



CS 330 MIP – Lab 04

多媒体信息处理介绍实验课 - 如何做项目3

Multimedia Information Processing Introduction - Projects

Jimmy Liu 刘江

2025-03-12

Homework 03

1

读至少3片项目相关文章，做个PPT讲述别人在这个方向的做法。

每组组长提交一份即可。



Team 1:基于MOOC的个性化教育智能体 1

Homework 02

基于MOOC的个性化教育智能体
助教：王星月老师

1



Homework 03

1

读至少3片项目相关文章，做个PPT讲述别人在这个方向的做法。

每组组长提交一份即可。

Homework 03

一、模型训练——预训练模型BERT

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova

Google AI Language

{jacobdevlin, mingweichang, kentonl, kristout}@google.com

Homework 03

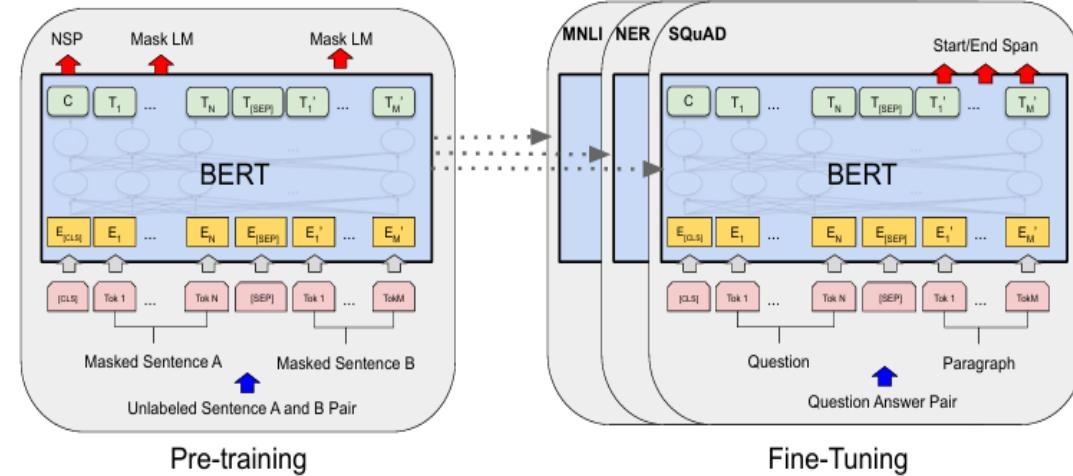
一、模型训练——预训练模型BERT

BERT (Bidirectional Encoder Representations from Transformers) 提出了一种新的预训练方法，通过遮蔽语言模型（Masked LM）和下一句预测（Next Sentence Prediction, NSP）任务，实现双向上下文建模。

Homework 03

一、模型训练——预训练模型BERT

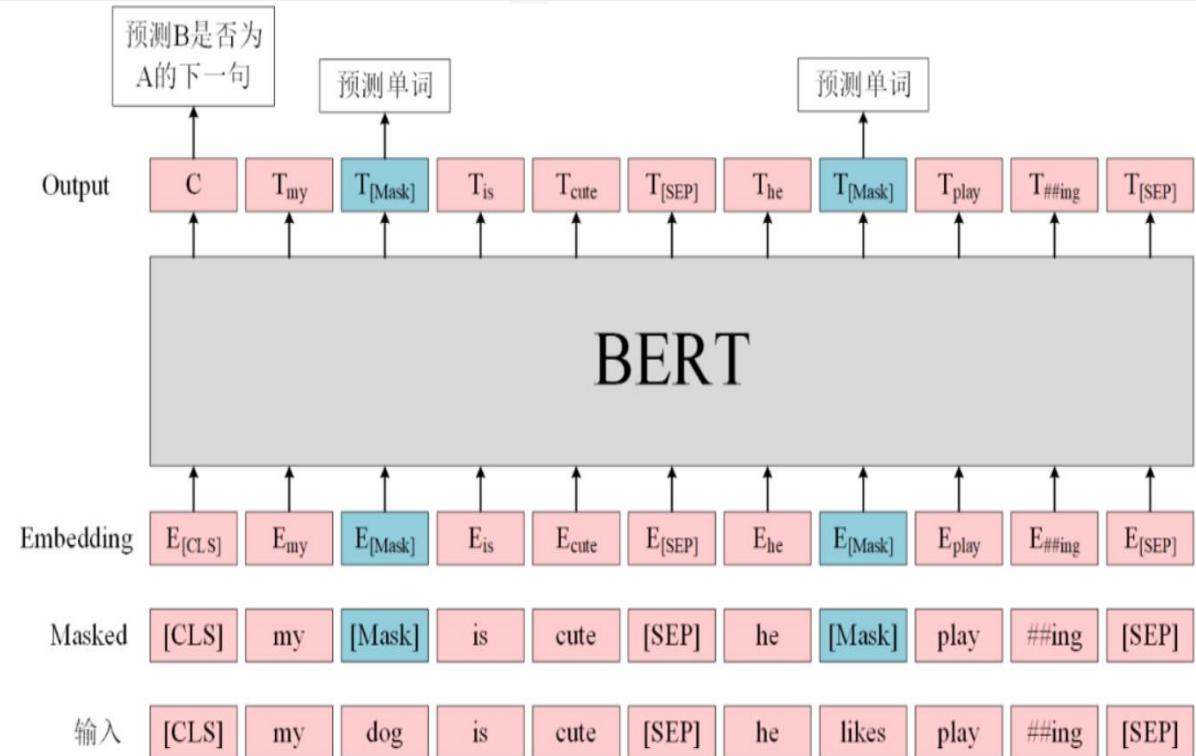
BERT的训练包含pre-train和fine-tune两个阶段。pre-train阶段模型是在无标注的标签数据上进行训练，fine-tune阶段，BERT模型首先是被pre-train模型参数初始化，然后所有的参数会用下游的有标注的数据进行训练



Homework 03

一、模型训练——遮蔽语言模型 (Masked LM)
随机遮蔽输入序列中的部分词，并让模型预测被遮蔽的词。

为了减少预训练和微调之间的不匹配，80%的时间用[MASK]替换遮蔽词，10%的时间用随机词替换，10%的时间保留原词。



Homework 03

一、模型训练——下一句预测（NSP）

Next Sentence Prediction (NSP) 的任务是判断句子B是否是句子A的下文。如果是的话输出'IsNext'，否则输出'NotNext'。

Homework 03

二、问题转义——文本匹配/语义匹配

先检索出一批相似问题再精排

1. 检索
 - a. BM25 (Best Matching 25)
 - b. SBERT
2. 精排
 - a. Cross-Encoders

Homework 03

二、问题转义——文本匹配/语义匹配

BM25 (Best Matching 25)

BM25 基于 TF-IDF (Term Frequency-Inverse Document Frequency) 的思想，但对其进行了改进以考虑文档的长度等因素。

BM25 算法的实现通常用于排序文档，使得与查询更相关的文档排名更靠前。在信息检索领域，BM25 已经成为一个经典的算法。

Homework 03

二、问题转义——文本匹配/语义匹配

BM25 (Best Matching 25)

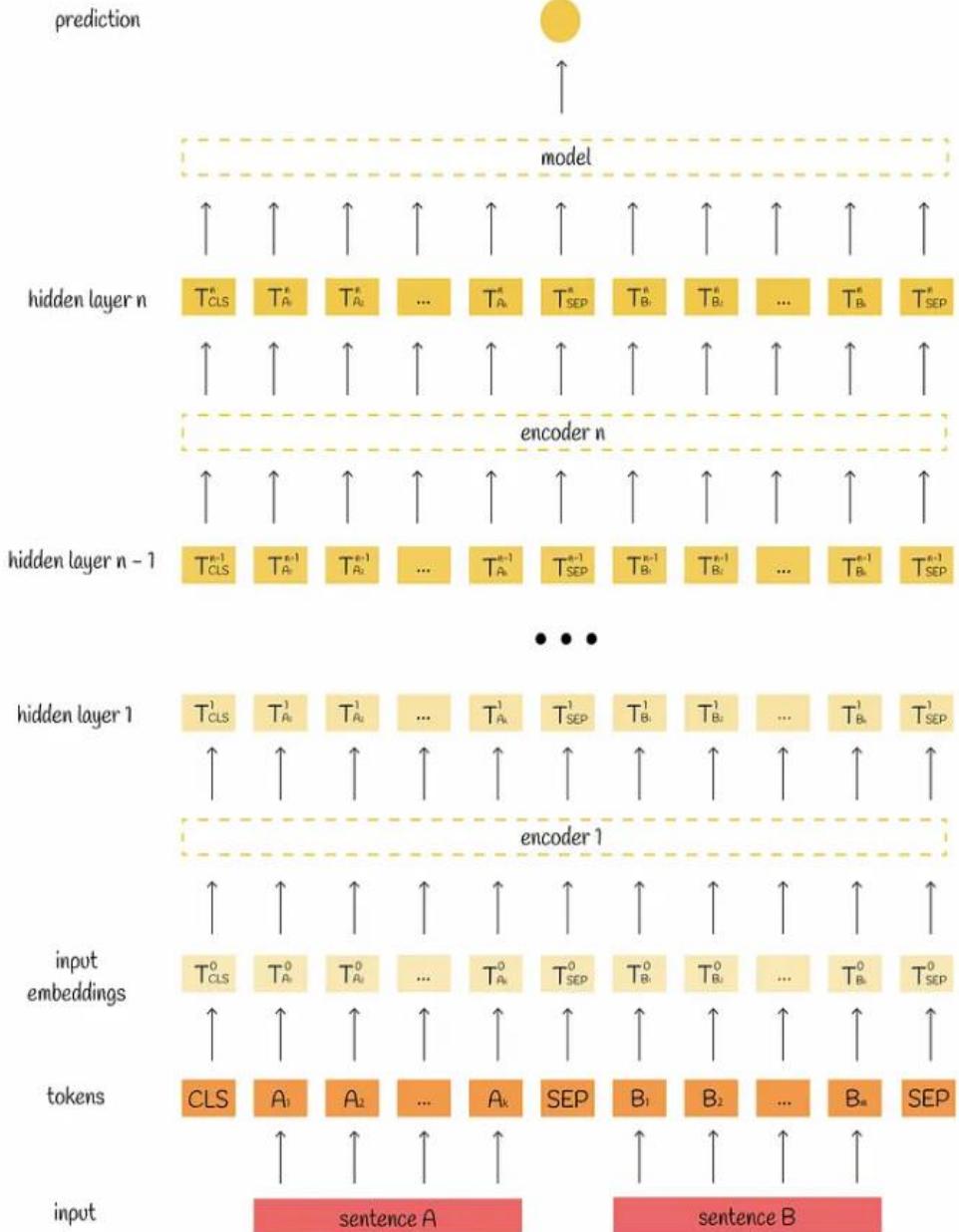
- n 是查询中的词项数。
- q_i 是查询中的第 i 个词项。
- $\text{IDF}(q_i)$ 是逆文档频率，计算方式通常是 $\log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}$ ，其中 N 是文档总数， $n(q_i)$ 是包含词项 q_i 的文档数。
- $f(q_i, D)$ 是词项 q_i 在文档 D 中的出现次数 (TF)。
- $\text{len}(D)$ 是文档 D 的长度。
- avg_len 是所有文档的平均长度。
- k_1 和 b 是调整参数，通常设置为 $k_1 = 1.5$ 和 $b = 0.75$ 。

$$\text{BM25}(D, Q) = \sum_{i=1}^n \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{\text{len}(D)}{\text{avg_len}}\right)}$$

Homework 03

二、问题转义——文本匹配/语义匹配 BERT

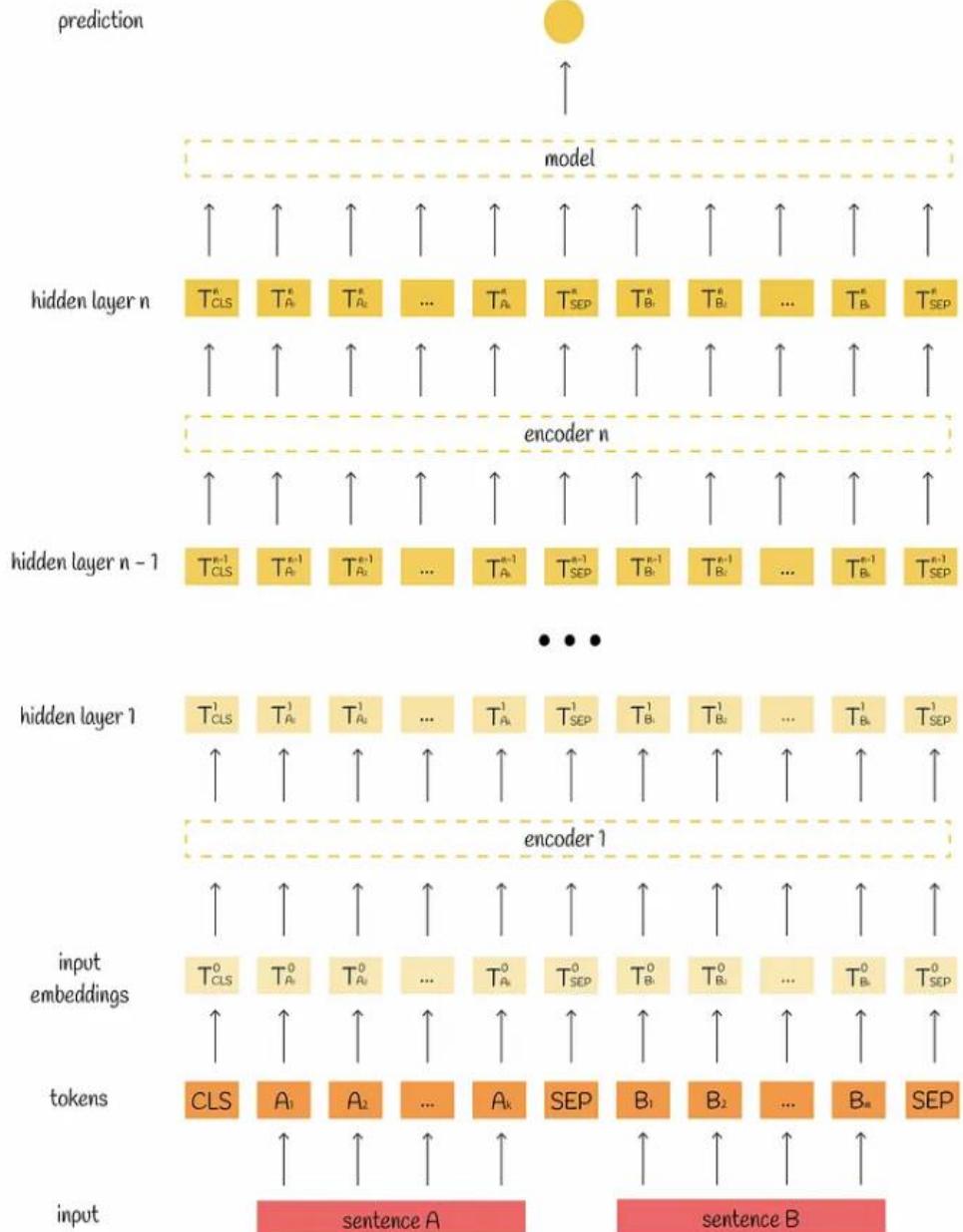
作为输入，它需要一个 [CLS] 标记和由特殊 [SEP] 标记分隔的两个句子。根据模型配置，该信息由多头注意力模块处理 12 或 24 次。然后，输出被聚合并传递到一个简单的回归模型以获得最终标签



Homework 03

二、问题转义——文本匹配/语义匹配 BERT

这会导致推理过程中出现二次复杂度。
例如，处理 $n = 10\,000$ 个句子需要
 $n * (n - 1) / 2 = 49\,995\,000$ 次推理
BERT 计算，这并不是真正可扩展的

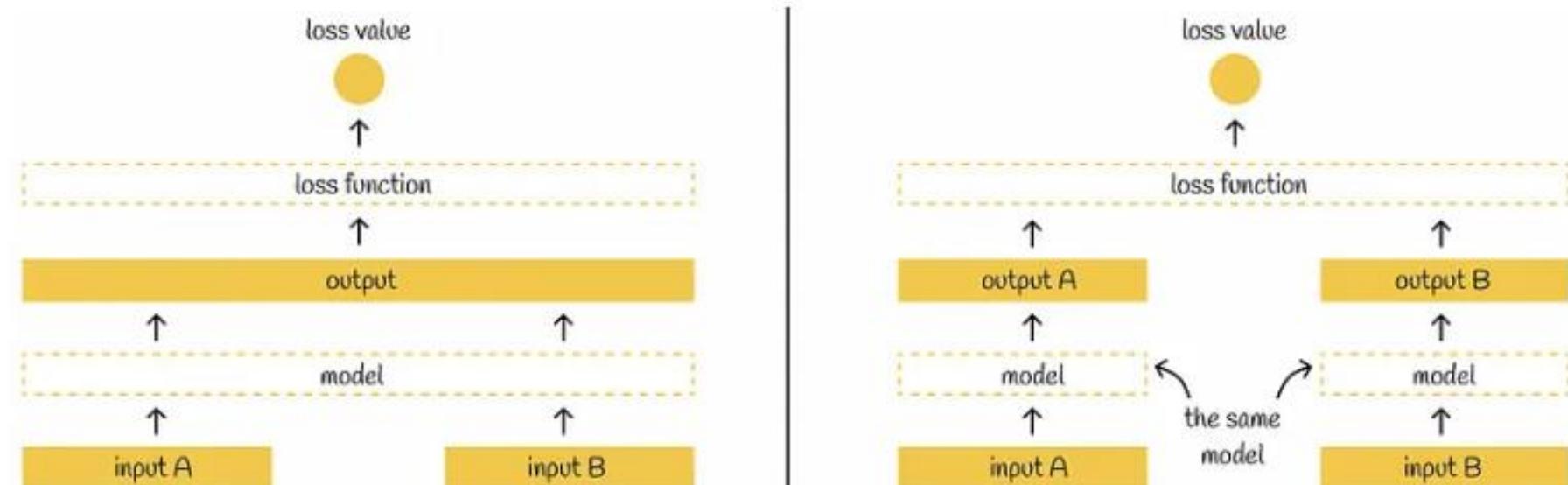


Homework 03

二、问题转义——文本匹配/语义匹配

SBERT

SBERT 引入了 Siamese 网络概念，这意味着每次两个句子都通过相同的 BERT 模型独立传递。

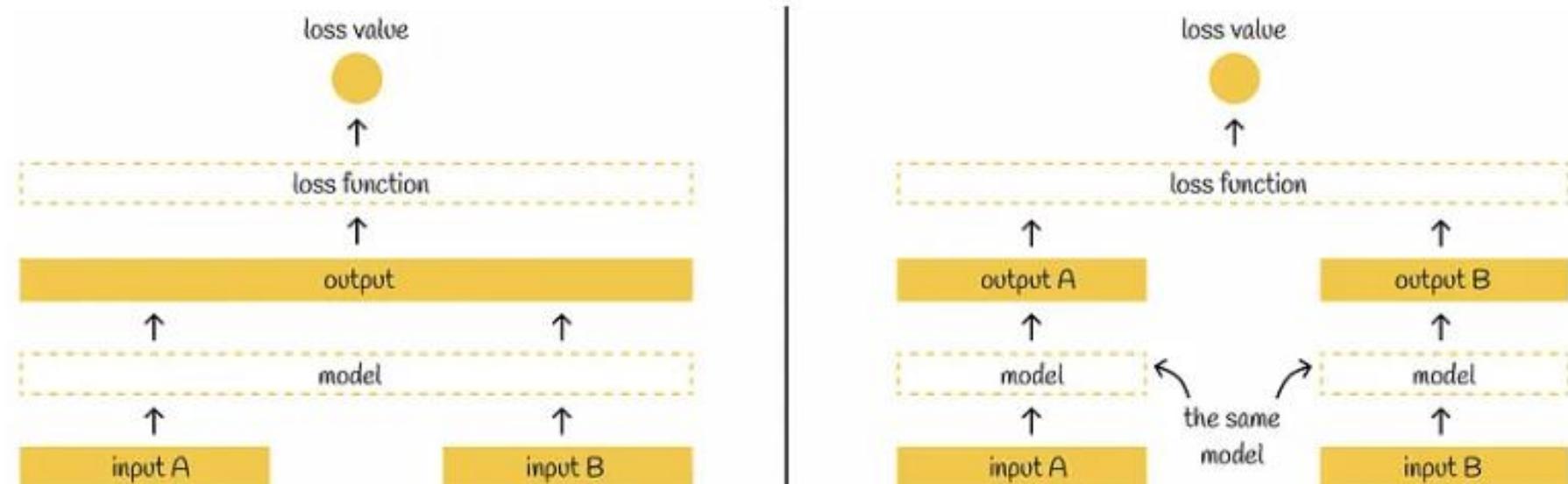


Homework 03

二、问题转义——文本匹配/语义匹配

SBERT

通过将 BERT 推理执行的二次次数减少为线性，SBERT 在保持高精度的同时实现了速度的大幅增长。

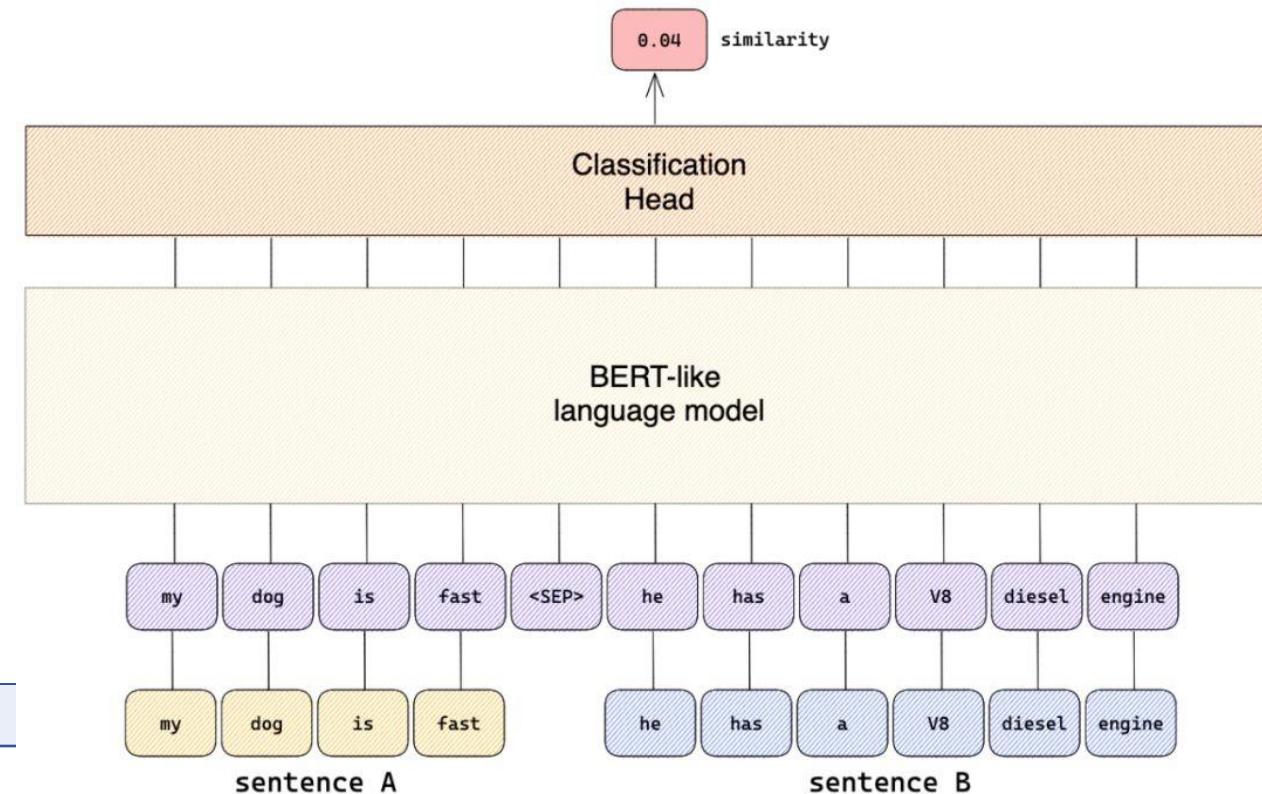


Homework 03

二、问题转义——文本匹配/语义匹配

Cross-Encoder

从本质上讲，交叉编码器所做的
是将两个句子通过分隔符<SEP>
拼接起来，并将其“喂进”一个语
言模型。在语言模型的顶部有一
个分类头，用以训练来预测一个
目标“相似度”数值。



Homework 03

二、问题转义——文本匹配/语义匹配

文献：A Hybrid Approach to Information Retrieval and Answer Generation for Regulatory Texts

A Hybrid Approach to Information Retrieval and Answer Generation for Regulatory Texts

Jhon Rayo

Universidad de los Andes
Bogotá, Colombia

j.rayom@uniandes.edu.co

Raúl de la Rosa

Universidad de los Andes
Bogotá, Colombia

c.delarosap@uniandes.edu.co

Mario Garrido

Universidad de los Andes
Bogotá, Colombia

m.garrido10@uniandes.edu.co

Homework 03

二、问题转义——文本匹配/语义匹配

文献：A Hybrid Approach to Information Retrieval and Answer Generation for Regulatory Texts

法规文本往往篇幅长、语言复杂 —— 传统的信息检索（IR）方法难以有效支持合规任务

BM25 —— 高效，但难以处理同义词、专业术语和语义匹配的问题
语义检索（Semantic Search）—— 能通过向量嵌入表示文本意义，
但也可能会忽略一些关键的词匹配信息

Homework 03

二、问题转义——文本匹配/语义匹配

文献：A Hybrid Approach to Information Retrieval and Answer Generation for Regulatory Texts

混合检索系统（Hybrid Retrieval System）

结合 BM25 和 微调的句子 Transformer 模型 以提高检索准确性，并使用 检索增强生成（RAG） 技术结合大模型生成合规答案

$$\text{Score} = \alpha \cdot \text{Semantic Score} + (1 - \alpha) \cdot \text{Lexical Score}$$

其中 $\alpha=0.65$ ，优先考虑语义匹配，同时保留一定的词法匹配能力。

Homework 03

二、问题转义——文本匹配/语义匹配

文献：基于知识图谱的问答系统研究与实现

南京邮電大學
专业学位硕士论文



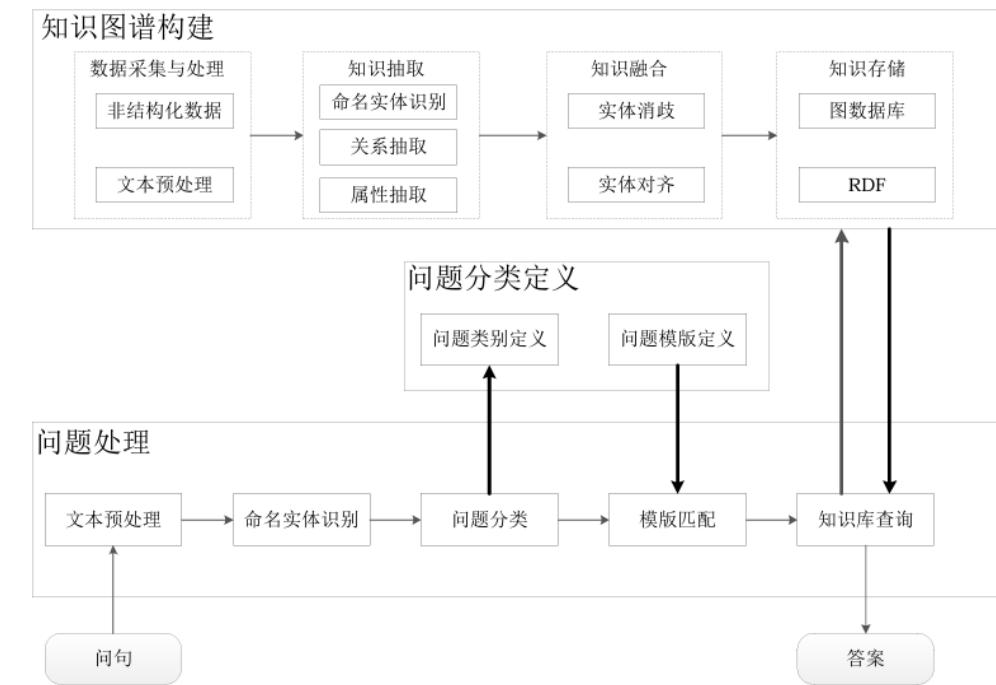
论文题目：基于知识图谱的问答系统研究与实现

学号 1219043607
姓名 李飞
导师 章韵
专业学位类别 工程硕士
类型 全日制
专业（领域） 计算机技术
论文提交日期 2022.03

Homework 03

二、问题转义——文本匹配/语义匹配 文献：基于知识图谱的问答系统研究与实现

- 命名实体识别 (NER)
- 问题分类

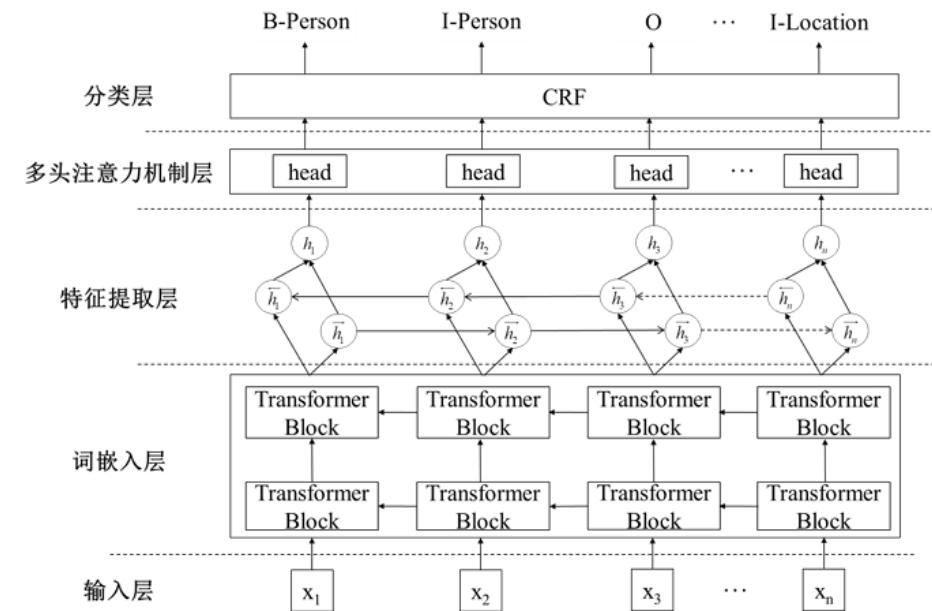


Homework 03

二、问题转义——文本匹配/语义匹配 文献：基于知识图谱的问答系统研究与实现

命名实体识别：

1. 输入层
文本预处理，分词并转化为字向量
2. 词嵌入层
使用BERT模型将字向量转化为词向量
3. 特征提取层
使用BiLSTM模型进行特征提取，
并为对应标签计算得分
4. 多头注意力机制层
将特征向量分配给不同注意力头，
最终使词向量表达能作用于整个句子
5. 分类层
使用CRF生成最符合的标注序列



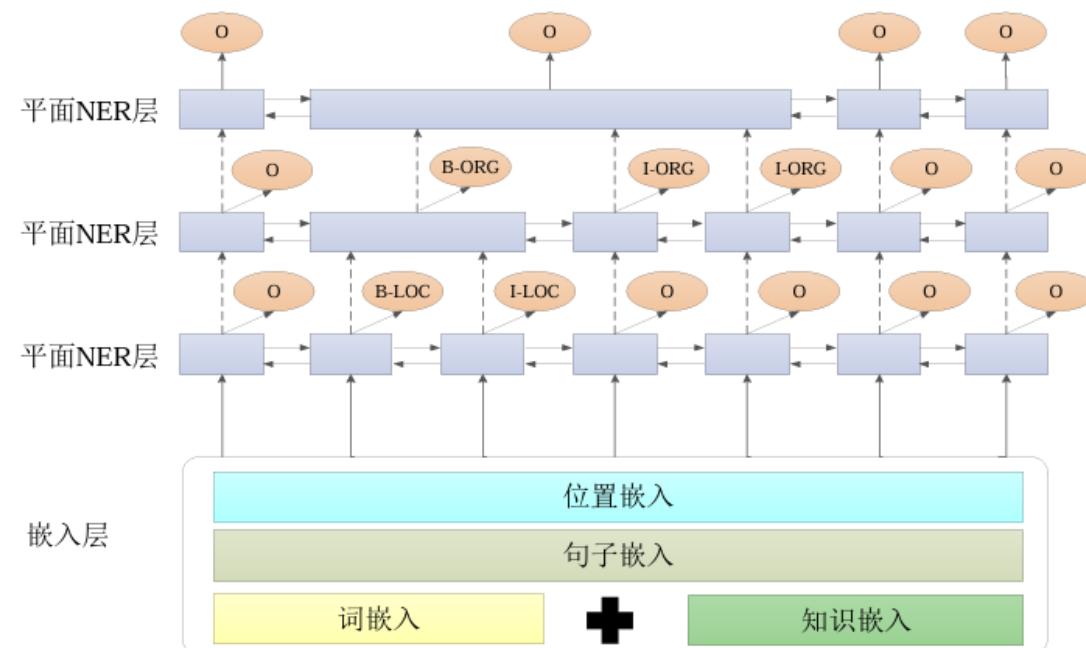
Homework 03

二、问题转义——文本匹配/语义匹配
文献：基于知识图谱的问答系统研究与实现

基于知识嵌入改进的多层次嵌套命名实体识别

解决传统命名实体识别方法无法处理多名词复合形成的嵌套名词实体的问题

- 知识嵌入：
把token对应的知识表示为低维度向量并与原始序列组合，通过同一个编码器构建联合特征空间，提取词法句法信息的同时携带知识信息
 - 多NER层迭代：
计算NER层输出时加强实体尾部字符权重，将尾部词语之前的实体视为与尾部词语拼接的一个整体传递给下一个NER层



Homework 03

二、问题转义——文本匹配/语义匹配

文献：基于知识图谱的问答系统研究与实现

问题分类

- 经过预处理和实体识别提取出关键问题信息，与预设的问题模板进行匹配，找到最相似的句型

- 问句的相似度计算：

通过TF-IDF进行问句的向量表示并计算距离
使用SVM构建分类器进行判断

| 用户输入问句 | 预处理和实体识别后的问句 |
|---------------|-----------------------------|
| 子宫肌瘤的症状有什么？ | [\$Disease] [\$Symptom] 有什么 |
| 甲亢能吃海带吗？ | [\$Disease] 能 吃 [\$Food] |
| 血浆纤维蛋白原能查出什么？ | [\$Check] 能 查出 什么 |

$$d = \frac{A * B^T}{|A| * |B|} = \frac{\sum_{i=1}^N a_i * b_i}{\sqrt{\sum_{i=1}^N a_i^T * a_i} * \sqrt{\sum_{i=1}^N b_i^T * b_i}}$$

Homework 03

三、视频匹配——视频自然语言定位

Curriculum-Listener: Consistency- and Complementarity-Aware Audio-Enhanced Temporal Sentence Grounding

Houlun Chen

chenhl23@mails.tsinghua.edu.cn
DCST, Tsinghua University

Xin Wang*

xin_wang@tsinghua.edu.cn
DCST, BNRist, Tsinghua University

Xiaohan Lan

lanxh20@tsinghua.org.cn
DCST, Tsinghua University

Hong Chen

h-chen20@mails.tsinghua.edu.cn
DCST, Tsinghua University

Xuguang Duan

duan_xg@outlook.com
DCST, Tsinghua University

Jia Jia*

Wenwu Zhu*
{jjia,wwzhu}@tsinghua.edu.cn
DCST, BNRist, Tsinghua University

Chen, Houlun, et al. "Curriculum-listener: Consistency-and complementarity-aware audio-enhanced temporal sentence grounding." Proceedings of the 31st ACM International Conference on Multimedia. 2023.

Homework 03

三、视频匹配——视频自然语言定位

视频片段定位（Temporal Sentence Grounding, TSG）的目的是根据自然语言查询，在一个未剪辑的视频中找到与之语义匹配的片段的起止时间戳，它要求方法具备较强的时序跨模态推理能力。

Homework 03

三、视频匹配——视频自然语言定位

Query: A person walks down stairs drinking from a glass bottle.



与此前其他方法相比，自适应双分支促进网络（ADPN）的特点在于能够高效利用视频中视觉和音频模态的一致性与互补性来增强视频片段定位性能。

Homework 03

三、视频匹配——视频自然语言定位

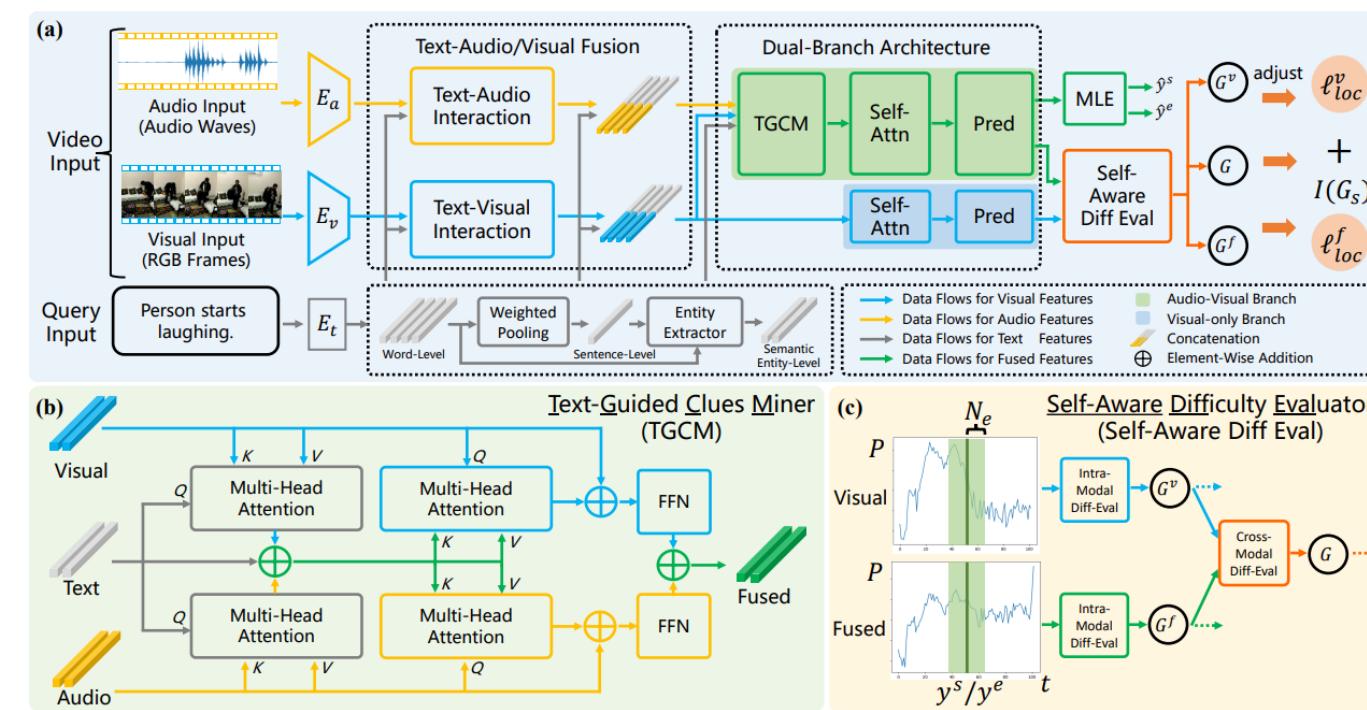
挑战：

音频和视觉模态的一致性和互补性是与查询文本相关联的，因此捕获视听一致性与互补性需要建模文本-视觉-音频三模态的交互。

音频和视觉间存在显著的模态差异，两者的信息密度和噪声强度不同，这会影响视听学习的性能。

Homework 03

三、视频匹配——视频自然语言定位 ADPN结构示意图：



三大设计：

双分支网络结构设计，音频视觉多模态学习同时，强化视觉信息

文本引导的线索挖掘单元 (TGCM)
分模态提取关联信息并集成，再传播到各自模态

课程学习优化策略，根据两个分支的预测输出差异评估样本难度，对训练过程的损失函数项进行重加权



Team 2:基于MOOC的个性化教育智能体 2

Homework 02

1

请列出组员、助教名字、项目名称，并附上本周课上拍的组员照片

项目名称：基于MOOC的教育智能体

助教：张佳璐，王星月

组员：杨祎勃，何逸沣，神远洋，刘玉林



Homework 03

参考文献：Nassiri, K., Akhloufi, M. Transformer models used for text-based question answering systems. *Appl Intell* **53**, 10602–10635 (2023). <https://doi.org/10.1007/s10489-022-04052-8>

1. 研究目的

本文系统综述了Transformer模型在文本问答系统（QA）中的应用，旨在：

- 分析Transformer架构的核心机制（如注意力、自注意力）及其在QA任务中的优势。
- 分类现有Transformer模型（编码器、解码器、编码器-解码器结构），并比较其性能。
- 探讨QA系统的实现方法、数据集与评估指标，提出未来研究方向。

2. 问题定义

传统QA系统的局限性：

- **领域依赖性**：特定领域模型难以泛化到其他领域，需频繁调整。
- **复杂问题处理**：对多跳推理、长答案生成（如“为什么/如何”类问题）支持不足。
- **数据与评估**：现有数据集多为特定任务设计（如SQuAD），缺乏开放域和复杂问题的覆盖；评估指标依赖词汇重叠而非语义理解。

Homework 03

3. 解决方法

模型架构与分类

- **编码器模型** (如BERT、RoBERTa、ELECTRA) :

- 通过预训练 (如MLM、NSP任务) 学习上下文表示。

- 优化策略: 参数共享 (ALBERT) 、动态掩码 (RoBERTa) 、知识蒸馏 (DistilBERT) 。

- **解码器模型** (如GPT系列) :

- 自回归生成答案, 适合开放域生成任务。

- GPT-3通过大规模参数和零样本学习实现广泛适应。

- **编码器-解码器模型** (如BART、T5) :

- 结合双向编码与生成能力, 支持文本到文本的统一框架。

关键技术

- **迁移学习**: 通过预训练 (如SQuAD) 和微调适应特定领域。

- **知识蒸馏**: 压缩大模型 (如TinyBERT) , 提升推理效率。

- **注意力优化**: 长文本处理 (Longformer) 、稀疏激活 (GLaM) 和多头注意力机制。

工具支持

- **Hugging Face库**: 提供预训练模型接口, 简化QA系统开发流程。

Homework 03

4. 局限性

1. **评估指标不足**:

- 现有指标（如EM、F1）依赖词汇重叠，忽视语义等效性。
- 缺乏对长答案生成质量的评估标准。

2. **模型可解释性差**:

- Transformer决策过程为黑箱，难以追溯答案生成逻辑。

3. **跨领域适应性弱**:

- 模型在实验室数据集表现优异，但实际场景（如医疗、法律领域）性能下降显著。

4. **复杂问题处理能力有限**:

- 多跳推理、时序依赖问题（如“气候变化的原因”）需多文档联合分析，当前模型效率低。

5. **数据依赖性**:

- 高质量标注数据稀缺，依赖人工构建（如考试题目），成本高且覆盖面有限。

Homework 03

A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models (大语言模型中幻觉缓解技术的全面调查)

<https://arxiv.org/pdf/2401.01313v1.pdf>

研究背景

大型语言模型（LLMs）在生成文本时容易产生幻觉（Hallucination），即生成看似合理但缺乏事实依据的内容（例如捏造细节或与输入矛盾的信息）。这一问题严重阻碍了LLMs在医疗、金融等敏感领域的实际应用。该论文系统梳理了32种缓解幻觉的技术，并提出分类框架。

幻觉缓解技术分类

1. 提示工程方法（Prompt Engineering）

检索增强生成（Retrieval-Augmented Generation, RAG）

- 生成前检索

- 技术案例：LLM-Augmenter、FreshPrompt
- 特点：在生成前从知识库获取证据链，整合到输入中

- 生成中检索

- 技术案例：D&Q框架、EVER
- 特点：实时验证每个生成句子的置信度，通过回溯搜索修正错误

- 生成后检索

- 技术案例：RARR、高熵词替换
- 特点：对输出进行后期验证和编辑

- 端到端集成

- 技术案例：Lewis et al., 2021
- 特点：预加载知识库以减少训练依赖

基于反馈的自我完善

- 可靠性分解（如Si et al., 2022）

- 通过提示优化模型的泛化性、社会偏见校准和事实性

- 矛盾检测与修正（如ChatProtect）

- 识别逻辑不一致的句子并重构

- 验证链（Chain-of-Verification, CoVe）

- 生成初始答案后通过独立验证问题修正错误

Homework 03

2. 模型开发方法 (Model Development)

新解码策略

- 上下文感知解码 (CAD)
 - 通过对比有/无上下文的输出分布增强一致性
- 对比层解码 (DoLa)
 - 利用模型不同层的语义差异调整生成概率

监督微调与知识编辑

- 知识图谱集成
 - 将结构化知识注入模型参数
- 合成任务训练 (SynTra)
 - 在合成数据上微调提升可靠性

关键技术与案例

| 技术名称 | 核心机制 | 适用场景 |
|---------------|--------------|--------|
| LLM-Augmenter | 迭代检索-生成-验证框架 | 黑盒模型优化 |

| 技术名称 | 核心机制 | 适用场景 |
|-----------------|--------------|---------|
| LLM-Augmenter | 迭代检索-生成-验证框架 | 黑盒模型优化 |
| FreshPrompt | 结合搜索引擎实时更新知识 | 动态信息场景 |
| D&Q框架 | 复杂问题拆解+子问题验证 | 多跳推理任务 |
| Self-Reflection | "生成-评分-改进"循环 | 医疗问答可靠性 |

挑战与未来方向

- 长文本生成的幻觉控制
 - 现有技术多针对短文本，长文本连贯性与事实性难以平衡
- 错误累积与动态知识更新
 - 检索增强方法可能因错误传递或知识库滞后导致新幻觉
- 多模态幻觉缓解
 - 视觉语言模型 (VLMs) 的跨模态对齐尚未充分研究
- 伦理与评估标准
 - 需建立更细粒度的检测基准 (如TruthfulQA、Med-HALT)

Homework 03

RAG: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

论文: Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.

1. 问题定义

针对传统大模型（如GPT-3）存在的两大瓶颈：

- **静态知识:** 无法动态整合最新知识（如MOOC课程更新）
- **幻觉风险:** 生成内容可能偏离事实（对教育场景尤其致命）

2. 方法论创新

提出 RAG (Retrieval-Augmented Generation) 框架，核心架构分为两阶段：

- **检索阶段:** 使用 Dense Passage Retrieval (DPR) 模型将用户问题编码为查询向量，从外部知识库（如教材/论文库）检索Top-K相关段落
- **生成阶段:** 将检索到的段落与原始问题拼接，输入生成模型（如BART）输出最终回答

技术亮点：

技术亮点：

- **双编码器设计:** 独立训练查询编码器（Query Encoder）和段落编码器（Passage Encoder），通过对比学习对齐语义空间
- **端到端优化:** 通过可微分检索机制（如Maximum Inner Product Search, MIPS）实现检索与生成联合训练

3. 实验结果

- 在开放域问答任务（NaturalQuestions, WebQuestions）中，RAG的EM值（Exact Match）比纯生成模型提升 **15.2%**
- 知识更新成本降低 **92%**（仅需更新检索库而非重新训练模型）
- 生成结果的事实准确性（Factual Correctness）提升 **37%**

4. 教育场景启示

- **动态知识整合:** 可将MOOC课程、教材PDF等作为检索库，实时更新模型知识
- **错误控制:** 通过检索约束减少生成幻觉（例如物理公式推导的严格性保障）
- **可解释性:** 展示检索到的参考段落，帮助学生验证答案来源（符合教学透明度需求）

5. 局限性与改进方向

- **检索依赖:** 若知识库未覆盖相关内容，性能显著下降 → 建议结合您的知识图谱补全策略
- **实时性:** 传统DPR检索延迟较高 → 可改用Faiss等高效向量数据库优化
- **领域适配:** 教育领域专业术语需微调编码器 → 使用课程语料继续预训练（Continue Pre-training）

Homework 03

参考文献：《Enhancing Chat Language Models by Scaling High-quality Instructional Conversations》(<https://arxiv.org/pdf/2305.14233.pdf>)

1. 角色分离与行为模拟

- **双模型协作**: 用两个独立ChatGPT API分别扮演用户和AI助手。
- **用户行为引导**: 通过特定Prompt指示用户模型模仿真实用户行为（如提问风格、话题多样性），避免其“越界”扮演AI角色。

示例指令：

"你是一个真实用户，请根据对话历史提出自然、多样的问题，避免机械式回复。"

2. 多样性增强机制

- **主题扩展**:
 - 从维基数据实体、30个元主题（如科技、艺术）生成问题，覆盖常识、创作、推理等场景。
 - 对生成指令进行多轮细化（如从“写故事”到“写科幻故事，包含时间旅行元素”）。
- **元信息驱动**: 结合C4语料库的文本类型（如文章、代码、剧本），生成针对性指令。

3. 多轮对话迭代生成

- **上下文传递**: 在每轮对话中，用户模型和AI模型交替生成内容，传递完整对话历史以保持连贯性。
- **任务聚焦**: 在创作类对话中，Prompt会反复强调目标（如“继续优化这篇科幻故事的第二段”），确保对话不偏离主题。

4. 质量控制与过滤

- **冗余过滤**: 自动移除用户回复中的“谢谢”“不客气”等机械性表达，提升真实感。
- **人工校验**: 通过ChatGPT评分（1-10分）筛选连贯性高、信息量大的对话样本。

Homework 03

参考文献：Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157-173.

1. 研究背景

当前的语言模型能够处理较长的上下文，但其在长文本中的信息利用仍然存在问题：

- 语言模型在长上下文中提取和使用信息的能力较差。
- 信息位置对模型性能有显著影响，模型更容易利用出现在输入开头或结尾的信息，而中间部分的信息容易被忽略（首位效应和新近效应，Primacy & Recency Bias）。

2. 实验验证

1. 多文档问答任务

- 任务输入：包含一个问题，以及多个文档，其中只有一个文档包含正确答案，其余为干扰文档。
- 通过调整文档顺序，使包含答案的文档出现在开头、中间或结尾，观察模型的表现变化。
- 采用来自NaturalQuestions-Open的数据集，每个问题配有Wikipedia段落作为文档来源。

Homework 03

2. 键值对检索任务

- 输入包含多个键值对，每个键和值都是随机生成的UUID。
- 任务要求模型根据查询的键，返回相应的值。
- 通过调整相关键值对的位置（开头、中间、结尾）评估模型的检索能力。

3. 其他影响因素分析

- 模型架构：研究Decoder-only（如GPT-3.5, Claude）与Encoder-Decoder（如Flan-T5）模型在长上下文信息利用方面的表现差异。
- 查询感知：测试在上下文前后都包含查询是否可以提高模型对长上下文的利用率。
- 指令微调：研究微调是否能提升模型对长文本信息的提取能力。

Homework 03

3. 实验结果

1. 信息位置显著影响性能

- 当相关信息出现在上下文开头或结尾时，模型表现较好；
- 当信息位于上下文中间时，性能显著下降，表现出U型曲线。

2. 仅仅增加上下文窗口长度并不能提高模型的信息利用能力

- GPT-3.5-Turbo (16K) 与标准 GPT-3.5 在相同上下文窗口下，表现基本一致。
- 这表明扩展上下文窗口的模型并不会自动学会更好地利用长上下文信息。

3. 模型架构影响

- Encoder-Decoder 模型相较于 Decoder-only 模型更擅长利用长上下文信息。

4. 查询感知

- 在上下文开头和结尾都包含查询，可以显著提高模型在键值对检索任务中的表现，但在多文档问答任务中的改进效果有限。



Team 3: 低视力视觉增强辅助: HoloLens MR 头显技术的创新应用



这是一行文字
这是一行文字
这是一行文字
这是一行文字
这是一行文字
这是一行文字
这是一行文字

这是一行文字
这是一行文字
这是一行文字
这是一行文字
这是一行文字
这是一行文字
这是一行文字
这是一行文字

低视力视觉增强辅助： HoloLens MR 头显技术的创新应用

Student Information:

谢景涛 12112010

金邦量 12110406

高胜寒 22510011

张阳 12110424

马鑫 12111644

助教学姐：席睿翎



我们的团队



ForeSee: A Customizable Head-Mounted Vision Enhancement System for People with Low Vision

Yuhang Zhao
Jacobs Technion-Cornell Institute
Information Science, Cornell University
yz769@cornell.edu

Sarit Szpiro
Jacobs Technion-Cornell Institute
Cornell Tech
sarit.szpiro@cornell.edu

Shiri Azenkot
Jacobs Technion-Cornell Institute
Cornell Tech
shiri.azenkot@cornell.edu

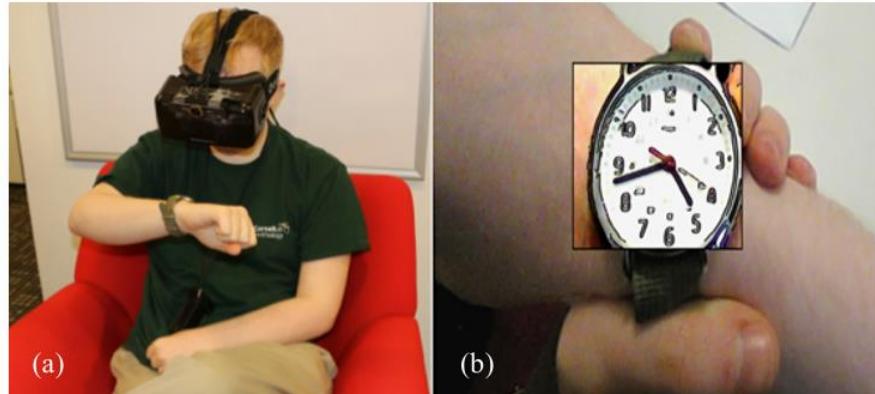


Figure 1. (a) A low vision person using ForeSee to check the time. (b) He sees an enhanced version of the watch using magnification, contrast enhancement, and edge enhancement in ForeSee's Window Display Mode.

In the United States, there are 19 million people with low vision, but only 1.3 million are legally blind. Millions have functional vision and prefer to see with their own eyes.

ForeSee

- Camera (captures the user's view)
- Embedded processor (enhances the captured video)
- Display (presents the enhanced video)

Display mode:

- Window display
- Full display

Enhancement Method:

- Magnification
- Contrast Enhancement
- Edge Enhancement
- Black/White Reversal
- Text Extraction

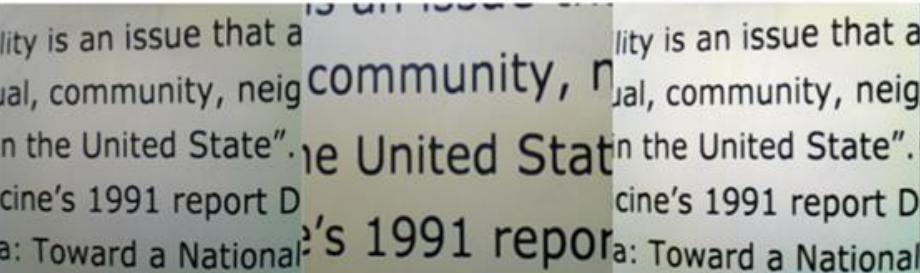
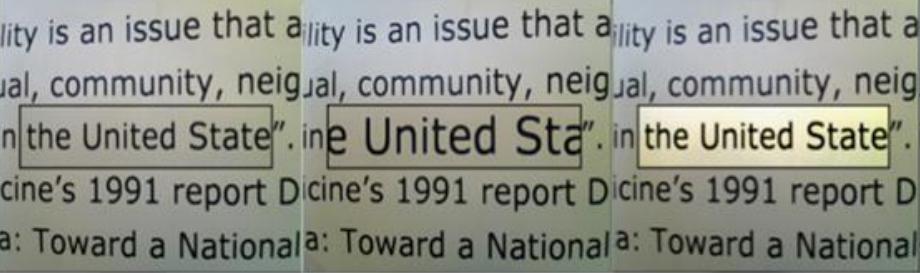
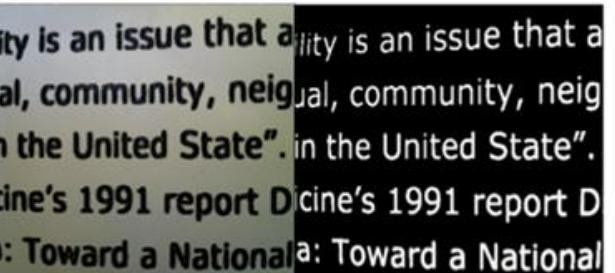
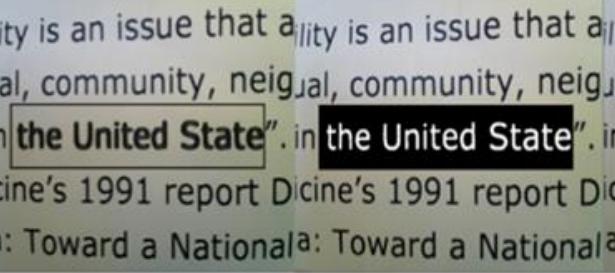
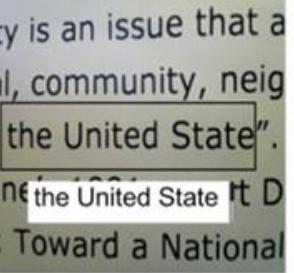
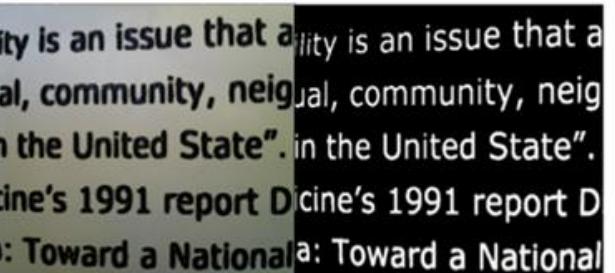
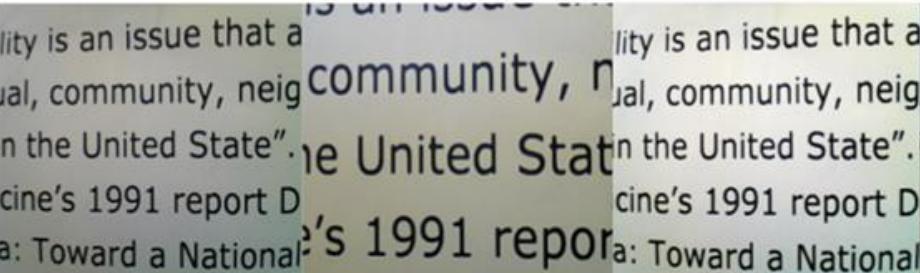
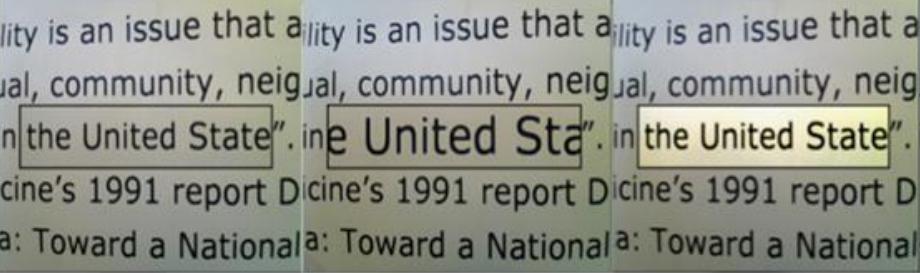
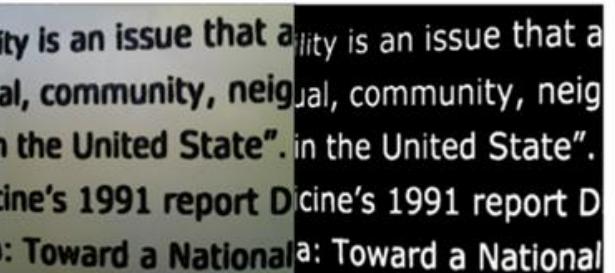
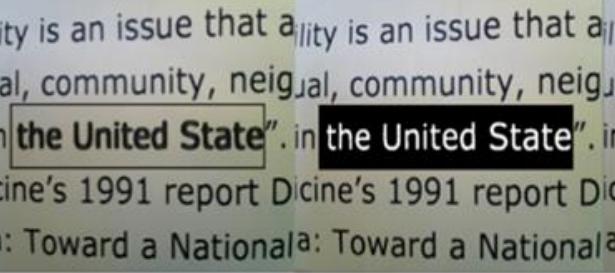
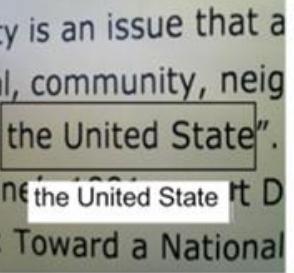
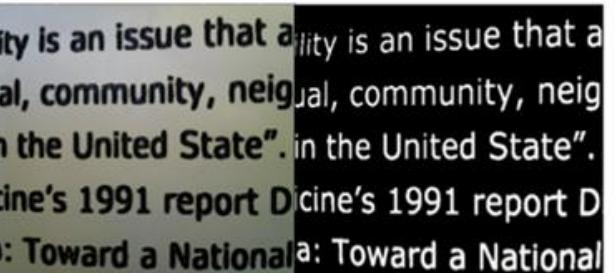
| | Full | Window | No enhancement | Magnification | Contrast Enhancement | Edge Enhancement | Black-White Reversal | Text Extraction |
|--------|--|---|---|--|---|--|---|---|
| Full |  |  |  |  |  |  |  |  |
| Window |  |  |  |  |  |  |  |  |

Figure 2. The visual effects of five enhancement methods: Magnification, Contrast, Edge Enhancement, and Black/White Reversal; in two display modes: Full Display Mode and Window Display Mode

Prototype

- camera: WideCam F100
- Processor: Laptop
- Display: Oculus Rift DK2

Users interacted with ForeSee with natural speech commands

Experiment:

19 people with low vision (different Visual Field, Acuity, Color Vision)
(use different enhancement method)

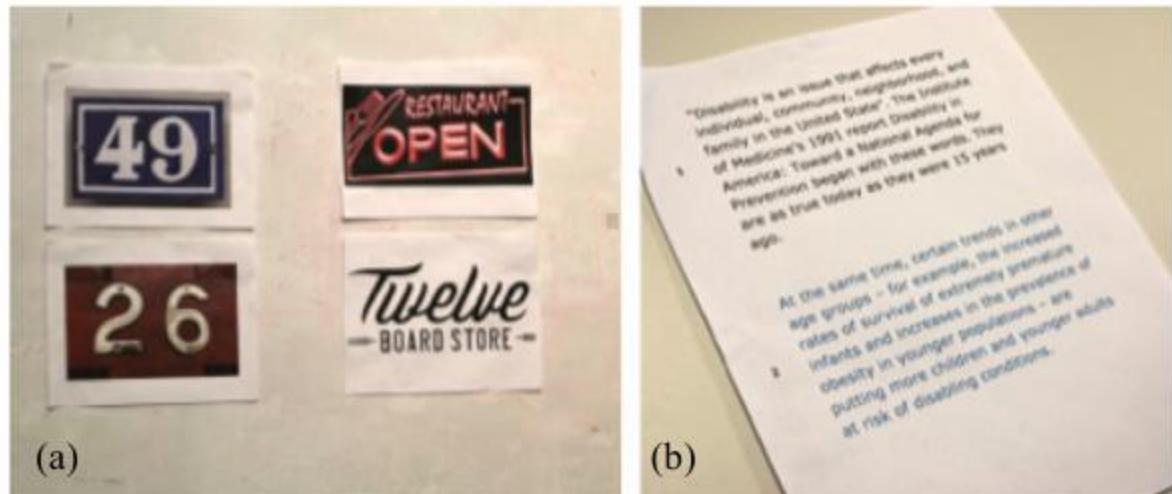


Figure 3. Experiment Materials: (a) four printed signs of numbers or writings hung on the wall (b) a handheld printed page

- Without using ForeSee
- Using ForeSee but without any enhancement methods
- Full Display Mode, with all enhancement methods
- Window Display Mode, with each of the five enhancement methods
- Using ForeSee with whatever enhancement methods (users customize)

Results

- Magnification is effective
- They need customization of enhancement methods and adjust them in different tasks.
- Window display is always useful especially in far distance task (it helped them concentrate on the target)

Limitation and future work:

- Too big for daily use
- Processing video caused a slight delay
- Low resolution
- Adaptive on changing the enhancement method and display mode

Augmented Reality Magnification for Low Vision Users with the Microsoft Hololens and a Finger-Worn Camera

Lee Stearns¹, Victor DeSouza¹, Jessica Yin³, Leah Findlater^{2,5}, Jon E. Froehlich^{1,4}

¹Department of Computer Science
²College of Information Studies
University of Maryland, College Park

³Poolesville High School
Poolesville, MD

⁴Computer Science and Engineering
⁵Human Centered Design and Engineering
University of Washington

lstearns@umd.edu, vdesouza@umd.edu, jessica.y.phs@gmail.com, leahkf@umd.edu, jonf@cs.umd.edu

finger camera + HoloLens

Background

Problem:

- magnification and enhancement of visual information

Traditional methods:

- display content on a large screen - not portable
 - closed-circuit television (CCTV) systems
 - desktop video magnifiers
- magnifying glasses - portable but not always available, limited visual area
 - magnifying glasses
 - smartphone-based magnifiers

Previous approaches:

- a head-worn display and a head-worn camera - need head movement

Prototype

Our system:

- a head-worn Microsoft Hololens unit
- an Awaiba NanEye Idule camera worn on the index finger via a custom ring

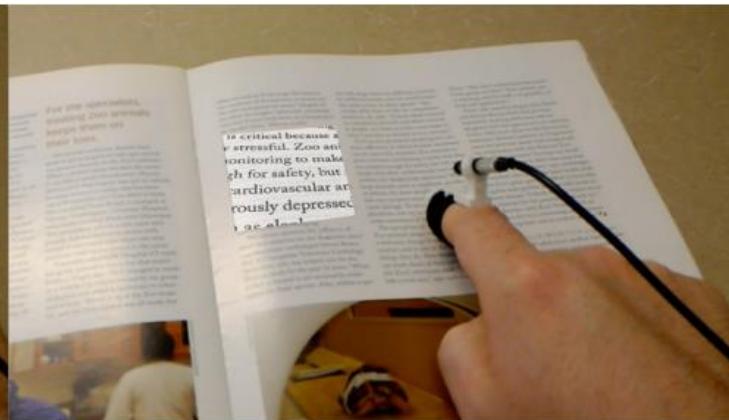
Three interface options:



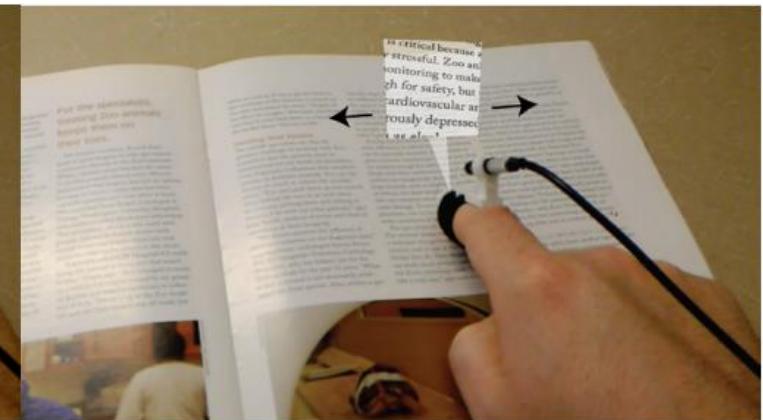
Figure 1. Prototype Hololens and finger-camera system: (a) reading a magazine article; (b) close-up view of the camera.



(a) Fixed position on 2D AR display (*i.e.*, fixed within the wearer's field of view).



(b) Fixed position on 3D surface or midair, in this case rendered flat on the magazine page.



(c) Dynamic finger tracking, where magnified view follows and hovers above the finger.

Figure 2. The three viewport positions in our prototype system, which vary in how they present the magnified view to the user.

Discussion

Limitations:

- Hololens
 - size
 - relatively narrow display area centered in the user's field of view
- finger-worn camera hardware
 - large form factor regular use
 - not fully wearable — it currently relies on a separate computer for power and streaming images to the display
- Interaction
 - should be able to customize the magnification level, position, text processing, and other settings
 - could be done via hand gestures or speech recognition(Future usability testing here is necessary to determine)

Future work:

- conduct an evaluation with low-vision participants comparing the designs in terms of usability and comprehension
- compare against the current status quo—a handheld smartphone magnifier

Envision:

- could change fonts or read the text
- content could be dynamically enhanced and placed over the original
- touchscreen text and image manipulation features such as highlighting or copy and paste

Designing AR Visualizations to Facilitate Stair Navigation for People with Low Vision

Yuhang Zhao¹, Elizabeth Kupferstein¹, Brenda Veronica Castro¹,
Steven Feiner², Shiri Azenkot¹

¹Jacobs Technion-Cornell Institute, Cornell Tech,
Cornell University, New York, NY, USA
{yz769, ek544, bvc5, shiri.azenkot}@cornell.edu

²Department of Computer Science, Columbia
University, New York, NY, USA
feiner@cs.columbia.edu



Figure 1: Our visualizations for (a) projection-based AR and (b) smartglasses to facilitate stair navigation for PLV.

Background & Problem

AR for Stair Navigation:

- 19M low-vision individuals in the US; most rely on residual vision.
- Stairs are a major mobility hazard due to poor depth perception and contrast sensitivity.
- Existing tools (white canes, audio feedback) are underutilized.

Methodology

AR Study (2019):

- Platforms:
 - Projection-based AR (highlights stairs).
 - Smartglasses (HoloLens, stage-based visualizations).
- Participants: 12 low-vision users per platform.
- Evaluation: Walking speed, psychological security, behavior observation.



Figure 7: Glow (a–d) and Path (e–g). Glow: (a) thin red glow on the landing; (b) thick cyan glow in the preparation area; (c) thick yellow glow in the alert area; (d) thin blue glow on the middle of the stairs. **Path:** (e) view of the Path on the landing; (f) view of the Path when getting close to the first stair; (g) view of the Path on the middle of the stairs.

Key Results

Projection-based AR:

- Increased walking speed (6.42% faster downstairs, 5.78% upstairs).
- Improved psychological security (mean score: 6.6/7).

Smartglasses:

- No significant speed improvement.
- High perceived safety (mean score: 6.1/7).

Session 3B: Accessibility

UIST '19, October 20–23, 2019, New Orleans, LA, USA

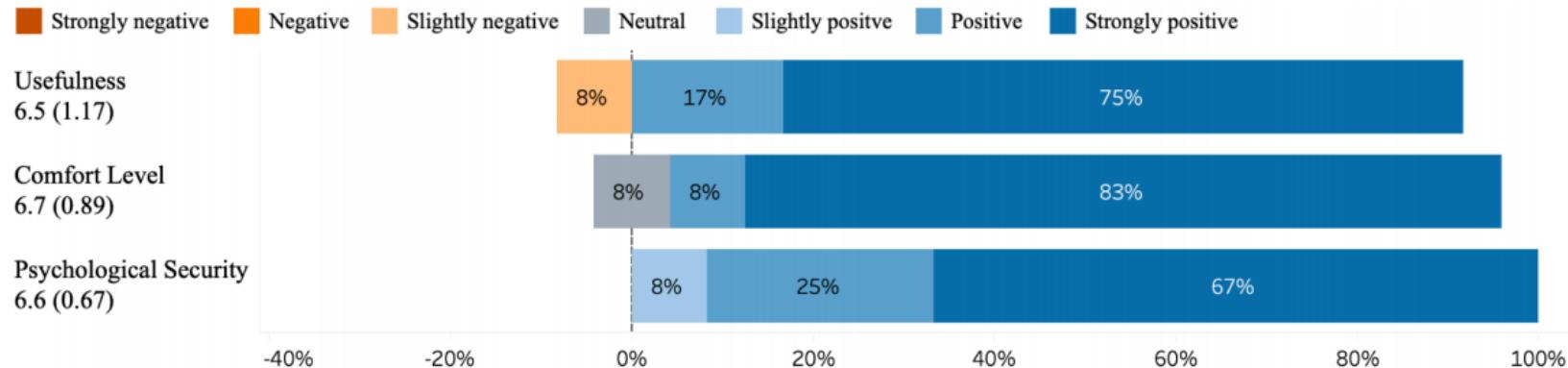


Figure 4: Diverging bars that demonstrate the distribution of participant scores (strongly negative 1 to strongly positive 7) for usefulness, comfort level, and psychological security when using visualizations on projection-based AR. We label the mean and SD under each category.

Conclusion

- AR visualizations enhance safety and confidence for low-vision users.
- Wearable systems require customization and portability.
- Future focus: Real-time adaptability, user-centered design.

Future Work

Technical challenges: Stair detection accuracy, environmental noise.

Smartglasses: Weight and limited FOV hindered usability.



Team 4:低视力人群室内导航系统

1.

- 组员：张立远，岳雪骋，李镓璇，李亭翰
- 助教：韩载道，张颖麟
- 项目名称：低视力人群室内导航系统



Homework 03

参考文献：[1]石胤斌.面向助盲导航设备的室内场景图文转换方法研究[D].北京化工大学,2024.DOI:10.26939/d.cnki.gbhgu.2024.001209.

这篇论文研究的工作也是助盲，但其主要方向在于标识牌和方向指引牌的识别和对周围环境的图像生成字幕，前者可以对室内导航工作进行辅助，进行一定的校验斟误或者辅助定位；后者则是一个新的启发，对周围环境图像直接生成文本描述或许可以使用户对当前的处境有更直观的感受，结合本人的判断可以减少失误的发生，或许可以做成一个额外的功能，当用户想知道周围的具体情况，可以通过启用这项功能来了解身边具体有哪些东西。

Indoor Positioning on Smartphones Using Built-In Sensors and Visual Images

在实际应用中，使用视觉图像进行位置估计很容易受到用户照片姿势的影响。该文提出了一种多传感器辅助视觉定位方法，该方法利用多个智能传感器构建机器学习分类器进行行人姿态估计，提高了检索效率和定位精度。该方法主要结合了视觉图像定位估计和基于多智能传感器的行人姿态估计的优点，并考虑了行人拍摄姿态对定位估计的影响。利用智能手机内置传感器作为行人姿态估计数据的来源，构成了一种可行的基于视觉信息的位置估计方法。

作者：李亭翰

引用

- J. An, D. H. Lee, H. H. Cho and O. H. Jeong, "Indoor Positioning System Using Smartphone and 360° Camera," 2021 IEEE International Conference on Smart Internet of Things (SmartIoT), Jeju, Korea, Republic of, 2021, pp. 342-343, doi: [10.1109/SmartIoT52359.2021.00062](https://doi.org/10.1109/SmartIoT52359.2021.00062).

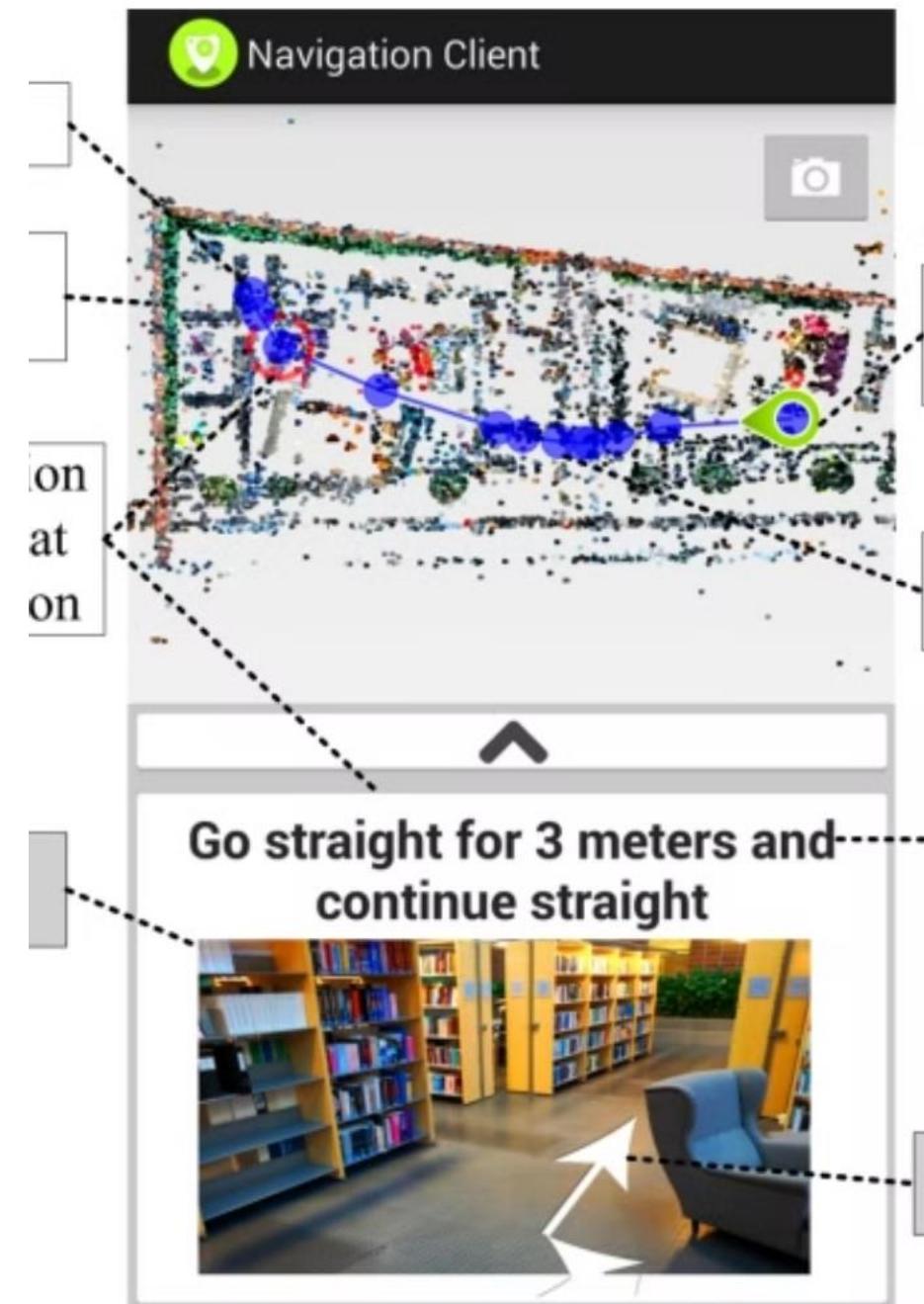
iMoon：基于智能手机的图像室内导航系统

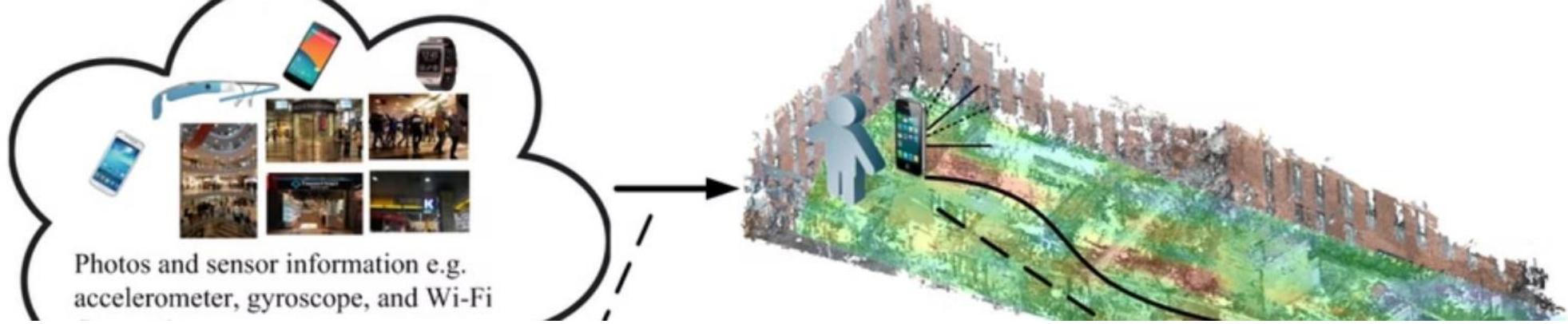
iMoon是一个创新的室内导航系统，它利用众包数据构建室内环境的传感器增强型3D模型。与传统导航系统不同，iMoon不需要预先创建的室内地图、预扫描的无线电地图或预安装的硬件，而是通过智能手机收集的照片和传感器数据来构建和不断更新3D模型。

该系统可以通过在新位置随机拍摄照片和收集传感器数据来引导。随着用户使用该应用程序移动和导航，系统会不断用新收集的照片和传感器数据更新3D模型。实际上，系统在某个位置使用得越多，其准确性就越高。

iMoon解决了众包数据质量参差不齐带来的技术挑战，并提供了直观的视觉导航指令，特别适合阅读地图有困难的人群使用。

雪骋 作者：雪骋 岳





构建传感器增强型3D室内模型

从照片构建3D模型

1

iMoon使用结构运动(SfM)技术从无序2D照片构建稀疏3D点云。该过程包括特征提取、特征匹配和捆绑调整三个步骤，生成包含3D点云和相机姿态估计的输出。

2

检测用户轨迹

系统收集智能手机的加速度计和陀螺仪读数，计算用户轨迹。这些轨迹通过沿途拍摄的照片进行自动校准，从而发现行人路径并将其集成到3D点云中。

3

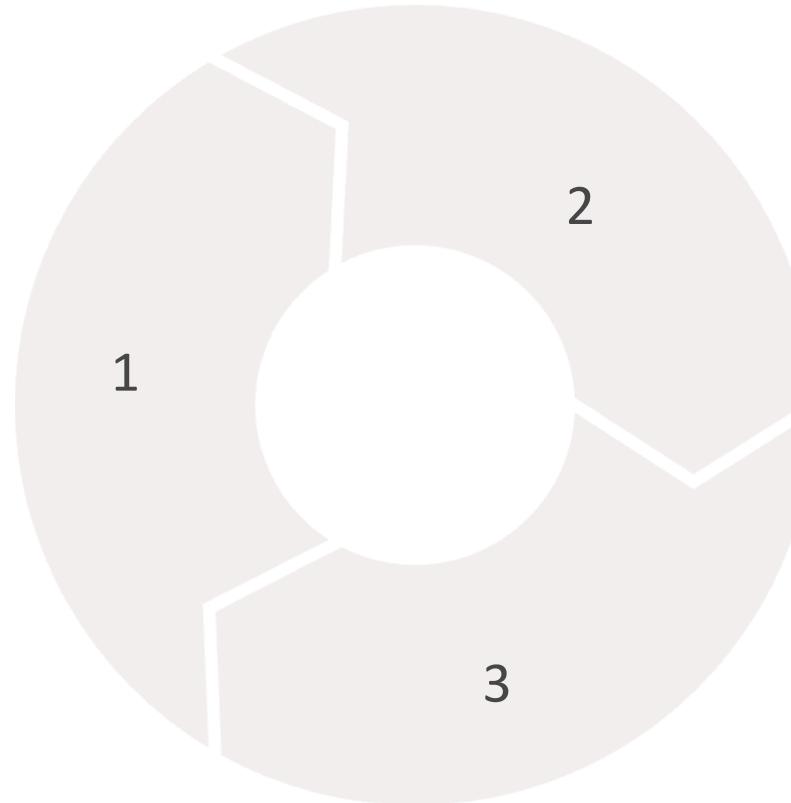
地理参考Wi-Fi指纹

在用户轨迹收集过程中，系统同时收集Wi-Fi指纹。这些指纹根据时间戳进行地理参考，并在用户轨迹自动校准后附加到3D点云上，用于基于指纹的3D模型分区选择。

基于3D模型的室内导航

室内定位

iMoon结合图像定位和Wi-Fi指纹方法实现快速定位。系统将3D模型分区，并使用Wi-Fi指纹选择分区进行特征匹配，大大减少响应延迟。



编译导航网格

系统从3D点云提取障碍物信息，并将从用户轨迹中提取的行人路径整合到导航网格中，解决众包照片可能无法覆盖空间每个角落的问题。

生成导航指令

iMoon创建包含三部分的导航指令：显示指令有效位置的照片、告诉用户直行或转弯的文本，以及指示下一步路径的箭头标志。

实验结果表明，在覆盖约1,100平方米的真实建筑中，使用从2,197张照片生成的3D点云，iMoon在大多数情况下可以在平均3.85秒内定位用户，位置误差小于2米，面向方向误差小于6度。



Team 5: BADAPPLE -BAsed on multimodality : A novel visual interpretation PiPeLinE

Homework 02

Leader: Shengding Liu

Group Member: Zhanwei Zhang, Jianan Xie, Kangrui Chen, Ran Wang

TA: Lingxi Zeng, Jiaqi Wei

Our project topic is **BADAPPLE -BAsed on multimoDality: A novel visual interpretation**

PiPeLinE

Group Photo:



Summary

1. Paper about Text-to-Image

- [ArtAug: Enhancing Text-to-Image Generation through Synthesis-Understanding Interaction](#)

- Adding Conditioning Control to Text-to-Image Diffusion Models

2. Paper about Image-to-Video

- [Animate Anyone Consistent and Controllable Image-to-Video Synthesis for Character Animation](#)

3. Paper about Text-to-Voice

- MARS 5

4. Research about Our Project

ArtAug: Enhancing Text-to-Image Generation through Synthesis-Understanding Interaction

Zhongjie Duan¹ Qianyi Zhao¹ Cen Chen¹ Daoyuan Chen² Wenmeng Zhou² Yaliang Li² Yingda Chen²

基于扩散模型的文本到图像生成技术（如Stable Diffusion、Latent Diffusion等）取得了巨大进展。但由于训练数据中往往包含大量低质量图像，很多预训练模型在没有额外引导的情况下，生成的图像在美学质量和细节上还难以令人满意。

ArtAug的整体流程可以分为三个主要模块：

- **交互算法 (Interaction Algorithm)**

首先利用基础的文本到图像模型生成初始图像 X 。随后利用一个图像理解模块（基于多模态大语言模型，如Qwen2-VL-72B）对生成的图像进行分析。该模块会自动识别图像中可以改进的细节（例如光影、细节、构图、色彩等），并以边界框+详细提示的形式输出修改建议。例如，对于图中花朵的部分，模型可能会给出“让花瓣更加鲜亮、增加光影对比”等描述。最后，将这些细粒度的提示词融入到图像生成过程，采用一种基于区域加权的生成方法，重新生成图像，使得改进建议能够精准地作用于相应区域，从而得到改进后的图像 X' 。

- **数据生成与过滤 (Data Generation and Filtering)**

通过上述交互算法，可以生成大量的图像对（原始图像 X 与改进后图像 X' ）。对于这些图像对，论文设计了严格的过滤流程：首先使用自动的美学评价模型（如Aesthetic score、PicScore、MPS等）筛选出美学得分提高的图像对；同时利用CLIP模型确保图像与文本提示的一致性。最终，经过自动和人工筛选，只保留那些在质量和语义上均有显著改进的图像对作为训练数据。

- **差分训练 (Differential Training)**

直接用过滤后的图像对对原模型进行微调容易导致过拟合，因为有效数据量较少。为此，作者提出了一种差分训练的方法，核心思想是学习原始图像与改进图像之间的“差异”。具体做法是：首先，对单个原始图像 X 用LoRA方法训练一个初步的LoRA模块，使得模型在输入相同提示词时始终生成 X 。然后，在这个基础上，对改进图像 X' 再训练一个新的LoRA模块，使得模型能生成改进后的图像 X' 。该第二个LoRA模块就捕捉到了图像质量提升的“差分信息”。最后，将这个差分LoRA模块融合回原始模型中，使得基础模型在不增加额外推理开销的情况下，直接具备了生成更高质量图像的能力。此外，这一过程是迭代进行的——每一次迭代都能进一步提升模型生成美学图像的能力，就像梯度下降中逐步更新参数一样。



Figure 1. Image examples improved by ArtAug. The base text-to-image model is FLUX.1[dev]. The ArtAug enhancement module is fused into the base model, without requiring additional computational resources.

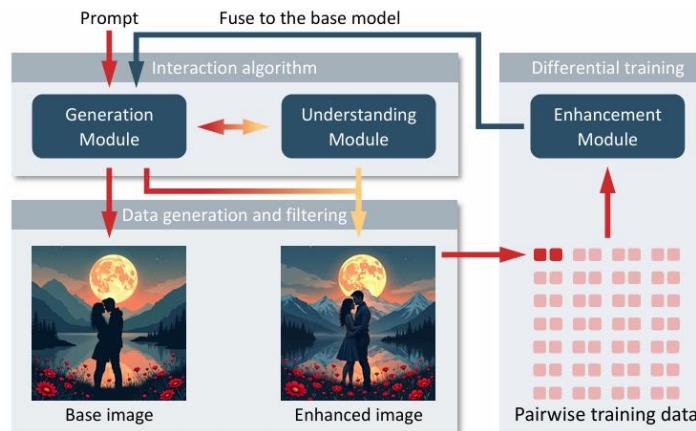


Figure 2. The framework of ArtAug encompasses three key components: the interaction algorithm, data generation and filtering, and differential training. This enhancement process can be iteratively applied to the model, facilitating iterative improvement.

Adding Conditional Control to Text-to-Image Diffusion Models

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala
Stanford University
{lvmin, anyirao, maneesh}@cs.stanford.edu



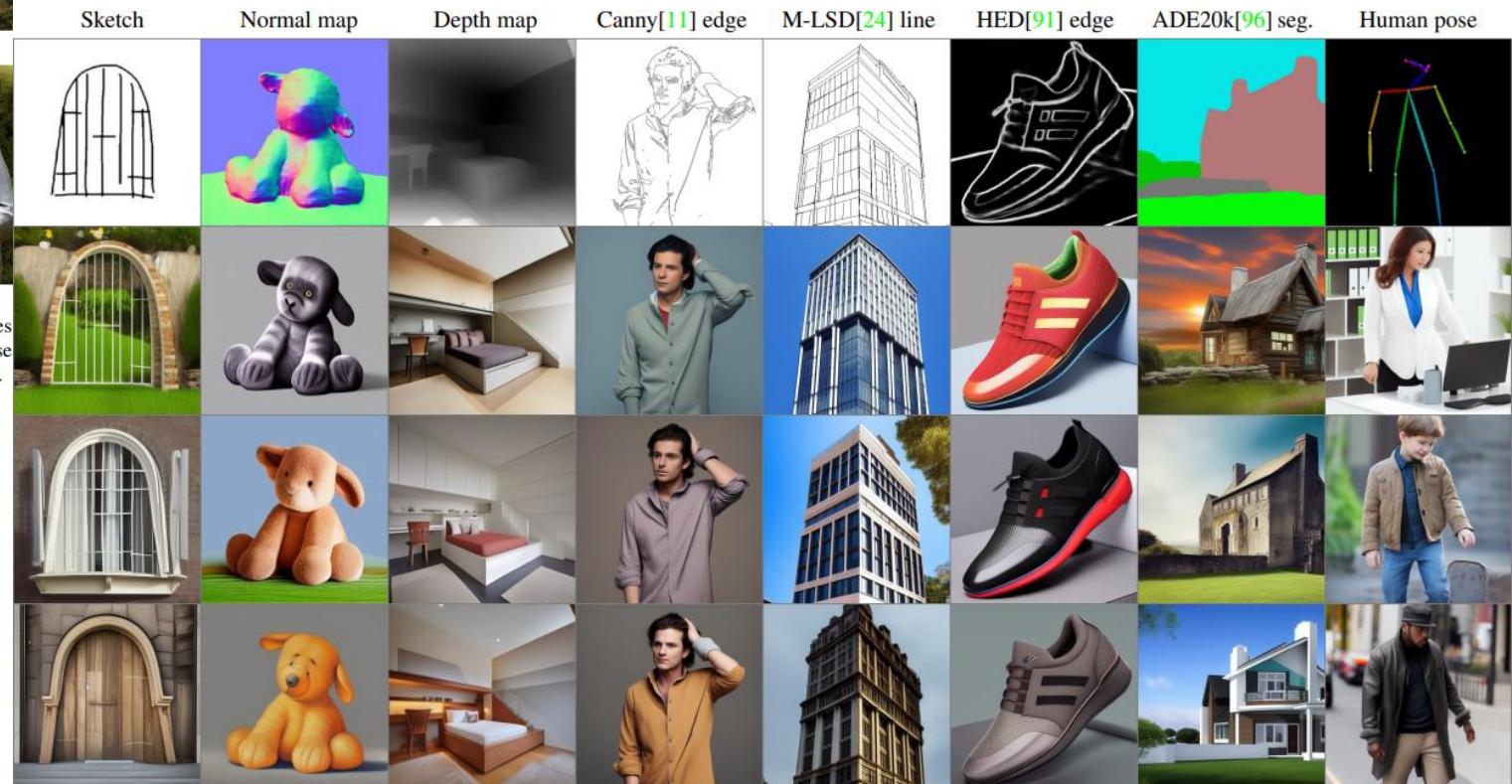
Figure 1: Controlling Stable Diffusion with learned conditions. ControlNet allows users to add conditions like Canny edges (top), human pose (bottom), etc., to control the image generation of large pretrained diffusion models. The default results use the prompt “a high-quality, detailed, and professional image”. Users can optionally give prompts like the “chef in kitchen”.

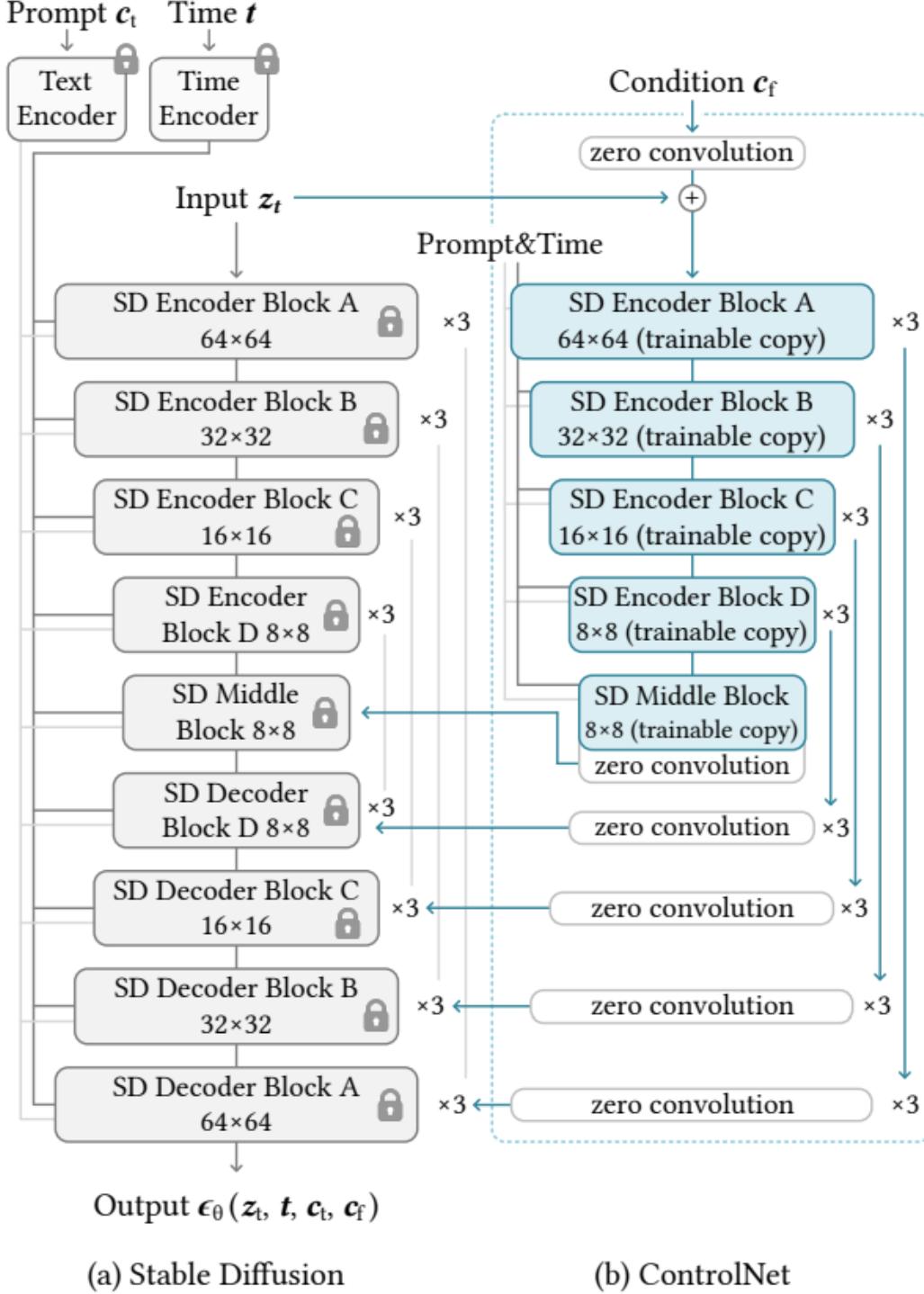
多种可用控制条件：

草稿、边缘、正交、深度、分割、姿势等

ControlNet: 条件控制SD一致性

Figure 7: Controlling Stable Diffusion with various conditions **without prompts**. The top row is input conditions, while all other rows are outputs. We use the empty string as input prompts. All models are trained with general-domain data. The model has to recognize semantic contents in the input condition images to generate images.





Structure

Keywords:

- Injection
- Frozen Original Model
- Adaptive
- Cumulative

Composing multiple ControlNets. To apply multiple conditioning images (*e.g.*, Canny edges, and pose) to a single instance of Stable Diffusion, we can directly add the outputs of the corresponding ControlNets to the Stable Diffusion model (Figure 6). No extra weighting or linear interpolation is necessary for such composition.

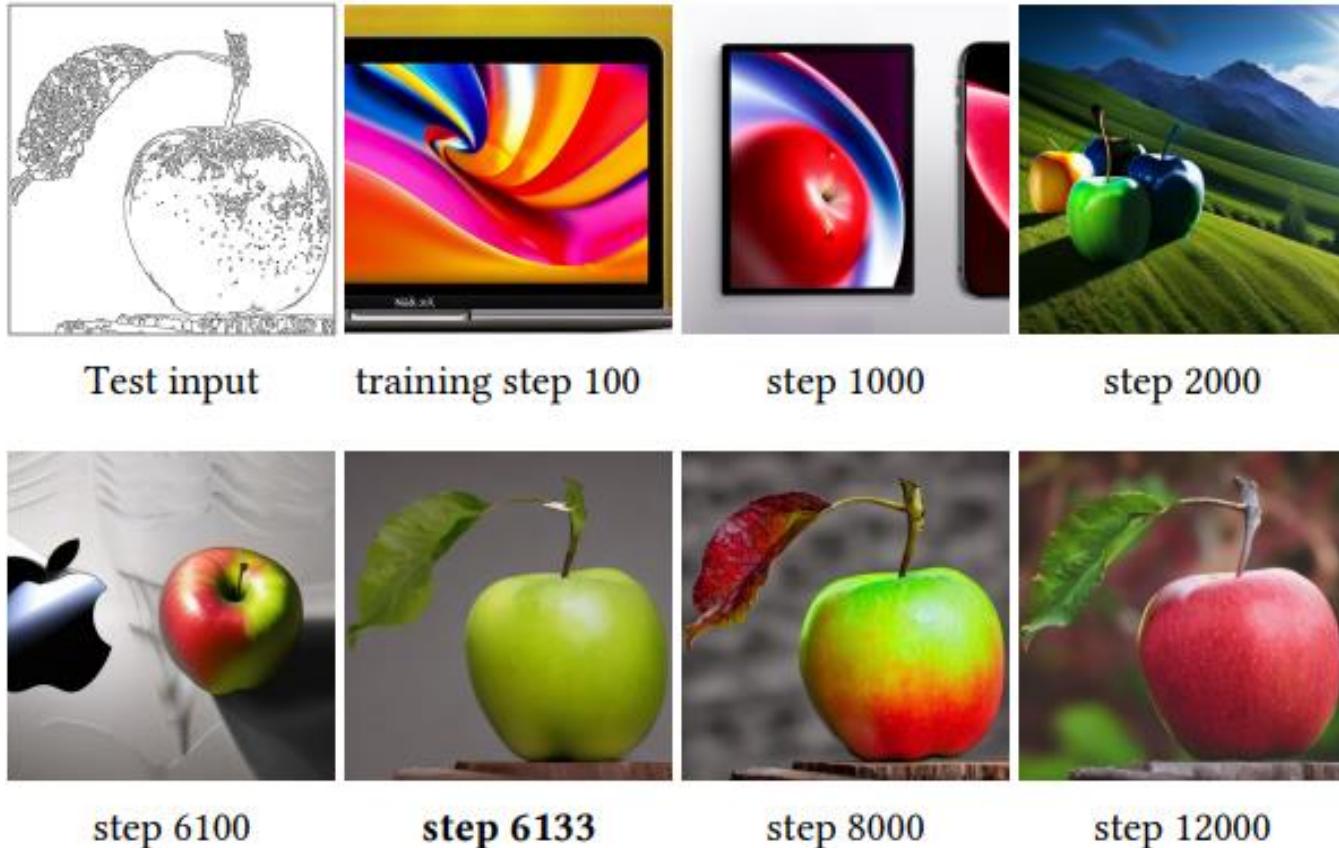


Figure 4: The sudden convergence phenomenon. Due to the zero convolutions, ControlNet always predicts high-quality images during the entire training. At a certain step in the training process (*e.g.*, the 6133 steps marked in bold), the model suddenly learns to follow the input condition.

Funny Observation:
"Sudden Convergence Phenomenon"

Ways to use ControlNet in BADAPPLE

- 如何保证AI每次画出的都同一张人脸? SD人物一致性教程!
- 如何让Stable Diffusion在不同场景保持人物形象的一致性

Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation

Li Hu

Institute for Intelligent Computing, Alibaba Group

hooks.hu@alibaba-inc.com

<https://humanaigc.github.io/animate-anyone/>



Figure 1. Consistent and controllable character animation results given reference image (the leftmost image in each group). Our approach is capable of animating arbitrary characters, generating clear and temporally stable video results while maintaining consistency with the appearance details of the reference character.

Motivation

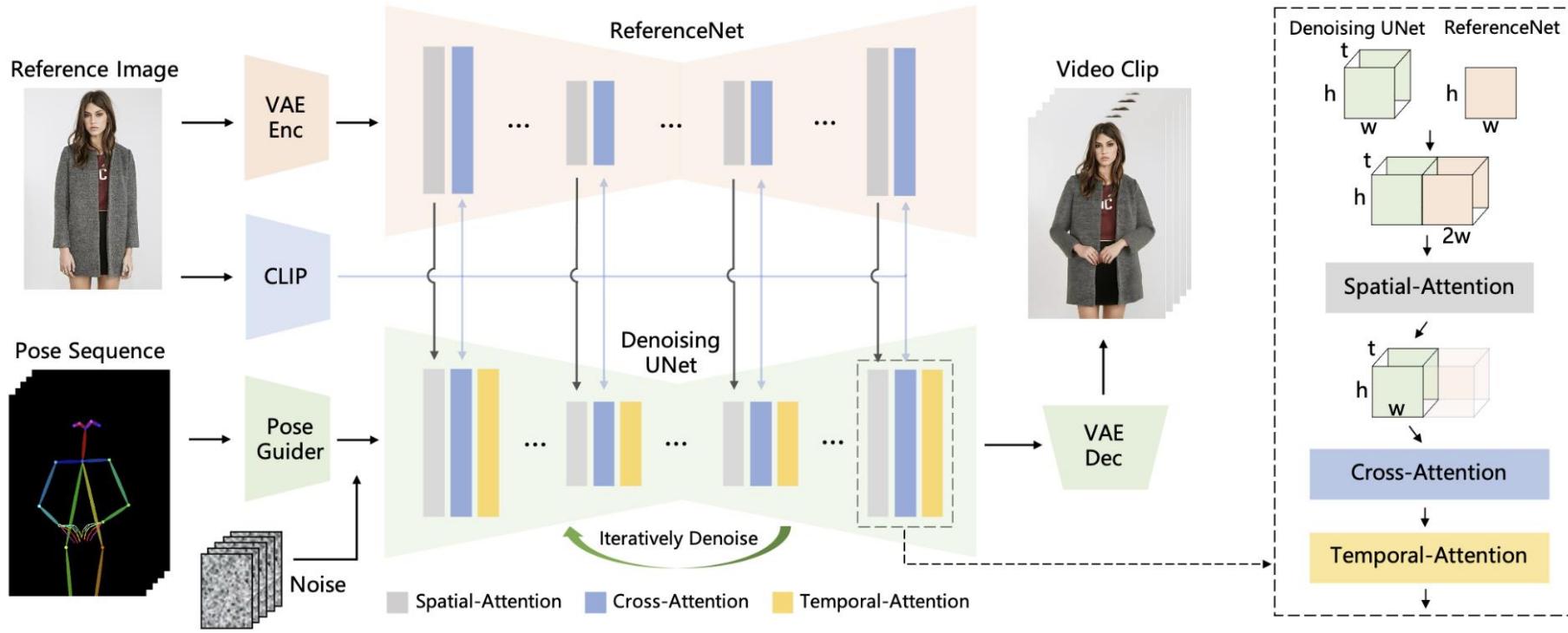
角色动画 (Character Animation) 的目标是通过驱动信号将静态角色图像生成为视频，具有广泛的应用前景，如在线零售、娱乐视频、艺术创作和虚拟角色等。

Challenge

尽管扩散模型在图像和视频生成方面表现出色，但在角色动画中，现有方法仍面临以下挑战：

- 1. 细节一致性：**在生成视频时，保持角色外观细节（如服装纹理、颜色等）与参考图像的一致性。
- 2. 时间稳定性：**生成的视频需要在时间上保持稳定，避免出现抖动或闪烁。
- 3. 姿态可控性：**需要能够根据目标姿态序列精确控制角色的动作。
- 4. 泛化能力：**现有方法大多针对特定任务或数据集，缺乏对任意角色图像的泛化能力。

openaccess.thecvf.com/content/CVPR2024/papers/Hu_Animate_Anyone_Consistent_and_Controllable_Image-to-Video_Synthesis_for_Character_Animation_CVPR_2024_paper.pdf



Pipeline

Animate Anyone 的整体流程如下：

1. **输入**：给定一个描述角色外观的**参考图像**和一个**目标姿态序列**。
2. **姿态编码**：使用**姿态引导器 (Pose Guider)**对姿态序列进行**编码**，并将其与**多帧噪声融合**。
3. **去噪过程**：通过**修改的去噪 UNet**进行视频生成。去噪 UNet 的计算模块包括**空间注意力 (Spatial-Attention)**、**交叉注意力 (Cross-Attention)** 和**时间注意力 (Temporal-Attention)**。
4. **特征整合**：
 1. 使用 ReferenceNet 提取参考图像的细节特征，并通过空间注意力将其整合到去噪 UNet 中。
 2. 使用 CLIP 图像编码器提取参考图像的语义特征，用于交叉注意力。
5. **时间建模**：通过时间层模拟多帧之间的关系，确保时间上的连续性和稳定性。
6. **输出**：最终通过 VAE 解码器将生成的特征解码为视频片段。

Experiments



泛化能力

Figure 3. Qualitative Results. Given a reference image (the leftmost image of each group), our approach demonstrates the ability to animate diverse characters, encompassing full-body human figures, half-length portraits, cartoon characters, and humanoid figures. The illustration showcases results with clear, consistent details, and continuous motion.



姿态控制，颜色一致性
Figure 5. Qualitative comparison between DisCo and our method. DisCo displays problems such as pose control errors, color inaccuracy, and inconsistent details. In contrast, our method demonstrates significant improvements in addressing these issues.

姿态预测，面部处理

Figure 7. Qualitative comparison with image-to-video methods, which struggle to generate substantial character movements and face challenges in maintaining long-term appearance consistency.



Figure 4. Qualitative comparison for fashion video synthesis. Other methods exhibit shortcomings in preserving fine-textured details of clothing, whereas our method excels in maintaining exceptional detail features.

纹理细节



Figure 8. Ablation study of different design. ReferenceNet ensures consistent preservation of details in character's appearance. 外观一致性

Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation

Li Hu

Institute for Intelligent Computing, Alibaba Group

hooks.h1@alibaba-inc.com

<https://humanaigc.github.io/animate-anyone/>



Figure 1. Consistent and controllable character animation results given reference image (the leftmost image in each group). Our approach is capable of animating arbitrary characters, generating clear and temporally stable video results while maintaining consistency with the appearance details of the reference character.

本文提出了一种名为 **Animate Anyone** 的框架，主要贡献包括：

1. **ReferenceNet**: 设计了一个对称的 UNet 结构，通过空间注意力 (Spatial-Attention) 将参考图像的细节特征整合到去噪 UNet 中，显著提高了外观细节的保留能力。
2. **姿态引导器 (Pose Guider)** : 引入了一个轻量级的姿态引导器，高效地将姿态控制信号整合到去噪过程中，实现了姿态的精确控制。
3. **时间建模**: 通过时间层 (Temporal Layer) 模拟多帧之间的关系，确保视频帧之间的平滑过渡，同时保持高分辨率细节。
4. **泛化能力**: 通过扩展训练数据，该方法能够动画化任意角色图像，优于现有的图像到视频方法。
5. **实验结果**: 在多个基准数据集 (如 UBC 时尚视频数据集、TikTok 数据集和 Ted-talk 数据集) 上取得了最先进的结果。

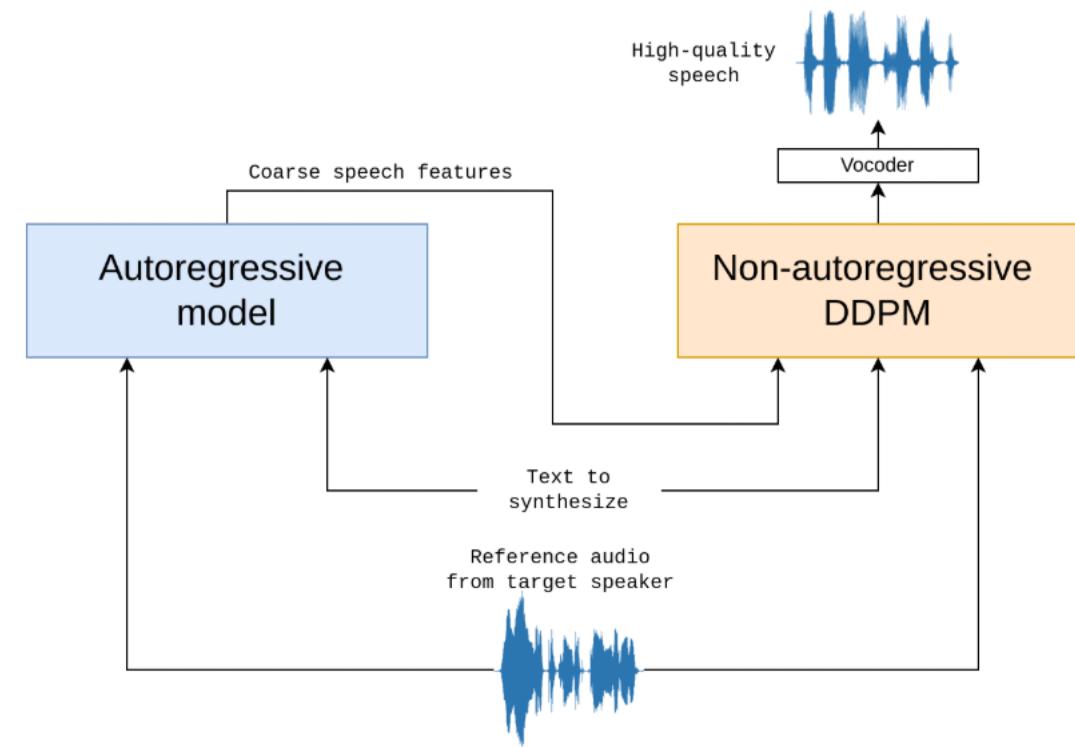
Conclusion:

Animate Anyone 提供了一种将静态角色图像转换为动画视频的有效方法，能够保持角色外观细节的一致性，并实现精确的姿态控制和时间稳定性。该方法不仅适用于一般角色动画，还在多个基准数据集上取得了优于现有方法的结果。尽管在手部动作的稳定性和生成未见部分时存在一些挑战，但其在角色动画领域展示了强大的潜力，并可能激发更多创新和创造性的应用。

MARS5

[Why MARS5?](#) | [Model Architecture](#) | [Samples](#) | [Camb AI Website](#)

Stars 2.6k 74 ONLINE HuggingFace [Join](#) Open in Colab



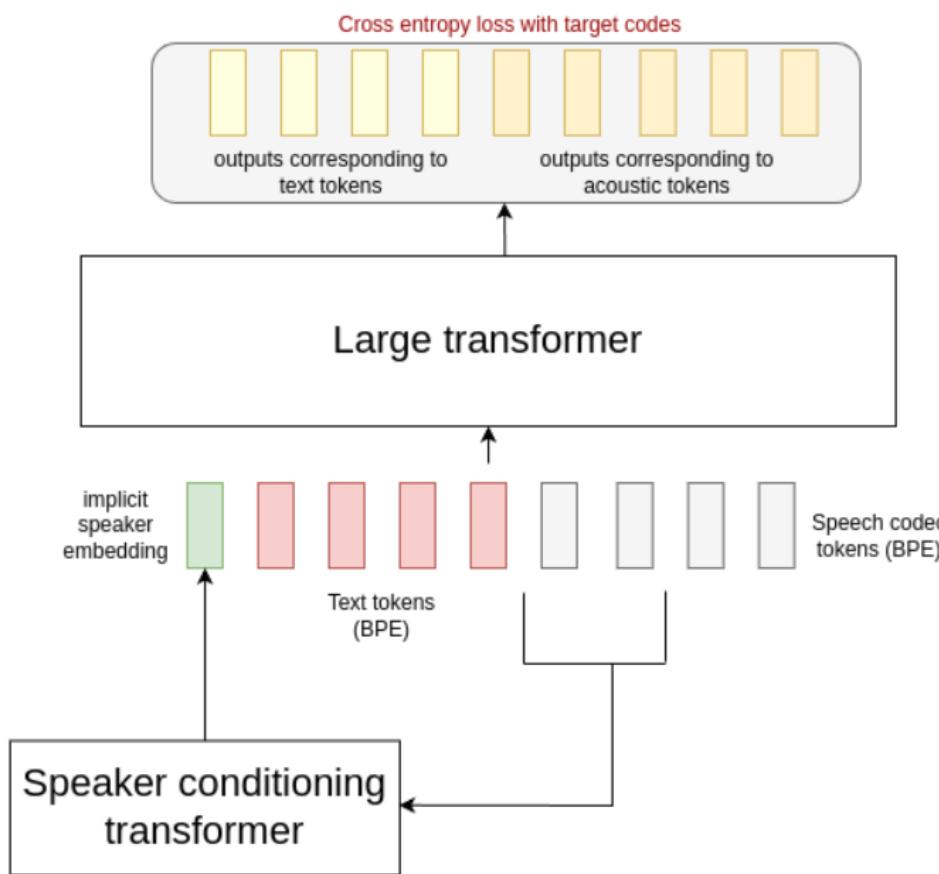
主要功能和优点：

- 开源：可本地部署集成到我们的pipeline中
- 多语言支持：支持140多种语言
- 复杂韵律处理：擅长处理体育解说、电影、动漫等具有复杂韵律的文本
- 参数引导：通过文本中的标点和大写等标记来引导语音的韵律和情感
- 多模式：快速克隆和深度克隆，根据自身需求选择生成速度或质量

◆ [项目官网](#)

◆ [Github仓库](#)

◆ [Demo](#)



a. Autoregressive component

AR模型使用Mistral-style的decoder-only transformer模型来预测Encodec L0编码

通过Next token prediction任务进行预训练

推理时通过浅克隆或者深克隆的方式进行采样生成L0 code

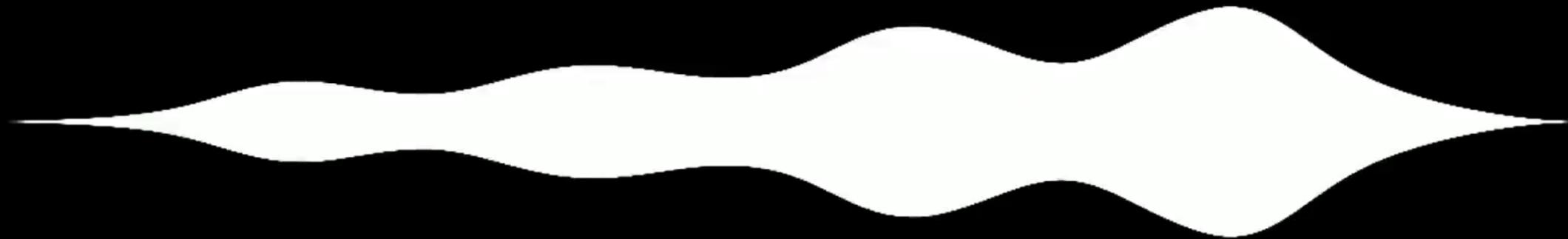
浅克隆：直接将参考音频的嵌入和目标文本拼接（推理更快但克隆效果没那么好）

深克隆：还将参考文本与目标文本拼接然后再将参考音频token与输入序列进行拼接再进行采样

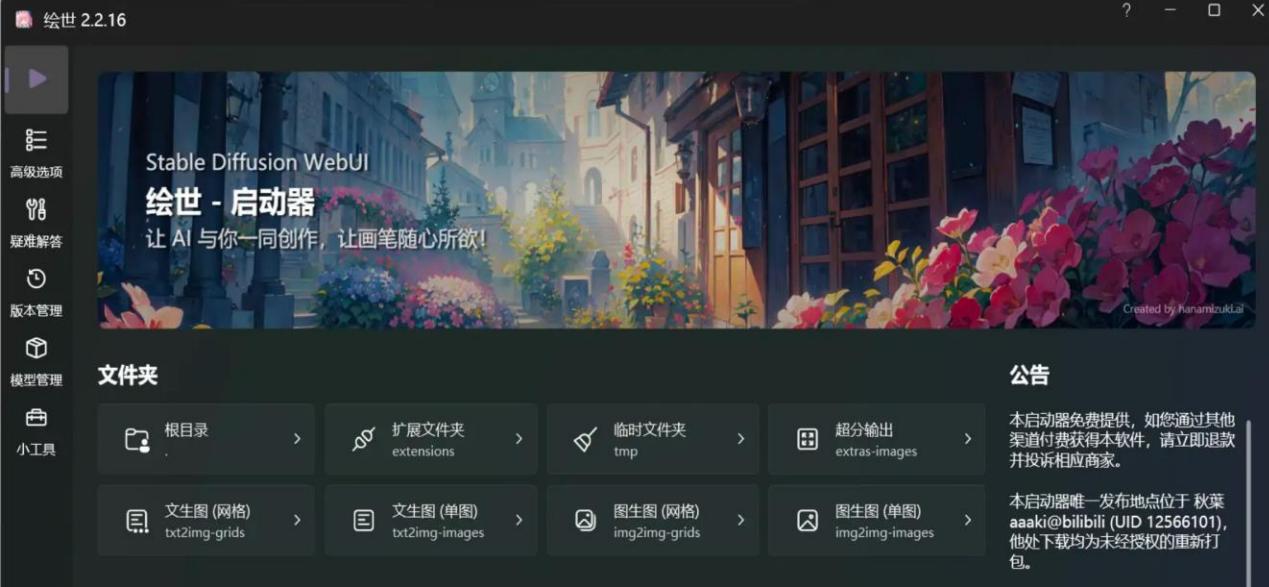


Ref: We actually haven't managed to meet demand.

Desired: Hi there, this is the test.



LADIES AND GENTLEMEN



AUTOMATIC1111 / stable-diffusion-webui Public

Code Issues 2.3k Pull requests 56 Discussions Actions Projects Wiki Security Insights

Home

w-e-w edited this page on Sep 10, 2023 · 37 revisions

Stable Diffusion web UI is a browser interface for Stable Diffusion based on Gradio library.

Setup

- Install and run on Nvidia GPUs
- Install and run on AMD GPUs
- Install and run on Apple Silicon
- Install and run on Intel Silicon (external wiki page)
- Install and run via container (i.e. Docker)
- Run via online services

Reproducing images / troubleshooting

- Seed breaking changes
 - Still can't reproduce results? Try this first.
- General troubleshooting

Pages 33

Setup

- Install and run on Nvidia GPUs
- Install and run on AMD GPUs
- Install and run on Apple Silicon
- Install and run on Intel Silicon (external wiki page)
- Install and run via container (i.e. Docker)
- Run via online services

Reproducing images / troubleshooting

- Seed breaking changes
 - Still can't reproduce results? Try this first.
- General troubleshooting

Usage

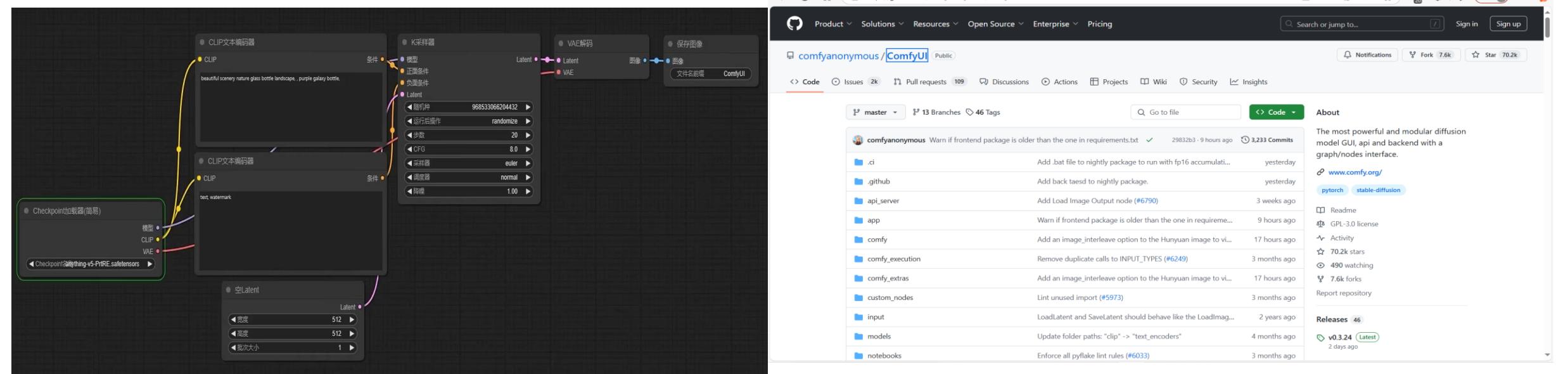
- Features

项目调研：

目前市面上有好用且开源的类似项目有两个，第一个就是大名鼎鼎的有哔哩哔哩up主秋葉aaaki，所整合的基于stable-diffusion-webui的一键整合包---绘世启动器。具体部署流程以及流水线可参考所发布的教程。

stable-diffusion-webui，是一种利用Matlab编程语言所实现的用于图像处理的数学计算模型。它是基于扩散方程和平衡反应的数学模型，可以进行多种类型的图像处理，包括图像去噪、边缘检测、图像增强、图像分割等。

这个模型可以应用于AI绘画软件中。AI绘画软件是一种基于深度学习和图像处理技术的绘画软件。通过应用stable-diffusion-webui模型，可以对图像进行处理和增强，提高绘画软件的效果和准确度。



项目调研：

第二个也是哔哩哔哩up主秋葉aaaki，所整合的基于ComfyUI的一键整合包。具体部署流程以及流水线也参考其所发布的教程。

ComfyUI，一款基于节点工作流稳定扩散算法的图形界面。通过将稳定扩散的流程巧妙分解成各个节点，成功实现了工作流的精准定制和可靠复现。**ComfyUI**工作流指的是一种基于节点式的工作流程，它通过将稳定扩散的流程分解成多个节点，实现了更加精细化的流程定制和更高的结果可重用性。这种工作流的设计使得用户能够通过直观的节点式界面设计和执行复杂的稳定扩散工作流程，无需编写任何代码。**ComfyUI**工作流不仅提高了工作效率，还使得复杂任务的处理变得更加直观和高效。在图像生成方面，**ComfyUI**相较于传统的WebUI具有更快的速度和更经济的显存占用，特别是在生成大图片时，不会导致显存爆满，而是通过切块运算来避免图片碎裂的问题。

稳定扩散：Stable Diffusion(稳定扩散)指的是一种文本到图像的人工神经网络模型，能够理解用户输入的描述并生成相应的图像。这种模型基于大量的数据进行训练，其作用是学习如何将输入的文本描述转化为图像内容。



CS 330 MIP – Lab 04

多媒体信息处理介绍实验课 - 如何做项目3

Multimedia Information Processing Introduction - Projects

Jimmy Liu 刘江

2025-03-12