扩散模型等关键技

术。

第三章为多模态手势识别算法研究，提出了一种可插拔的多策略解耦和语义集成网络 (MDSI)，通过解耦视觉特征提取与多模态语义集成，有效解决了 RGB-D 手势识别中的信息冗余和信息缺失的挑战。

第四章为协同手势生成算法研究，提出了一种基于描述驱动的协同手势生成框架 (CoordSpeaker)，通过引入可控潜在扩散模型与多粒度手势描述策略，实现了手势生成过程中语义和节奏的协同精确控制。

第五章为交互式手语学习助手设计与实现，基于前述算法，构建了一个可交互的手语辅助学习系统，实现了手语学习、辅助练习、实时反馈等功能。

第六章对本研究的主要工作进行归纳，阐述了研究的创新贡献，并探讨了后续的研究方向。

第 2 章 相关技术

2.1 计算机视觉基础

2.1.1 卷积神经网络

卷积神经网络 (CNN) 是深度学习中处理网格结构数据的重要模型。它利用局部感受野、参数共享和降采样机制来提取图像特征。CNN 主要由三类层次构成: 卷积层负责特征提取, 池化层实现特征压缩, 全连接层完成特征映射。这种结构设计使 CNN 在计算机视觉领域获得了广泛应用。

2.1.2 3D 卷积神经网络

相比于处理单帧图像的 2D CNN, 3D CNN 在时间维度上引入了额外的卷积操作, 能够同时对视频数据的空间和时间特征进行建模。如图 2.1 所示, 3D 卷积层使用三维卷积核在输入体积上进行滑动, 提取时空联合特征。这种结构设计让 3DCNN 可以有效提取视频中的时序动态特征, 为视频分析和理解提供了重要的技术支持。在 3D CNN 的发展进程中, C3D^[54] 是一个重要的里程碑。该模型首次将 3D 卷积应用于大规模视频分类任务, 通过端到端的训练方式学习视频的时空特征表示。

figures/3dcnn.png

图 2.1 2D 和 3D 卷积示意图^[54]。2D 卷积在图像或视频帧上运算得到特征图, 而 3D 卷积可在视频序列上同时提取时空特征。

2.1.3 视觉变换器

随着 Transformer 在自然语言处理领域取得巨大成功, 研究者们开始探索将其应用于计算机视觉任务。视觉变换器 (Vision Transformer, ViT) 是这一探索的代表性成果。不同于 CNN 基于局部感受野的特征提取方式, ViT 将输入图像分割成固定大小的图像块 (patches), 并将这些图像块序列化后输入 Transformer 进行处理。通过自注意力机制, ViT 能够建模图像块之间的全局依赖关系, 为视觉特征提取提供了新的范式。

鉴于视觉 Transformer 在图像领域的成功, 一些工作将其拓展到视频领域, 提出了视频变换器 (Video Transformer)。典型的 Video Transformer 结构包括时空分

离 (Spatiotemporal Separated) 和时空联合 (Joint Spatiotemporal) 两种策略。前者如 TimeSformer^[73], 通过独立的时间和空间注意力进行计算, 减少计算复杂度; 后者如 ViViT^[74], 直接在 3D 令牌 (Token) 上施加全局注意力, 以获得更强的时空建模能力。视频变换器在动作识别、视频理解等任务上已经展现出强大的性能。

figures/vt.png

图 2.2 Video Transformer 不同设计选择的可视化^[72]。数据标记采用浅灰色 (如果使用标记则采用黑色描边), 而增强标记采用深灰色; 白色标记是初始化的可学习标记; [CLS] 标记用“C”表示 (增强后填充黑色)。从侧面流入 (T) 变压器的数据用于交叉注意。

2.2 生成模型基础

2.2.1 变分自编码器

变分自编码器 (VAE)^[112] 是一类通过概率建模实现数据生成的深度学习模型。与传统自编码器不同, VAE 的核心思想是对潜在变量进行概率建模, 学习数据的概率分布, 而不是直接学习数据到数据的映射。这种基于概率的建模方式使得 VAE 能够生成多样化的样本, 并且具有良好的插值性质。

VAE 假设观测数据由某个潜在变量生成, 其生成过程可表示为 $p_{\theta}(\mathbf{x} | \mathbf{z})$, 其中 \mathbf{z} 服从某一先验分布 (通常为标准正态分布)。然而, 直接计算后验分布 $p_{\theta}(\mathbf{z} | \mathbf{x})$ 通常是不可行的, 因此 VAE 采用变分推断, 引入一个近似分布 $q_{\phi}(\mathbf{z} | \mathbf{x})$ 来进行估计。模型的优化目标是最大化数据的对数似然, 由于难以直接计算, 通常转而最大化其证据下界 (ELBO):

$$\log p_{\theta}(\mathbf{x}) \geq \mathbb{E}_{q_{\phi}(\mathbf{z} | \mathbf{x})} [\log p_{\theta}(\mathbf{x} | \mathbf{z})] - D_{\text{KL}}(q_{\phi}(\mathbf{z} | \mathbf{x}) \parallel p(\mathbf{z})). \quad (2.1)$$

其中, 第一项鼓励模型在给定 \mathbf{z} 的情况下能够准确重构数据, 第二项则约束潜在变量的分布, 使其接近先验分布, 以保证生成的合理性。

在实际实现中, VAE 采用神经网络参数化编码器和解码器, 其中编码器将输入数据映射为潜在变量的均值和方差, 并通过重参数化技巧 $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$ (其中 $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$) 实现可微分的采样。最终, 通过梯度优化训练, 使 VAE 既能够进行数据重构, 又能生成高质量的新样本。

2.2.2 扩散模型

扩散模型 (Diffusion Model) ^[113] 是一类生成模型, 通过逐步去噪的方式学习从噪声分布到数据分布的映射。该模型由前向 (扩散) 过程和反向 (去噪) 过程组成, 其中前向过程逐步向数据添加噪声, 而反向过程学习从噪声中逐步恢复数据分布, 实现样本生成。扩散模型的一个重要优势是其生成结果多样, 生成质量高。通过精心设计的噪声调度, 模型能够逐步学习数据分布的不同尺度特征。

2.2.2.1 扩散模型的前向与反向过程

前向过程通过在 T 个时间步内逐步向数据样本 $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ 添加高斯噪声, 形成数据序列 $\{\mathbf{x}_t\}_{t=1}^T$, 其定义如下:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}), \quad (2.2)$$

其中, $\beta_t \in (0, 1)$ 为控制噪声水平的方差调度参数。通过重参数化技巧, 该分布可直接表示为 \mathbf{x}_0 的函数:

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I}), \quad (2.3)$$

其中 $\bar{\alpha}_t = \prod_{i=1}^t (1 - \beta_i)$ 。此公式允许从任意中间时间步高效采样。

反向过程旨在建模后验分布 $q(\mathbf{x}_{t-1} | \mathbf{x}_t)$, 但该分布通常难以直接求解。为此, 使用可学习的神经网络 p_θ 进行近似:

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)). \quad (2.4)$$

在大多数实现中, 协方差 $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)$ 可以固定或学习, 而均值 $\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$ 通常由神经网络建模。

训练过程中, 模型学习预测加入的噪声 ϵ , 训练目标是最小化重建误差:

$$\mathcal{L} = \mathbb{E}_{t, \mathbf{x}_0, \epsilon} [\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t)\|_2^2] \quad (2.5)$$

在推理阶段, 模型通过从标准高斯分布 $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 逐步去噪, 最终生成符合数据分布的样本。通过这种设计, 扩散模型能够在保证生成质量的同时, 实现稳定的训练过程。

2.2.2.2 条件扩散模型

扩散模型通过在去噪过程中引入条件信息, 可以实现对生成过程的精确控制。目前条件控制生成主要有两种范式: 基于分类器引导的事后修改 (Classifier-Guidance) ^[114] 和无分类器引导的事前训练 (Classifier-Free) ^[115]。

基于分类器引导的方法 (Classifier-Guidance) 首先训练一个无条件扩散模型,

然后利用额外的分类器在推理阶段调整生成过程以实现条件控制。这种方法的优点是训练成本较低，可以复用已有的预训练模型；但其缺点是推理计算开销大，且对生成结果的细节控制能力有限。此外，分类器的质量也会直接影响生成效果。

无分类器引导的方法（**Classifier-Free-Guidance**）则直接在扩散模型的训练过程中加入条件信号，使模型能够在生成过程中自然地融入条件信息，而无需依赖显式的分类器。这种方法能够实现更精细的条件控制，生成质量也更有保障。但其主要缺点是需要较大的训练开销，对计算资源和训练数据的要求较高。在实际应用中，需要根据具体场景的资源约束和性能需求来选择合适的条件控制策略。

第3章 基于多模态的动态手势识别算法研究

针对动态手势识别中“信息冗余 (IR)”和“信息缺失 (IA)”的挑战，如图 3.1 所示，我们提出了一种用于 RGB-D 手势识别的新型可插拔方法，称为多策略解耦和语义集成网络 (MDSI)。所提出的可插拔式 MDSI 方法的总体框架如图 3.1 所示。为了解决信息冗余 (IR) 问题，我们引入了多策略解耦网络 (MDN)，通过“姿势-运动”和“空间-时间-通道”解耦来强调不同尺度与不同维度的特征信息 (第 3.1 节)。为了解决信息缺失 (IA) 问题，我们提出了一个语义集成网络 (SIN)，它有两个关键组件：语义滤波器 (SF) 和语义标签平滑 (SLS)，以补充手势识别中缺失的信息 (第 3.2 节)。

3.1 多策略手势特征解耦网络

在本节中，我们介绍了一种多策略解耦网络 (MDN)，它配置为双分支架构，包括姿势-运动解耦模块 (PMD) 和空间-时间-通道解耦模块 (STCD)。如图 3.2 所示，我们首先利用 PMD 将手势视频 \mathbf{V} 解耦为细粒度姿势 \mathbf{h} 和粗粒度运动 \mathbf{m} 。原始视频 \mathbf{V} 以及解耦结果 (\mathbf{h}, \mathbf{m}) 通常分别输入到 MDN 的全局分支 (\mathcal{G}) 和分解分支 (\mathcal{D}) 进行视觉时空特征提取。此外，可插入 STCD 插入编码器 $\epsilon_{\mathcal{G}}$ 和 $\epsilon_{\mathcal{D}}$ 的各个阶段，以实现有效的与维度无关的解耦和注意。

figures/MDSI_v1.2.pdf

图 3.1 我们的可插入式多策略语义集成解耦 (MDSI) 框架概述。MDSI 可以无缝集成到基本编码器 ϵ 中，从两个方面增强手势识别性能：i) 对于信息冗余，MDN (图 3.2) 通过 PMD 和 STCD 强调不同维度和尺度的特征信息 (图 3.3)。ii) 对于信息缺失，SIN 通过 SF 将自然语言建模与 SLS 一起集成到手势识别中 (图 3.4)。

figures/MDN_v1.pdf

图 3.2 多策略解耦网络 (MDN)。MDN 配置为双分支视频编码器。i) 全局分支 (\mathcal{G}) 将原始视频 \mathbf{V} 作为输入并对全局特征进行编码。ii) 解耦分支 (\mathcal{D}) 利用 PMD 模块同时捕获解耦的细粒度姿势 \mathbf{h} 和粗粒度手部运动 \mathbf{m} 。iii) STCD 模块插入编码器 $\epsilon_{\mathcal{G}}$ 和 $\epsilon_{\mathcal{D}}$ 的不同阶段，以执行与维度无关的解耦和注意。

figures/PMD.pdf

figures/STCDv1 .pdf

(a)

(b)

图 3.3 多策略解耦管道。(a) 姿态-运动解耦 (PMD)，(b) 空间-时间-通道解耦 (STCD)。

3.1.1 姿态-运动解耦 (PMD)

如图 3.3(a) 所示，PMD 主要包括三个阶段：i) 姿势估计、ii) 深度校准和 iii) 姿势-运动解耦。对于 RGB 数据，我们首先使用在 COCO-WholeBody 数据集^[116]上预训练的姿势估计器 HRNet^[110]来提取并保存每帧 2D 身体骨架^①。对于深度数据，我们提前进行视频校准以确保配准^[120]。具体来说，对于 IsoGD 数据集^[17]，我们利用 Liu 等人提供的对齐深度视频数据^[120]，并进一步调整 Narayana 等人提供的深度数据映射^[11]，使用以下变换： $D' = \frac{D-7}{0.93} - 14$ 。对于 THU-READ 数据集，深度数据已提前校准。对齐良好的深度数据可以采用从 RGB 视频中提取的相同骨架进行后续操作。

我们利用提取的骨架数据将原始视频 \mathbf{V} 分解为手部姿势 \mathbf{h} 和手部运动 \mathbf{m} 。首先，给定第 i 帧的骨架数据 ($i \in 0, 1, \dots, T$)，我们分别过滤每只手的 22 个关键点并

① 请注意，姿势估计器在训练期间处于冻结状态，确保它不会引入任何额外的计算成本。此外，使用手部检测器检测和分割手部是现场的常见做法^[117-119]，这是一个完全公平的设置。

描绘出最小手部边界框。随后，我们裁剪缩放后的边界框（比例=1.2）以获得表示细粒度姿势 \mathbf{h}^i 的裁剪手部图像。为了公式化粗粒度运动向量 $\mathbf{m}^i \in \mathbb{R}^8$ ，我们将边界框的8个位置属性堆叠如下： $\mathbf{m}^i = [e^i, x_{min}^i, y_{min}^i, x_{max}^i, y_{max}^i, w^i, h^i, r^i]^\top$ ，其中 $e^i \in \{0, 1\}$ 表示相应的手是否出现在第 i 帧， $\{x_{min}^i, y_{min}^i\}$ 和 $\{x_{max}^i, y_{max}^i\}$ 为边界框左上角和右下角顶点的坐标， w^i, h^i 分别对应宽和高， $r^i = w^i/h^i$ 表示长宽比。为了区分左右手，上述操作分别在每只手上独立执行，然后连接起来，得到最终的解耦特征： $\mathbf{h}^i = [\mathbf{h}_l^i, \mathbf{h}_r^i]$ ， $\mathbf{m}^i = [\mathbf{m}_l^i, \mathbf{m}_r^i]$ 。具体来说， \mathbf{h} 和 \mathbf{m} 分别在宽度方向和通道方向连接。

此外，为了减轻姿势估计不准确的影响，我们设计了一个分割验证机制：检测到的关键点置信度低且关键点数量较少的骨架被标记为错误检测，并且相应的姿势特征 \mathbf{h}^i 被分配零值。

3.1.2 空间-时间-通道解耦 (STCD)

鉴于手势识别中存在紧密耦合的时空冗余^[6]，我们引入了 STCD（图 3.3(b)），它配置了三个包含注意机制的独立于维度的模块：i) 空间解耦模块 (SDM)、ii) 时间解耦模块 (TDM) 和 iii) 通道解耦模块 (CDM)。以 ε 中提取的输入中间特征图 $\mathbf{f}_{in} \in \mathbb{R}^{C \times T \times H \times W}$ 作为输入，该框架帮助网络有效滤除各个维度上的冗余信息，输出优化的特征图 $\mathbf{f}_{out} \in \mathbb{R}^{C \times T \times H \times W}$ ，从而帮助训练。

以 SDM 为例，其目的是识别每帧手势特征中的关键空间区域。首先，给定 \mathbf{f}_{in} ，我们使用 Conv3d 来导出查询、键和值特征，分别表示为 \mathbf{Q}^S 、 \mathbf{K}^S 和 \mathbf{V}^S 。我们合并时间和通道维度以获得空间查询、键和值向量： $\mathbf{q}^S \in \mathbb{R}^{HW \times CT}$ 、 $\mathbf{k}^S \in \mathbb{R}^{HW \times CT}$ 、 $\mathbf{v}^S \in \mathbb{R}^{CT \times HW}$ 。然后，利用 \mathbf{q}^S 和 \mathbf{k}^S 计算注意矩阵 $\mathbf{A}^S \in \mathbb{R}^{HW \times HW}$ ，如下所示：

$$\mathbf{A}^S = \text{softmax} \left(\mathbf{q}^S (\mathbf{k}^S)^\top \right). \quad (3.1)$$

随后，空间注意力向量 $\mathbf{f}^S \in \mathbb{R}^{CT \times HW}$ 可以通过以下公式计算：

$$\mathbf{f}^S = \mathbf{v}^S \mathbf{A}^S. \quad (3.2)$$

我们恢复时间和通道维度以获得空间注意特征图 $\mathbf{f}'^S \in \mathbb{R}^{C \times T \times H \times W}$ 。此外，为了促进网络收敛并调节网络注意力水平，我们引入了可学习的残差连接。最终的空间优化特征 \mathbf{f}_{out}^S 可以按如下方式计算： $\mathbf{f}_{out}^S = \gamma^S \mathbf{f}'^S + \mathbf{f}_{in}$ ，其中 γ^S 是 SDM 的可学习残差权重。

TDM 和 CDM 分别关注手势特征中的关键帧和通道，其操作与 SDM 类似。此外，STCD 的结果输出被表达为三个解耦模块输出的聚合。具体而言，我们设计了

figures/SIGRv1 .pdf

图 3.4 语义整合网络 (SIN) 结合了语义滤波器 (SF) 和语义标签平滑 (SLS)，以促进语义知识整合。

一个并行范式 (图 3.3(b)) STCD 的结果输出 f_{out} 可以表示如下: $\mathbf{f}_{out} = \mathbf{f}_{out}^S + \mathbf{f}_{out}^T + \mathbf{f}_{out}^C$ 。

3.1.3 双分支视频编码器

如图 3.2 所示, MDN 配置有两个分支: i) 全局分支 (\mathcal{G}) 将原始视频 \mathbf{V} 作为特征提取和识别的输入; ii) 解耦分支 (\mathcal{D}) 利用解耦后的特征 (\mathbf{h}, \mathbf{m}) 作为输入, 同时捕获细粒度的手势姿势变化和粗粒度的手部运动。我们为每个分支配置一个视频编码器 ϵ 。此外, 我们在 \mathcal{D} 分支中引入了一个长短期记忆 (LSTM) 模型, 用于运动建模 (图 3.2)。特征提取过程可以表述如下:

$$\begin{aligned}\mathbf{f}_v^{\mathcal{G}} &= \Phi_{\mathcal{G}}(\mathbf{V}), \\ \mathbf{f}_v^{\mathcal{D}} &= \Phi_{\mathcal{D}}(\mathbf{h}, \mathbf{m}).\end{aligned}\tag{3.3}$$

其中 $\Phi_{\mathcal{G}}$ 表示 \mathcal{G} 分支网络, $\Phi_{\mathcal{D}}$ 表示 \mathcal{D} 分支网络。得到的视觉特征 $\mathbf{f}_v^{\mathcal{G}}, \mathbf{f}_v^{\mathcal{D}}$ 将分别输入到下面的 SIN 网络中进行语义集成。

3.2 多模态手势语义集成网络

手势识别中的信息缺失挑战 (第 1 节) 激励我们结合自然语义信息作为指导。所提出的语义集成网络 (SIN) (如图 3.4 所示) 由两个分支组成: i) *semantic* 分支, 以语义嵌入和语义滤波器 (SF) 为特色; ii) *vision* 分支, 以语义标签平滑 (SLS) 为辅助。

3.2.1 语义嵌入

如图 3.4 所示, 鉴于手势数据集^[62,121-122] 通常使用类索引作为标签, 我们首先生成与手势标签相对应的语义提示。具体来说, 我们编译语义注释并设计提示

(prompt), 以强调每个手势数据集的独特特征。然后, 我们利用预训练的 CLIP 的文本编码器^[80] 和适配器^[82] 生成和细化深度语义嵌入 $\mathbf{f}_s \in \mathbb{R}^{N \times 1024}$ (N 表示类数)。

3.2.2 语义滤波器 (SF)

从动态卷积技术中汲取灵感^[123-124], 我们引入了语义滤波器 (SF), 用于通过卷积混合多模态特征。

我们引入一个线性 *Filter Learner* (图 3.4) 来导出语义滤波器。设语义滤波器的数量为 n , 以相应的 n 个语义嵌入 $\mathbf{f}_s \in \mathbb{R}^{n \times 1024}$ 为输入, 通过线性映射输出语义滤波器参数 $\mathbf{f}'_s \in \mathbb{R}^{n \times N_p}$, 然后将这些参数重塑为 n 个语义滤波器 $\boldsymbol{\Theta}_s \in \mathbb{R}^{n \times c \times k^{d_k}}$, 其中每个滤波器 $\boldsymbol{\Theta}_s^i$ 代表一个卷积核。第 i 个滤波器 $\boldsymbol{\Theta}_s^i$ 的参数数量定义为: $N_p = c \times k^{d_k}$, 其中 c 表示视觉特征 \mathbf{f}_v 的输入通道 (Eq.3.3), k 表示滤波器核大小 (经验上设置为 3), d_k 表示卷积维数 (例如, 对于 3d 卷积核, 设置 $d_k = 3$)。

随后, 我们将每个 $\boldsymbol{\Theta}_s^i$ (对应第 a 个手势类别) 作为过滤核, 将视觉特征 \mathbf{f}_v (对应第 b 个手势类别) 作为输入, 执行深度卷积。得到的每个混合语义特征 \mathbf{f}_{SF}^i 可以表示为:

$$\mathbf{f}_{SF}^i = \text{Convolution}(\mathbf{f}_v, \boldsymbol{\Theta}_s^i), \quad i \in 1 \cdots n. \quad (3.4)$$

这些混合语义特征被进一步激活和扁平化, 通过全连接 (FC) 层生成语义分支识别结果。

为了监督语义分支的优化, 我们将^[13] 每个语义标签 a 与样本视觉标签 b 混合。具体来说, 语义分支的混合标签 $\mathbf{y}_{SF}^i \in \mathbb{R}^N$ 可以表示为:

$$y_{SF}^i[k] = \begin{cases} 1 & \text{如果 } k = a = b, \\ 0.5 & \text{如果 } k = a \text{ 或 } k = b, a \neq b, \\ 0 & \text{否则,} \end{cases} \quad (3.5)$$

其中 $y_{SF}^i[k]$ 表示 \mathbf{y}_{SF}^i 的第 k 个条目。我们利用交叉熵损失来优化语义滤波器:

$$L_s^i = -\frac{1}{N} \sum_{k=1}^N y_{SF}^i[k] \log(\hat{y}^i[k]), \quad (3.6)$$

同样, 我们根据选定的 n 个语义滤波器形成 n 个混合标签, 并获得相应的 n 个语义预测。整体语义损失 L_s 是 n 个交叉熵损失的平均值: $L_s = \frac{1}{n} \sum_{i=1}^n L_s^i$ 。

如图 3.4 所示, 为了捕获 SF 中嵌入的语义知识以进行推理, 我们引入了一个与视觉滤波器 $\boldsymbol{\Theta}_v$ 类似的卷积核, 遵循我们的视觉网络 MDN, 并将语义分支的知识集成到视觉分支中。具体而言, $\boldsymbol{\Theta}_v$ 和 FC_v 分别通过其参数的加权和以及相应的 $\boldsymbol{\Theta}_s$ 和 FC_s 的参数进行更新。

3.2.3 语义标签平滑 (SLS)

与普通标签平滑^[125]相比,我们利用手势的语义相似性来生成有偏平滑标签,从而增强视觉分支识别视觉相似手势的能力。具体来说,我们首先通过 CLIP 的文本编码器获取语义嵌入 $\mathbf{f}_s \in \mathbb{R}^{N \times 1024}$, 然后使用余弦相似度^[13]构建提示相似度表 \mathbf{S} : $\mathbf{S} = \|\mathbf{f}_s\|_2 \|\mathbf{f}_s\|_2^\top \in \mathbb{R}^{N \times N}$ 。对于第 b 个手势类别的训练样本, 建议的平滑标签 y_{SLS} 可以表示为:

$$y_{SLS}[k] = \begin{cases} 1 - \sigma & \text{如果 } k = b, \\ \sigma \times \text{softmax}(\mathbf{S}[b][k]/\tau) & \text{否则,} \end{cases} \quad (3.7)$$

其中 $\sigma = 0.2$, $\tau = 0.5$ 。类似地, 交叉熵用于计算视觉分支损失:

$$L_v = -\frac{1}{N} \sum_{k=1}^N y_{SLS}[k] \log(\hat{y}[k]). \quad (3.8)$$

3.2.4 整体损失

整体损失是 L_s 和 L_v 的加权和:

$$Loss = w_s \times L_s + w_v \times L_v. \quad (3.9)$$

请注意, 本文仅将 SIN 网络应用于每个全局视觉特征 f_v^G (Eq. 3.3), 以确保视觉特征和语义特征之间交互的完整性。因此, 本文采用常用的分数融合方法来结合每个分支和模态的预测。

3.3 实验结果与分析

3.3.1 实验设置

3.3.1.1 数据集

我们在两个公共 RGB-D 手势数据集上评估了我们的方法: IsoGD^[17] 和 THU-READ^[122]。它们都被广泛使用, 并且包含具有挑战性的场景。IsoGD 数据集^[17]收录了 47933 段 RGB-D 手势视频, 由 21 位受试者完成 249 类手势动作。本文图 1.2 展示了该数据集的示例样本。此数据集中样本的背景多样性和手势相似性使其成为评估我们的方法对信息冗余 (IR) 和信息缺失 (IA) 影响的良好基准。THU-READ 数据集^[122]包含 1920 个 RGB-D 自我中心手势视频, 包含 8 个受试者执行的 40 个不同动作, 由于细微的类内差异和背景噪音, 这仍然具有挑战性。为了进行评估, 我们采用提供的留一法交叉验证协议^[7]。

3.3.1.2 编码器骨干

我们实现了两个版本的 MDSI: *MDSI-CNN* 和 *MDSI-Transformer*, 分别将我们的可插拔方法集成到基于 CNN 的视频编码器^[9] 和基于 Transformer 的视频编码器^[10] 中。这突出了 MDSI 的可插拔特性, 使其能够集成到不同的编码器架构中。我们将这两个版本与最先进的方法进行比较, 以验证我们的方法。

3.3.1.3 实现细节

我们使用 Adam 作为优化器。训练时, 每个视频随机抽取 32 帧连续序列。原始输入序列随机裁剪至 224×224 像素, 解耦的手部序列调整至 112×112 像素。数据增强包括空间裁剪、旋转、翻转和 ShuffleMix^[6]。与^[6,10,12,118] 类似, 我们所有的实验都是在 20BN Jester V1 数据集上进行预训练的^[126]。在推理过程中, 输入序列被中心裁剪为 224×224 。我们根据视觉特征形状分别将 MDSI-CNN 和 MDSI-Transformer 的 SF 中的卷积维度 d_k 设置为 3 和 1。除非另有说明, 所有消融研究均在 IsoGD 数据集的 RGB 数据上进行。

3.3.2 与最先进方法的比较

我们将我们的 MDSI 方法与 IsoGD 和 THU-READ 数据集上的其他最佳方法进行了比较。需要注意的是, 这里报告了每种方法的最佳性能。

IsoGD 上的性能 表 3.1 显示了在 IsoGD 数据集上与最先进方法的性能比较。我们的 MDSI-CNN 和 MDSI-Transformer 在所有当前方法中取得了最佳和第二好的性能, 与最先进方法相比分别提高了 2.48% 和 2.12%, 表明取得了显著的进步。除了在 RGB-D 融合结果中的强劲表现外, MDSI-CNN 和 MDSI-Transformer 在每个单一模态中也都排名前两位。具体而言, MDSI-CNN 在 RGB 数据中表现出色, 超过 SOTA 9.11%, 而 MDSI-Transformer 在深度数据方面表现出色, 超过 SOTA 3.38%。这一额外的发现表明, 不同类型的编码器可能对不同的数据模态具有不同的优势, 这表明这是一个潜在的进一步探索领域。

THU-READ 上的性能 表 3.2 展示了在 THU-READ 数据集上与最先进方法的性能比较。报告的结果是根据^[131] 在 CS 协议下对所有 4 个分割取平均值得出的。我们的 MDSI-Transformer 在所有方法中都实现了最先进的性能, 在简单的分数融合下, RGB-D 结果的最高性能达到 89.36%, 凸显了 MDSI 的优势。同时, 我们的 MDSI-CNN 在 RGB、深度和 RGB-D 模态上分别比现有的最佳 CNN 方法有显著的改进, 分别提高了约 4.5%、1% 和 1%, 进一步验证了我们方法的有效性。

表 3.1 与 IsoGD 数据集上最先进的方法的性能比较。最佳和第二佳方法通过 **加粗**和 下划线 标注。

Modality	Backbone	Methods	Accuracy(%)
RGB	CNN	ConvLSTMForGR ^[8]	57.42
		NAS ^[12]	58.88
		RAAR3DNet ^[118]	62.66
		MSA-3D ^[119]	<u>62.73</u>
		Ours	72.79
	Transformer	Decouple+Recouple ^[10]	60.87
		MFST-Large ^[127]	61.26
		UMDR (32 frame) ^[6]	<u>63.68</u>
		Ours	68.32
Depth	CNN	ConvLSTMForGR ^[8]	54.18
		NAS ^[12]	55.68
		RAAR3DNet ^[118]	60.66
		MSA-3D ^[119]	<u>61.72</u>
		Ours	65.99
	Transformer	Decouple+Recouple ^[10]	60.17
		MFST-Large ^[127]	61.29
		UMDR (32 frame) ^[6]	<u>64.62</u>
		Ours	68.00
RGB-D	CNN	ConvLSTMForGR ^[8]	61.05
		NAS ^[12]	65.54
		RAAR3DNet ^[118]	66.62
		MSA-3D ^[119]	<u>68.15</u>
		Ours	75.09
	Transformer	Decouple+Recouple ^[10]	66.79
		MFST-Large ^[127]	68.47
		UMDR (32 frame) ^[6]	<u>72.61</u>
		Ours	74.73

表 3.2 与 THU-READ 数据集上最先进的方法的性能比较。最佳和第二佳方法通过 **加粗** 和 下划线 标注。

Modality	Backbone	Methods	Accuracy(%)
RGB	CNN	VGG ^[128]	41.90
		SlowFast ^[129]	69.58
		NAS ^[12]	71.25
		TSN ^[130]	<u>73.85</u>
		Ours	78.34
	Transformer	Trearr ^[7]	80.42
		Decouple+Recouple ^[10]	81.25
		UMDR (32 frame) ^[6]	<u>82.50</u>
		Ours	85.30
Depth	CNN	VGG ^[128]	34.06
		TSN ^[130]	<u>65.00</u>
		Ours	65.94
	Transformer	Trearr ^[7]	76.04
		Decouple+Recouple ^[10]	<u>77.92</u>
		UMDR (32 frame) ^[6]	79.59
		Ours	74.87
RGB-D	CNN	SlowFast ^[129]	76.25
		NAS ^[12]	<u>78.38</u>
		Ours	82.71
	Transformer	Trearr ^[7]	84.90
		Decouple+Recouple ^[10]	87.04
		UMDR (32 frame) ^[6]	<u>88.09</u>
		Ours	89.36

表 3.3 MDN 各个子分支的性能比较。“Trans.”表示 Transformer。

Branches	RGB ^G	RGB ^D	Depth ^G	Depth ^D	Accuracy(%)	
					MDSI-CNN	MDSI-Trans.
(0) Ours-RGB ^G	✓	✗	✗	✗	68.93	65.79
(1) Ours-RGB ^D	✗	✓	✗	✗	50.47	51.75
(2) Ours-Depth ^G	✗	✗	✓	✗	61.34	66.01
(3) Ours-Depth ^D	✗	✗	✗	✓	47.89	47.99
(4) Ours-RGB	✓	✓	✗	✗	72.58	68.32
(5) Ours-Depth	✗	✗	✓	✓	65.32	68.00
(6) Ours-RGB-D	✓	✓	✓	✓	75.09	74.73

讨论 总体而言,表 3.1 和 3.2 中的结果证明了 MDSI 在不同主干设置中的稳健性,在两个数据集的两种配置下均实现了最先进的性能。虽然 MDSI 始终优于其他方法,但我们观察到某些偏差。具体而言,不同的编码器主干表现出对不同数据集的偏好: CNN 在 IsoGD 上表现最佳,而 Transformer 在 THU-READ 上表现出色。这可能归因于数据集特征的差异,例如自我中心与第三人称视角,或对超参数和预处理技术的敏感性(例如,高级姿势估计可能会进一步提高性能)。未来的工作将进一步探索这些偏差。

3.3.3 消融研究

3.3.3.1 姿势-运动解耦 (PMD) 的影响

如表 3.3 所示,我们首先分析 MDN 各子分支的性能,以验证姿势-运动解耦 (PMD) 的影响。以 MDSI-CNN 为例:将 \mathcal{G} 分支与 \mathcal{D} 分支融合后,与单独使用单个 \mathcal{G} 和 \mathcal{D} 分支相比,准确率显著提高(RGB: $\uparrow 3.65\%$ 和 $\uparrow 22.11\%$,深度: $\uparrow 3.98\%$ 和 $\uparrow 17.43\%$)。这一实质性的增强凸显了 PMD 的有效性,并表明全局信息和解耦信息都至关重要,因为它们共同促进了识别性能。此外,与单模态(RGB-D: $\uparrow 2.01\%$ 和 $\uparrow 9.27\%$)相比,多模态融合带来了另一项改进,证实了该模型强大的拟合能力。在我们的 MDSI-Transformer 中也观察到了这种趋势,突显了我们的方法在不同编码器架构中的一致优势。

此外,我们评估了基于 MDSI-CNN 的 PMD 模块中姿势和运动表征的影响。结果表明,仅使用姿势特征(12.71%)或运动特征(20.99%)是不够的,因为每个特征仅代表部分手势信息。然而,将这些特征集成到 PMD 模块中可显著提高识别准确率至 50.47%。这进一步证明了 PMD 的有效性,突出了姿势和运动信息可以同

时捕捉细微的手势特征，并在有效结合时增强识别能力。

3.3.3.2 时空通道解耦 (STCD) 的影响

如表 3.4 所示，我们分析了时空通道解耦模块 (STCD) 的有效性。集成 STCD 的模型分别比基础 MDSI-CNN 和 MDSI-Transformer 模型实现了 4.17% 和 2.16% 的改进，这反映了时空通道解耦纠缠特征有助于学习判别信息。值得注意的是，我们的 STCD 模块既可插入又轻量级，参数最小增加 0.5M（表 3.6）。为了减轻额外信息的干扰，本研究未启用 SIN。

表 3.4 THU-READ(CS4) 上 STCD 模块的消融研究。“Trans.” 表示 Transformer。

Variants	Accuracy (%)	
	MDSI-CNN	MDSI-Trans.
Ours- <i>w/o</i> STCD	72.08	80.83
Ours- <i>w/</i> STCD	76.25	82.99

表 3.5 SIN 组件的消融研究：语义过滤器 (SF) 和语义标签平滑 (SLS)。SF_{*n*} 表示使用 *n* 个语义过滤器混合一个视觉特征（SF₀ 表示视觉过滤器是随机初始化的，没有集成语义过滤器）。

Variants	SLS	SF	FilterNum	Accuracy(%)
Base Model			/	67.93
Ours-SLS	✓		/	68.64
Ours-SF		✓	1	68.65
Ours-SF ₀			0	68.41
Ours-SF ₁			1	68.79
Ours-SF _{<i>b</i>}	✓	✓	16	<u>68.86</u>
Ours-SF _{<i>N</i>}			249	68.93

3.3.3.3 语义整合网络 (SIN) 的组成部分

表 3.5 展示了 SIN 网络的组成部分分析：语义过滤器 (SF) 和语义标签平滑 (SLS)。首先，SLS 有助于提高性能，将准确率提高 0.71%。此外，SF 将准确率提高了 0.72%，证明了语义过滤器整合的有效性。最终，利用 SF 和 SLS 可实现最佳性能 (↑ 1.00%)，显著超越基础模型。

我们进一步研究了 SF 中使用的语义过滤器的数量 *n* 对识别准确率的影响。我们设计了三种 SF 模式：1) SF₁，其中每个视觉特征都与相应的单个语义特征集成；

2) SF_b , 其中每个视觉特征都与其批次对应的 b 个语义特征集成; 3) SF_N , 其中每个视觉特征都与所有 N 个语义特征集成。如表 3.5 所示, 将过滤器的数量从 1 增加到 16 再增加到 249, 分别带来了 0.07% 的改进。这一趋势表明, 语义过滤器的数量越多, 集成的语义信息就越丰富, 从而更有效地增强了视觉网络的性能。此外, 我们的研究结果表明, 当过滤器随机初始化时 (如在 SF_0 设置中), 与集成语义信息的过滤器相比, 识别性能下降了 0.52%。这一观察进一步强调了整合语义信息的必要性和有效性。

3.3.3.4 参数评估

为了验证 MDSI 的效率, 我们评估了每个子模块中可训练参数的数量。如表 3.6 所示, MDSI 非常轻量, 仅为 Transformer (38.15M) 主干添加了 2.61M 个参数, 为 CNN (12.03M) 主干添加了 2.69M 个参数, 额外计算成本很低 (分别为 6.84% 和 22.36%)。这证明了 MDSI 的效率。值得注意的是, PMD 和 SLS 模块无需训练, 因此将可插拔 MDSI 集成到任何编码器主干中非常有利。

表 3.6 每个子模块的可训练参数, 其中 MDSI 的总参数被加粗。“Trans.”表示 Transformer。

Networks	Modules	Params (M)	
		MDSI-CNN	MDSI-Trans.
MDN	PMD	0	0
	STCD	0.50	0.33
SIN	SF	2.21	2.28
	SLS	0	0
MDSI	-	2.69	2.61
Backbone	ϵ	12.03	38.15
Total	-	14.72	40.76

3.3.3.5 定量和可视化结果

为了进一步阐明我们方法的有效性, 我们在图 3.5 中展示了 t-SNE 可视化和混淆矩阵。此外, 我们在图 3.6 中提供了统计结果, 以突出 MDSI 在解决信息冗余 (IR) 和信息缺失 (IA) 挑战方面的改进。

在图 3.5(a) 和 3.5(b) 中, 我们展示了基础模型和我们提出的 MDSI 方法的 t-SNE 可视化。我们评估了与基础模型相比准确率提高最大的 10 个最具挑战性的类别。基线模型 (图 3.5(a)) 的视觉模式以相互交织的类别表示为特征, 与我们的 MDSI (图 3.5(b)) 实现的良好分离的聚类形成鲜明对比, 这表明我们的 MDSI 有效

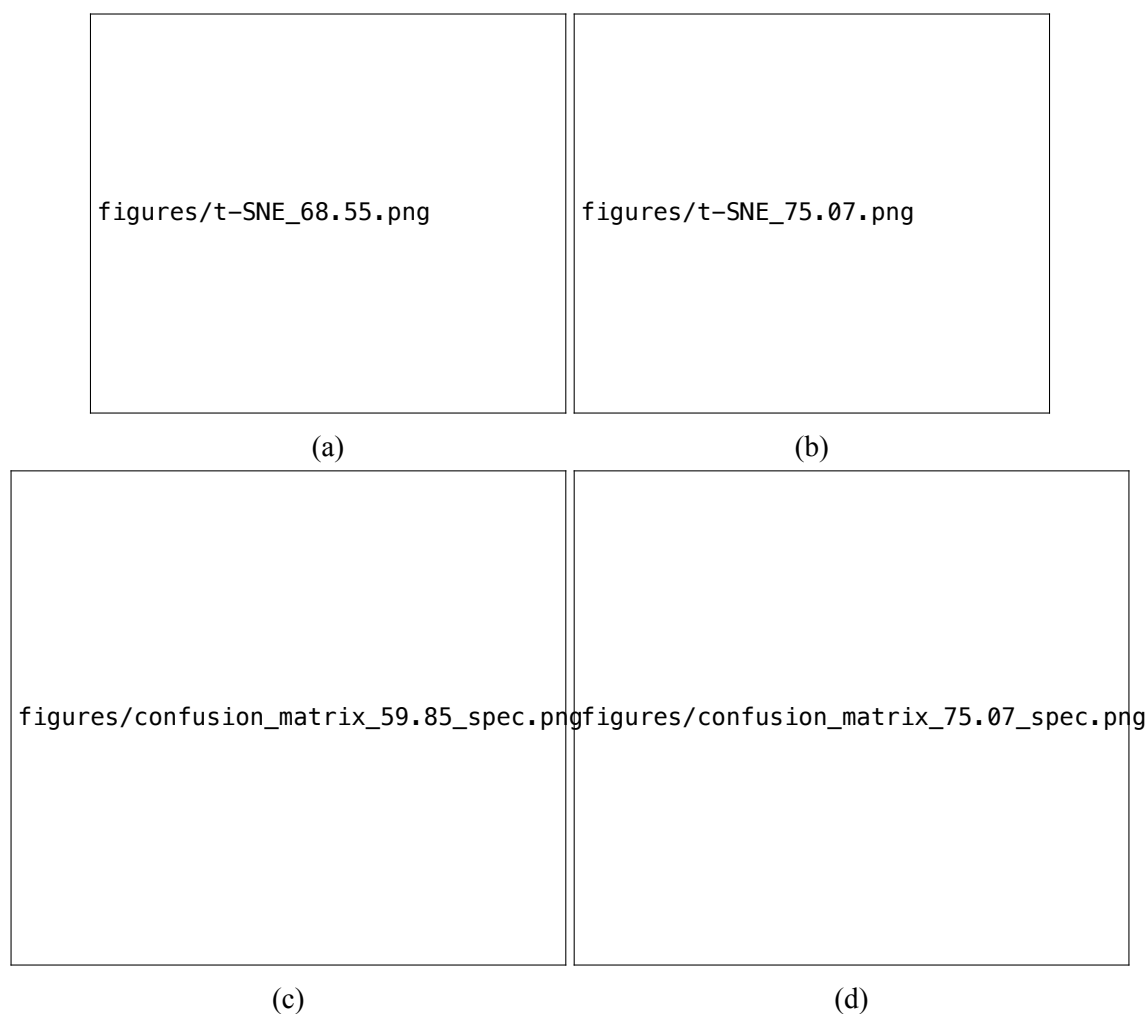


图 3.5 可视化所提方法的增强效果。(a) 基础模型的特征分布可视化。(b) 所提出的 MDSI 的特征分布可视化。(c) 基础模型的混淆矩阵。(d) 所提出的 MDSI 的混淆矩阵。

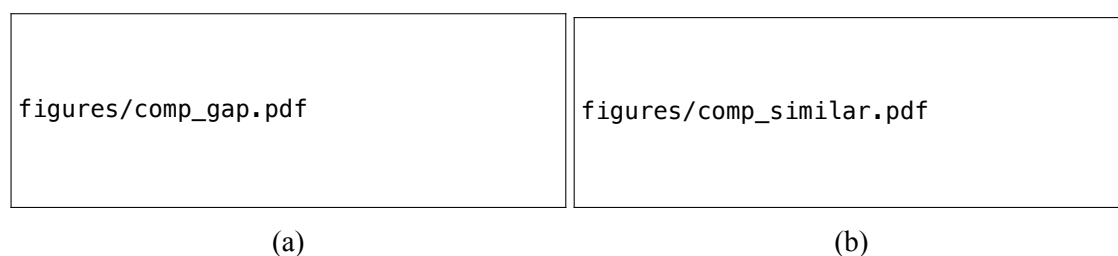


图 3.6 阐明所提出的 MDSI 在解决信息冗余 (IR) 和信息缺失 (IA) 挑战方面的增强功能。(a) IR 场景下的性能改进 (对应于图 1.2(a) 中的样本)。(b) IA 场景下的性能改进。(对应于图 1.2(b) 中的样本)

地增强了特征表示的判别能力。在图 3.5(c) 和 3.5(d) 中, 我们展示了基线模型和我们提出的 MDSI 的混淆矩阵。矩阵显示, 使用我们的方法, 错误分类明显减少, 说明了它通过最小化无关信息干扰来应对手势识别复杂性的卓越能力。这种增强表明, 所提出的方法不仅有效地增强了识别的鲁棒性和泛化能力, 而且还增强了模型对关键手势特征的感知。

同时, 我们对模型在信息冗余 (IR) 和信息缺失 (IA) 场景下的性能提升进行了详细分析。考虑了图 1.2 中描绘的两类样本对, 图 3.6(a) 和图 3.6(b) 分别说明了所提方法的影响。MDSI 带来的改进值得注意: i) 对于遭受 IR 的样本类别, MDSI 将准确率提高了 58%; ii) 对于遭受 IA 的样本类别, MDSI 将准确率提高了 28.57%, 并将视觉相似类别的误分类率降低了 23.81%。这些示例表明, 我们的 MDSI 显著减轻了 IR 和 IA 对手势识别的影响, 确保了稳健和通用的性能。

3.4 本章小结

本章提出了一种可插拔的多模态动态手势识别算法, 称为多策略解耦与语义集成 (MDSI), 旨在解决 RGB-D 手势识别中的信息冗余 (IR) 和信息缺失 (IA) 挑战。首先, 第 3.1 节引入了多策略解耦网络 (MDN), 通过解耦”姿态-运动”和”时空-通道”特征来减轻冗余信息。随后, 第 3.2 节提出了语义集成网络 (SIN), 该网络通过语义过滤和标签平滑增强语义理解, 有效地指导视觉相似手势的区分。MDSI 的可插拔特性使其能够以最小的计算开销无缝集成到各种视频编码器架构中, 展示了其在动作识别和手语识别等相关领域的扩展潜力。大量实验 (第 3.3 节) 表明, MDSI 在两个广泛认可的基准测试中超越了先前的最先进方法。

第4章 基于手势描述的协同手势生成算法研究

本章将详细介绍一种描述驱动的协同手势生成框架（如图 4.1 所示）。首先，我们提出了一种统一的运动表示方法，将不同来源的运动数据嵌入到紧凑的潜在空间中（第 4.1 节）。在此基础上，我们设计了一种可控的手势潜在扩散模型（第 4.2 节），该模型包含两个关键组件：(1) 一个手势变分自编码器，用于学习手势的低维潜在表示；(2) 一个在该潜在空间中高效运行的分层条件扩散模型。为了实现语义和节奏的协同精确控制，我们提出了一种手势描述框架（第 4.3 节），该框架利用运动-语言模型为手势数据生成描述性文本标注，填补了手势数据描述性文本标注的空白。此外，我们还设计了一种多粒度描述控制机制（第 4.4 节）以确保精确的语义注入。

figures/teaser.png

图 4.1 **CoordSpeaker** 支持 手势字幕和定制的 协同的说话者动作生成，既能与字幕保持语义一致，又能与音频保持节奏同步。例如，在演讲场景中，我们的方法允许说话者在讲话时自然地向前走并鞠躬，无缝地做出结束手势。

4.1 运动表示

为了将不同来源的运动数据统一到紧凑的潜在空间中，我们实现了一种紧凑的运动表示方法。首先将各种数据格式转换为统一的 **HumanML3D** 格式表示，然后通过我们的运动编码器 \mathcal{E} （在第 4.2 节中描述）将其映射到一个具有代表性的潜在空间。

4.1.1 统一运动表示

为了利用额外人体运动数据集中的语义先验知识，我们采用 **HumanML3D**^[93] 并将不同的运动数据统一到一个一致的特征空间中。参照^[30,93]，第 i 帧运动被表示为 $x^i = \{\dot{r}^a, \dot{r}^x, \dot{r}^z, r^y, \mathbf{j}^p, \mathbf{j}^v, \mathbf{j}^r, \mathbf{c}^f\}$ ，其中 $\dot{r}^a, \dot{r}^x, \dot{r}^z, r^y$ 表示根关节的角速度、线速度和高度， $\mathbf{j}^p, \mathbf{j}^v, \mathbf{j}^r$ 表示关节位置、速度和旋转， \mathbf{c}^f 表示足部接触特

figures/model_v1_zh.png

图 4.2 协同手势生成模型概览。我们的条件潜在扩散模型（第 4.2 节）由两个关键组件组成：（1）手势变分自动编码器，可学习统一的低维潜在表示，从而实现紧凑的跨数据集运动建模，以及（2）分层控制的降噪器，确保分层条件注入并在该学习到的潜在空间中有效运行。

征。虽然原始 HumanML3D 格式遵循 SMPL^[132] 骨架的 22 个关节，但本文将其扩展到 55 个关节以更好地适应手势数据。因此，每一帧运动被表示为 659 维特征向量，记为 $x \in \mathbb{R}^{T \times 659}$ ，其中 T 为序列长度。

4.1.2 潜在表示

近期研究表明，潜在表示在神经运动建模中发挥着关键作用 []。本文采用基于 Transformer 的 VAE 模型将运动序列编码到紧凑且信息丰富的潜在空间中，该模型由编码器 \mathcal{E} 和解码器 \mathcal{D} 组成，并通过长跳跃连接来保持运动细节。具体而言，运动序列 $x^{1:L}$ 首先通过编码器 \mathcal{E} 编码为潜在向量 $z \in \mathbb{R}^{n \times d}$ ，其中 d 表示潜在维度。编码器接收逐帧运动特征和可学习的分布令牌作为输入，生成运动潜在空间的高斯分布参数 μ 和 σ 。这些参数通过标准 VAE 采样过程重参数化得到潜在向量。对于运动解码， \mathcal{D} 采用带有交叉注意力机制的 Transformer 架构。它以零运动令牌作为查询，潜在向量 z 作为键和值，生成重建的运动序列 $\hat{x}^{1:L}$ 。整个 VAE 通过均方误差（MSE）和 Kullback-Leibler（KL）散度的组合进行训练：

$$L_{\text{VAE}} = \|x^{1:L} - \hat{x}^{1:L}\|_2^2 + \beta \text{KL}(q_\phi(z|x^{1:L})\|p(z)) \quad (4.1)$$

其中 β 平衡重建和正则化项。这种潜在表示方法在保持运动保真度和多样性的同时实现了高效的手势合成。

4.2 可控手势潜在扩散模型

基于 VAE 学习的紧凑潜在表示，我们提出了一种可控手势潜在扩散模型，用于生成高质量手势。与以前直接在原始运动序列上执行扩散的方法不同，我们的模型通过对潜在向量进行操作，显著降低了计算要求，同时保持了生成质量。

4.2.1 潜在扩散过程

潜在空间中的扩散过程遵循马尔可夫链，逐步向潜在向量 $\mathbf{z} \in \mathbb{R}^{n \times d}$ 添加高斯噪声。对于噪声步骤 $t \in [1, T]$ ，前向过程定义为：

$$q(\mathbf{z}_t | \mathbf{z}_{t-1}) = \mathcal{N}(\sqrt{\alpha_t} \mathbf{z}_{t-1}, (1 - \alpha_t) \mathbf{I}), \quad (4.2)$$

其中 $\alpha_t \in (0, 1)$ 是常数方差调度，且 $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 。为了生成高质量手势，我们采用基于 Transformer 的去噪器 ϵ_θ 来迭代预测和去除噪声。从随机噪声 \mathbf{z}_T 开始，去噪器逐步恢复潜在向量 $\hat{\mathbf{z}}_0$ ，然后通过 \mathbf{D} 解码为手势运动。去噪过程同时受音频特征 \mathbf{A} 和描述嵌入 \mathbf{C} 的条件约束。

4.2.2 分层条件注入

为了对手势生成的精确多模态控制，我们的模型通过精心设计的嵌入和分层注入过程整合多个条件。对于字幕编码，使用预训练的 CLIP 文本编码器^[133]提取语义特征 $\mathbf{C} \in \mathbb{R}^{512}$ 。使用 WavLM 编码器^[134]提取音频特征 $\mathbf{A} \in \mathbb{R}^{T \times 1133}$ ，捕获丰富的声学信息，包括韵律和节奏模式。然后通过特定于模态的嵌入层处理这些原始特征，以形成条件嵌入 \mathbf{C} 和 \mathbf{A} 。此外，为了实现更灵活和精确的生成控制，我们提出了一种分层条件注入机制，包含两个阶段（图 4.2）：首先，初始条件 \mathbf{c}_1 与噪声潜在向量和时间步嵌入串联后输入去噪器 ϵ_θ ，提供基本指导。其次，细化条件 \mathbf{c}_2 通过交叉注意力机制在去噪器内部整合，实现精细的控制调整。

4.2.3 无分类器指导

为了增强生成手势的质量和可控性，我们在训练和推理过程中采用无分类器指导^[115]。在训练过程中，随机屏蔽 10% 的条件输入以同时学习有条件和无条件分布。在推理过程中，噪声预测通过不同指导的预测加权组合计算：

$$\begin{aligned} \epsilon_\theta^s(\mathbf{z}_t, t, \mathbf{c}) = & s_1 \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c} = \{\emptyset, \mathbf{A}\}) \\ & + s_2 \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c} = \{\mathbf{C}, \emptyset\}) \\ & + (1 - s_1 - s_2) \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c} = \{\emptyset, \emptyset\}), \end{aligned} \quad (4.3)$$

其中 s_1 、 s_2 分别是音频和描述的指导尺度，且 $s_1, s_2 > 1$ 可以增强其效果。这种方法允许对描述和音频分别应用指导，在手势生成过程中实现更精确的条件控制。

4.2.4 训练与推理

潜在扩散模型使用简单的 ℓ_2 目标进行训练^[97,113]：

$$\mathcal{L}_{\text{Diff}} = \mathbb{E}_{\epsilon, t} [\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})\|_2^2], \quad (4.4)$$

figures/gesture_captioning_v1_zh.png

图 4.3 手势描述生成框架。我们的手势描述生成框架包含两个主要组件: 运动分词器和运动感知语言模型。运动分词器将手势序列编码为离散的运动 **token** 序列, 运动感知语言模型则基于这些 **token** 和提示模板生成对应的手势描述。

其中 $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ 且 $z_0 = \mathcal{E}(x^{1:L})$ 。在推理过程中, 我们的模型首先通过 T 个去噪步骤预测潜在向量 \hat{z}_0 , 然后通过解码器 \mathcal{D} 的单次前向传播将其解码为运动序列 $\hat{x}^{1:L}$ 。

4.3 手势描述

手势数据缺乏描述性文本标注显著限制了通过文本提示控制手势合成的能力^[31]。为了解决这个问题, 我们提出了一种新颖的手势描述方法, 利用运动-语言模型为手势生成描述性标注。这种方法为手势数据集的语义标注稀缺问题提供了一个高效的解决方案, 同时通过生成的描述实现精细的语义控制 (详见第 4.4 节)。

我们的手势描述框架 (图 4.3) 包含两个主要组件: 运动分词器和运动感知语言模型 (*MotionLLM*)。运动分词器基于^[1]中使用的 VQ-VAE 架构, 由编码器 $\mathcal{E}_{\mathcal{M}}$ 和解码器 $\mathcal{D}_{\mathcal{M}}$ 组成, 用于生成离散的运动令牌。运动感知语言模型采用基于 Transformer 的架构, 具有统一的文本-运动词表 $\mathbf{V} = \{\mathbf{V}_t, \mathbf{V}_m\}$, 使其能够在单个模型中灵活地联合建模文本和运动。具体而言, 手势序列首先转换为第 4.1.1 节中描述的统一运动表示, 然后输入我们的手势描述框架。给定一个包含 M 帧的手势运动序列 $m^{1:M} = \{x^i\}_{i=1}^M$, 运动分词器首先将其编码和量化为离散的运动令牌序列 $z^{1:L} = \{z^i\}_{i=1}^L$ 。精心设计的提示模板被分词为文本令牌 $w^{1:N} = \{w^i\}_{i=1}^N$ 。随后, 离散运动令牌和文本令牌被混合并共同输入运动感知语言模型, 生成相应的手势描述 $\hat{w}^{1:L} = \{w^i\}_{i=1}^L$ 。在实践中, 我们利用^[92]提供的预训练模型, 并在推理过程中保持其参数冻结。

实现细节 表 4.1 展示了我们的手势描述框架中使用的提示模板集合。这些经过精心设计的模板会被多次随机采样, 并与不同的手势片段配对以生成多样化的手势描述。这些模板的设计受到了最新运动-语言建模研究的启发^[92]。

表 4.1 手势描述框架中使用的提示模板示例。

Task	Input	Output
Gesture-to-Text	Give me a summary of the motion being displayed in [motion] using words.	[caption]
	Explain the motion illustrated in [motion] using language.	
	Describe the action being represented by [motion] using text.	
	What kind of action is being demonstrated in [motion]?	
	Describe the movement demonstrated in [motion] in words.	
	Generate a sentence that explains the action in [motion].	
	Please describe the movement depicted in [motion] using natural language.	
	Provide a description of the motion being displayed in [motion] using language.	
	Give me a brief summary of the movement depicted in [motion].	
	Describe the movement demonstrated in [motion] using natural language.	

此外，由于我们使用的 MotionLLM 是在 22 个关节点的人体运动数据集上预训练的，在进行描述生成推理之前，我们需要先将手势特征从 $x \in \mathbb{R}^{T \times 659}$ 转换为 $\mathbb{R}^{T \times 263}$ 的相似格式，这是通过仅保留前 22 个关节点的数据来实现的。这种转换不可避免地会导致一些细微的手指动作细节丢失。如果能有一个包含明确手指动作标注的更精细数据集，将能显著缓解这一局限性。

4.4 多粒度描述控制

由于手势的时序动态性和语义信息的不同粒度，有效利用手势描述进行精确控制具有挑战性。为了解决这个问题，我们提出了一种多粒度手势控制机制，实现了跨多个时间和语义尺度的精细控制。

给定一个手势序列，我们将其分段为 $m^{1:M} = \{m^{i:i+K-1}\}_{i=1}^{M-K+1}$ ，其中 K 表示分段长度。每个分段通过我们的手势描述框架生成局部描述 $\hat{w}^{1:L} = \{w^i\}_{i=1}^L$ ，同时通过用分隔符 <SEP> 连接局部描述形成全局描述。基于这种分层描述结构，我们设计了三种互补的控制策略：(1) 常规控制。该策略应用均匀时间分段生成局部描述，在保持时间一致性的同时实现对手势细节的精确控制。(2) 动态控制。为了增强灵活性和鲁棒性，该策略在训练过程中引入可控的随机性。分段被随机采样，局部描述通过随机选择（从候选描述中随机选择）或加权聚合（基于时间重叠程度组合描述）进行装配。这种方法帮助模型学习处理不同的时间尺度和描述组合。(3) 分层控制。在前述策略的基础上，我们将分层描述与分层条件注入（第 4.2 节）相结合：局部描述、音频和时间步嵌入被串联并作为 c_1 注入去噪器编码器，确保与局部手势分段的精确语义和节奏同步；然后，全局描述作为 c_2 通过交叉注意力注入去噪器解码器，增强手势的整体连贯性和语义相关性。这种多粒度控制机制

在确保时间一致性的同时，实现了灵活的多尺度语义控制。

4.5 实验结果与分析

在本节中，我们通过定量和定性分析从四个方面评估我们的方法：(1) **协同手势生成**。与最先进的基线相比，评估模型在联合语音和字幕控制下生成语音同步、语义相关的全身手势的能力。(2) **文本驱动运动生成**。与最先进的文本到运动方法进行比较，以评估语义理解和非自发手势生成能力。(3) **手势字幕**。评估生成的手势字幕的质量和多样性标志着弥合手势数据集中语义差距的第一种方法。(4) **消融研究**。分析我们方法中关键组件的影响。

4.5.1 实验设置

数据集 本文利用语音到手势数据集 BEAT^[84]和文本到运动数据集 HumanML3D^[93]进行训练。这种跨数据集学习策略使我们能够引入额外的语义运动先验。HumanML3D 数据集包含 14,616 个基础运动序列和 44,970 个对应的文本描述，为运动生成任务提供了丰富的语义先验。BEAT 数据集包含约 76 小时的语音-手势对齐的多模态序列，涵盖了多样化的手势类型和说话场景，为语音驱动的手势生成提供了重要支持。参照^[84]，我们使用四位英语演讲者的手势数据。

评估指标 我们从三个方面评估生成结果：(1) 重建质量：使用抖动度和加速度指标^[135]评估动作的平滑度和自然度。(2) 语音到手势生成：使用 FGD^[85]评估身体手势的真实性。通过计算不同手势片段间的平均 L1 距离评估多样性^[23]，使用 BC 评估语音与动作的同步性。(3) 文本到运动生成：使用 Frechet Inception Distance (FID) 和 Diversity (DIV) 评估生成运动的真实性和多样性，使用 motion-retrieval precision (R Precision) 和 Multi-modal Distance (MM Dist) 评估运动与文本描述的匹配程度^[97]。

实现细节 我们基于 Transformer 的 VAE 和降噪器 ϵ_θ 均由编码器 E 和解码器 D 组成，每个包含 9 层和 4 个注意力头，具有 GELU 激活和残差连接。潜在维度设置为 $z \in \mathbb{R}^{1 \times 512}$ 。语义和音频嵌入都通过线性层投影到 512 维空间中，然后输入到降噪器中。对于训练，我们使用 AdamW 优化器，学习率为 $1e^{-4}$ ，批处理大小为 128。VAE 训练了 6000 个 epoch，而扩散模型训练了 2000 个 epoch。在训练过程中，我们使用 1000 个扩散步骤和 50 个推理步骤，噪声方差 β_t 从 8.5×10^{-4} 线性缩放到 0.012。在 VAE 阶段，KL 损失权重 β 设置为 0.0001（公式 4.1）。在推理过

figures/visualization_v1.png

图 4.4 协同手势生成的定性比较。**红色**框突出显示语义不一致，**黄色**框表示不自然的动作，**绿色**框表示协同良好的自然手势。

程中，对于无分类器指导，音频和字幕指导比例默认设置为 $s_1 = 7$ 和 $s_2 = 1$ ，以平衡不同的条件贡献。

为平衡训练过程中不同数据集的数据分布，我们对数据加载器采用加权随机采样策略。参照^[30]，所有运动序列重采样至 20 FPS，并截断或填充至 180 帧。对于 HumanML3D 数据集，仅使用长度在 40 至 180 帧之间的序列。

4.5.2 协同手势生成

4.5.2.1 定性对比

我们首先评估我们的方法在生成协同的语音同步、语义相关的运动方面的表现。如图 4.4 所示，我们将其与仅有的两部作品^[30-31]进行了比较，这两部作品都涉及相关的协同生成任务。参照^[31]，我们使用语音输入和四种不同的文本描述展示结果：“*raising both hands up*”、“*sitting*”、“*holds a cup of tea*”和“*walks straight forward*”。结果表明，我们的方法有效地产生了节奏一致且语义一致的协同手势，同时实现了更自然的外观。相比之下，FreeTalker 无法生成语义动作，而仅专注于语音驱动的手势。虽然 SynTalker 实现了一定的协同性，但它表现出细节上的不一致和不自然的僵硬性。例如，在传达“*holding ... in left hand*”时，我们的方法指示说话者将左手保持在腰部附近的稳定位置，这是 FreeTalker 和 SynTalker 结果中都缺失的细节。在表达“*raising both hands up*”时，我们的方法通过同时抬起双臂保

表 4.2 与基线模型和消融研究进行比较的定量结果。‘→’表示越接近真实运动越好。每个指标均在 20 次运行的 95% 置信区间下报告。我们报告 $BC \times 10^{-1}$ 和 Top-1 R-Precision。

Methods	Reconstruction		Audio-to-Gesture			Text-to-Motion			
	Jerk→	Accel.→	FGD↓	BC↑	L1Div↑	FID↓	MM-Dist↓	Div→	R-Precision↑
GT	1.165 \pm .000	0.043 \pm .000	-	-	-	-	6.205 \pm .043	5.512 \pm .114	0.140 \pm .008
FreeTalker ^[30]	0.611 \pm .013	0.030 \pm .000	2.101 \pm .026	1.147 \pm .028	11.332 \pm .025	0.761 \pm .048	6.737 \pm .051	5.396 \pm .127	0.102 \pm .008
Ours	1.190 \pm .015	0.039 \pm .001	3.173 \pm .123	1.327 \pm .049	10.861 \pm .066	1.118 \pm .061	6.814 \pm .056	5.558 \pm .126	0.100 \pm .008
Ours-w/o mo.	1.005 \pm .014	0.032 \pm .000	2.654 \pm .041	2.627 \pm .043	19.600 \pm .089	3.911 \pm .163	7.664 \pm .050	4.070 \pm .117	0.043 \pm .004
Ours-w/o hcd.	1.201 \pm .017	0.038 \pm .001	2.302 \pm .061	1.910 \pm .004	12.781 \pm .044	1.260 \pm .063	6.872 \pm .058	5.303 \pm .107	0.102 \pm .010
Ours-w/o mgc.	1.239 \pm .013	0.039 \pm .000	3.123 \pm .139	2.256 \pm .045	15.363 \pm .103	2.568 \pm .099	7.031 \pm .044	5.447 \pm .150	0.082 \pm .006

持与文本的一致性，而 FreeTalker 未能抬起双手，SynTalker 则在举起后放下一只手。此外，在表现 “*sitting*” 和 “*walking straight forward*” 时，我们的方法生成了更自然的说话者动作，相比之下 SynTalker 的结果显得较为僵硬。值得注意的是，与 SynTalker 不同，我们的方法不需要额外的预训练或推理成本，并且可以实现更快的推理，如第 4.5.5 节所示。

4.5.2.2 定量结果

由于只有 FreeTalker^[30] 和 SynTalker^[31] 与本文工作研究范围相似，且 SynTalker 不支持多模态条件的定量评估方法，因此我们复现 FreeTalker 作为对比基线。

如表 4.2 所示，我们的方法在所有评估指标中取得了可比或优越的结果，超过了 FreeTalker 的重建质量，节拍一致性 (BC 0.180 ↑)，多样性 (Div 0.162 ↑)，同时维持有竞争力的文本匹配度。值得注意的是，这些结果突出了我们方法在保持语音同步和语义相关性之间良好平衡的独特能力，证实了其在生成自然且可控的协同手势方面的有效性。

4.5.3 文本驱动的运动生成

我们进一步将我们的方法与最先进的文本到运动方法进行比较，以验证其在捕获语义和生成非自发运动方面的能力。为了公平比较，我们仅使用文本条件报告 HumanML3D 测试集的结果。表格 4.3 显示，与 SOTA 模型相比，我们的方法实现了高级性能，实现了最佳文本对齐 (MM-Dist 3.584) 和第二好的生成保真度 (FID 0.405)。这强调了我们的分层控制降噪器在集成细粒度语义方面的有效性。值得注意的是，我们的方法明显优于类似的协同方法 SynTalker^[31]，突出了其在协同生成和增强语义控制方面的优势。

表 4.3 与 HumanML3D^[93] 测试集上的最新方法进行比较。我们按照^[97] 计算标准度量。‘→’ 表示越接近真实运动越好。每个度量均在 20 次运行的 95% 置信区间下报告。

Methods	FID↓	MM-Dist↓	Div→	R-Precision↑		
				Top-1	Top-2	Top-3
GT	0.001±.001	3.378±.007	10.471±.083	0.490±.003	0.682±.003	0.783±.003
MDM ^[89]	1.390±.088	4.599±.037	10.704±.066	0.363±.007	0.553±.008	0.662±.007
T2M-GPT ^[136]	0.564±.012	3.867±.008	10.558±.083	0.433±.003	0.615±.002	0.716±.003
MLD ^[97]	0.963±.029	3.898±.012	10.401±.096	0.429±.003	0.613±.003	0.717±.002
MoMask ^[137]	0.222±.007	3.620±.011	10.621±.096	0.461±.002	0.657±.003	0.760±.002
SynTalker ^[31]	4.385±.034	4.499±.012	9.374±.073	0.375±.003	0.564±.003	0.681±.002
Ours	0.405±.012	3.584±.012	9.109±.235	0.424±.003	0.601±.003	0.702±.003

figures/visualization_caption_v1.png

图 4.5 手势描述生成结果示例。生成的描述准确地描述了整体运动模式和细粒度的手势细节。

4.5.4 手势描述生成

如图 4.5 所示，所提出的手势描述框架展示了在生成描述性手势标注方面的实用能力。结果表明，我们的方法不仅能准确描述手势运动分布内的细粒度手部动作（如 “*moving both hands and forearms at chest height*”），还能捕捉粗粒度的全身动作（如 “*moves their waist from ...*”）。此外，由于手势运动通常持续较长时间，所提出的多粒度描述方法在捕捉长时间窗口内的连续复杂组合动作方面表现出优势（如 “*a person <motion 1>, then <motion 2> as <motion 3>*”）。然而，我们也注意到对于较长的手势序列，模型在时序位置感知方面存在困难，偶尔会混淆动作顺序。这一限制可以通过使用我们提出的细粒度描述策略来管理动作复杂度。未来的改进可能涉及引入更强的位置编码或时序建模。



图 4.6 定性消融研究。结果是使用语音音频和单字幕 “*A person is raising both hands up while talking*” 生成的。

4.5.5 消融实验

为了进一步评估不同组件对协同生成的影响，我们在此任务下进行了消融研究。定量结果如表 4.2 所示，定性结果如图 4.6 所示。我们使用音频和单个字幕条件 (“*A person is raising both hands up while talking*”) 进行了定性评估。如图 4.6(a) 所示，我们的完整模型有效地捕捉了音频节奏和语义指令，生成自发的语音手势，同时遵循字幕指导产生非自发的手势（向上举起双手）。

手势描述 图 4.6(b) 展示了移除手势描述后的影响。没有描述的语义指导，模型只能生成与语音同步的基本协同手势，无法产生有意义的非自发运动。这表明手势描述在提供语义指导和补充缺失的手势文本标注方面发挥着关键作用。

统一运动表示 移除统一的运动表征（图 4.6(c)）显著限制了模型生成除典型同语动作之外的多样且语义上有意义的手势的能力。虽然细微的向上手部动作仍然存在，但说话者未能完全执行标题中描述的举起双手的预期手势。表 4.2 (*Ours-w/o mo.*) 中的定量结果进一步证实了这种退化，显示语义相关性大幅下降，反映在明显更差的 FID (2.793 ↑)、MM-Dist (0.850 ↑) 和 R-Precision (Top-1 0.057 ↑) 中。

分层控制降噪器 如表 4.2 所示，删除分层控制的降噪器 (*Ours-w/o hcd.*) 会导致重建质量下降 (Jerk 0.011 ↑) 和条件协同性减弱。具体而言，该模型表现出更强的偏向于重建同语手势 (BC 0.583 ↑)，同时无法保持语义相关性 (MM-Dist 0.058 ↑)。这些发现强调了我们的分层控制降噪器在确保有效的多模态条件协同方面的关键作用。

表 4.4 多粒度字幕策略的消融研究。“Reg.”表示常规字幕策略，“Dyn.”表示动态字幕策略，“Hie.”表示分层字幕策略。每个指标均在 20 次运行的 95% 置信区间下报告。我们报告 $BC \times 10^{-1}$ 和 Top-1 R-Precision。

Methods	Reconstruction		Audio-to-Gesture			Text-to-Motion			
	Jerk→	Accel.→	FGD↓	BC↑	L1Div↑	FID↓	MM-Dist↓	Div→	R-Precision↑
GT	1.165±.000	0.043±.000	-	-	-	-	6.205±.043	5.512±.114	0.140±.008
Ours-Reg.	1.201±.017	0.038±.001	2.302±.061	1.910±.004	12.781±.044	1.260±.063	6.872±.058	5.303±.107	0.102±.010
Ours-Dyn.	1.189±.013	0.038±.000	2.866±.106	1.943±.037	14.471±.110	1.404±.049	6.955±.044	5.440±.114	0.095±.006
Ours-Hie.	1.190±.015	0.039±.001	3.173±.123	1.327±.049	10.861±.066	1.118±.061	6.814±.056	5.558±.126	0.100±.008

多粒度字幕 如表 4.2 所示，删除多粒度字幕 (*Ours-w/o mgc.*) 会导致生成的手势的语义相关性显著下降。这凸显了我们的多粒度字幕机制在促进精确和上下文感知的语义注入方面的重要性。在这种情况下，音频作为初始条件，而本地字幕提供精细条件。

多模态无分类器指导 图 4.6(d) 展示了仅对文本模态应用无分类器指导的效果。由于训练时缺乏语音模态的无条件指导，模型在推理时难以融合多模态信号，过度依赖语音输入，导致手势多样性有限且语义对齐性差。这些发现突显了组合多模态指导在实现平衡手势生成中的重要性，有助于整合语音节奏和语义约束。

多粒度描述控制 表 4.4 进一步展示了本文方法在不同粒度描述控制策略下的性能。分层控制策略 (分层控制) 在所有指标上实现了最优平衡：与常规和动态控制策略相比，它表现出更好的语义相关性 (MM-Dist 更低，6.814 vs 6.872 和 6.955)、更好的运动质量 (FID 降低 11.3% 和 25.6%)，同时有效维持了语音同步性和运动多样性。此外，动态策略在协同手势同步性 (BC: 1.943) 和多样性 (L1Div: 14.471) 方面表现出优势，这可能归因于其自适应采样机制增强了训练数据的多样性，使模型能够产生更丰富的节奏手势模式。这些综合结果表明，我们的多粒度描述控制策略有效协同了多模态信号，在运动自然度、语音同步性和语义对齐性之间实现了平衡的权衡。分层控制策略在细粒度语义注入方面表现尤为出色，导致生成手势具有更强的语义相关性。

推理时间 尽管扩散模型表现出色，但其较长的推理时间仍是主要瓶颈。为评估推理效率，我们在单个 NVIDIA RTX 3090 GPU 上测量每句平均推理时间 (AITS)^[97]，批量大小设为 1，不包括模型加载时间。结果显示我们的方法实现了 $0.842 \pm .002$ 秒的 AITS，比 FreeTalker^[30] ($6.632 \pm .044$ s) 和 SynTalker^[31] ($5.804 \pm .044$ s) 快 6 倍以上。

这种显著的加速可以归因于我们高效的分层去噪架构和优化的潜在扩散过程，使我们的方法更适合实际应用。

4.5.6 更多可视化结果

更多协同生成结果 如图 4.7 所示，我们使用相同的音频和不同的文本标题提供了更加协同的生成结果。这些结果进一步证实了我们提出的方法在联合语音字幕控制下生成协同语音自发手势和字幕驱动的非自发运动的有效性。

更多手势字幕结果 我们在图 4.8 中展示了额外的手势字幕结果，进一步证明了我们的方法在准确地将手势映射到文本方面的有效性。正如彩色框中突出显示的那样，该模型在描述复杂、连续的动作时成功捕捉了细粒度的手部动作和粗粒度的全身动作。

4.6 本章小结

本章提出了一种基于描述驱动的协同手势生成框架 (CoordSpeaker)，旨在解决手势生成中的语义标注缺失和多模态协同控制挑战。首先，第 4.1 节引入了统一的运动表示方法，通过将不同来源的运动数据映射到紧凑的潜在空间，为跨数据集手势生成奠定基础。随后，第 4.2 节提出了可控手势潜在扩散模型，该模型通过手势 VAE 学习紧凑的潜在表示，并利用分层条件扩散模型实现高效的多模态条件生成。为了解决手势数据缺乏描述性文本标注的问题，第 4.3 节设计了一个手势描述框架，利用动作-语言模型为手势数据生成语义标注。第 4.4 节提出的多粒度描述控制机制进一步增强了对生成过程的精确控制。最后，第 4.5 节通过大量定性与定量实验验证了本文方法的有效性。大量实验表明，本文方法不仅能生成语义连贯、节奏同步的协同手势动作，而且在推理效率上相比现有方法提升了 6 倍以上，展现了在实际应用中的潜力。

figures/visualization_supply.png

图 4.7 协同手势生成的更多视觉结果。

figures/visualization_caption_supply.png

图 4.8 更多手势字幕结果。彩色框突出显示了手势和文本字幕之间的精确映射。

第 5 章 交互式手语学习助手设计与实现

手语是听障人士进行交流的主要方式。然而，由于手语学习资源有限、学习成本高等原因，听障人士与普通入之间存在较大的交流障碍。随着人工智能技术的发展，基于计算机视觉的手语识别与生成技术为解决这一问题提供了新的可能。本章将基于前文提出的多策略解耦和语义集成网络 (MDSI) 以及协同手势生成算法 (CoordSpeaker), 设计并实现一个交互式手语学习助手系统。该系统旨在为手语学习者提供实时的动作评估与反馈, 帮助提升手语学习效果。

5.1 系统总体设计

5.1.1 设计目标

本系统旨在通过人工智能技术辅助手语学习过程，为学习者提供实时、准确的动作评估与反馈。系统的核心目标是实现高效的手语学习体验，这要求系统具备实时的动作识别能力和自然的人机交互界面。基于这一目标，系统需要在准确性、实时性、交互性和适应性等多个维度上达到较高水平。在准确性方面，系统基于本文提出的 MDSI 算法，通过多模态特征的解耦与融合实现高精度的手语动作识别。该算法能够有效处理手语动作中的细微差异，为学习评估提供可靠的技术支持。在实时性方面，系统采用并行计算技术，确保手语动作的捕获、识别和反馈过程能够实时完成，从而保证良好的交互体验。在交互性方面，系统设计了直观的可视化界面，通过实时对比展示标准动作与用户动作的差异，并结合定量评估指标提供针对性的改进建议。在适应性方面，系统支持根据用户的学习进度和掌握程度动态调整练习内容和难度，实现个性化的学习体验。

5.1.2 系统架构

为实现上述设计目标，本系统采用模块化的软件架构设计，如图 5.1 所示。系统整体划分为手语识别、手语生成、交互反馈三个核心功能模块，各模块通过统一的消息总线进行通信协作。这种模块化设计不仅提高了系统的可维护性和可扩展性，也便于各个功能模块的独立优化和升级。手语识别模块作为系统的核心感知单元，负责实时捕获和识别用户的手语动作。该模块基于 MDSI 算法实现，能够有效处理手语动作中的复杂特征。手语生成模块则基于 CoordSpeaker 协同手势生成算法，负责生成标准的手语动作示范，支持基于文本描述的精确控制，为学

习者提供准确的参考样本。交互反馈模块通过比对用户动作与标准动作，实时生成评估结果和改进建议。



figures/sys.drawio.png

图 5.1 交互式手语学习助手系统架构

5.2 核心功能模块设计

5.2.1 手语识别模块

手语识别模块是系统的核心组件之一，负责实时捕获并识别用户的手语动作，进行评估打分。该模块基于第三章提出的 MDSI 算法，通过 RGB-D 相机实时采集用户手语动作的 RGB 图像序列与深度信息。为提升系统实时性能，本模块采用可插拔的高效 MDSI 算法，基于 GPU 加速，有效降低了系统延迟。

5.2.2 手语生成模块

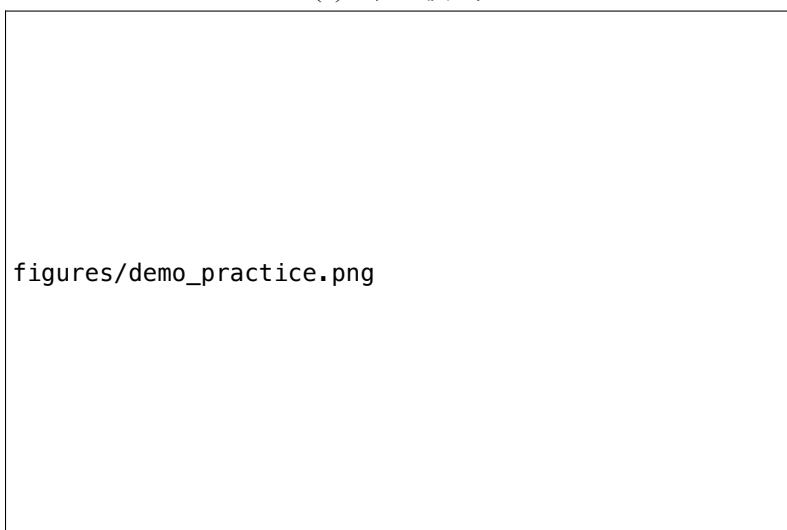
手语生成模块负责生成标准手语动作示范，以供学习/练习使用。基于第四章提出的 CoordSpeaker 协同手势生成算法实现。该模块支持两种工作模式：离线生成模式和实时生成模式。离线生成模式用于预先生成标准动作库，包含常用手语词汇和短语的标准示范动作。实时生成模式则支持根据文本输入即时生成手语动作示范，为用户提供更灵活的学习参考。

5.2.3 交互反馈模块

交互反馈模块通过实时比对用户动作与标准动作，生成评估得分与改进建议。系统通过可视化方式直观展示用户动作与标准动作的差异，并结合自然语言描述提供具体的改进建议。此外，模块还会记录用户的学习轨迹，支持学习进度的追踪与分析。



(a) “学习模式”



(b) “练习模式”

图 5.2 用户界面设计

5.3 人机交互设计

5.3.1 手语学习模式

系统支持手语学习、练习评估和自由练习三种主要交互模式。在手语学习模式下，系统首先展示标准手语动作示范，用户通过摄像头采集练习动作，系统实时提供评估反馈。练习评估模式则提供多组不同难度系统化的练习题目，用户需要根据题目完成指定手语动作，系统给出评分和改进建议。自由练习模式支持用户自主选择练习手语词汇，系统生成相应的示范动作，供用户自主练习，并给出评估反馈。

5.3.2 用户界面设计

系统界面采用分区布局设计，主要包含主视图区、评估反馈区和功能控制区三个主要部分。以“学习模式”为例(如图 5.2 所示)：主视图区采用左右分屏布局，左侧显示标准手语动作示范，右侧实时显示用户的动作画面，便于用户直观对比和模仿。评估反馈区位于界面中部，在用户完成当前手语动作后，显示动作评估分数、关键改进建议和动作要点提示。功能控制区则集中展示学习模式选择、难度调节等操作按钮，保证操作的便捷性。练习模式与学习模式类似，区别在于反馈区位于上下，分别显示题目和评估的分数，用户根据题目完成指定手语动作后，系统给出评分和改进建议，用户可进一步查看答案示范。

5.3.3 交互引导设计

如图 5.3 所示为帮助用户更好地理解系统功能，系统在主界面及各个模式中设计了交互引导功能。当用户首次启动系统时，会显示可选择的学习模式与系统功能。当首次进入子模式时，系统会通过弹窗提示用户进行相关操作，如开启摄像头，并引导用户进行手语学习。在手语学习过程中，系统支持用户根据的学习进度和掌握程度，自主调整练习内容和难度，确保用户能够持续进步。

5.4 系统实现与部署

5.4.1 开发环境与技术栈

本系统采用 Python 作为主要开发语言，基于 PyQt5 构建图形用户界面。系统的核心算法模块采用 PyTorch 深度学习框架实现，通过 CUDA 加速实现 GPU 并行计算。系统开发环境和主要依赖包括：Python 3.13、PyQt5 5.15、OpenCV 4.11 等。在系统架构方面，采用基于消息队列的模块间通信机制，确保各功能模块之间的

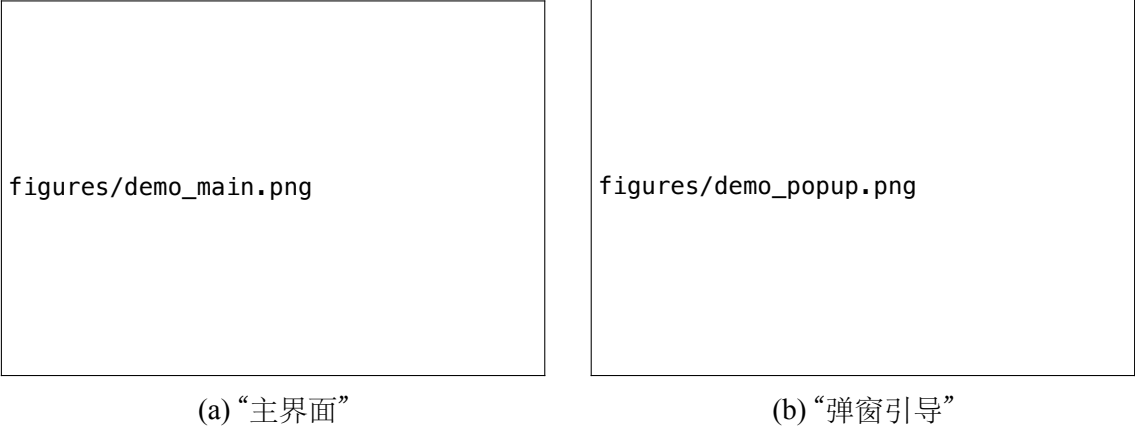


图 5.3 交互引导设计

解耦与协同。系统部署采用容器化方案，使用 **Docker** 实现环境隔离与快速部署。为保证系统的实时性能，手语识别与生成模块均采用 **GPU** 加速，并通过多线程技术实现并行执行。

5.4.2 关键技术实现

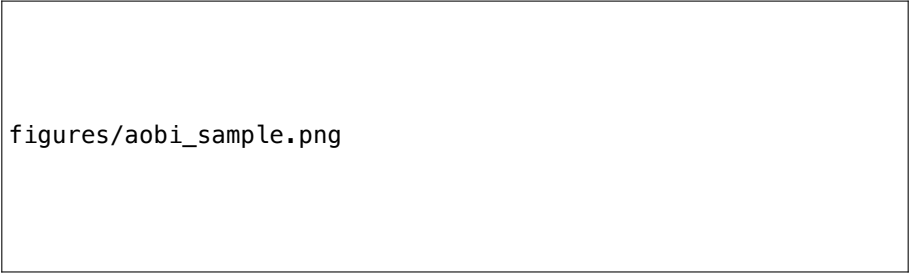
系统的关键技术实现主要包括以下几个方面：首先，在手语动作捕获方面，系统通过 **OpenCV** 实现 **RGB-D** 相机的实时数据采集，采用多线程技术将数据采集与处理解耦，有效降低系统延迟。其次，在实时手语识别方面，系统基于 **MDSI** 算法实现了高效的特征提取与动作识别。在手语动作生成方面，系统实现了基于 **CoordSpeaker** 协同手势生成算法的动作合成引擎。通过预计算与缓存机制，系统显著降低了动作生成的延迟。同时，实现了动作平滑过渡算法，确保生成动作的连贯性与自然性。在交互反馈方面，系统基于 **MDSI** 算法实现了高效准确的手语动作评估，通过分类置信度实时给出评估得分，并基于 **CoordSpeaker** 算法生成参考动作答案，同时给出改进建议与学习激励。同时支持学习进度追踪，用户可查看历史学习记录，并根据进度调整学习计划。

5.5 系统测试与评估



图 5.4 奥比中光 RGB-D 视觉传感器 Gemini 2 正视实物图。

识别准确率 为了评估系统的识别准确性，我们使用商业级奥比中光 RGB-D 相机（图 5.4）构建了自采数据集，并基于此评估了我们的方法。共采集了 12 类数据，每类包含 25 个样本（图 5.5）。采集过程遵循^[17]的设置，采集分辨率为 320×240 ，平均帧数 32，每帧同时包含 RGB 图像和 Depth 图像。




figures/aobi_sample.png

图 5.5 商业 RGB-D 相机自采数据集样本示例

结果显示，所提出的识别算法，在 RGB-D 模态的识别准确率达到了 99.27%，验证了所提出方法在实际应用场景中的高效性与可靠性。

识别推理速度 我们在单张 RTX2080 显卡上对手语识别模块进行了推理速度测试，对同一样本进行了五次测试。如图 5.6 所示，平均推理时间为 83.9ms，快于基线方法^[138]所公布的 96ms。这表明我们的算法模块具有高效的推理能力，能够满足实时交互需求。



figures/time.png

图 5.6 手语识别模块推理时间（单位：ms）

用户感知研究 我们进行了用户研究以评估系统生成的手部动作质量。我们招募了 20 名参与者评估 10 对 9 秒的结果。每个生成结果从两个方面进行评估：(i) 自然度：生成的动作与人类手势相比的真实性和自然度；(ii) 匹配度：生成的动作对给定文本描述的反映准确度。在每次评估环节中，参与者观看不同模型生成的视频片段，并针对每个方面选择表现最佳的方法。如表 5.1 所示，我们的方法具有

表 5.1 手部动作生成结果的用户偏好胜率 (%)。结果表明我们生成的结果被认为更加真实和可控，在自然度和匹配度方面分别优于之前的工作^[31]4.65% 和 1.87%。

	自然度	匹配度
基线 1 ^[30]	20.93	19.38
基线 2 ^[31]	37.21	39.38
本文方法	41.86	41.25

优胜的用户偏好。

5.6 本章小结

本章详细介绍了交互式手语学习助手系统的设计与实现。系统基于本文提出的 MDSI 手势识别算法和 CoordSpeaker 协同手势生成算法，实现了实时手语动作评估与反馈功能。通过模块化设计和高效的系统架构，成功构建了一个具有实时性、准确性和良好交互体验的手语学习辅助系统。未来工作将进一步优化系统性能，扩展学习内容库，提升系统的实用性和适用范围。

第6章 总结与展望

6.1 工作总结

本文针对手语学习中存在的教学资源匮乏、实时评估困难等问题，基于深度学习技术开展了手势识别与生成算法研究，并设计实现了一个交互式手语学习助手系统。主要研究工作及其创新成果如下：

(1) 在多模态手势识别方面，本文提出了一种可插拔的多策略解耦和语义集成网络(MDSI)。该方法通过“姿势-运动”和“时空-通道”特征解耦，有效缓解了RGB-D手势识别中的信息冗余问题。同时引入语义滤波器和标签平滑机制，增强了语义理解能力，实现了对视觉相似手势的精确区分。在IsoGD和THU-READ两个主流基准测试中，MDSI实现了最先进的识别准确率，其中MDSI-CNN相比现有最优方法分别提升了2.48%和4.33%。此外，通过可插拔设计，模型在保持性能的同时，参数量仅占主干网络的6.84%。

(2) 在协同手势生成方面，本文设计了一种基于描述驱动的手势生成框架(CoordSpeaker)。该框架首次引入手势描述生成模块，创新性地解决了手势数据缺乏描述性文本标注的问题。通过统一的运动表示方法和可控的潜在扩散模型，实现了语义和节奏的协同精确控制。大量定量与定性实验结果表明，该方法能够有效生成高质量的(Jerk 0.179 \rightarrow)、语音同步(BC 0.057 \uparrow)、语义相关(MM-Dist: 6.814)的协同手势运动，优于现有方法。此外，通过优化的潜在扩散过程，模型的平均推理时间(AITS)仅为0.842秒，较现有方法提升了6倍以上，显著增强了系统的实用性。

(3) 基于上述算法创新，本文设计并实现了一个交互式手语学习助手系统。该系统采用模块化架构设计，集成了实时手语识别、标准动作生成和交互反馈等功能模块。系统支持手语学习、练习评估和自由练习三种主要交互模式，并提供了清晰的分区布局与友好的交互界面引导。基于商业级RGB-D相机开展的实验证明，系统能够实现高效、可靠的识别评估(12类数据识别准确率>99%，推理时间<0.1s)，并能生成高质量的、用户满意的手部动作(自然度偏好高于同类方法4.65%)。系统的实现验证了本文提出算法的实用价值，为解决手语教学资源匮乏的问题提供了新的技术途径。

本文的主要创新点体现在算法研究、应用实践两个层面。首先，在算法层面，针对手势识别中的“信息冗余”与“信息缺失”挑战提出了“多策略解耦与语义集成手势识别(MDSI)”网络，提升了手势识别的效率与准确性；针对手势生成中的描述

注释缺失和多模态协同控制困难的挑战，设计了基于描述驱动的协同手势生成框架 (CoordSpeaker)，实现了高效与高质量的协同手势生成。其次，在应用层面，本文首次实现了基于深度学习的交互式手语学习系统，针对手语教学中的资源匮乏和实时评估挑战，提出了有效的算法解决方案，为听障人士的手语学习提供了新途径。

6.2 研究展望

尽管本文在手势识别与生成方面取得了一定的研究成果，但仍存在诸多值得深入探讨和改进的方向。未来的研究工作可以从以下几个方面展开：

(1) 手势识别算法方面，目前采用的多分支特征融合策略仍较为简单，主要依赖于分数级别的融合方法。未来可以探索更复杂和精细的多模态联合建模机制，如注意力引导的特征交互、跨模态对比学习等方法，以充分挖掘不同模态信息之间的互补性，进一步提升识别系统的性能和鲁棒性。

(2) 手势生成算法方面，虽然本文提出的手势描述生成模块在一定程度上缓解了手势生成中的语义控制问题，但模型对精细手部动作（如手指细节）的描述和生成能力仍有待提升。未来可以探索基于人体骨架的分层次生成策略，将全身动作、手臂运动和手指姿态分别建模，并设计合理的协同机制，以提升生成结果的精确度和自然度。同时，引入更丰富的上下文信息和情感特征，有望进一步增强生成手势的表现力。

(3) 手语学习系统方面，当前系统仅支持单个手语词汇的识别和生成，在处理复杂手语句子方面仍有较大局限性。未来研究可以围绕手语语法规则建模、上下文语义理解等关键技术展开深入探索，结合自然语言建模技术，逐步实现对连续手语句子的准确识别和流畅生成。此外，系统的交互体验和个性化学习支持也有待进一步优化，如引入自适应学习策略、更丰富的学习模式等，为用户提供更加智能和沉浸式的学习环境。

参考文献

- [1] Guo L, Lu Z, Yao L. Human-machine interaction sensing technology based on hand gesture recognition: A review[J]. IEEE Transactions on Human-Machine Systems, 2021, 51(4): 300-309.
- [2] 伍杰. 基于视觉的实时手势识别方法研究[D]. 大连理工大学, 2019.
- [3] Desai S, Desai A. Human computer interaction through hand gestures for home automation using microsoft kinect[C]//Proceedings of International Conference on Communication and Networks: ComNet 2016. Springer, 2017: 19-29.
- [4] Strickland M, Tremaine J, Brigley G, et al. Using a depth-sensing infrared camera system to access and manipulate medical imaging from within the sterile operating field[J]. Canadian Journal of Surgery, 2013, 56(3): E1.
- [5] Rautaray S S, Agrawal A. Vision based hand gesture recognition for human computer interaction: a survey[J]. Artificial intelligence review, 2015, 43: 1-54.
- [6] Zhou B, Wang P, Wan J, et al. A unified multimodal de-and re-coupling framework for rgb-d motion recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.
- [7] Li X, Hou Y, Wang P, et al. Trear: Transformer-based rgb-d egocentric action recognition[J]. IEEE Transactions on Cognitive and Developmental Systems, 2021, 14(1): 246-252.
- [8] Zhu G, Zhang L, Yang L, et al. Redundancy and attention in convolutional lstm for gesture recognition[J]. IEEE transactions on neural networks and learning systems, 2019, 31(4): 1323-1335.
- [9] Zhu G, Zhang L, Shen P, et al. Continuous gesture segmentation and recognition using 3dcnn and convolutional lstm[J]. IEEE Transactions on Multimedia, 2018, 21(4): 1011-1021.
- [10] Zhou B, Wang P, Wan J, et al. Decoupling and recoupling spatiotemporal representation for rgb-d-based motion recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 20154-20163.
- [11] Narayana P, Beveridge R, Draper B A. Gesture recognition: Focus on the hands[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5235-5244.
- [12] Yu Z, Zhou B, Wan J, et al. Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition[J]. IEEE Transactions on Image Processing, 2021, 30: 5626-5640.
- [13] Zuo R, Wei F, Mak B. Natural language-assisted sign language recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 14890-14900.
- [14] Kevin N, Ranganath S, Ghosh D. Trajectory modeling in gesture recognition using cybergloves and magnetic trackers[C]//TENCON 2004. IEEE Region 10 Conference. 2004: 571-574.
- [15] Softkinetic iisu sdk[M/OL]. 2012. <http://www.softkinetic.com/Solutions/iisuSDK.aspx>.
- [16] Li Y, Wei G, Desrosiers C, et al. Decoupled and boosted learning for skeleton-based dynamic hand gesture recognition[J]. Pattern Recognition, 2024, 153: 110536.

-
- [17] Wan J, Zhao Y, Zhou S, et al. Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2016: 56-64.
 - [18] Yang S, Wu Z, Li M, et al. Diffusestylegesture: Stylized audio-driven co-speech gesture generation with diffusion models[A]. 2023.
 - [19] Yang S, Wang Z, Wu Z, et al. Unifiedgesture: A unified gesture synthesis model for multiple skeletons[C]//Proceedings of the 31st ACM International Conference on Multimedia. 2023: 1033-1044.
 - [20] Xu Z, Lin Y, Han H, et al. Mambataalk: Efficient holistic gesture synthesis with selective state space models[J]. Advances in Neural Information Processing Systems, 2025, 37: 20055-20080.
 - [21] Zhi Y, Cun X, Chen X, et al. Livelyspeaker: Towards semantic-aware co-speech gesture generation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 20807-20817.
 - [22] Pang K, Qin D, Fan Y, et al. Bodyformer: Semantics-guided 3d body gesture synthesis with transformer[J]. ACM Transactions on Graphics (TOG), 2023, 42(4): 1-12.
 - [23] Liu H, Zhu Z, Becherini G, et al. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 1144-1154.
 - [24] Qi X, Liu C, Li L, et al. Emotiongesture: Audio-driven diverse emotional co-speech 3d gesture generation[J]. IEEE Transactions on Multimedia, 2024.
 - [25] Qi X, Pan J, Li P, et al. Weakly-supervised emotion transition learning for diverse 3d co-speech gesture generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 10424-10434.
 - [26] Ao T, Zhang Z, Liu L. Gesturediffuclip: Gesture diffusion model with clip latents[J]. ACM Transactions on Graphics (TOG), 2023, 42(4): 1-18.
 - [27] Ghorbani S, Ferstl Y, Holden D, et al. Zeroeggs: Zero-shot example-based gesture generation from speech[C]//Computer Graphics Forum: Vol. 42. Wiley Online Library, 2023: 206-216.
 - [28] Yang S, Xue H, Zhang Z, et al. The diffusestylegesture+ entry to the genea challenge 2023[C]//Proceedings of the 25th International Conference on Multimodal Interaction. 2023: 779-785.
 - [29] Cheng Q, Li X, Fu X. Siggester: Generalized co-speech gesture synthesis via semantic injection with large-scale pre-training diffusion models[C]//SIGGRAPH Asia 2024 Conference Papers. 2024: 1-11.
 - [30] Yang S, Xu Z, Xue H, et al. Freetalker: Controllable speech and text-driven gesture generation based on diffusion models for enhanced speaker naturalness[C]//ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024: 7945-7949.
 - [31] Chen B, Li Y, Ding Y X, et al. Enabling synergistic full-body control in prompt-based co-speech motion generation[C]//Proceedings of the 32nd ACM International Conference on Multimedia. 2024: 6774-6783.

-
- [32] Kaâniche M. Gesture recognition from video sequences[D]. Université Nice Sophia Antipolis, 2009.
- [33] Ottenheimer H J, Pine J M. The anthropology of language: An introduction to linguistic anthropology[M]. Cengage Learning, 2018.
- [34] Oudah M, Al-Naji A, Chahl J. Hand gesture recognition based on computer vision: a review of techniques[J]. *Journal of Imaging*, 2020, 6(8): 73.
- [35] 解迎刚, 王全. 基于视觉的动态手势识别研究综述[J]. *Journal of Computer Engineering & Applications*, 2021, 57(22).
- [36] Sigal L, Sclaroff S, Athitsos V. Skin color-based video segmentation under time-varying illumination[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, 26(7): 862-877.
- [37] Chen Q, Georganas N D, Petriu E M. Real-time vision-based hand gesture recognition using haar-like features[C]//2007 IEEE instrumentation & measurement technology conference IMTC 2007. IEEE, 2007: 1-6.
- [38] Tekin B, Bogo F, Pollefeys M. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4511-4520.
- [39] Pun C M, Zhu H M, Feng W. Real-time hand gesture recognition using motion tracking[J]. *International Journal of Computational Intelligence Systems*, 2011, 4(2): 277-286.
- [40] Jiang S, Sun B, Wang L, et al. Skeleton aware multi-modal sign language recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 3413-3423.
- [41] Crowley J, Berard F, Coutaz J, et al. Finger tracking as an input device for augmented reality [C]//International Workshop on Gesture and Face Recognition. 1995: 195-200.
- [42] Argyros A A, Lourakis M I. Real-time tracking of multiple skin-colored objects with a possibly moving camera[C]//Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part III 8. Springer, 2004: 368-379.
- [43] Kalman R E. A new approach to linear filtering and prediction problems[Z]. 1960.
- [44] Pérez P, Hue C, Vermaak J, et al. Color-based probabilistic tracking[C]//Computer Vision—ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part I 7. Springer, 2002: 661-675.
- [45] Wang X, Li X. The study of movingtarget tracking based on kalman-camshift in the video[C]//The 2nd International Conference on Information Science and Engineering. IEEE, 2010: 1-4.
- [46] Burges C J. A tutorial on support vector machines for pattern recognition[J]. *Data mining and knowledge discovery*, 1998, 2(2): 121-167.
- [47] Thirumuruganathan S. A detailed introduction to k-nearest neighbor (knn) algorithm[EB/OL]. 2010. <http://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/>.

-
- [48] Liang R H, Ouhyoung M. A sign language recognition system using hidden markov model and context sensitive search[C]//Proceedings of the ACM symposium on virtual reality software and technology. 1996: 59-66.
 - [49] Corradini A. Dynamic time warping for off-line recognition of a small gesture vocabulary[C]//Proceedings IEEE ICCV workshop on recognition, analysis, and tracking of faces and gestures in real-time systems. IEEE, 2001: 82-89.
 - [50] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014: 1725-1732.
 - [51] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos [J]. Advances in neural information processing systems, 2014, 27.
 - [52] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]//European conference on computer vision. Springer, 2016: 20-36.
 - [53] Zhou B, Andonian A, Oliva A, et al. Temporal relational reasoning in videos[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 803-818.
 - [54] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2015: 4489-4497.
 - [55] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset [C]//proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6299-6308.
 - [56] Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 6450-6459.
 - [57] Xie S, Sun C, Huang J, et al. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 305-321.
 - [58] Zhang Y, Shi L, Wu Y, et al. Gesture recognition based on deep deformable 3d convolutional neural networks[J]. Pattern Recognition, 2020, 107: 107416.
 - [59] Jozefowicz R, Zaremba W, Sutskever I. An empirical exploration of recurrent network architectures[C]//International conference on machine learning. PMLR, 2015: 2342-2350.
 - [60] Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: Deep networks for video classification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 4694-4702.
 - [61] Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 2625-2634.
 - [62] Molchanov P, Yang X, Gupta S, et al. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 4207-4215.

-
- [63] Cao C, Zhang Y, Wu Y, et al. Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules[C]//Proceedings of the IEEE international conference on computer vision. 2017: 3763-3771.
- [64] Shi X, Chen Z, Wang H, et al. Convolutional lstm network: A machine learning approach for precipitation nowcasting[J]. Advances in neural information processing systems, 2015, 28.
- [65] Zhu G, Zhang L, Shen P, et al. Multimodal gesture recognition using 3-d convolution and convolutional lstm[J]. Ieee Access, 2017, 5: 4517-4524.
- [66] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.
- [67] Zhang L, Zhu G, Shen P, et al. Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition[C]//Proceedings of the IEEE international conference on computer vision workshops. 2017: 3120-3128.
- [68] Zhang L, Zhu G, Mei L, et al. Attention in convolutional lstm for gesture recognition[J]. Advances in neural information processing systems, 2018, 31.
- [69] Lin C, Wan J, Liang Y, et al. Large-scale isolated gesture recognition using a refined fused model based on masked res-c3d network and skeleton lstm[C]//2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, 2018: 52-58.
- [70] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
- [71] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[A]. 2021. arXiv: 2010.11929.
- [72] Selva J, Johansen A S, Escalera S, et al. Video transformers: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.
- [73] Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding? [C]//ICML: Vol. 2. 2021: 4.
- [74] Arnab A, Dehghani M, Heigold G, et al. Vivit: A video vision transformer[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 6836-6846.
- [75] Neimark D, Bar O, Zohar M, et al. Video transformer network[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 3163-3172.
- [76] Khan S, Naseer M, Hayat M, et al. Transformers in vision: A survey[J]. ACM computing surveys (CSUR), 2022, 54(10s): 1-41.
- [77] Guo F, He Z, Zhang S, et al. Normalized edge convolutional networks for skeleton-based hand gesture recognition[J]. Pattern Recognition, 2021, 118: 108044.
- [78] Liu J, Liu Y, Wang Y, et al. Decoupled representation learning for skeleton-based gesture recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 5751-5760.
- [79] Bigalke A, Heinrich M P. Fusing posture and position representations for point cloud-based hand gesture recognition[C]//2021 International Conference on 3D Vision (3DV). IEEE, 2021: 617-626.

-
- [80] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PMLR, 2021: 8748-8763.
- [81] Li J, Li D, Xiong C, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//International Conference on Machine Learning. PMLR, 2022: 12888-12900.
- [82] Gao P, Geng S, Zhang R, et al. Clip-adapter: Better vision-language models with feature adapters[J]. International Journal of Computer Vision, 2024, 132(2): 581-595.
- [83] Nyatsanga S, Kucherenko T, Ahuja C, et al. A comprehensive review of data-driven co-speech gesture generation[C]//Computer Graphics Forum: Vol. 42. Wiley Online Library, 2023: 569-596.
- [84] Liu H, Zhu Z, Iwamoto N, et al. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis[C]//European conference on computer vision. Springer, 2022: 612-630.
- [85] Yoon Y, Cha B, Lee J H, et al. Speech gesture generation from the trimodal context of text, audio, and speaker identity[J]. ACM Transactions on Graphics (TOG), 2020, 39(6): 1-16.
- [86] Yang S, Wu Z, Li M, et al. Qpgesture: Quantization-based and phase-guided motion matching for natural speech-driven gesture generation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 2321-2330.
- [87] Yi H, Liang H, Liu Y, et al. Generating holistic 3d human motion from speech[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 469-480.
- [88] Alexanderson S, Nagy R, Beskow J, et al. Listen, denoise, action! audio-driven motion synthesis with diffusion models[J]. ACM Transactions on Graphics (TOG), 2023, 42(4): 1-20.
- [89] Tevet G, Raab S, Gordon B, et al. Human motion diffusion model[A]. 2022.
- [90] Zhang M, Cai Z, Pan L, et al. Motiondiffuse: Text-driven human motion generation with diffusion model[J]. IEEE transactions on pattern analysis and machine intelligence, 2024, 46(6): 4115-4128.
- [91] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 10684-10695.
- [92] Jiang B, Chen X, Liu W, et al. Motiongpt: Human motion as a foreign language[J]. Advances in Neural Information Processing Systems, 2024, 36.
- [93] Guo C, Zou S, Zuo X, et al. Generating diverse and natural 3d human motions from text[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 5152-5161.
- [94] Guo C, Zuo X, Wang S, et al. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts[C]//European Conference on Computer Vision. Springer, 2022: 580-597.
- [95] Tevet G, Gordon B, Hertz A, et al. Motionclip: Exposing human motion generation to clip space[C]//European Conference on Computer Vision. Springer, 2022: 358-374.

-
- [96] Jiang B, Chen X, Zhang C, et al. Motionchain: Conversational motion controllers via multi-modal prompts[C]//European Conference on Computer Vision. Springer, 2024: 54-74.
- [97] Chen X, Jiang B, Liu W, et al. Executing your commands via motion diffusion in latent space [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 18000-18010.
- [98] Dai W, Chen L H, Wang J, et al. Motionlcm: Real-time controllable motion generation via latent consistency model[C]//European Conference on Computer Vision. Springer, 2024: 390-408.
- [99] Takano W, Nakamura Y. Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions[J]. The International Journal of Robotics Research, 2015, 34(10): 1314-1328.
- [100] Yamada T, Matsunaga H, Ogata T. Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions[J]. IEEE Robotics and Automation Letters, 2018, 3(4): 3441-3448.
- [101] Luo M, Hou R, Li Z, et al. M³ gpt: An advanced multimodal, multitask framework for motion comprehension and generation[A]. 2024.
- [102] Al Mahmud J, Das B C, Shin J, et al. 3d gesture recognition and adaptation for human-robot interaction[J]. IEEE Access, 2022, 10: 116485-116513.
- [103] Parada-Loira F, González-Agulla E, Alba-Castro J L. Hand gestures to control infotainment equipment in cars[C]//2014 IEEE Intelligent Vehicles Symposium Proceedings. IEEE, 2014: 1-6.
- [104] Manawadu U E, Kamezaki M, Ishikawa M, et al. A hand gesture based driver-vehicle interface to control lateral and longitudinal motions of an autonomous vehicle[C]//2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2016: 001785-001790.
- [105] Melnick K. Prototype ar app translates sign language in real time[EB/OL]. <https://vrscout.com/projects/prototype-ar-app-translates-sign-language/>.
- [106] Borghino D. Augmented reality glasses perform real-time language translation[EB/OL]. <https://newatlas.com/language-translating-glasses/23494/>.
- [107] Taylor J, Bordeaux L, Cashman T, et al. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences[J]. ACM Transactions on Graphics (ToG), 2016, 35(4): 1-12.
- [108] Zeng J, Sun Y, Wang F. A natural hand gesture system for intelligent human-computer interaction and medical assistance[C]//2012 Third Global Congress on Intelligent Systems. IEEE, 2012: 382-385.
- [109] Starner T, Weaver J, Pentland A. Real-time american sign language recognition using desk and wearable computer based video[J]. IEEE Transactions on pattern analysis and machine intelligence, 1998, 20(12): 1371-1375.
- [110] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5693-5703.

-
- [111] 智东西. 昇腾 AI 的温度: 关爱超 2700 万听障者, 破解手语学习难题[EB/OL]. https://mp.weixin.qq.com/s?__biz=MzA4MTQ4NjQzMw==&mid=2652752384&idx=1&sn=6be9fbb333354425292f7776f0069856&chksm=847d5d0eb30ad418ee4bd601a7810e108eacaf3ab518a4e9ce31c6c347c86a36648c861c9021&scene=27#wechat_redirect.
 - [112] Kingma D P, Welling M, et al. Auto-encoding variational bayes[M]. Banff, Canada, 2013.
 - [113] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.
 - [114] Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis[J]. Advances in neural information processing systems, 2021, 34: 8780-8794.
 - [115] Ho J, Salimans T. Classifier-free diffusion guidance[A]. 2022.
 - [116] Jin S, Xu L, Xu J, et al. Whole-body human pose estimation in the wild[C]//Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. Springer, 2020: 196-214.
 - [117] Avola D, Cinque L, Fagioli A, et al. 3d hand pose and shape estimation from rgb images for keypoint-based hand gesture recognition[J]. Pattern Recognition, 2022, 129: 108762.
 - [118] Zhou B, Li Y, Wan J. Regional attention with architecture-rebuilt 3d network for rgb-d gesture recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 35. 2021: 3563-3571.
 - [119] Chen H, Li Y, Fang H, et al. Multi-scale attention 3d convolutional network for multimodal gesture recognition[J]. Sensors, 2022, 22(6): 2405.
 - [120] Liu Z, Chai X, Liu Z, et al. Continuous gesture recognition with hand-oriented spatiotemporal feature[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. 2017: 3056-3064.
 - [121] Wan J, Lin C, Wen L, et al. Chalearn looking at people: Isogd and congld large-scale rgb-d gesture recognition[J]. IEEE Transactions on Cybernetics, 2020, 52(5): 3422-3433.
 - [122] Tang Y, Tian Y, Lu J, et al. Action recognition in rgb-d egocentric videos[C]//2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017: 3410-3414.
 - [123] Yang B, Bender G, Le Q V, et al. Condconv: Conditionally parameterized convolutions for efficient inference[J]. Advances in neural information processing systems, 2019, 32.
 - [124] Chen Y, Dai X, Liu M, et al. Dynamic convolution: Attention over convolution kernels[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11030-11039.
 - [125] He T, Zhang Z, Zhang H, et al. Bag of tricks for image classification with convolutional neural networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 558-567.
 - [126] Materzynska J, Berger G, Bax I, et al. The jester dataset: A large-scale video dataset of human gestures[C]//Proceedings of the IEEE/CVF international conference on computer vision workshops. 2019: 0-0.

-
- [127] Ma Y, Zhou B, Wang R, et al. Multi-stage factorized spatio-temporal representation for rgb-d action and gesture recognition[C]//Proceedings of the 31st ACM International Conference on Multimedia. 2023: 3149-3160.
- [128] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [A]. 2014.
- [129] Feichtenhofer C, Fan H, Malik J, et al. Slowfast networks for video recognition[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6202-6211.
- [130] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]//European conference on computer vision. Springer, 2016: 20-36.
- [131] Tang Y, Wang Z, Lu J, et al. Multi-stream deep neural networks for rgb-d egocentric action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 29 (10): 3001-3015.
- [132] Loper M, Mahmood N, Romero J, et al. Smpl: A skinned multi-person linear model[M]//Seminal Graphics Papers: Pushing the Boundaries, Volume 2. 2023: 851-866.
- [133] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PMLR, 2021: 8748-8763.
- [134] Chen S, Wang C, Chen Z, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing[J]. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6): 1505-1518.
- [135] Kucherenko T, Hasegawa D, Henter G E, et al. Analyzing input and output representations for speech-driven gesture generation[C]//Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents. 2019: 97-104.
- [136] Zhang J, Zhang Y, Cun X, et al. Generating human motion from textual descriptions with discrete representations[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 14730-14740.
- [137] Guo C, Mu Y, Javed M G, et al. Momask: Generative masked modeling of 3d human motions [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 1900-1910.
- [138] 范桂双. 基于 S3D+ BiConvLSTM+ MobileNet 的动态手势识别算法研究[D]. 哈尔滨工业大学, 2020.

致 谢

感谢导师杨文明副教授对本人的精心指导。

感谢同门的帮助。

感谢学校、学院和中心提供的平台和环境。

感谢我的父母对我的爱和支持。

感谢我爱的人和爱我的人。

感谢生命。

感谢我自己。

声 明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签 名：_____ 日 期：_____

个人简历、在学期间完成的相关学术成果

个人简历

2000 年 08 月 29 日出生于山东省济南市。

2018 年 9 月考入厦门大学信息学院软件工程系数字媒体技术专业，2022 年 6 月以优秀毕业生身份毕业，获得工学学士学位。

2022 年 9 月免试进入清华大学深圳国际研究生院 Open Fiesta 攻读电子信息工程硕士至今。

在学期间完成的相关学术成果

学术论文：

- [1] Fang F, Liao Z, Kan Z, Yang W, et al. MDSI: Pluggable Multi-strategy Decoupling with Semantic Integration for RGB-D Gesture Recognition [J]. Pattern Recognition. (Minor Reivision)
- [2] Fang F, Yang S, Yang W. CoordSpeaker: Exploiting Gesture Captioning for Co-ordinated Caption-Empowered Co-Speech Gesture Generation [C]. In International Conference on Computer Vision (ICCV). (Under Review)

指导教师评语

论文提出了……

答辩委员会决议书

论文提出了……

论文取得的主要创新性成果包括：

1. ……

2. ……

3. ……

论文工作表明作者在 ××××× 具有 ××××× 知识，具有 ×××× 能力，论文 ××××，
答辩 ××××。

答辩委员会表决，（× 票/一致）同意通过论文答辩，并建议授予 ×××（姓名）
×××（门类）学博士/硕士学位。