

面向教育内容生成的动画多镜头音视频数据集构建

摘要

面向情景化教育视频生成，现有音视频数据集在多镜头结构、动画域适配和音画一致性方面仍存在不足。构建动画多镜头音视频数据集 Edu-Anime-MultiShot，并形成“场景—镜头”层次组织、分阶段质量过滤和“全局—局部”结构化标注相结合的半自动构建方法。方法重点设计动画多镜头数据治理规则与双阈值门控，并通过场景级角色原型和统一角色 ID 约束降低跨镜头指代漂移。与 TalkCuts 相比，该数据集规模较小，但增加双层结构、对白—角色映射和角色音色信息。人工抽样显示，画外音比例由 24.6% 降至 1.2% 以内。结果表明，该流程可改善数据结构化程度和音画一致性，为教育内容生成提供可复用的数据治理方案。

关键词

多模态数据集；多镜头；数据治理；音视频对齐；动画风格

中图分类号: TP391.41

文献标志码: A

Construction of an animated multi-shot audio-visual dataset for educational content generation

Abstract

Contextualized educational video generation requires training data with multi-shot structure, animation-domain adaptation, and reliable audio-visual correspondence. To address these requirements, Edu-Anime-MultiShot was constructed, and a semi-automatic construction method integrating scene-shot hierarchical organization, staged quality filtering, and global-local structured annotation was developed. Instead of modifying backbone models, task-specific governance rules and threshold design were emphasized. Visual-subject coverage within each speech interval was used as a local constraint, and the proportion of qualified speech duration was used as a global constraint; a two-level gate was adopted to balance sample retention and audio-visual consistency. Scene-level character prototypes and shot-level unified character IDs were used to reduce cross-shot reference drift. Compared with TalkCuts, which contains more than 500 h of multi-shot human speech videos, the constructed dataset is smaller but provides a scene-shot hierarchy, dialogue-character mapping, and character-level voice attributes for animated educational scenarios. Manual sampling showed that off-screen speech decreased from 24.6% in raw candidates to below 1.2% after governance. The resulting dataset provides a reusable data-governance and structured-annotation framework for educational animation generation, while the platform results should be interpreted as observational evidence rather than causal proof of learning improvement.

Key words

Multimodal Dataset, Multi-Shot, Data Governance, Audio-Visual Alignment, Animation Style

0 引言

近年来，随着多模态大语言模型（multimodal large language models, MLLMs）与生成式人工智能（generative artificial intelligence, GenAI）在教育人工智能（artificial intelligence in education, AIED）领域的深入应用，自动生成情景化、多模态的教学内容已成为重要趋势。近期多项综述与研究也指出，大语言模型在赋能个性化教育与重塑多模态教学辅助模式方面展现出了巨大的应用潜力^{[1][2]}。特别是在第二语言教育领域，生成式人工智能通过自动构建高度情境化的多模态学习资源，正在深刻重塑传统的语言教学与评估范式^[3]。在第二语言学习、幼儿启蒙及科普教育等场景中，利用特定风格（如 Disney 动画风格）的虚拟人物进行知识讲解与场景模拟，能够突破传统枯燥的文本教学，显著提升学习者的沉浸感、情感联结与学习动机。然而，训练具备高表现力、能够进行长时序叙事的多镜头动画人物音视频生成模型，极度依赖于高质量、结构化的底层训练数据。

现有开源音视频数据集大多聚焦于单镜头、自然真人场景或开放域描述，难以同时满足动画域、多镜头结构和音画一致性要求。本文面向教育内容生成任务构建 Edu-Anime-MultiShot，并提出半自动构建方法。需要明确的是，本文不将基础模型结构作为创新点，技术贡献主要集中在三方面：（1）构建 52.4 h 的动画多镜头音视频数据，并采用“场景—镜头”双层组

织，以补充 TalkCuts 等真人多镜头数据在动画教育场景上的覆盖不足；（2）提出“语音区间局部覆盖率—视频级合格语音时长占比”的双阈值治理规则，将语音有效性、视觉主体性与音画一致性组织为可执行的分层门控流程；（3）采用场景级角色原型、统一角色 ID 和镜头级对白约束形成“全局—局部”标注结构，并从分阶段过滤、人工抽样质量核验和应用观察三个层面进行评估。

1 相关工作

1.1 多模态视频与音视频数据集构建

近年来，多模态学习的发展推动了视频及音视频数据集的快速演进，研究重心也逐步从单纯追求规模转向兼顾数据质量、语义对齐和任务适配性。

在视频数据集方面，近年来研究逐步从弱监督构建转向高质量筛选与对齐。例如，HD-VILA 通过多阶段过滤提升视频与文本之间的语义一致性^[4]，InternVid、Panda-70M 等数据集进一步在规模与质量控制方面进行优化^{[5][6]}。在音视频数据集方面，M³AV 提供语音、视觉及文本的联合数据^[7]，CASTLE 2024 关注多视角同步采集^[8]。TalkCuts 则面向多镜头人物说话视频，包含约 16.4 万个片段、超过 500 h 视频和 1 万余个身份，并提供文本描述、2D 关键点与 3D SMPL-X 运动标注^[9]，代表了当前多镜头真人说话数据在规模和运动标注上的重要进展。

与 TalkCuts 相比，Edu-Anime-MultiShot 的 52.4 h 数据规模明显更小，

且不以身份数量和人体运动标注覆盖为主要目标；其差异主要体现在结构复杂度和应用场景上：一是面向风格化动画而非真人视频，需处理夸张口型、角色形变和镜头切换造成的域偏移；二是显式组织“场景—镜头”层级，并增加对白—角色映射、跨镜头角色 ID 与角色音色等字段；三是面向情景化教育视频、虚拟教师和第二语言教学等任务。因此，本文的定位并非以规模超越 TalkCuts，而是针对动画教育场景补充一套可复用的治理规则与结构化标注方案。

1.2 多镜头视频结构建模与跨镜头一致性问题

视频的镜头与场景结构是组织复杂叙事逻辑与高级语义的重要基础。镜头边界检测是视频结构建模的前提工作之一。近年来，基于深度学习的方法已在镜头切分任务上取得较高性能，其中 TransNet V2^[10]提出了一种高效且准确的镜头转场检测网络，为大规模视频档案的镜头级分析提供了有效工具。这类方法为将长视频拆分为结构化镜头单元提供了技术基础。

在更高层次的场景建模方面，研究逐步从单纯的镜头切分转向跨镜头的语义聚合与长时序建模。ImageBind 通过构建图像、文本、音频等多模态统一嵌入空间，为跨模态、跨片段的相似性建模提供了新路径^[11]。此外，如 Video-LLaVA 等最新前沿工作通过在特征投影前进行模态对齐，显著增强了视频大语言模型（Video-LLMs）对复杂视觉特征的统一表征能力^[12]。这类统一表征方法使得相邻镜头之间的语义关联分析成为可能，也为镜头聚合到场景级单元提供了表示基础。

近期，研究重心加速向“长时序”与“跨镜头一致性”转移。例如，

TimeExpert 等专用 Video-LLMs 通过解耦时间戳预测与语义描述，实现了对视频内在时间结构的精准对齐^[13]。在多镜头生成领域，MultiShotMaster 则提出了一种可控的多镜头生成框架，着重解决跨镜头的人物与场景一致性问题^[14]。

教育动画通常同时包含对话、角色动作和情景转换，若跨镜头角色指代不稳定，将直接影响后续生成模型对人物一致性和教学情境连续性的学习效果。因此，本文在镜头切分基础上，进一步强调场景级语义组织和跨镜头角色一致性约束，以适配教育内容生成场景对长时序结构化数据的需求。

1.3 自动化数据治理与质量过滤方法

在大规模多模态模型训练中，自动化数据治理是保证训练数据可用性和稳定性的关键环节。数据治理关注原始数据中的噪声剔除、有效样本保留以及跨模态一致性控制问题。已有研究表明，仅依赖原始网络视频数据会显著引入低质量样本，从而削弱模型学习效果，因此多阶段过滤与质量控制已成为高质量视频数据集构建中的常见策略^[15]。

在音频处理方面，鲁棒的自动语音识别是实现大规模语音内容对齐与治理的重要基础。Whisper^[16]等大规模自动语音识别模型基于大规模弱监督语音数据训练，展现出较强的多语种识别和零样本泛化能力，为复杂场景下的语音转写和时间戳定位提供了可靠支撑。在视觉主体筛选方面，目标检测模型被广泛用于识别视频中的有效主体区域。虽然 YOLOv8^[17]本身是通用实时检测框架，但其高效检测能力使其成为构建特定场景视觉主体筛选模块的重要基础。除了视觉和语音层面的基础工具，

生成式音频的质量控制与真实性评估也逐渐成为多模态数据治理的重要议题。相关综述系统梳理了深度伪造音频的生成机制、检测方法及其应用挑战，为理解音频可靠性、生成内容真实性及多模态数据处理中音频质量评估问题提供了有益参考^[18]。

总体来看，现有研究已经为自动化过滤、语音识别和视觉筛选提供了较成熟的技术基础，但针对动画多镜头场景中“语音有效性—视觉主体性—跨镜头一致性”协同治理的问题，仍缺乏面向特定任务的数据治理范式。本文正是在这一背景下，提出面向教育内容生成的分层数据治理流程，以提升数据集在复杂动画场景中的可用性与稳定性。

1.4 多模态标注与视频语义结构化建模

除数据过滤外，如何将原始音视频数据转化为可供模型学习的结构化语义表示，也是多模态数据集构建中的核心问题。随着大语言模型和视觉语言模型的发展，自动化语义标注逐渐从粗粒度描述扩展到包含动作、角色、语境及时间边界的细粒度结构化输出。Qwen2.5-VL 等新一代视觉语言模型在视觉识别、长视频理解和跨模态交互方面表现出更强能力，为复杂视频的层次化描述提供了新的可能性。近期研究如 ShareGPT4Video^[19]证实，通过多模态大模型生成的高质量、细粒度视频描述，能够显著增强生成模型对复杂时空细节的掌控力。

在视频理解领域，统一嵌入空间和长时序建模方法的发展，使“全局语义—局部镜头”两层级标注思路逐渐具有可操作性。场景级标注有助于建立视频的整体角色关系、环境背景与情节主线，而镜头级标注则适合细化动作、对白及时间边界。

相比仅提供单条视频描述的传统方案，层次化结构化标注更适合支持长视频生成、人物一致性建模以及高精度时序对齐等任务。

然而，现有研究中多数自动化标注框架仍主要面向自然图像或短视频的整体描述生成，对于多镜头动画场景中角色连续性、对白—角色映射以及跨镜头一致性控制的研究仍较少。特别是在教育内容生成场景中，标注不仅要回答“画面中有什么”，还需要描述“谁在说话、如何说、在什么情境中说”，这对多模态标注的结构化程度提出了更高要求。因此，本文在语音识别和视觉语言模型协同基础上，构建了面向场景级与镜头级的双层结构化标注体系，以提升数据对教育内容生成任务的适配性。

1.5 教育内容生成与相关多模态资源构建

教育内容生成已成为教育人工智能的重要研究方向之一。相关综述指出，大语言模型和生成式人工智能正在推动教学资源生成、个性化辅导和多模态教学支持方式的发展^[20]。与此同时，教育技术领域已有较多关于动画教学代理和多媒体学习的实证研究：教学代理的指向手势与视线提示能够改善学习者对关键信息的注意分配，并在部分实验中提升保持与迁移成绩^[21]；在英语作为外语的学习环境中，教学代理的具身程度与信息模态会影响学习结果^[22]；具身动画教学代理还被报告与更好的学习结果及学习过程指标相关^[23]。相关元分析进一步表明，教学代理手势对保持与迁移具有小到中等效应，但效果受到任务和设计因素调节^[24]。

从资源形态看，现有教育多模态资源更多集中于课件、讲座视频、文本问答对

或课堂互动数据，而对动画角色驱动、多镜头情景化叙事和高精度音画对应的底层数据资源关注不足。尤其在面向教育内容生成的视频生成任务中，角色连续性、对白时序、场景组织以及教学语境的一致性会直接影响生成结果的教学可用性和沉浸感。

因此，面向教育内容生成的数据集构建不仅是通用多模态数据建设问题的延伸，更具有明确的任务导向和场景特殊性。本文从教育动画生成需求出发，将视频结构建模、自动化治理和结构化标注有机结合，旨在为情景化教学视频生成、虚拟教师驱动和第二语言教学等任务提供更有针对性的数据基础。

2 数据集架构与采集

为满足多镜头教育动画生成模型对时空对齐与复杂叙事逻辑的要求，本研究设计了一套面向多模态联合建模的层次化数据架构，并制定了严格的数据采集与规范

化标准。

2.1 层次化数据架构

高质量的多镜头视频生成本质上是一个从“宏观语义控制”到“微观时空执行”的生成过程。为此，本文数据集以“场景(Scene) — 镜头(Shot)”为核心骨架，构建了如图1所示的双层结构表示机制：

全局语义层 (Global Level – Scene): 以“场景”为最大粒度的组织单元。由于原始长视频包含多个不连贯的故事情节，本架构通过镜头聚合技术将物理相邻且语义连贯的多个镜头组合为一个场景。在数据表示上，全局层负责锚定高维的上下文信息，包括：核心出场角色、环境背景以及整体故事主线。

微观时空层 (Local Level – Shot): 以“镜头”为最小粒度的对齐单元。在每个场景内部，包含若干个视觉连续、无剪辑跳跃的单镜头。在数据表示上，微观层负责细粒度的多模态锚定，包含：特定角色的具体动作、镜头运动语言、高精度时间戳

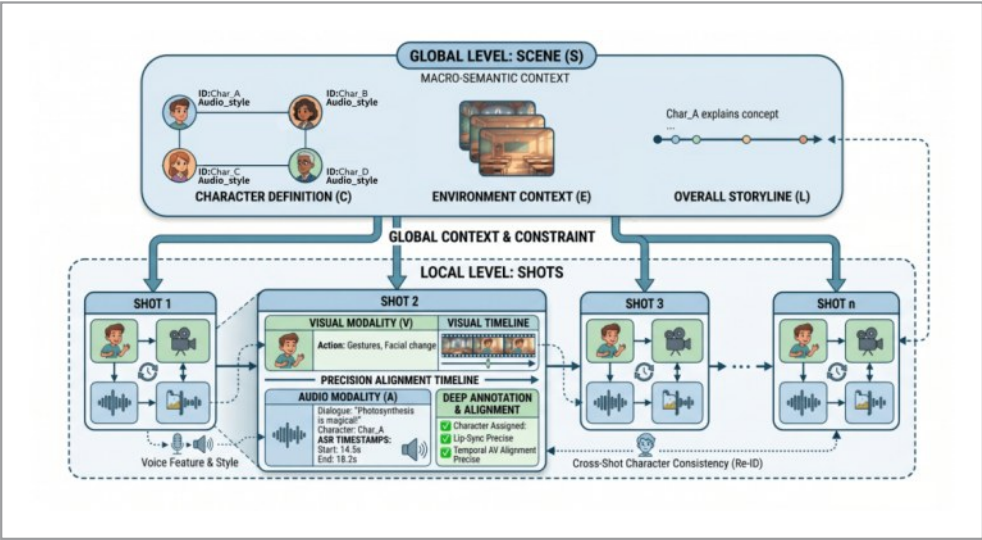


图1 “场景(Scene)—镜头(Shot)”双层数据结构

以及严格对齐的语音对白。

通过这种层次化架构，数据集不仅实现了音视频特征在物理时间轴上的精准绑定，还建立起了跨镜头间角色与场景特征的逻辑关联，为后续模型学习长时序连贯性提供了结构化的数据支撑。

2.2 源数据采集

在数据获取阶段，本文主要从公开的视频平台定向采集具有情节对话的高质量动画视频。区别于泛领域的自然场景数据，本数据集的采集严格遵循以下约束：

题材选择：聚焦于具备高表现力的三维或高质量二维商业动画风格素材。该类素材通常具有较丰富的面部表情、口型变化和肢体动作，适合用于虚拟教师或教学角色驱动研究。公开数据版本不以具体商业品牌或角色名称作为分类标签。

质量筛选：剔除分辨率低于 720p、帧率低于 24 fps 的视频源；为兼顾叙事完整性与计算效率，优先保留总时长为 1~5 min 的完整情节片段。

2.3 数据合规与伦理声明

数据合规贯穿采集、处理与发布全过程。源数据在内部记录来源页面、权利状态和处理用途，原作品版权仍归原权利人所有。对于商业动画素材，公开发布版本仅纳入已获得研究使用/再分发许可或许可条款明确允许公开再分发的内容；对无法确认再分发许可边界的原始视频和音频，仅用于内部方法验证，不对外提供可还原的媒体片段，仅发布不可逆统计特征、时间索引或不含原始媒体内容的结构化标注。发布前去除平台账号、上传者昵称、原始 URL、抓取时间等可识别来源元数据，内部以哈希化 source_id 保留审计与溯源关

系；角色名称统一映射为 Character ID，避免直接使用商业角色名称；连续对白文本按照任务所需最小化保留，避免形成对原作品的替代性再传播。许可记录、剔除清单及数据版本信息在内部留档，以支持后续合规审计。

3 多阶段层级数据治理方法

由于网络采集的原始视频源包含大量不具有代表性或无法对齐的高噪样本，直接应用会严重削弱模型的生成效能。因此，本章设计并实施了一套数据治理范式，从时序结构化、语音有效性及视觉主体性等维度对数据进行深度筛选与治理，以逐层收缩的质量控制机制，彻底清洗了来源多样、质量参差不齐的噪音数据。

3.1 场景与镜头分割

原始视频通常包含复杂的剪辑结构及冗余内容，难以直接用于多镜头建模。为此，本文首先在物理层面对视频进行镜头级分割处理。

第一阶段利用 Transnet v2 镜头边界检测模型进行镜头切割，对视频进行时间轴切分，将原始视频划分为一系列时间连续且视觉语义相对一致的镜头片段 (shot)。每个镜头片段保留原始音频轨道，从而形成基本的音视频单元。在此基础上，进一步根据镜头间的语义连续性，使用 ImageBind 对相邻镜头进行合并，构建候选多镜头场景片段 (scene)。该过程以镜头间视觉相似性及时间邻接关系为依据，确保合并后的片段能够表达完整的对话或动作过程。在合并之前对镜头片段进行简单的筛选，目的是过滤掉坏死的空壳

镜头切片。

为保证数据质量，对候选片段施加进一步的约束：剔除时长过短（低于3秒）或过长（超过15秒）的片段、镜头频繁切换的片段和仅包含单个镜头的片段，以确保数据具备多镜头结构特征的同时镜头切换更稳定。通过上述处理，构建得到结构化的候选数据集，为后续多模态过滤提供基础。

3.2 多层级数据过滤

在完成候选多镜头片段构建后，本文进一步设计了一种基于多模态协同的分层过滤机制，从语音、视觉及跨模态一致性等多个维度对数据进行精细化筛选。该机制以逐层收缩的方式对数据进行质量控制，逐步剔除噪声数据并保留高一致性样本。显著提升数据在语音有效性、主体一致性及音画同步方面的整体质量，为多镜头音视频生成模型的训练提供高可靠的数据基础。

3.2.1 语音有效性过滤

针对多镜头人物说话任务，语音信息的有效性与主导性是数据可用性的前提。本文基于语音活动检测（Voice Activity Detection, VAD），将连续音频划分为若干语音区间：

$$S = \{(s_i, e_i)\}_{i=1}^N \#(l)$$

其中 (s_i, e_i) 表示第 i 段语音的起止时间。

该过程用于定位视频中的“真实说话区间”，为后续音画对齐提供时间约束基础。在此基础上，引入语音占比对候选片段进行筛选。对于每个片段，计算语音区间在整体时长中的占比，并结合语音段的平均长度与分布密度，过滤掉语音稀疏或

高度碎片化的样本。

3.2.2 视觉主体性筛选

在视觉层面，本文关注视频中是否存在稳定且可感知的表达主体。考虑到源数据的动画特性，本文采用基于YOLOv8的卡通人脸检测模型对视频帧进行低频采样检测。具体而言，对视频按约2 FPS进行均匀采样，并对每一采样帧执行人脸检测操作，记录检测到人脸的时间点集合。若整个视频中未检测到任何人脸，则判定该样本不具备有效视觉主体，直接过滤。

该步骤用于剔除以下类型数据：仅包含背景或场景切换的视频、片头/片尾非角色画面、无人物出现的动画片段，从而保证数据集中所有样本均具备明确的角色表达基础。

3.2.3 双阈值音画一致性判定

在自然真人视频处理中，SyncNet^[25]常用于估计唇音同步置信度。对于风格化动画，其绝对置信度会受到非刚性面部形变、夸张口型以及镜头切换等域偏移影响。

动画域偏移：风格化角色的嘴部形态和表情运动与真人训练域存在差异，SyncNet的绝对分数可能发生系统性偏移，难以直接解释为“同步/不同步”的可靠判据。

多镜头不连续性：机位切换、景别变化和遮挡会破坏连续的人脸轨迹，使像素级同步检测更容易产生假阴性。基于上述原因，本文采用与动画域特征更匹配的规则化一致性筛选。

因此，本文摒弃了直接依赖像素级唇音同步工具的路线，转而设计了一种基于“语音活性-视觉主体约束”的音画对齐判定机制。为实现语音与视觉内容的精确对

应，在每个语音区间内，对视觉检测结果进行时间匹配分析。在语音区间 (s_i, e_i) 内统计采样帧中检测到人脸的比例：

$$R_i = \frac{N_i^{face}}{N_i^{sample}} \#(2)$$

其中， N_i^{face} 为语音区间内检测到人脸的帧数， N_i^{sample} 为该语音区间内采样帧总数。当满足：

$$R_i > \tau_1 \#(3)$$

则认为该语音片段与视觉主体存在有效对应关系(本文设定局部阈值 $\tau_1 = 0.3$)。

进一步地，为评估整个视频的音画同步质量，定义全局一致性指标：

$$R = \frac{\sum_{i=1}^N \mathbb{I}(R_i > \tau_1) \cdot (e_i - s_i)}{\sum_{i=1}^N (e_i - s_i)} \#(4)$$

其中，分子表示满足音画对齐条件的语音时长，分母表示总语音时长。

当满足：

$$R \geq \tau_2 \#(5)$$

则认为该视频在整体层面具备稳定的音画同步关系（本文设定全局阈值 $\tau_2 = 0.5$ ）。否则判定存在以下问题并予以过滤：画外音、无人物回应的旁白片段或镜头切换导致的视觉主体缺失。

本文的治理规则采用局部—全局两级门控。局部指标 R_i 衡量单个语音区间内检测到视觉主体的帧比例，设置 $\tau_1 = 0.3$ ，用于排除长时间只有语音、缺少可见说话主体的区间；全局指标 R 统计满足局部条件的语音时长占总语音时长的比例，设置 $\tau_2 = 0.5$ ，用于避免少量有效片段掩盖大段画外音或旁白。两个阈值基于开发集抽样与人工校验确定，目标是在样本保留率和音画一致性之间取得经验性平衡。该设计的技术贡献在于把语音活性、视觉主体存

在性和视频级质量控制转化为可解释的双阈值门控规则，而不是修改基础检测模型。阈值并非跨域通用常数，迁移到其他动画风格、虚拟数字人或真人教学视频时需重新标定；对阈值附近样本仍建议人工抽检。

4 多模态结构化标注方法

为进一步提升数据集在多模态生成任务中的可用性，本文构建了一套基于语音模型与多模态大模型协同的自动化标注框架。需要说明的是，本文所采用的自动化标注框架以预训练模型为核心支撑，其输出质量在一定程度上受模型训练数据、跨域泛化能力以及对动画场景适配性的影响。尤其是在角色外观变化明显、镜头切换频繁、多人交互密集或语音信号复杂的场景下，模型可能出现描述偏差、角色指代不稳定或对白绑定不准确等问题。因此，本文将模型作为半自动构建流程中的候选标注生成器，而非无条件可信的最终决策器，并通过层级提示约束、规则校正和一致性控制等方式对模型输出进行进一步约束。

该框架从音频、视觉及语义三个层面对数据进行结构化标注，形成如图2统一的“全局——镜头”层级表示，实现了数据资产从无序的物理像素向有序的深层结构化语义转变。

4.1 基于语音模型的时序对白标注

在多镜头教育动画中，对白是驱动角色口型与情感表达的核心信号。在音频标注阶段，本文首先对视频进行音频解耦与标准化处理，利用音视频处理工具将视频中的音轨提取为 16 kHz、单声道的无损音频，以满足语音识别模型的输入要求。随

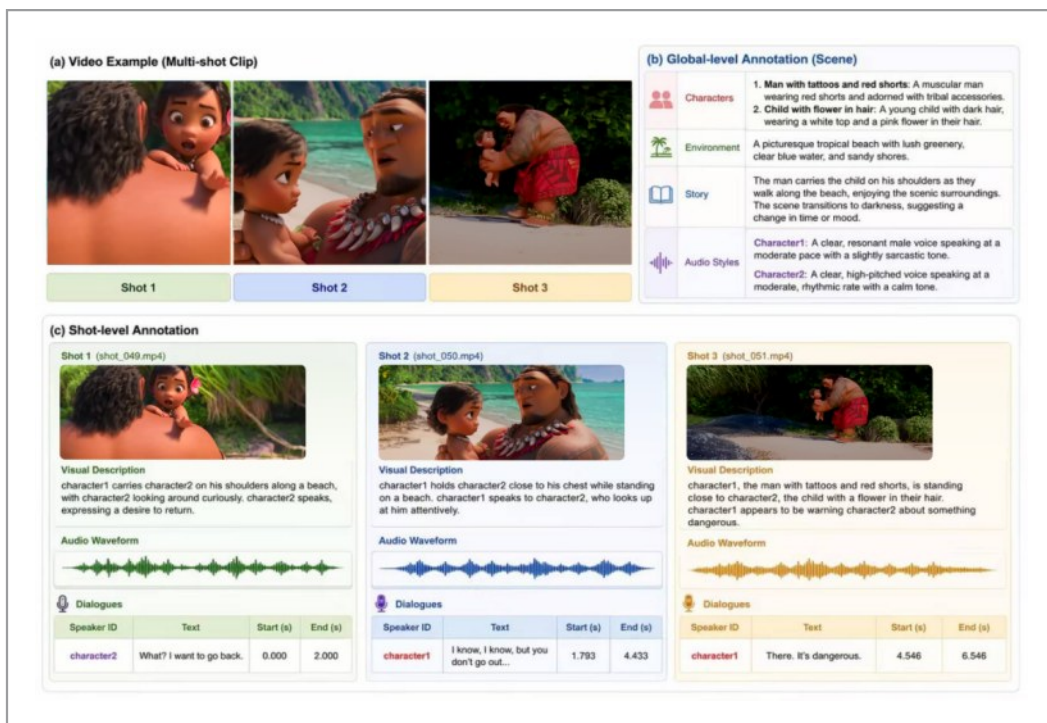


图2 “全局——镜头”多层次多模态标注示例

后，采用 whisper 自动语音识别模型对音频进行转录，获得带时间戳的语音片段序列：

$$A = \{(t_i^{start}, t_i^{end}, w_i)\}_{i=1}^M \#(6)$$

其中， t_i^{start} 和 t_i^{end} 表示语音片段的起止时间， w_i 表示对应文本内容。

为适配多镜头结构，本文在镜头级别对音频进行切分，并分别进行语音识别，从而获得每个镜头内部的对白及其时间对齐信息。进一步对识别结果进行清洗，去除空文本及异常片段，并统一时间精度。该过程为后续多模态标注提供了精确的语音时间锚点，使视觉内容与语义信息能够在时间轴上对齐。

需要指出的是，语音识别结果仍受背景音乐、多人重叠发声、角色夸张配音风格及音频质量波动等因素影响，因此本文在后续标注环节中仅将其作为对白候选与

时间锚点来源，并结合镜头级语义校验进行修正。

4.2 基于多模态大模型的视觉语义标注

为了解决长视频生成中常见的角色身份跳变与场景失真问题，本文设计了一种层次化提示策略，驱动 Qwen2.5-VL 模型进行深度描述^[26]，该过程分为全局标注与镜头级标注两个层次。需要说明的是，尽管视觉语言模型在复杂视频理解方面表现出较强能力，但其输出仍可能受到模型先验知识、上下文长度以及动画风格适配性的影响，进而出现角色命名漂移、局部描述不一致或细节幻觉等现象。因此，本文通过全局角色定义、统一角色 ID 约束和镜头级输出归并等机制，对模型生成结果进行结构化约束与后处理。

4.2.1 全局语义建模

首先对完整场景（Scene）进行全局表征。将完整视频输入多模态模型，生成全局语义描述，包括：

角色信息，包括角色的视觉特征（发型、服饰）与 ID 唯一化分配（如 Character1）。

环境描述，界定场景所处的物理空间及整体基调。

故事情节，概括该片段的主要人物之前的情节主线。

通过预定义结构化提示模板，引导模型输出标准化文本，并解析为结构化字段。该过程实现了视频级语义抽象，为后续镜头级理解提供上下文约束。

4.2.2 镜头级细粒度对齐标注

在镜头级别，本文结合视频片段与对应的语音识别结果，构建多模态输入，生成细粒度标注信息。具体而言：

模型输入：单镜头视频片段，语音识别对白（含时间戳）和全局角色定义信息。

模型输出：视觉描述（动作、镜头语言等）和对白——角色映射关系。

为保证标注一致性，本文设计严格的约束策略：仅允许使用统一的角色 ID（如 Character1），且对白必须来源于语音识别结果，此外仅在可确定说话者时进行对白绑定。

最终，生成的结构化镜头标注：

$$\mathcal{V}_k = \left(d_k, \{(c_{k,j}, w_{k,j}, t_{k,j}^{start}, t_{k,j}^{end})\}_{j=1}^{N_k} \right) \# (7)$$

其中， d_k 表示第 k 个镜头的视觉描述， $(c_{k,j}, w_{k,j}, t_{k,j}^{start}, t_{k,j}^{end})$ 表示第 k 个镜头中第 j 条对白的说话角色、文本内容及时间区间， N_k 为该镜头内对白数量。该过程实现了视觉内容、语音内容与角色信息的三者对齐，

有效解决了动画场景中多个角色共存时，对白与说话者难以匹配的问题。

4.2.3 跨镜头角色一致性校正

在多镜头动画场景中，由于景别变化、遮挡、角色表情夸张变形以及镜头切换频繁，单纯依赖镜头级独立标注容易出现角色指代漂移的问题。为缓解这一问题，本文在“全局—局部”标注框架基础上，进一步采用“场景级角色原型约束—镜头级输出归并”的一致性校正策略。

具体而言，首先在场景级生成统一的角色集合，并为每个角色建立包含外观特征、服饰信息和身份描述的角色原型；随后，在镜头级标注过程中，限制模型优先从该角色集合中进行引用，并结合镜头邻接关系、视觉描述相似性以及对白上下文，对可能冲突的角色标识进行归并修正。同时，强制使用统一的角色 ID 在各个分镜头中进行指代，要求同一场景内相同角色始终对应同一标识，不因镜头切换、景别变化或动作状态变化而重复生成新的角色名称或局部别名，从而降低跨镜头角色漂移带来的标注不一致问题。对于说话者难以明确判断的片段，不强制绑定角色，以减少错误传播。

该策略能够在一定程度上提升场景内角色标识的一致性，但需要指出的是，当前方法仍依赖预训练模型的视觉理解能力与提示约束效果。在远景镜头、多人同时出镜、角色外观变化较大或转场较快的场景下，跨镜头角色一致性仍存在进一步优化空间。后续可结合更稳健的角色重识别机制和人工复核回流策略，以进一步提升标注质量。

4.3 基于音频理解的角色音色建模

为进一步增强数据的多模态表达能力，本文引入音频理解模型对角色语音特征进行建模，具体采用 Qwen2.5-Omni^[28]。

首先基于镜头级标注结果，收集同一角色在不同镜头中的语音片段，并在时间轴上进行拼接，形成角色级语音集合：

$$\mathcal{S}_c = \{(t_i^{start}, t_i^{end}) | speaker_i = c\} \#(8)$$

$$\mathcal{U}_c = \bigcup_{(t_i^{start}, t_i^{end}) \in \mathcal{S}_c} [t_i^{start}, t_i^{end}] \#(9)$$

其中， $\mathcal{S}_c$ 表示属于角色 c 的所有语音时间区间集合， $\mathcal{U}_c$ 表示拼接后的角色级语音序列。随后，输入多模态模型进行语音风格分析。模型输出包括：音色特征、音高、语速以及情绪语气。最终，将上述信息压缩为简洁的语音风格描述，并作为角色级属性加入数据集中：

$$audio_style(c) = f_\theta(\mathcal{U}_c) \#(10)$$

其中， f_θ 表示音频理解模型， $\mathcal{U}_c$ 为角色 c 的语音输入，输出为对应的语音风格描述。

该步骤实现了从“语音内容”到“语音风格”的扩展，使数据集不仅包含“说了什么”，还包含“如何说”，实现了从

“语义对白”到“表现力特征”的跨模态闭环。

5 数据集质量评估

为了验证本治理框架的有效性，Edu-Anime-MultiShot 数据集在生成任务中的高质量属性，本章从治理效能分析、标注质量校验、音画同步评估以及面向教育场景的代理任务验证四个维度进行系统评估。

5.1 分阶段过滤结果

本数据集从约 282.5 h 的原始动画视频出发，依次执行基础时长过滤、语音有效性过滤、视觉主体筛选和音画一致性治理，最终保留 52.4 h 数据。表 1 报告的是串行治理流程的分阶段存留统计：每一阶段均以前一阶段输出为输入，阈值在统计前固定，表中“存留率”均相对于原始采集数据计算。该结果用于说明各阶段造成的数据收缩及主要噪声类型。

表1 分阶段数据治理统计

治理阶段	剩余时长 (h)	存留率 (%)	核心解决噪声类型
原始采集 (Raw)	282.5	100%	—
基础时长过滤	258.6	91.5%	过短/过长镜头、物理坏死切片、镜头频繁切换
语音有效性过滤	119.8	42.4%	语音稀疏
视觉主体筛选	84.2	29.8%	空镜头、无角色背景、片头片尾
音画一致性治理	52.4	18.5%	画外音、旁白、音画异处

5.2 多模态标注质量验证

为验证自动化标注框架的准确性，本研究采用专家采样评估法评估标注精度。我们从数据集中随机抽取500个镜头片段，由三名具有语言学背景的标注人员进行双盲打分（1-5分）。需要说明的是，本轮实验记录仅保留了汇总结果，因此该部分定位为人工抽样质量核验而非严格的标注者一致性实验。

语音转写精度：Whisper 模型在本项目领域下表现稳健，字错率（WER）为4.2%。经人工校验，其对白时间戳与视觉口型变化的平均偏移量 $\Delta t < 120ms$ ，达到了“肉眼无感延迟”的高标准。

视觉描述准确性：评估 Qwen2.5-VL 生成的“镜头语言”与“角色动作”描述。实验结果显示，其在角色识别准确率上达到 92.4%，在描述复杂镜头语言（如 Medium Close-up）方面的准确性得分（MOS）为 4.38。

跨镜头角色一致性：验证角色 ID（Character ID）在多镜头间的分配准确率。实验表明，得益于“全局-局部”协同标注体系，跨镜头角色映射准确率达到 85.8%。

5.3 音画一致性评估

如 3.2.3 节所述，SyncNet 在动画域中的绝对置信度受域偏移影响，不能作为本文治理规则的直接过滤标准。因此，本节以人工画外音核验作为主要证据，并保留 SyncNet 相对变化作为辅助代理指标，用于观察治理前后数据是否呈现同向趋势。

SyncNet 辅助代理指标：在采用同一 SyncNet 模型、相同预处理方式和同类动画数据进行治理前后对比时，治理后样本的 Confidence Score 均值相对提高

32.5%。由于 SyncNet 与动画域存在训练分布差异，该数值不能解释为“唇音同步准确率提高 32.5%”，也不能单独证明治理有效；它仅说明在相同偏差来源下，代理分数呈现与人工核验一致的相对变化方向。

人工画外音检出率对比：在随机抽取的 1000 个原始样本中，画外音占比高达 24.6%；而在经本文治理后的最终数据集中，该比例下降至 1.2% 以内。这一结果支持本文治理流程能够明显减少“有语音但缺少对应视觉主体”的样本。受抽样方式和动画域范围限制，该结果用于数据质量核验，不作为对所有音画同步错误类型的完整测量。

6 教育平台应用验证

为检验 Edu-Anime-MultiShot 在实际教育内容生成流程中的可用性，本文将治理后数据用于图生视频模型训练，并接入“水杉中文”国际中文教育平台进行场景化应用观察。本章关注工程可用性和用户侧描述性指标，不将平台结果作为数据集导致教学效果提升的严格因果证据。

平台预设 6 名具有不同语言背景与角色设定的虚拟人物，用于对话演示和语法讲解。治理后多镜头数据用于训练图生视频模型，输入教学文本后生成包含口型、表情和镜头切换的动画片段。生成的动画内容被集成至“水杉中文”国际中文教育平台，用于情景对话和语法讲解等模块。图 3 展示平台学习页面示例。

6.1 教学内容的自动化生成与赋能

在国际中文教育场景中，学习者往往需要高度情景化的多模态素材来辅助语言

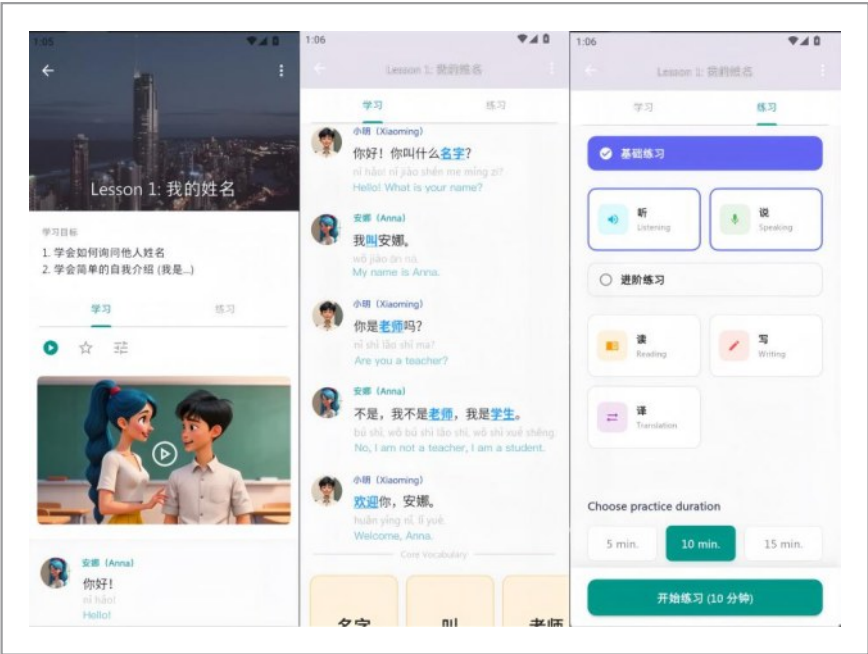


图3 “水杉中文”的学习页面

习得。本数据集通过驱动高性能生成模型，为“水杉中文”平台提供了以下核心支撑：

情景化教学视频定制：在训练阶段，模型充分学习了本数据集中“场景-镜头”的层级结构与叙事逻辑。在生成阶段，平台能够驱动模型自动化生成由多个动画分镜组成的情景对话视频，确保了跨镜头角色的一致性。

高表现力发音示范：得益于治理过程中对音画对齐精度的严苛控制，生成模型在角色口型精准度上表现优越。这为学习者提供了高质量的视觉发音参考，尤其是在声调与吐字细节的呈现上，有效解决了传统动画素材中“唇音不符”带来的干扰。

多模态交互式学习支架：平台利用数据集中丰富的视觉描述与角色属性标注，构建了交互式学习支架。模型生成的实时反馈内容在音色与视觉风格上与教学视频保持高度统一，学习者在与动画角色进行对话练习时，能够获得极具沉浸感的交互体验。

6.2 平台应用效果评价

生成的音视频内容上线后，该平台对教学内容更新前后学习者的参与度进行了对比分析。我们收集了主题学习模块的访问频率以及配套练习的正确率统计数据。结果显示，学习者对该模块的参与度从约58% 提升至 71%，表明新的教学形式显著提高了学习者的参与度。同时，练习的平均正确率从 63.8% 提升至 72.4%。同时，通过平台上的问卷收集了学习者的主观反馈。问卷包含 12 个问题，涵盖四个维度：学习体验、对视听内容的评价、对理解和记忆的贡献以及继续使用该模块的意愿。大多数问题为选择题，辅以少量开放性问题。结果表明，56% 的学习者认为新的教学形式“显著”或“大幅”提升了他们的学习体验；74% 的学习者认为它有助于他们理解真实的语言使用场景；63% 的人表示，与纯文字内容相比，更容易记住相关的词汇和表达；81% 的人表示愿意继续使

用这种教学方法。教师反馈认为，由高质量数据集支撑生成的动画教学素材在一定程度上缓解了传统合成视频中常见的角色不稳定与音画错位问题，有助于降低学习者的认知负荷。需要指出的是，上述结果主要用于说明本文数据集在教育场景中的应用潜力，而非对下游生成模型性能提升作严格因果归因。

6.3 方法可迁移性讨论

本文提出的数据集构建流程具有较强的模块化特征，其中镜头切分、场景聚合、语音有效性过滤、视觉主体筛选、音画一致性判定以及“全局—局部”结构化标注等步骤，并不局限于教育动画数据本身，因此对其他相关多模态视频数据集构建任务也具有一定参考价值。从应用场景看，该方法可进一步迁移至儿童启蒙动画、科普讲解类视频、虚拟数字人教学视频、剧情化知识讲解视频以及第二语言情景对话视频等场景。这些任务同样需要兼顾角色表达、音视频对齐和多镜头叙事组织，因此在数据治理流程上与本文具有较高相似性。但在迁移过程中，相关模块仍需结合目标场景进行适配。例如，不同视频类型下的视觉主体检测器、音画一致性阈值、镜头聚合规则以及标签体系均需重新标定；对于真人教学视频，传统唇音同步检测方法可能重新适用，而对于虚拟数字人或风格化动画场景，则需进一步关注角色模板一致性和渲染风格变化带来的影响。总体而言，本文工作的意义不仅在于构建了一个面向教育内容生成的动画多镜头音视频数据集，也在于为相关场景的多模态数据资源建设提供了一种可复用的流程框架。

7 结论

本文面向教育内容生成中的动画多镜头音视频数据需求，构建了 Edu-Anime-MultiShot，并形成由场景—镜头组织、分阶段质量过滤和全局—局部标注组成的半自动流程。本文的贡献重点在数据治理规则、阈值设计与结构化组织，而非 TransNet V2、ImageBind、Whisper、YOLOv8 或 Qwen2.5 系列基础模型本身。与超过 500 h 的 TalkCuts 相比，本数据集规模较小，但面向动画教育场景补充了双层叙事结构、对白—角色映射与角色音色等字段。

数据治理方法：通过语音有效性过滤、视觉主体筛选以及局部 $\tau_1=0.3$ 、全局 $\tau_2=0.5$ 的双阈值音画一致性规则，对原始候选数据进行串行筛选。分阶段统计表明，282.5 h 原始数据最终保留 52.4 h。

结构化标注：采用场景级角色原型、统一角色 ID、镜头级对白和时间边界约束组织多模态标注，以降低多镜头角色指代漂移。自动标注结果通过人工抽样进行质量核验。

应用验证：平台端更新前后指标和问卷结果显示该类动画内容具有实际使用潜力；由于缺少同期对照组和随机化设计，本文仅报告描述性观察，不将其解释为数据集对教学效果的因果提升。

本文仍存在若干限制：预训练模型在复杂转场、远景、多人交互和角色形变场景下可能产生识别或指代误差；双阈值由当前开发集经验标定，跨动画风格迁移时需重新校准；人工评估缺少可追溯的一致性系数；平台应用缺少同期对照组；商业动画数据的公开发布还受到许可范围约束。

参考文献：

[1] Wang S, Xu T L, Li H, et al. Large lan -

guage models for education: A survey and outlook[J]. arXiv preprint arXiv: 2403.18105, 2024.

[2] Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education[J]. Learning and Individual Differences, 2023, 103: 102274.

[3] Vajjala S, et al. Exploring the role of generative AI in second language education: Insights for instruction, learning, and assessment[J]. Language Learning & Technology, 2024, 28(1): 1–15.

[4] Xue H, Hang T, Zeng Y, et al. Advancing high-resolution video-language representation with large-scale video transcriptions[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 5036–5045.

[5] Wang Y, He Y, Li Y, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation[J]. arXiv preprint arXiv:2307.06942, 2023.

[6] Chen T S, Siarohin A, Menapace W, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 13320–13331.

[7] Chen Z, Liu H, Yu W, et al. M3AV: A multimodal, multigenre, and multipurpose audio-visual academic lecture dataset[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024: 9041–9060.

[8] Rossetto L, Bailer W, Dang-Nguyen D T, et al. The castle 2024 dataset: Advancing the art of multimodal understanding[C]//Proceedings of the 33rd ACM International Conference on Multimedia. 2025: 12629–12635.

[9] Chen J, Wang Z, Zeng A, et al. Talkcuts: A large-scale dataset for multi-shot human speech video generation[J]. arXiv preprint arXiv:2510.07249, 2025.

[10] Soucek T, Lokoc J. Transnet v2: An effective deep network architecture for fast shot transition detection[C]//Proceedings of the 32nd ACM International Conference on Multimedia. 2024: 11218–11221.

[11] Girdhar R, El-Nouby A, Liu Z, et al. Imagebind: One embedding space to bind them all[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 15180–15190.

[12] Lin B, Zhu B, Ye Y, et al. Video-LLaVA: Learning united visual representation by alignment before projection[J]. arXiv preprint arXiv:2311.10122, 2023.

[13] Yang Z, Yu Y, Zhao Y, et al. Timeexpert: An expert-guided video llm for video temporal grounding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025: 24286–24296.

[14] Wang Q, Shi X, Li B, et al. Multishot-master: A controllable multi-shot video generation framework[J]. arXiv preprint arXiv:2512.03041, 2025.

[15] Xue H, Hang T, Zeng Y, et al. Advancing high-resolution video-language representation with large-scale video transcriptions[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 5036–5045.

[16] Radford A, Kim J W, Xu T, et al. Robust speech recognition via large-scale weak supervision[C]//International conference on machine learning. PMLR, 2023: 28492–28518.

[17] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779–788.

[18] 曾志平,张旭龙,瞿晓阳,等. 深度伪造音频生成与鉴伪技术综述[J]. 大数据, 2025, 11(5): 130–151.

Zeng Z, Zhang X, Qu X, et al. Review of deepfake audio generation and detection technologies[J]. Big Data Research, 2025, 11(5): 130–151.

[19] Chen L, Li J, Dong X, et al. Share-GPT4Video: Improving video understanding and generation with better captions[J]. arXiv preprint arXiv: 2406.01463, 2024.

[20] Wang S, Xu T, Li H, et al. Large language models for education: A survey and outlook[J]. IEEE Signal Processing Magazine, 2026, 42(6): 51–63.

[21] Li W, Wang F, Mayer R E. How to guide learners’ processing of multimedia lessons with pedagogical agents[J]. Learning and Instruction, 2023, 84: 101729.

[22] Carlotto T, Jaques P A. The effects of animated pedagogical agents in an English-as-a-foreign-language learning environment[J]. International Journal of Human-Computer Studies, 2016, 95: 15–26.

[23] Li W, Wang F, Mayer R E, et al. Animated pedagogical agents enhance learning outcomes and brain activity during learning[J]. Journal of Computer Assisted Learning, 2022, 38(3): 621–637.

[24] Davis R O. The impact of pedagogical agent gesturing in multimedia learning environments: A meta-analysis[J]. Educational Research Review, 2018, 24: 193–209.

[25] Chung J S, Zisserman A. Out of time: automated lip sync in the wild[C]//Asian conference on computer vision. Cham: Springer International Publishing, 2016: 251–263.

[26] Bai J, Bai S, Yang S, et al. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond[J]. arXiv preprint arXiv: 2308.12966, 2023.

[27] Bai S, Chen K, Liu X, et al. Qwen2.5-VL Technical Report[J]. arXiv preprint arXiv:2502.13923, 2025.

[28] Xu J, Guo Z, He J, et al. Qwen2.5-Omni Technical Report[J]. arXiv preprint arXiv:2503.20215, 2025.

录用日期: XXXX-XX-XX

通信作者:

基金项目:

Foundation Items:

