ADX3：高质量无障碍音频描述的协作工作流

Lana Do
东北大学，圣何塞，加利福尼亚州 95113
do.ng@northeastern.edu

Shasta Ihorn旧金山，加利福尼亚州 94132
sihorn@sfsu.edu

Charity Pitcher-Cooper史密斯-凯特尔韦尔眼科研究所，旧金山，加利福尼亚州 94115 cpc@ski.org

Juvenal Francisco Barajas旧金山州立大学，旧金山，加利福尼亚州 94132
jbarajas8@mail.sfsu.edu

Gio Jung旧金山州立大学 旧金山，加利福尼亚州 94132 gjung1@sfsu.edu

Xuan Duy Anh Nguyen
旧金山州立大学 旧金山，加利福尼亚州 94132
anguyen96@sfsu.edu

Sanjay Mirani旧金山州立大学，旧金山，加利福尼亚州 94132
smirani1@mail.sfsu.edu

Ilmi Yoon
东北大学，圣何塞，加利福尼亚州 95113
i.yoon@northeastern.edu

arXiv:2602.02684v1 [cs.HC] 2 Feb 2026

摘要

音频描述（AD）让盲人及低视力受众能够获取视频内容，但制作高质量描述需要投入大量资源。自动化音频描述具有可扩展性，先前研究表明，在回路编辑和用户查询能有效改善旁白。我们提出了 ADx3，这是一种新颖的框架，整合了三个模块：GenAD，利用基于无障碍需求的提示，引导现代视觉语言模型（VLM）升级基线描述生成；RefineAD，通过包容性界面支持盲人及低视力用户和明眼用户查看并编辑草稿；以及 AdaptAD，支持按需用户查询。我们开展了一项研究来评估 GenAD，由七位无障碍专家依据专业指南审核视觉语言模型生成的描述。研究结果表明，经过针对性提示，视觉语言模型能够生成符合基本标准的“良好”描述，但要达到“优秀”水平，则需要人工编辑（RefineAD）和互动（AdaptAD）。ADx3 展示了用于无障碍内容创作的协作工作流程：各组件相互促进，推动持续改进；编辑内容可指导未来的基线描述生成，而用户查询则能揭示 AI 生成和人工撰写描述中的不足。

关键词 视觉语言　　言模型 · 人在回路 ·交互式查询 · 音频描述 · 盲人和低视力　　n
受众

1 引言

数字视频主导着教育、娱乐和传播，但对于全球估计 2.85 亿盲人和低视力（BLV）人士而言，其无障碍程度仍然很低。音频描述（AD）——对视觉元素进行口述旁白——已被证明能够显著提升盲人和低视力用户的理解和参与度，但其制作仍然耗费大量人力，在专业环境之外也十分罕见 Schmeidler 等人（2000）；Naraine 等人（2018）。目前，YouTube、TikTok 和 Instagram 等短视频用户生成内容平台上仍普遍无法获得音频描述，而这些平台如今已占据数字视频消费的绝大多数，由此造成严重的无障碍鸿沟，并对盲人和低视力用户获取当代数字文化的机会产生不成比例的影响。

YouDescribe YouDescribe（[n. d.]）、Rescribe Pavel 等人（2020）和 LiveDescribe Branje 和 Fels Structured（[n.d.]）等由志愿者驱动的系统，提供简化的创作界面和协作编辑支持，以降低

创建音频描述的门槛，并推动描述创作民主化。这些人工方法虽有价值，却无法扩展到与数字平台每天产生的内容量相匹配的规模。近年来，使用视觉语言模型（VLM）的自动化方法展现出可扩展性，但在捕捉使音频描述发挥作用的语境意识、连贯旁白和精准时机方面仍然不足。

Yuksel 等人提出了一种人在回路（HITL）方法，将自动生成与人工编辑相结合。他们针对盲人和低视力用户及新手描述者开展的研究表明，这种方法提升了描述质量，同时降低了新志愿者参与的门槛 Yuksel 等人（2020）。即使依赖较早的图像描述生成模型，这些结果依然十分显著。机器翻译（MT）领域的既有研究也表明，更高质量的输出能够减少编辑工作量并提高效率 Green 等人（2013）；Läubli 等（[n.d.]），这说明音频描述中更强的人工智能基线同样有望扩大产能。在此基础上，我们通过超越 Pythia 与外部光学字符识别相结合的图像描述生成流程等方法，推进基线描述生成 Stangl 等（2023）。我们测试了三个现代视觉语言模型——Qwen2.5-VL Bai 等人（2025）、Gemini 1.5 Pro Google DeepMind（2024）和 GPT-4o OpenAI（2024）——它们能够在单个模型中生成视觉描述并识别屏幕文字。

虽然人在回路编辑降低了描述撰写者的门槛，但其他研究致力于让盲人和低视力用户在这一过程中拥有更多自主权，并开发了交互式系统，使他们能够请求与自身观看目标最相关的细节。Stangl 等人在针对盲人和低视力参与者的研究中，展示了交互式查询工具与自动旁白相结合的价值 Stangl 等（2023），凸显了根据用户需求定制音频描述的重要性。综合来看，这些研究表明自动化、编辑和交互都颇具潜力，但尚无系统将它们全面整合为一个连贯的工作流程。

我们提出 ADx3，这是一个端到端框架，将 AI、人工编辑和盲人和低视力受众纳入协作工作流程。该系统由三个模块构成：(1) GenAD 利用现代视觉语言模型和提示策略生成基线描述；(2) RefineAD 支持描述撰写者开展无障碍协作编辑；(3) AdaptAD 让盲人和低视力受众能够提问并请求补充细节，从而根据其目标定制音频描述。这个过程具有迭代性，因为编辑和用户互动会反馈到后续的提示和微调中。
为评估 ADx3，我们提出了两个研究问题（RQ）：

> 1. **研究问题1**。 现代视觉语言模型生成基线描述的效果如何？更具体地说，在我们的系统中，人工智能生成的描述（GenAD）有哪些优势和不足？2. **研究问题2**。 整合 GenAD、RefineAD 和 AdaptAD，如何提升旁白质量和用户体验？

我们通过一场评估研讨会探讨这些问题。在研讨会上，七位无障碍专家依据既定的音频描述指南，评估了由 Qwen2.5-VL、Gemini 1.5 Pro 和 GPT-4o 生成的匿名描述。此前的研究通过对盲人和低视力参与者开展广泛研究，证实了人在回路编辑和用户按需查询的价值（Yuksel 等人（2020）；Stangl 等（2023））。作为补充，本研究邀请无障碍专家评估升级后的基线描述生成（GenAD）；这项评估既需要无障碍知识，也需要核验视觉细节的能力。我们还通过形成性演示，展示 RefineAD 和 AdaptAD 如何在此基线之上扩展，构建一个连贯的系统。研究结果表明，尽管视觉语言模型能够生成流畅、细节丰富的草稿，但有针对性的润色仍不可或缺，才能将尚可的草稿提升为高质量描述；此外，以用户为主导的交互对于使描述适应多样化情境至关重要。我们的贡献包括：（1）评估多个视觉语言模型生成基线描述的能力，并指出其优势与局限；（2）提出统一的系统设计，将人在回路编辑和按需查询纳入可扩展的流程，以生成高质量的无障碍音频描述。

2 相关工作

2.1 音频描述的无障碍指南与志愿者驱动平台

长期以来，电影和广播一直要求提供音频描述，这得益于专业框架和行业指南的逐步形成。网页内容无障碍指南2.0 AA级要求所有预录同步媒体提供音频描述 200（2008年）。描述与字幕媒体项目（DCMP）为教育和教学场景提供了广受引用的指导，强调学习环境中的精准性和一致性 DCMP（2024年）。WGBH教育基金会旗下的国家无障碍媒体中心（NCAM）是一家总部位于波士顿的公共媒体机构，率先开展音频描述研究，并制定了广播和电影行业的标准，确立了有关时机、风格和整合方式的规范，至今仍具有重要影响力国家无障碍媒体中心（2017年）。在商业领域，3Play Media 等公司已将音频描述工作流程纳入更广泛的无障碍服务，把描述、图像描述生成和字幕翻译整合进专业制作流程。这些来源共同确立了行业基准，塑造了专业描述员和志愿贡献者的实践方式，并影响着各领域对质量、风格和时机的期望。

随着数字视频和用户生成视频的快速增长，音频描述的需求已远远超出电影院和电视，扩展到 YouTube、TikTok 和 Instagram 等平台；这些平台以大规模生产内容和较短的制作周期为特点 Schmeidler 等人（2000年）；Naraine 等人（2018年）。为此，早期的创作工具和众包平台试图通过降低非专业人士的参与门槛来扩大无障碍服务，将专业无障碍指南加以调整，以支持基于社区的音频描述创作。LiveDescribe 为非专业人士提供了简单的界面，用于录制音频描述并使其与视频时间线对齐 Branje 和 Fels Structured（[n. d.]），而 Rescribe 则将脚本编写、时机对齐和音频编辑整合到一个协作工具中，从而简化了创作流程 Pavel 等人（2020年）。YouDescribe 展示了众包音频描述的可行性，允许志愿者应盲人和低视力用户的请求，为 YouTube 视频添加描述 YouDescribe（[n. d.]）。这些方法扩大了音频描述的可用性，但由于依赖志愿者劳动，难以大规模持续开展。

尽管由志愿者驱动的平台扩大了无障碍音频描述的覆盖范围，但它们依赖人工劳动，因此难以跟上在线内容的增长速度，也难以维持稳定的质量。满足这一需求，需要更先进的自动化技术，在扩大现有指南和社区贡献的影响力的同时，确保符合无障碍标准。

2.2 自动化音频描述与人在回路的优化

多模态学习的进步拓展了自动化音频描述的潜力。Tiresias 等系统通过层次注意力，将基于事件的音频描述与显著性线索及音频对齐（Wang 和 Liang，2021）。后续模型如 BLIP-2（Li 等人，2023）通过适配器结合冻结的视觉与语言模型，使图像描述生成和视觉问答等生成任务成为可能，但对深层时间推理的支持有限。近期的工作流程利用这些模型生成音频描述——例如，LLM-AD 使用 GPT-4V，结合场景分割和角色跟踪（Chu 等人，2024）；MMAD 则将 CLIP 与 LLaMA2 相结合，整合字幕和声音线索，以保持叙事连贯性（Ye 等人，2024）。然而，单靠自动化音频描述往往难以达到理想效果：系统会认错人物或加入幻觉细节（Bergin 和 Oppegaard，2025），生成冗长或重复的描述，不但无助于理解，反而造成困扰（Wang 和 Liang，2021）；即使在图像描述生成指标上得分很高，其结果也可能与人类对音频描述质量的判断不一致（Rohrbach 等人，2017）。

为提升清晰度和易用性，先前研究探索了人在回路系统，让人类编辑人工智能生成的草稿。Yuksel 等人发现，这种方法生成的描述在质量和理解方面获得了盲人和低视力用户更高的评价，同时也减轻了描述者的工作负担 Yuksel 等人（2020）。然而，该研究仅涉及新手描述者，聚焦于无对白视频，并依赖图像描述生成模型（如 Microsoft Video Insight、Pythia），这些模型的输出在准确性方面被评为“中等”。尽管如此，参与者表示，与从头撰写相比，这种方法减少了他们的工作量，说明即使基线描述并不完美，也能有效减轻工作负担。
机器翻译领域的研究进一步支持了这一观点：专业译者编辑高质量神经机器翻译输出时，修改次数更少，工作速度也显著快于修改质量较低的输出 Green 等人（2013）；Läubli 等人（[n.d.]）。这些发现表明，音频描述中更可靠的人工智能生成基线同样可以减少认知和编辑工作量，从而提升无障碍描述创作的能力。我们将基线描述生成从图像描述生成模型升级为三种最先进的视觉语言模型，包括 Qwen2.5-VL Bai 等人（2025）、Gemini 1.5 Pro Google DeepMind（2024）和 GPT-4o OpenAI（2024），并在相同的流程条件下采用无障碍引导提示。我们根据一位经验丰富的顾问提供的反馈反复改进提示策略。该顾问拥有与盲人和低视力用户合作以及培训新手描述者的经验，使我们的输出在清晰度和专业实践契合度方面持续提升。随后，我们在一项研究中邀请无障碍专家评估生成的描述，涵盖多种视频类型，包括对白密集和文字密集的内容。本研究拓展了先前以新手描述者为对象的、以盲人和低视力人士为中心的研究，分析专家希望改进人工智能生成描述的哪些方面，以及他们如何利用系统实现这些改进。

2.3 交互性与人在回路系统

虽然人在回路优化能够提升清晰度和准确性，但盲人和低视力受众的需求也多种多样，且因情境而异，仅靠固定旁白无法满足。研究表明，用户偏好会随观看目标和内容体裁而变化（Jiang Crescentia Jung Mahika Phutane 等人，[n. d.]）。例如，在社交或家庭视频中，有些用户更喜欢丰富的角色描述和情感背景，以便更好地理解人际互动。相较之下，观看教育视频时，用户通常更喜欢简洁、重点突出的旁白，突出关键教学内容，避免不必要的细节。真正无障碍的系统必须允许盲人和低视力用户调整信息类型和详细程度，以契合各自的目标和观看情境。

越来越多的研究强调交互式系统，使盲人和低视力用户能够更大程度地自主决定如何获取视觉信息。例如，DescribeNow Cheema 等（2024）允许观众调整旁白风格，在详细描述与简明摘要之间切换；而 ShortScribe Van Daele 等（2024）则支持在不同详细程度之间切换，从逐镜头旁白到长篇摘要和高层次概览。

SPICA Ning等人（2024）通过支持利用帧字幕进行时间导航，以及在关键帧内对对象进行空间探索，进一步扩展了交互性。这些系统共同展示了交互性如何超越静态、千篇一律的旁白，让用户能够有意义地控制描述性信息的层次、类型和时机。相关研究还表明，将基础旁白与交互式代理结合起来也具有益处。Bodi等人 Bodi 等（2021）和 Stangl等人 Stangl 等（2023）发现，当盲人和低视力用户能够请求补充细节或提出具体问题时，他们在理解、享受度和满意度方面的反馈更高。

现有研究表明，仅靠自动化并不足够，基线草稿需要润色，旁白也必须适应个人目标。这些要素仍分散在各个独立系统中。本研究将这些要素整合到一个协作工作流程中：GenAD 利用现代视觉语言模型和有针对性的提示策略生成基线描述；RefineAD 通过人工编辑提升清晰度和一致性；最重要的是，AdaptAD 让盲人和低视力用户能够直接自主决定何时以及如何接收描述。为明确 AdaptAD 最能发挥价值之处，我们的研究邀请无障碍专家评估生成的音频描述基线（GenAD），以此补充此前以盲人和低视力用户为中心的评估（Bodi 等（2021）；St angl 等（2023））， 从而揭示遗漏和歧义的模式。除了满足用户的即时需求，AdaptAD 还建立了一个反馈渠道：通过分析用户提出的问题，以及这些请求是否集中在角色识别、空间关系或频繁出现的屏幕文本事件上，系统便能揭示 AI 描述和人工编辑中存在的系统性缺漏。

3 系统

图 1 展示了完整的 ADx3 系统，该系统包含三个模块：用于生成初始描述的 GenAD、用于与用户交互的 AdaptAD（两者均由视觉语言模型提供支持），以及提供人工编辑层的 RefineAD。
由于 GenAD 和 AdaptAD 依赖视觉语言模型，因此选择恰当的模型至关重要。Video-MME 是一个大规模基准测试，用于评估 50 个视觉语言模型在多模态推理、时间定位以及各类视频任务中的一致性方面的表现 Fu et al. (2024)。在其最新排行榜上，Gemini 总体排名第一，Qwen-VL 排名第五，GPT 排名第六。我们选择这些模型，不仅因为它们的表现位居前列，也因为它们便于用于研究：Qwen2.5-VL 在 Hugging Face 上开源，而 Gemini 1.5 Pro 和 GPT-4o 则可通过商业API使用。为进行受控比较，我们分别使用这三个模型运行整个流程，同时保持其他所有组件不变。

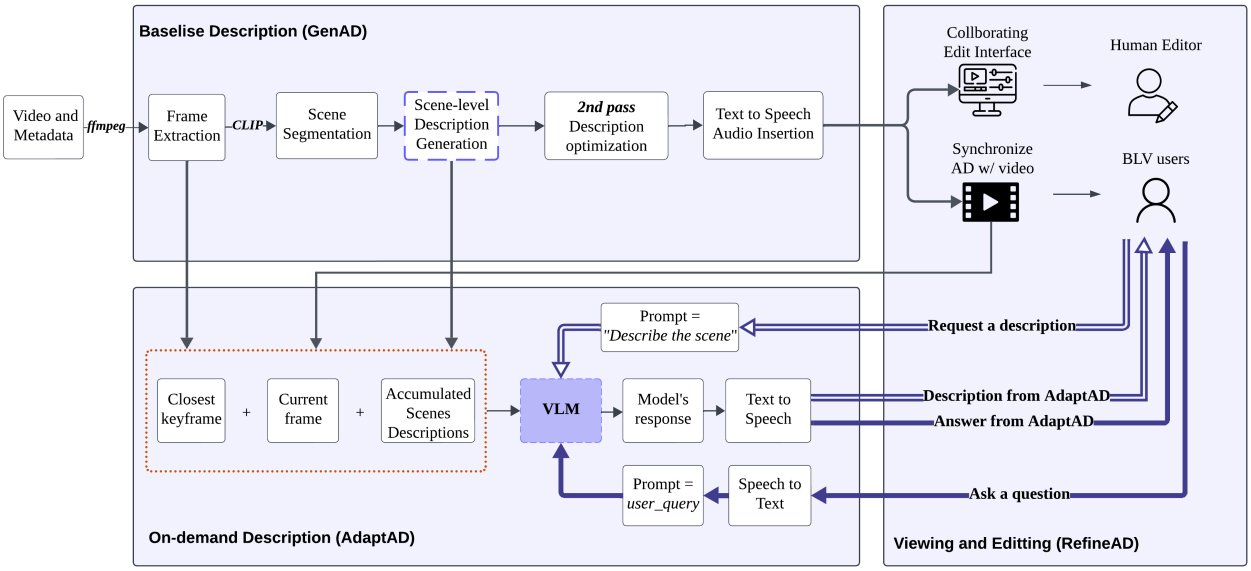


图 1：ADx3 系统架构，展示了三个集成模块：用于基线描述生成的 GenAD、用于人工编辑的 RefineAD，以及用于用户与同步音频描述交互的 AdaptAD

3.1 基线描述生成：GenAD

基线描述旨在向盲人和低视力受众传达视频中的关键视觉细节，包括显著物体、人物及其动作和互动、可见文字，以及场景和服饰等环境线索。

我们的系统支持两种呈现方式：内嵌式旁白，即将描述插入对话中的自然停顿；以及扩展式 旁白，即短暂暂停播放以补充更多细节。传统音频描述实践主要侧重于内嵌式传递，这既反映了广播的限制，也体现了避免打断内容的偏好。然而，在对白密集型视频或需要精确描述的内容（例如教育材料 Fun（[n. d.]））中，仅靠内嵌描述往往并不足够。在这些情况下，扩展旁白可以提供原本可能被省略的关键信息。该系统在技术上支持这两种模式，使其能够灵活调整呈现策略，在不同视频类型中兼顾内容忠实度、时机和无障碍需求。

3.1.1 视频输入与场景分割

工作流程首先使用 yt-dlp 获取目标视频。yt-dlp 是一款命令行工具，可从 YouTube 下载视频及相关元数据。元数据包括视频标题、描述、类别，以及有字幕时的字幕。首先使用 FFmpeg（一种命令行多媒体框架）提取视频帧（TomarSuramya，2006）。为将视频分割成连贯的场景，我们使用 OpenCLIP 的视觉编码器提取帧级嵌入向量（Schuhmann 等，2022）。我们计算相邻帧之间的余弦相似度，当相似度低于预设阈值时，便标记一个场景边界。此方法使系统能够将视觉上保持一致的片段归为一组，并有助于减少生成描述中的冗余。

3.1.2 场景级描述生成

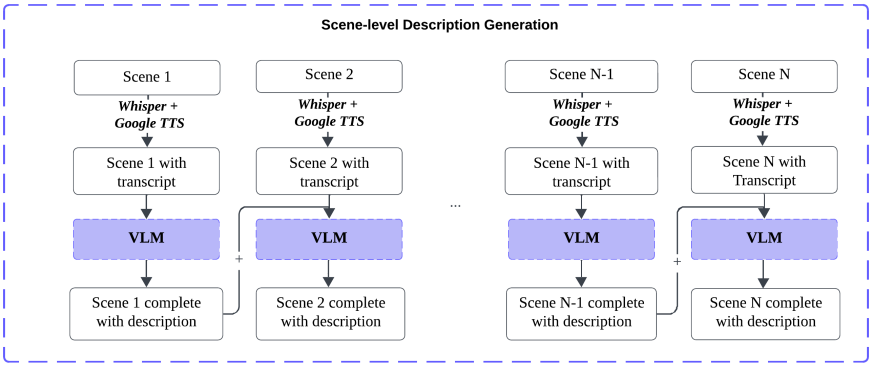


图2：利用音频转录文本和累积的先前场景描述进行上下文提示，以生成场景级描述。

对于每个分割后的场景，早期实验仅用场景画面提示视觉语言模型，生成的内容看起来更像字幕，而非连贯的音频描述。这些内容冗长、重复对话，且常常未能清楚描述人物或交代背景；先前的研究和指南也记录了这些问题（Sarhan 和 Hegelich，2023；Stangl 等（2023）；DCMP（2024））。我们将这些草稿交给一位经验丰富、同时负责培训志愿者的音频描述员审阅时，对方指出，视觉语言模型缺乏上下文参照和连贯流畅的表达，其风格类似于初级描述撰写者，他们会“试图做得太多”，把每个可见细节都叙述出来，包括那些可以从音轨中推断的信息。

为解决这些问题，我们引入了两项关键策略。首先，加入角色提示，将模型设定为“专业音频描述员。”随后，我们通过引导式自检环节进一步强化这一设定：向模型提供一组简短的音频描述指南，这些指南通常推荐给人工描述员，并要求模型在生成内容前确认理解。指南强调简洁、事实准确性和语气匹配，同时提醒避免过度描述、主观解读和过早剧透等问题（完整提示词见附录）YouDescribe（[n.d.]）。其次，我们通过在提示词中补充视频的元数据（标题和摘要）、场景的音频转录文本以及此前场景积累的描述，改善了上下文依据。这有助于保持不同片段间的连贯性，避免使用含糊或泛泛的措辞。

准确的音频转录对于上下文生成和交付优化都至关重要：没有可靠的文字转录，模型就无法生成精准的描述；没有准确的时间戳，系统就无法优化内嵌式传递。为确保这两点，我们采用了集成方法：Whisper 通过动态时间规整生成词级时间戳，Google 语音转文字则用于交叉核验，以减少幻觉，只保留可靠内容。图2展示了这一场景级生成流水线的概览。

最后，提示词指定了输出格式，要求每个场景包含两种带标签的事件类型——屏幕文字 和视觉，并注明每个事件的明确起始时间，从而与视频时间线对齐，并提升后续优化的可用性。

3.1.3 描述优化

即使生成效果有所改善，输出内容仍往往过于冗长，难以流畅呈现。在顾问持续提供的反馈指导下，我们改进了提示词和旁白策略，以更好地符合盲人和低视力用户的偏好：采用简洁叙述，尽量减少打断（DCMP，2024年；国家无障碍媒体中心，2017年；3Play Media，2020年）。为此，我们进行了第二轮基于提示的优化。内嵌式优化会压缩旁白，使其适应可用的对话间隙；同时使用重试提示进一步缩短描述，并优先保留关键信息。只有在无法内嵌式传递、但省略重要信息会造成影响时，才采用扩展旁白。在此过程中，筛选提示将模型设定为“无障碍专家”，并指示模型仅在描述传达无声动作、新出现的视觉元素或角色辨识信息时，才将其标记为“必要”。对于教程类视频，屏幕文字和视觉动作的旁白若支离破碎，可能会造成干扰，因此合并提示会将两者整合为一个连贯的句子，同时保留精确的测量值。这一策略体现了程序性内容音频描述的最佳实践，并通过顾问的反复反馈得到了强化。通过准确转录、依据指南编写提示词、结合上下文以及迭代润色，GenAD 将初始视觉语言模型输出转化为简洁连贯的旁白，体现专业音频描述实践，并满足盲人和低视力受众对清晰度和质量的期望。

3.1.4 提示词设计带来的改进示例

图3展示了我们在迭代设计过程中观察到的若干定性改进。通过场景级生成，基于指南的提示和上下文提示生成了更具体的旁白：原始视觉语言模型输出将画面描述为“一名女子站在沙漠中”，而优化后的版本则认出了她是“蕾伊”（图 3a ）。通过描述优化，提示词减少了冗长内容，并提升了连贯性。图 3b 展示了如何将一段过长、原本需要扩展式音轨的草稿，压缩为适合自然停顿的内嵌描述，在保持流畅的同时不遗漏细节。图3c展示了合并提示如何将屏幕文字与同时发生的动作整合为一句流畅的话，仿照人类描述者自然传达场景的方式。这些初步观察表明，我们的策略缓解了不使用无障碍引导提示时运用视觉语言模型的常见不足，并为下一节更全面、细致的专家评估奠定了基础。



✗ **Naive output (without guidelines → overly detailed):** A young woman with brown hair tied back stands in a vast desert landscape. She is wearing a beige, sleveless output wiht a brown belt and arm wraps. The woman looks determined as she gazes into the distance. Her expression is focused. The woman reaches for a cylindrical object attached to her belt.
➜ **With guidelines (concise but generic)**: A woman stands in a vast desert, dressed in a beige outfit. She is scanning the horizon, cylindrical object attached to her belt.
✓ **With guidelines + context (concise and specific): Rey** stands in a vast desert, scanning the horizon, lightsaber hilt at her waist.

（a）《星球大战》预告片——从简单的 VLM 提示逐步发展为由指南与上下文引导的提示。



First pass: A lush, expansive forest stretches across rolling hills under a bright blue sky dotted with fluffy white clouds.(**Extended**)

Optimized: Lush forest covers rolling hills, sky dotted with clouds. (**Inline**)

First pass: A lush green tree canopy fills the frame, with leaves gently swaying in the breeze. The sky above is clear and bright. A chimpanzee appears on the right side of the frame, partially obscured by the foliage. It moves its arms, possibly reaching for something or interacting with its environment (**Extended**)

Optimized : Green canopy sways. Chimpanzee moves, arms reaching through leaves.(**Inline**)

(b) 简·古道尔片段——优化将冗长的草稿精简为适合自然停顿的描述。



First Pass:
{"type": "Text on Screen", "text": "SPICY:"},
{"type": "Text on Screen", "text": "CARAWAY SEED 1 tsp"},
{"type": "Text on Screen", "text": "DRY MUSTARD 1/2 tsp"},
{"type": "Visual", "text": "Stirs the mixture in the pot vigorously."}
Optimized: SPICY. Stirs pot. Adds 1 tsp caraway, 1/2 tsp dry mustard.

(c) 《制作泡菜的三种方法》——优化将文本与视觉线索合并为一条简洁的描述。

图 3：不同视频中提示词与优化改进的示例：(a) 根据指南和上下文改进提示词，(b) 将冗长的草稿精简至适应自然停顿，(c) 将文本与视觉线索合并为流畅的描述。

3.1.5 文本转语音音频插入

优化后的描述随后会转换为语音，并与视频时间线同步。内嵌式旁白会插入自然停顿期间，而扩展旁白则会在播放时短暂停顿，以传达重要细节，同时避免与现有音频重叠。这种插入策略确保描述起到增强作用，而不会干扰现有音频。最终文本使用 Google文本转语音（TTS）合成为语音。为便于区分不同类型的内容，视觉描述采用女声朗读，屏幕文字条目则采用男声朗读——这一做法参考了顾问的建议；顾问指出，反复听到“屏幕文字”这句话可能会让盲人和低视力用户感到厌烦。

随后，这些音频片段会与原视频同步，并呈现在交互界面中，供用户播放和润色。

3.2 人在回路（HITL）工作流程

用于查看生成的描述的界面如图4所示。为确保无障碍，平台采用了 Web 无障碍倡议——无障碍富互联网应用（WAI-ARIA）这一技术规范，该规范通过提供语义角色、状态和属性，使辅助技术（如屏幕阅读器，例如 JAWS、NVDA、VoiceOver）能够感知交互式 Web 内容（WAI，[n.d.]）。结合高对比度模式和全面的键盘导航，这些基础功能使平台能够满足不同设备和用户的需求，保持包容性。

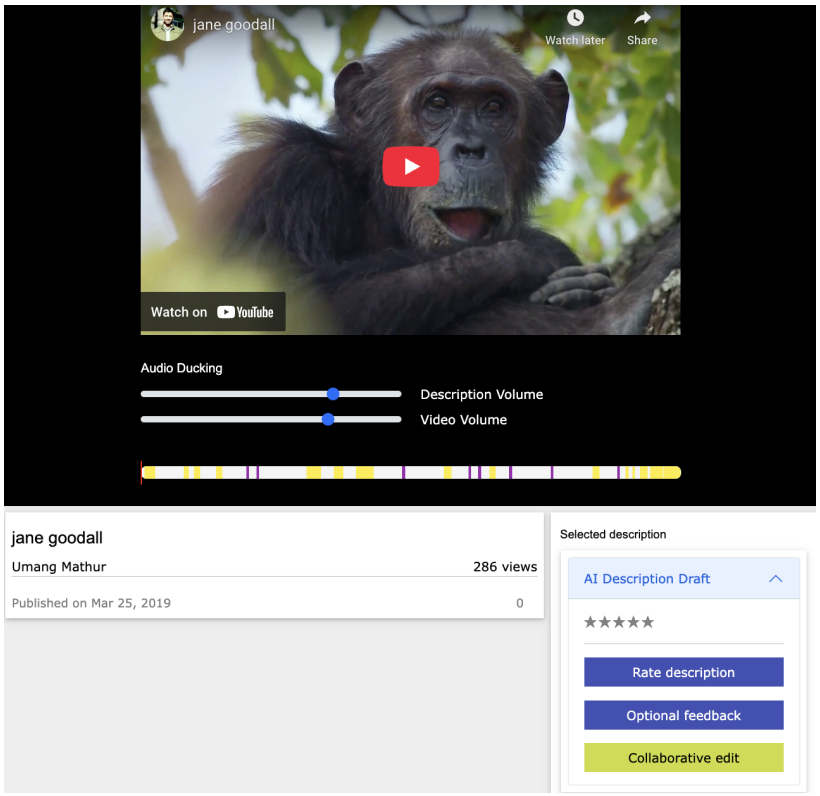


图4：用户可预览人工智能生成的草稿，其中时间轴栏包含内嵌式和扩展式音频描述轨道。按钮可用于评分、提供可选反馈，并且（默认情况下）启用协作编辑以打开编辑器。

在观看界面中，内嵌式旁白以黄色片段的形式嵌入时间轴，而扩展旁白则以紫色标记表示；播放会短暂停顿，以便呈现较长的细节。这种颜色编码让视力正常的用户一眼就能区分旁白类型。音频闪避控制可进一步平衡旁白与原始音频，全面的键盘导航也为播放提供支持。用户可以为人工智能生成的描述评分并发表评论，从而建立反馈循环，识别有用的草稿，并突出显示需要润色的内容。

如需进行更实质、更细致的编辑，用户可点击黄色的“协作编辑”按钮，打开编辑界面（图 5）。描述草稿会预先按类型（视觉或屏幕文字）和呈现方式（内嵌式或扩展式）标注。视频播放且 AI 草稿同步显示时，当前片段会以绿色边界框标出。编辑者也可以利用备注区规划修订。在该界面中，用户可以修改文本、切换呈现方式、对齐片段、添加新轨道，以及删除不需要的 AI 生成轨道。用户可以直接在时间戳框中调整时机，也可以使用**微调工具**进行精细调整，按帧级精度向左或向右移动提示。编辑界面采用了与观看界面相同的无障碍功能，帮助盲人和低视力用户及视力正常的志愿者直接纠正错误并改进音频描述。

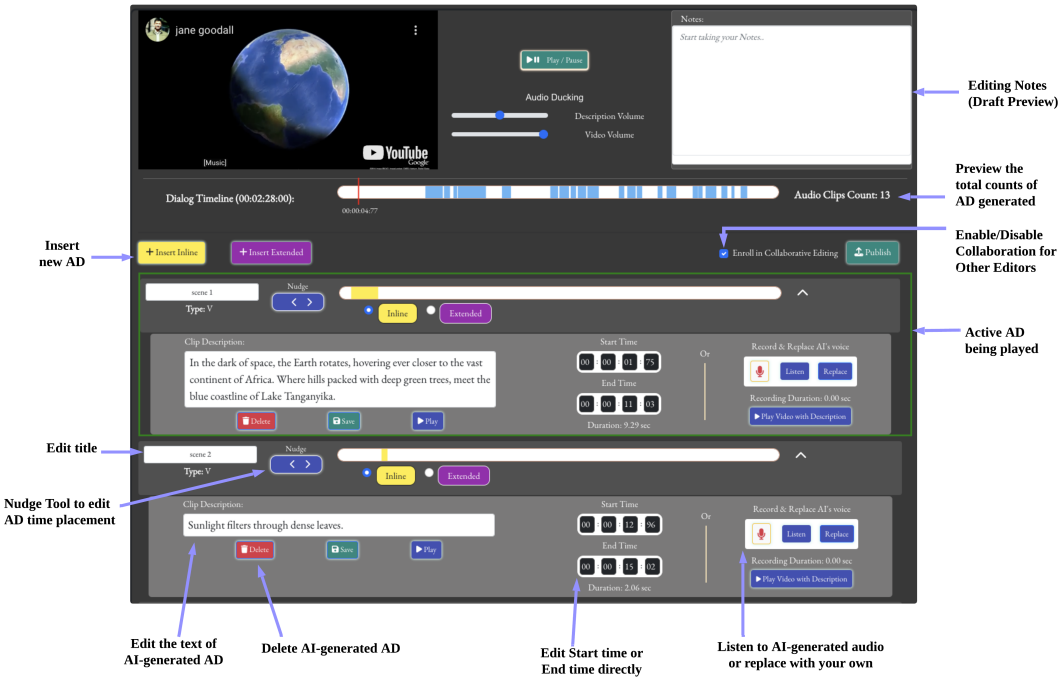


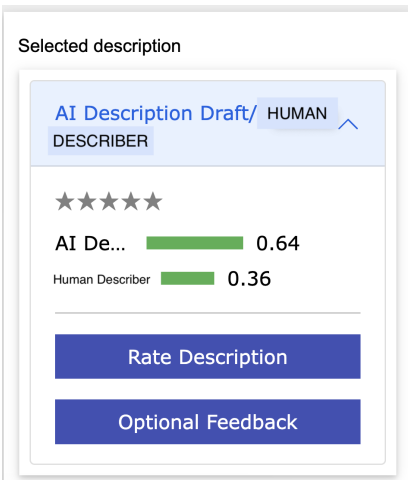
图5：编辑者借助文本修订、切换呈现方式、音轨对齐、使用微调工具编辑时间戳以及添加/移除轨道等功能完善音频描述草稿。无障碍支持（屏幕阅读器、高对比度、键盘导航）使盲人和低视力用户能够直接参与。

为使呈现方式更具吸引力、更亲切，编辑者也可以用编辑者录制的旁白替换合成语音音轨。研究表明，盲人及低视力受众始终更青睐人工旁白，认为其清晰度更高、情感表达更丰富；尤其是在戏剧性或快节奏内容中，合成语音可能显得平淡或疏离 Fernández-Torné（2015年）；Walczak 和 Fryer（2018年）；RNI（[n.d.]）。这一功能体现了以人为本的设计理念，在合成语音的可扩展性与录制旁白的真实感和表现力之间取得平衡。

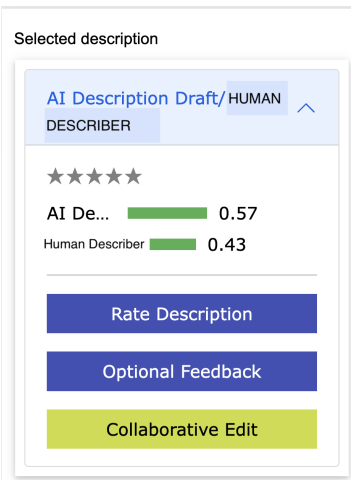
发布前，作者可以选择启用启用协作编辑。默认情况下，仅初始的人工智能生成的草稿会启用此设置，以鼓励人工审核。完成任何编辑后，协作编辑便成为可选项：描述撰写者可以保持启用状态，以便其他编辑者参与；也可以将其关闭，这样只有原作者可以编辑草稿。这种模式允许编辑者邀请他人协作，与此前的研究相符。例如，ViScene 的研究表明，新手描述者与专家（无论是视力正常的专业人士，还是具有亲身经验的盲人和低视力用户）合作，可以提高旁白质量，同时减少所有贡献者的工作量（Natalie 等人，2020年）。所有进度都会自动保存，描述撰写者可以随时返回继续编辑，或在需要时取消发布。

发布最终版描述音轨，并展示 AI 与人工之间以及不同作者之间的贡献比例。所有修订记录都会自动记录（图 6），从而确保责任可追溯，并支持迭代改进。贡献采用编辑距离法进行量化：将每次修订与前一版本之间的莱文斯坦距离按内容长度归一化，以此按比例计入贡献。这种方法支持多作者贡献，并能体现人工参与的程度和顺序。

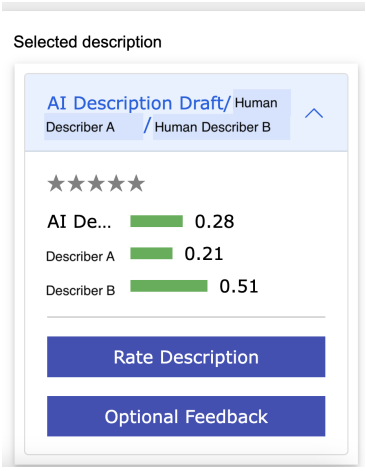
通过提供可编辑的草稿，而不是要求从头开始创建描述，该系统借助结构化的人类监督提升自动化音频描述的质量，降低参与门槛，并扩大可作出贡献的人群。人工智能草稿可作为起点，由人工编辑逐步润色，在保持质量与问责性的同时，支持音频描述的规模化。随着草稿质量提高，所需的编辑次数减少，从而降低编辑人员的工作量。该界面注重易用性，具备快速时间调整等功能，可简化传统上较为复杂的任务。通过将视觉语言模型与无障碍引导提示及功能丰富的界面相结合，该系统推动了无障碍工作的发展：不仅为盲人和低视力用户提供更多音频描述，也通过编辑润色，生成更高质量的旁白。



（a）禁用协作编辑时的人工智能与人类贡献分解。不会显示邀请他人的选项。



(b) AI–人类贡献分解-下，在协作编辑启-用时，允许进一步贡献。



(c) 协作编辑视图，含来自多位人类编-一份草稿上的编辑者。

图6：创作界面功能：(a) 显示禁用协作编辑时人工智能与人类贡献的分解；(b) 显示启用协作编辑并允许更多贡献者参与时的分解；(c) 显示多位人工编辑共同编辑一份草稿的协作编辑视图。

3.3 按需描述：AdaptAD

AdaptAD 允许用户在视频播放的任何时刻请求更多信息。启用后，视频会自动暂停，让用户能够查看视觉内容，而不会错过后续动作。为了生成准确的回答，AdaptAD 会结合当前暂停帧与最近的关键帧进行分析，尤其是在暂停帧可能模糊或视觉信息不完整时。视觉语言模型会接收这些帧、场景转录文本以及 GenAD 先前生成的描述输出，并据此生成既反映当前视觉状态、又保持叙事语境的回答。

在咨询无障碍专家并参考此前针对盲人和低视力用户开展的研究（Stangl 等（2023））后，我们将盲人和低视力用户偏好简洁的一句话回答这一特点纳入设计，而非采用视觉语言模型常生成的冗长输出。AdaptAD 中的提示词经过专门设计，以体现这一偏好，引导模型生成语境丰富而简洁的回答，始终围绕用户的查询展开，并突出最相关的细节。

3.3.1 即时描述请求

按下 *Option* + *D*，用户即可请求即时描述视频当前显示的内容。生成的描述会通过 Google TTS 转换为语音并朗读。此功能让盲人和低视力用户能够立即获取原始音频描述中可能未包含的视觉细节。

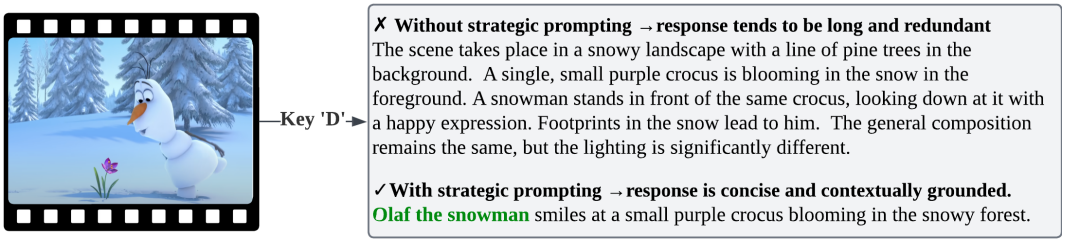
3.3.2 问答

按下 *Option* + *Q*，用户即可针对暂停的视频画面提出口头问题。系统会使用 Google STT（语音转文本）转录问题，然后将其与前文所述的相关视觉和上下文输入一并传递给视觉语言模型。随后，模型的回答会通过 TTS 转换为语音并播放给用户。此功能让用户能够主动探索和澄清场景的具体方面，从而提高参与度和自主性。

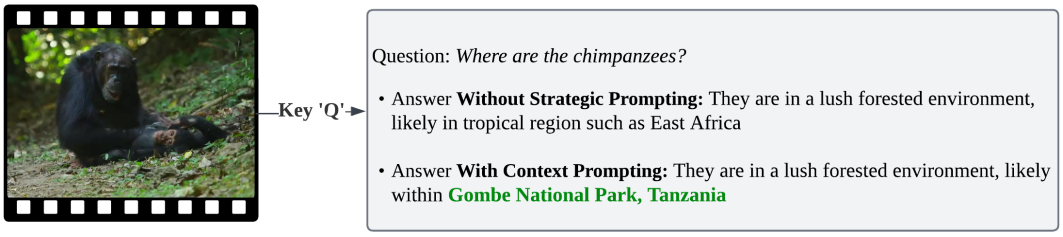
3.3.3 初步观察

AdaptAD 与 GenAD 结合使用时，采用了类似的提示策略——鼓励视觉语言模型生成简洁的回答，这也是我们的顾问和既有研究所建议的做法——并融入了额外的上下文信息，例如场景转录文本以及截至暂停帧为止累积的描述。生成的内容会从 GenAD 中复用，从而降低响应延迟并提高相关性。初步测试表明，与 GenAD 一样，策略性提示能够生成更具上下文关联性的回答，并符合盲人和低视力用户对简洁旁白的偏好。

图 7 展示了两个案例。在《冰雪奇缘》预告片（图 7a）中，我们暂停在某一画面，然后按下 *D*以请求更多描述。在没有采用策略性提示的情况下，系统返回的回答冗长且泛泛而谈。然而，采用提示后，输出变得简短且结合了上下文，并正确说出了雪人的名字：“雪宝。”在简·古道尔—贡贝国家公园视频（图 7b）中，按下 *Q* 询问“黑猩猩在哪里？”后，最初得到的只是一个更为笼统的区域性回答。不过，在补充了此前场景描述中的上下文后，AdaptAD 给出了准确答案：“坦桑尼亚贡贝国家公园。”



（a）《冰雪奇缘》预告片——未采用策略性提示时，描述冗长且笼统；采用提示后，描述变得简洁，并且更贴合具体情境。



（b）贡贝国家公园视频——未采用情境提示时，模型只能模糊猜测所在地区；采用情境提示后，模型能准确指出地点是坦桑尼亚的贡贝国家公园。

图 7：AdaptAD 在采用和未采用无障碍引导提示时的响应示例。

先前研究已表明，交互式按需描述对盲人和低视力用户很有价值，让他们能够以仅靠固定音频描述无法实现的方式探索内容。基于这些证据，我们尝试使用视觉语言模型和策略性提示生成更贴合情境的回答，同时将 AdaptAD 与升级后的 GenAD 模块集成。尽管这些初步观察令人鼓舞，但仍需对盲人和低视力参与者开展进一步研究，以确认这些回答在多大程度上满足用户需求，并找出改进机会。

4 GenAD 的形成性专家评估

为探究视觉语言模型在音频描述方面的现有优势与不足，我们邀请七位无障碍顾问参加了一场评估研讨会。这些专家拥有丰富经验，曾在专业和倡导场景中与盲人和低视力用户合作，包括教授新手描述者、设计无障碍媒体资源、培训盲人和低视力用户使用辅助技术，以及支持社区无障碍倡议。由于事实准确性是评估维度之一，因此需要由视力正常的专家核查描述是否忠实呈现了屏幕内容。尽管盲人和低视力用户未直接参与这一阶段，无障碍顾问仍以盲人和低视力人士为中心的原则和标准为评估提供依据，同时核查准确性等需要视力正常者审阅的质量维度。作为一项设计研究，此次研讨会为了解 GenAD 的能力与不足提供了形成性见解。

4.1 视频选择



（a）《星球大战：天行者崛起》— 预告片



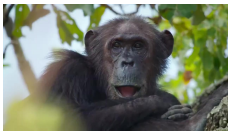
（b）《伯德小姐》| 官方预告片 高清版 |A24



(c) 《冰雪奇缘》预告片（2013）- 迪士尼动画电影画电影



（d）《精灵》片段 - Buddy 意识到自己是人类（2003



(e) Jane 古道尔 - 贡贝国家公园



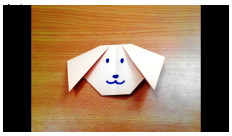
(f) 非牛顿流-体：Crash Course Kids



(g) 白头海雕 | Ani-儿童动物 | 动物 TV



(h) 自制泡菜的三种方法 自制泡菜



(i) 如何制作折纸狗脸



(j) 快速，Easy 5分钟化妆教程

图8：评估中使用的十个 YouTube 视频，涵盖娱乐、教育以及操作与风格类别。副标题显示了 YouTube 视频标题。

我们从娱乐、教育和教程类与风格三个类别中选取了十个视频，以涵盖不同类型惯例、受众和描述挑战。

娱乐类包括四段广为人知的片段：《星球大战》、《伯德小姐》、《精灵》和《冰雪奇缘》。这些片段涵盖了动作驱动的场景、青少年剧情片、家庭喜剧和儿童动画。它们不断变化的语气、叙事密度，以及节奏、视觉效果和氛围等电影技巧，使其对细致描述角色互动和情绪提出了特别高的要求。

教育类包括 Crash Course 制作的非牛顿流体讲解视频、一段关于白头海雕的儿童野生动物视频，以及一段简·古道尔在贡贝国家公园的纪录片节选。每部作品都体现了不同的教学风格：Crash Course 将复杂的科学知识压缩为快速的旁白和丰富多彩的视觉内容；白头海雕视频通过引人入胜的画面，为儿童简化相关概念；而简·古道尔的节选则体现了更为传统的纪录片模式，采用以保护为重点的旁白。在这一背景下，描述的难点在于兼顾事实准确性与无障碍性，同时避免干扰口头对话；视频中的大部分对话由讲师和旁白完成。

操作与风格包含三段日常教程类短视频：腌黄瓜食谱、化妆教程和折纸狗演示。与叙事驱动的内容不同，这些视频需要精准描述细微、连续的手部动作、工具和材料。短视频风格节奏较快，因此尤其具有挑战性，必须格外注重时机和清晰度，才能让盲人和低视力观众不受困扰或认知负担过重地跟上逐步进行的操作。

这些视频平均时长约为三分钟（范围：1:30–5:04），并且曾由 YouDescribe 等社区驱动的音频描述平台提出需求（YouDescribe [n.d.]）。这些视频的入选反映了盲人和低视力受众的真实需求，而非随意挑选。总体而言，这组视频为评估视觉语言模型如何适应不同体裁、受众和视觉模态提供了有意义的测试场景，涵盖动作密集的预告片、喜剧表演、任务导向型教程、动画科学解说视频，以及盲人和低视力受众积极寻找并要求进行描述的观察式纪录片素材。

4.2 工作坊设计

每个视频都使用相同的流程（第 3.1 节）生成了三个音频描述版本，唯一变量是所用模型：Qwen2.5-VL、Gemini 1.5 Pro 和 GPT-4o。为尽量减少偏见，我们在三个层面进行了匿名化和随机化：（1）将版本重新标记为 A、B 和 C，并为每个视频随机分配标签；（2）针对每位顾问，将十个视频的播放顺序随机化；（3）随机调整每个视频中各版本的呈现顺序。这些措施减少了顺序效应，避免标签被识别，并鼓励参与者进行独立评估。

顾问依次观看了每段视频的三个版本，并在观看每个版本后立即填写评估表。所有十段视频均重复这一流程，因此每位顾问完成了 30 份评估（共 210 份）。虽然表格是在

每次观看后提交的，但顾问在本次环节中可以回看并修改先前的回答——这既能确保首次印象独立形成，也允许之后重新考虑。

4.2.1 评估量规

评估量规以描述与字幕媒体项目（DCMP）指南为依据，该指南是教育类和娱乐音频描述领域被引用最广泛的标准 DCMP（2024年）。从 DCMP 中提炼出五个主要的内容维度：准确、优先排序、恰当、一致性和平等性。然而，DCMP 对呈现方式格式的指导较为有限。随着众包平台现已同时支持内嵌式和扩展式旁白，且放置位置正成为音频描述流程的重要组成部分，我们纳入了两个以呈现方式为重点的维度，其依据来自 NCAM 广播指南国家无障碍媒体中心（2017年）以及 3Play Media 的行业实践 3Play Media（2020年）。

由此形成的量规 w　被组织为七个维度，共同评估音频描述的内容质量与呈现方式质量
（详见附录中的　　向顾问提供的完整文档）。

1. 准确：描述在事实上正确且无错误。2. 优先排序：突出显著的视觉细节；省略琐碎元素。3. 恰当：语言和细节程度与类型及目标受众相匹配（例如，教育注重清晰；娱乐注重参与度）。4. 一致性：术语、风格和语气始终保持连贯。5. 平等性：描述使用客观、无偏见的语言；避免个人解读。6. 呈现方式的策略性使用（内嵌式/扩展式）选择恰当：当自然停顿足够时采用内嵌式；当关键细节无法以其他方式容纳时采用扩展式。7. 时机与位置：描述与视觉事件对齐，并避免掩盖关键音频。

顾问按1–5分制对每个维度进行评分（1 = 重大违规； 3 = 基本合格但存在轻微问题； 5 = 堪称典范）。附录中提供了包含行为锚定示例的完整评分标准。这确保了评判标准化，同时兼顾文本准确性以及与视听材料的时间整合。

4.3 定量结果

标准	Qwen		Gemini		GPT	
	Mean	SD	Mean	SD	Mean	SD
总体	3.78	1.00	4.01	1.02	**4.05**	0.97
准确	3.63	1.13	**3.94**	1.08	**3.94**	1.13
优先排序	3.43	1.06	3.61	1.11	**3.76**	1.10
恰当	3.90	1.06	**4.16**	1.08	**4.16**	1.06
一致性	4.03	0.97	**4.31**	1.02	4.27	0.96
平等性	4.41	0.95	**4.56**	0.88	4.41	1.06
呈现方式的策略性运用	3.37	1.01	3.67	1.15	**3.84**	0.98
音轨位置	3.69	0.94	3.79	1.01	**3.97**	0.87

表1：总体与 p　　er-criteria 人工智能评分，并分别列出均值和标准差。粗体表示表现最佳的模型 pe　r row.

4.3.1 总体均值评分

如表 1 所示，各模型的均值分数相对接近，GPT-4o 为 4.05，Gemini 1.5 Pro 为 4.01，Qwen2.5-VL 为 3.78。尽管绝对差异不大，但图 9 所示的分布揭示了各模型不同的方差模式。Qwen2.5-VL 的分数覆盖了完整的 1-5 范围，而 GPT-4o 和 Gemini 1.5 Pro 的分布更为集中，大多数分数落在 3-5 范围内。这些分布模式表明，通义千问产生的输出质量波动更大，而 GPT 和 Gemini 在其表现分数上表现出更高的一致性。

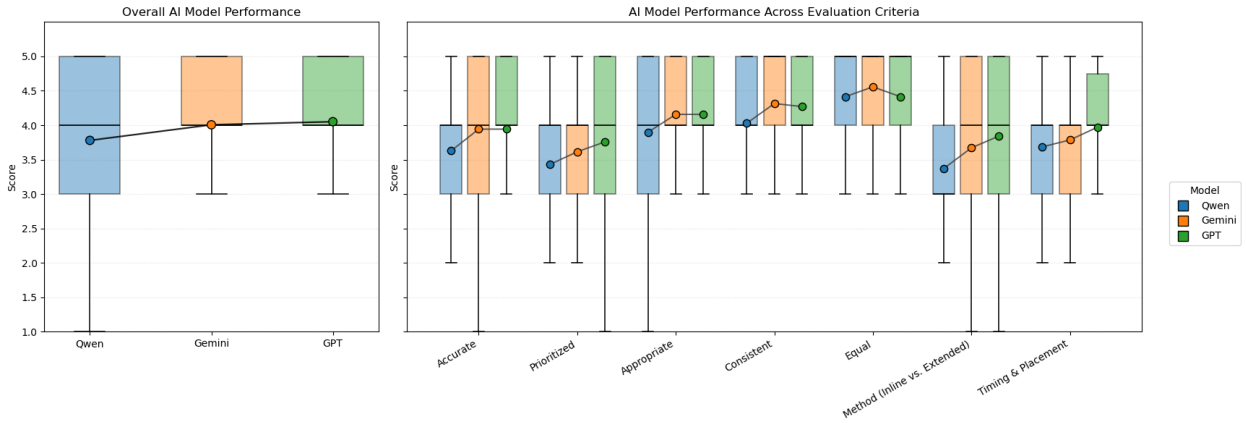


图 9：通义千问、Gemini 和 GPT 在总体指标及七项标准上的评分分布。箱线图显示评分的变化范围；圆点表示均值，并在各类别内相连

4.3.2 各维度表现

维度层面的分数（表 1、图 9）显示，一致性、平等性和恰当的平均值介于 4.0 与 4.3 之间。这些维度的评分主要落在 4–5 范围内，低于 4 的评估相对较少。这一模式表明，在各视频中，评估者总体上认为这些模型保持了术语稳定性、中立性以及与体裁风格的契合。准确维度总体获得中等分数（均值为 3.6–3.9），但各模型的分布模式各不相同。通义千问的均值较低（3.63），大多数评分介于 3 与 5 之间，少数降至 2。Gemini 与 GPT 共享较高的均值（3.94），但其分布有所不同；GPT 的评分集中在 4 附近，而 Gemini 覆盖了完整的 1–5 量表，表明变异性更大。优先排序维度在所评估的维度中均值分数最低，其中包括内容和格式，平均值介于 3.3 与 3.7 之间。GPT 在该方面的评分覆盖范围更广，而通义千问和 Gemini 更常被评为 3 到 4 之间，且从未达到最高分。由于优先级排序本质上具有主观性，取决于评估者对不同类型内容中何为相关的判断，分数总体上往往更低且更具波动性。

两个格式维度——传递方式的策略性运用（内嵌式/扩展式）和时间安排与音轨位置——关注的是描述如何融入视听流，而非描述的内容。除了优先排序这一维度外，这两个维度的平均分普遍低于内容维度：准确、一致性、平等性和恰当。传递方式的策略性运用这一维度的旁白得分介于 3.4 至 3.8 之间，评分范围较广。这种差异可能反映了评估者对扩展旁白必要性的判断不同：虽然通常更倾向于采用内嵌式传递，但对于何时应在播放过程中插入扩展描述以呈现重要视觉细节，评估者意见不一。时间安排与音轨位置的得分介于 3.7 至 4.0 之间，评分分布的跨度与传递方式的策略性运用相近。这可能反映出描述与视频时间安排及对话对齐方式上的差异。

总体而言，内容相关维度与格式相关维度之间存在差异。一致性、平等性和恰当性的评分分布相对集中，大多数评分聚集在量表的较高端。准确的得分略低，偶尔出现下降；而优先排序在所有维度中的得分最低，且分布范围更广。两个格式维度的评分也低于大多数内容维度（优先排序除外），并呈现出更宽的分布范围，表明对呈现方式选择以及与视听流同步性的评价存在更大的差异。鉴于样本量较小，这些发现应谨慎解读，但它们凸显了进一步提升 GenAD 表现的机会，尤其是在生成具有优先排序且格式良好的描述方面，从而为盲人和低视力用户提供更有意义的音频描述。

4.3.3 视频类别层面的表现

不同体裁之间的表现有所差异（表2）。在三个模型中，教育类片段的均值得分最高，其次是教程类与操作和风格类片段，娱乐类片段的评分最低。这一排序（教育类 > 教程类与操作和风格类 >娱乐类）不受模型影响，始终保持一致。在每个类别中，GPT 和 Gemini 的得分都高于通义千问；其中，GPT 在教育类和教程类与操作和风格类中略高于 Gemini，而 Gemini 在娱乐类中略高于 GPT。

总之，这些结果表明，各模型的相对排名在不同体裁中保持稳定：GPT 和 Gemini 的平均得分通常较高，而通义千问则始终较低。结果还表明，体裁对总体

视频类别	Qwen	Gemini	GPT
娱乐（星球大战、伯德小姐、《精灵》、《冰雪奇缘》）	3.66	**3.91**	3.87
教育类（Crash Course Kids、白头海雕）	4.01	4.14	**4.27**
教程类（腌黄瓜、折纸、化妆）	3.76	4.12	**4.14**

表2：各视频类别的人工智能评分均值。粗体表示每行中的最佳模型。

评分也有影响；以事实为导向的教育类内容比以任务为导向的教程类内容或以叙事为主的娱乐类片段获得了更高的评价。娱乐类内容得分较低，可能反映了描述电影素材时面临的额外复杂性：丰富密集的视觉细节和风格元素，会让人更难判断应纳入哪些方面，才能既清晰又能吸引观众。鉴于每个类别中的视频数量有限，且评估者人数较少，这些规律只能视为探索性观察。

4.4 定性讨论

顾问的反馈既凸显了人工智能生成的音频描述的优势，也指出了其局限。几位评分者将一些轨道描述为“非常好”或“很出色，”，甚至有人评论说他们“好奇这是人类还是人工智能。”一位同时培训有视力志愿者和盲人和低视力用户使用无障碍工具的顾问指出，“人工智能已经取得了长足进步”，并且部分输出“堪比”新手描述者的描述。与此同时，顾问们强调，这些描述仍缺少让盲人和低视力受众得以充分享受所需的“最后那*10%*”，并且在某些情况下与广播媒体和流媒体平台上通常遇到的专业音频描述内容“根本无法相比”。几位顾问指出，人工智能往往会犯与初学者相似的判断错误，例如过度描述视觉细节、讲话压过重要的音乐或音效、过度使用提示语或过渡语，以及依赖重复或笼统的用词，这凸显了编辑把关和润色的必要性。

最常被提及的缺点之一与时机和位置有关。评分者描述了一些情况：描述内容“早得离谱”，在词语中间打断对话，令他们觉得“恼人”，或被压缩成一连串“语速很快、难以跟上”的片段。例如，在电影预告片中，时间对齐不佳源于人工智能试图在有限的停顿间隙内塞入大量视觉信息，往往与对话和语音节奏相冲突。另一个例子是操作指南视频：即使没有对话，其本身节奏很快且强调易于理解，因此仓促的描述尤其难以跟上。一位评估者指出，采用扩展旁白或许能缓解这一问题，因为这样可以留出更多时间跟上内容。

然而，过度依赖扩展旁白并非总是有效，因为优先级排序也成为一个反复出现的问题。尽管人工智能生成的描述往往准确，评分者仍指出其中有过多“多余的描述”，而且倾向于描述那些已能从对话中推断出的信息。由于优先级排序本质上具有主观性，一位观众认为重要的内容，对另一位观众而言可能并不重要。顾问建议，用户驱动工具或许有助于弥补这一不足，让盲人和低视力用户能够自主选择按需获取更多细节。
几位评估者观察到，人工智能生成的音频描述往往处于<style id='1'>“不错，但不够优雅”的水平。一位顾问总结道：<style id='3'>“我认为，只需稍作调整，就能让它从非常好提升到卓越。”这一主题凸显了人在回路工作流的价值：由人工描述员对草稿进行完善，在优先级排序和格式方面做出细微调整，就能将尚可的描述转变为自然流畅、衔接无缝的描述。

值得注意的是，评估者区分了日常观看和教学用途所需的合格标准。盲人和低视力受众认为，一些人工智能生成的输出<style id='1'>“看起来已经够用了”，但这些输出未达到教授新手描述者所需的标准。一位评估者说道：<style id='3'>“我绝不会把这段视频用作教学工具；我会自己制作。”这表明评估标准存在分歧：采用消费者视角的顾问对不完美之处更为宽容，而从教学视角出发的评估者则要求具备专业音频描述所应有的精准性、克制性和时间安排上的优雅。
总体而言，评分者对输出内容的流畅性和信息量印象深刻，但他们始终发现在优先级排序、内嵌式与扩展式旁白的使用，以及时机与位置方面存在不足，这些模式与定量结果相呼应。这些差距，例如过度描述、与对话衔接不佳或重点错位，在对话密集或快节奏视频中最具干扰性。研究结果表明，将自动化与润色及交互性相结合存在机会：人在回路工作流可强化编辑决策与呈现方式，而AdaptAD则能让盲人和低视力用户更好地掌控他们希望优先关注的细节。

5 RefineAD 实践

正如一位专家所言，“只需稍加调整，就能从良好变为出色，”我们邀请他们通过编辑界面（图5）试用该人在回路工作流程，以润色人工智能生成的描述。由于时间和预算限制，我们无法要求所有顾问完成全部视频的任务。相反，该专家选择为四段视频各编辑一个人工智能生成的版本：《冰雪奇缘》预告片（2013）、简·古道尔——贡贝国家公园、自制泡菜的三种方法，以及如何折纸狗脸。这些选择是任意的，并非基于视频类型、AI系统或初始质量。该顾问强调，尽管编辑仍需付出一定努力，但该工作流程比从零开始既更快也更容易——这些优势对新手描述者而言可能更为显著。

简·古道尔的视频展示了描述的呈现方式仍然是一项挑战。如图 10 所示，口述旁白本身已极具感染力——描绘出一幅生动的画面：“黑猩猩、狒狒和猴子，鸟儿与昆虫，生机勃勃的森林中繁盛的生命，以及永不停息的水流涌动。”在这种情况下，扩展描述会暂停视频并插入诸如“两只狒狒的特写……”或“一只黑猩猩在葱郁的棕榈树树冠间荡过。”之类的细节，从而打断旁白带来的沉浸感。尽管这些描述准确无误，专家指出，它们重复了口述音轨已经暗示的信息，最终破坏了旁白的节奏。通过协作编辑界面，编辑者可以轻松删除此类音轨，从而得到更精炼的版本，更好地保留原始音频的流畅性。

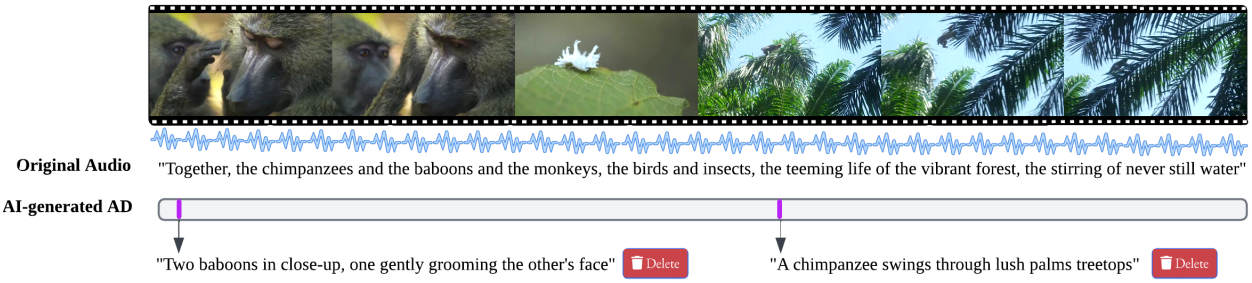


图 10：简·古道尔视频示例。为视觉细节生成的扩展描述打断了简·古道尔旁白的流畅性。通过编辑界面，可以删除此类音轨，以保持原始音频自然的节奏。

专家反馈指出，HITL 组件能够带来“最后 10%”的提升，《冰雪奇缘》预告片就是一个鲜明的例子（见图 11）。在这个场景中，AI 生成的描述准确地指出，斯文跳到了雪宝身旁。然而，专家通过选用动词“蹦跳”并将斯文比作小狗，为描述增添了细腻之处，同时还指出雪宝“挠了挠头”。这些选择不仅描绘了动作，还突显出其俏皮、孩子气的特质，与面向儿童的动画电影语气十分契合。同样，专家将“雪宝看着自己拿着斯文毛皮的树枝手臂，显得很惊讶”改成了简洁得多的“毛皮挠得雪宝鼻子痒”。这一修改以更生动、更精炼的方式传达了相同的视觉信息，强化了幽默感和角色特点，同时避免了不必要的细节。



AI-generated AD: Olaf smiles as Sven jumps playfully, landing beside him. Olaf looks at his hand, which is holding Sven's fur, and seems concerned.

👉 **Expert's Edit:** Sven hops like a puppy and Olaf sratches his head. Fur tickles Olaf's nose.

图 11：《冰雪奇缘》预告片示例。专家的修改通过加入俏皮的比喻并精简措辞，将直白的 AI 草稿改写得更吸引人，也更符合目标年龄段。

在操作指南视频中，时机和优先排序的细节至关重要——专家指出，人工智能生成的描述往往在这两个方面存在不足。有几行需要进行位置调整。正如专家所观察到的，一些提示的呈现“过早，”而当动作被描述得“过快，”时，盲人和低视力受众将难以复现这些步骤。通过调整开始时间，或使用协作编辑界面中的**微调工具**，这些问题很容易得到纠正

（图5）。将旁白与屏幕上的事件对齐，而不是提前预告，提供了更清晰的节奏，使盲人和低视力受众有足够的时间跟上每一步并复现这些动作。

一个清晰的例子可以说明专家编辑如何补充关键细节，见《自制泡菜的三种方法》片段（图 12）。人工智能生成的描述识别出了主要食材并概述了基本动作，但表述相当笼统。专家将这些内容扩展为精确的程序步骤，明确点名工具，例如使用“夹钳”提起罐子，或用“漏斗”倒入盐水，并将各项动作按逻辑顺序排列。这一转变把宽泛的概述变成了受众能够切实照做的、可操作的分步说明。通过强调工具的使用，此次编辑还突出了触觉元素——例如拿取热罐子或对准漏斗——这些对于复现至关重要，对盲人和低视力用户尤其如此，但在自动生成的初稿中却常常被忽略。



AI-generated AD: A cutting board displays Kirby cucumbers, a clove of garlic, and fresh dill next to a pot of brine, while a jar is filled with the brine solution. The jar is sealed with a lid, indicating it's ready for storage.

👉 **Expert's Edit:** With **tongs**, remove one of the jars from the water bath, add one large, peeled garlic clove, a small bunch of dill, and enough cucumber spears to fill the jar. Placing a **large mouth canning funnel** on top of the jar, pour in the hot brine until the jar is full. Screw on the sterilized top.

图 12：来自《自制泡菜的三种方法》片段的示例。专家编辑通过指明工具并澄清操作程序，扩展了人工智能生成的动作，使步骤更加实用且可复现。

在《如何折纸狗脸》视频（图13）中，AI生成的旁白准确地捕捉了绘制狗脸的基本步骤，但专家用具体的空间参照丰富了这些说明。例如，专家着重标出了“耳朵的内缘”和“狗脸的中央折痕，”，提供了清晰的参照点来引导观众。胡须也被描述为“曲线*W*”形，这一简洁的视觉提示使说明清晰且富有吸引力。临近结尾时，AI将指导者的动作按字面描述为“把它放在桌上，稍微抬起，然后再放回去。”虽然正确，但这种表述可能会让观众感到困惑，因为这些手势只是在展示成品，而非引入新的步骤。专家则改为补充：“恭喜！你现在拥有一只耳朵折得可爱的折纸狗。”这一替代说法传达了完成之意，并为旁白赋予了温暖、自然的语气，而这正是AI草稿所缺乏的。



AI-generated AD: Aligns paper with a vertical crease, draws two dots for eyes, then adds a curved line below for a mouth. Adds small dots above the mouth for whiskers, then completes the dog's face on the origami. Rotate the origami dog face to show its front, then place it flat on the table. Lift it slightly, then set it back down.

👉 **Expert's Edit:** Using the inner edges of the ears, and the landmark crease in the center of the dog's face as boundaries, Take a marker and draw two dots for eyes. And below that a larger dot for a nose. Just below the nose make a shallow curved W to form the mouth. Congrats! You now have an origami dog with adorable folded ears.

图13：折纸片段示例。专家的修改强调了空间方位，并增加了对话式的结尾，从而提高了清晰度与参与度

这些修改共同展示了人在回路工作流如何将尚可但参差不齐的草稿转化为更精炼、更清晰且与视听流更加契合的成果。通过删减冗余、明确程序步骤，

并调整时机，这位顾问证明，即使是适度的干预，也能显著提升人工智能生成的描述的实用性和流畅度。

6 AdaptAD：打造用户驱动体验

音频描述没有放之四海而皆准的方法。尽管可以依据既定指南评估“优质”音频描述，但正如一位专家所指出的：“音频描述的最终评判者，是那些观看内容的盲人和低视力用户。”偏好不仅因人而异，也会因视频类型和具体情境而异，并受到个人经历和观看目标的影响（Jiang Crescentia Jung Mahika Phutane 等人，[无日期]）。这一点在我们的研讨会定量和定性讨论中均有所体现：GenAD 经常难以确定优先级和呈现方式，尤其是在选择内嵌式旁白还是扩展式旁白时。由于扩展式旁白通常被定义为“仅在必要时使用”，关键难题便在于判断哪些内容不可或缺，而这在很大程度上取决于主观判断。一些盲人和低视力观众更喜欢简洁的音频轨道，以尽量减少冗余；另一些人则希望获得更多阐述，以满足好奇心或加深参与度。

AdaptAD 旨在通过将控制权交给用户来应对这一挑战。AdaptAD 不再依赖单一的编辑判断，而是支持盲人和低视力观众按需请求更多细节。对三个示例的初步演示表明，这项功能可以通过填补传统音频描述系统留下的信息空白来提升无障碍体验。重要的是，这些情境由具有直接服务盲人和低视力用户经验的顾问设计，他们分享了观众经常提出的问题示例。

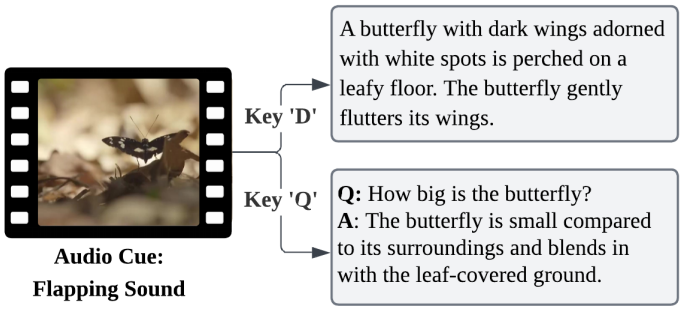
在《简·古道尔——贡贝国家公园》的蝴蝶场景中（图14a），简·古道尔的旁白已对场景作了丰富交代，因此部分观众可能觉得对蝴蝶的扩展描述并无必要，甚至会干扰体验。另一些观众则可能注意到音频中的拍打声，想知道它来自鸟还是蝴蝶。AdaptAD支持每位用户自行决定：按下*D*，即可获得诸如“一只蝴蝶振翅飞过画面。”的描述。仍感好奇的用户可按下*Q* 提出后续问题，例如“这是哪种蝴蝶？”并获得详细解答。如此，AdaptAD既能兼顾两种偏好，又不会把单一的编辑选择强加给所有听众。

教学视频也带来类似挑战。在《自制泡菜的三种方法》中（图14b），有些观众可能更喜欢简洁旁白，以免被过多细节淹没，而另一些观众则需要精确信息才能复现食谱。按下*D*可获得场景概览，例如：“新鲜莳萝、蒜瓣和柯比黄瓜条……”若需更细致的信息，用户可按下*Q* 并提问：“黄瓜块有多大？”从而得到答案：“它们被纵向切成四条，每条约*4–5*英寸长。”AdaptAD确保寻求更清晰教学说明的用户能够获取所需信息，同时让希望更高效观看的用户不必被不必要的细节所累。

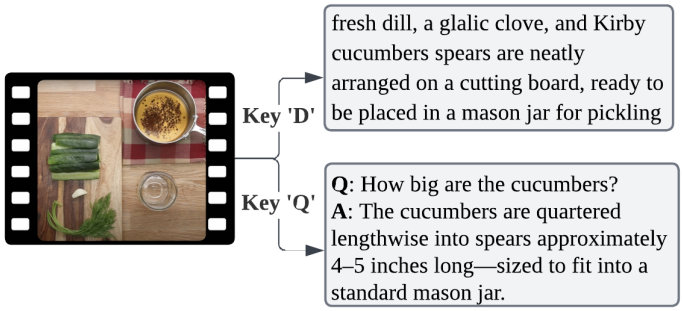
娱乐类内容进一步凸显了人们对音频描述偏好的高度主观性。在《星球大战：天行者崛起》预告片（图14c）中，有些观众可能希望进一步了解世界构建和场景，另一些观众则更关注人物，甚至是扮演他们的演员。顾问指出，盲人和低视力用户经常会询问演员的外貌和风格。例如，观众可能会按下*Q*来提问：“蕾伊长什么样？”AdaptAD 可以给出客观描述，例如：“蕾伊是一位年轻女性，头发扎成三个发髻，身穿一套朴素的白色服装。”这样，对角色外貌更感兴趣的用户便能按需获取相关信息。

这些例子说明，音频描述中的优先级排序无法通过单一的编辑决策来解决：对一位观众来说多余的内容，对另一位观众而言可能至关重要。AdaptAD 让盲人和低视力受众能够快速请求澄清或更深入的补充说明，从而使音频描述更灵活、更个性化，也更能响应个人需求。顾问还指出，挑战不仅在于平衡不同偏好，也在于填补固定旁白可能留下的空白。例如，一位顾问回忆说，曾参与制作一段较长、没有对话的 Baby Gorilla Cam 视频，描述撰写者直到节目很后面才解释大猩猩究竟长什么样。一位盲人质量审核员发现了这一疏漏，促使团队修改描述。此类情况表明，明眼描述者往往“依赖视觉”，可能忽略他们认为理所当然、但对盲人和低视力受众至关重要的信息。借助 AdaptAD，这些信息缺口可以得到弥补，甚至消除，因为用户只需提问，例如“大猩猩长什么样？”或“大猩猩和人类相比是什么样的？”。

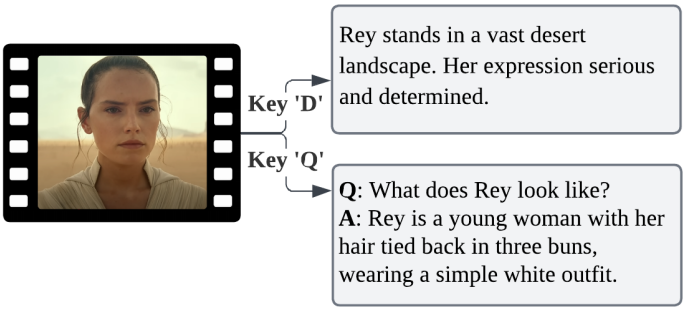
AdaptAD 的设计参考了兼具亲身经历和专业知识的无障碍顾问所提供的意见，也借鉴了此前涉及盲人和低视力参与者的研究成果（Bodi 等（2021）；Stangl 等（2023））。尽管这些演示颇具前景，接下来的首要任务是与盲人和低视力用户一起评估集成系统，重点关注他们使用*D*和*Q* 功能的频率、促使他们使用这些功能的因素，以及他们提出的问题类型。这些数据有助于发现持续存在的不足、指导改进，并揭示盲人和低视力受众最看重哪些信息。



（a）简·古道尔——贡贝国家公园中的蝴蝶场景（教育）



（b）自制泡菜的三种方法（操作与风格）



（c）《星球大战：天行者崛起》预告片（娱乐）

图14：由音频提示或用户查询触发的 AdaptAD 响应示例，可按需提供额外的场景细节。

与我们从无障碍专家研究中获得的启示一致，AdaptAD 的设计不仅将其作为音频描述工作流程的重要组成部分，也将其作为一种让用户主动塑造这一过程的机制。盲人和低视力观众不必被动地接收描述，而是可以指出旁白在哪些地方不够完整，以及哪些细节最有意义。展望未来，这项功能还可以赋能盲人和低视力人士成为内容创作者，让他们录制并发布自己的音频描述版本，既贴合他们认为重要的内容，也考虑到他们预计其他盲人和低视力用户所看重的内容。如此一来，AdaptAD 不仅有望让内容更具无障碍性，也能使其更具包容性、更能响应需求，并更多地由社群推动。

7 讨论

7.1 对研究问题的反思

我们的研究旨在回答两个研究问题：（1）现代视觉语言模型生成基线描述的效果如何？（2）整合 GenAD、RefineAD 和 AdaptAD 后，如何提升旁白质量和用户体验？

对于**研究问题1**，我们的初步分析显示，采用无障碍引导提示和上下文线索（例如音频转录文本、先前的场景描述、视频元数据）能够生成更切题、更连贯且更符合盲人和低视力用户偏好的描述。专家评估表明，通义千问、Gemini 和 GPT 等视觉语言模型能够生成合格的草稿，在恰当性、一致性和平等性方面得分较高，准确性得分则处于中等水平。然而，在信息优先级排序、呈现方式的选择（内嵌式或扩展式）以及时机方面仍存在明显不足，而这些都是音频描述实践的核心。这些结果表明，尽管视觉语言模型能够提供有用的起点，但要构建一个完全无障碍的系统，仍需进一步投入设计工作。

对于**研究问题2**，既有研究已证明编辑和交互式工具的价值，即使使用较早期的图像描述生成模型也是如此（Yuksel 等人（2020）；Bodi 等（2021）；Stangl 等（2023））。我们与无障碍专家开展的研讨会和形成性演示表明，RefineAD 和 AdaptAD 能够精准针对 GenAD 的弱点。RefineAD 借助人工判断，对优先级排序、呈现方式和音轨位置进行把控；AdaptAD 则让盲人和低视力用户能够请求对他们最相关的细节，直接弥补固定旁白无法填补的信息缺口。

7.2 局限性与未来工作

本研究存在若干局限性。尽管与无障碍专家开展的评估研讨会带来了宝贵见解，但仍应谨慎解读定量结果。视频集范围较窄，仅涵盖三种体裁，因此研究发现仍属描述性结果，无法支持有力的统计结论。目前尚无法确定所测试的三种视觉语言模型——Gemini、GPT 或通义千问——哪一种最适合投入生产，也无法断言其性能差异具有普遍性。要开展具有统计意义的基准测试，还需要规模更大、更多样化的数据集，并纳入更多评估者。

评估也面临挑战。大规模获取专家反馈并非易事，因为评估完整的视频描述既耗费人力，又需要关注语言和时间细节，成本也很高。为此，我们将探索结合专家判断与自动化评估的混合方法。近期研究表明，大语言模型可以充当评估者，其评分往往与人工评分一致 Zheng等人（[n. d.]）；Kocmi和Federmann（[n. d.]）。我们旨在使此类自动化方法符合无障碍领域专家标准，从而实现可靠且可扩展的评估。要让人工智能生成的评估获得信任，严格的质量控制至关重要。

一个更显著的局限是缺乏针对盲人和低视力参与者开展的直接研究。尽管 RefineAD 在设计时考虑了无障碍需求，纳入了 WAI-ARIA 以确保屏幕阅读器兼容性、完整的键盘导航以及其他包容性功能，但目前尚未对盲人和低视力受众进行全面、系统的评估。先前研究指出，人在回路编辑和按需交互式音频描述能够提升系统质量和用户体验 Yuksel 等人（2020年）；Stangl 等（2023）。在此基础上，我们升级了 GenAD，并预期改进后的基线能够减少编辑工作量、提高满意度。然而，最终而言，盲人和低视力观众才是评判音频描述质量的黄金标准。我们计划开展受控的可用性研究和纵向部署，考察盲人和低视力受众如何使用该系统，以及他们的编辑和偏好与明眼描述者相比有何不同。

解决这些局限将开启几个前景广阔的研究方向。涵盖更多体裁且变化更丰富的数据集，对于检验 GenAD 在专业人士和社区撰写的音频描述上的表现至关重要。扩大评估范围，纳入更多类型的受访者，包括专家、专业描述撰写者、新手描述者，尤其是盲人和低视力参与者，将确保评估能够反映最依赖音频描述的人群的观点。

我们的发现表明，有必要重新审视评估框架。总体而言，顾问认为评估模型清晰易用，并且符合专业指南。有人指出，与一致性等风格特征相比，以呈现方式为重点的维度（例如内嵌式旁白与扩展式旁白，以及音轨位置）或许值得得到更多重视。如何界定和衡量音频描述中的“质量”，尤其是针对完整视频，仍是一项悬而未决的挑战。为了确保 GenAD 生成有意义且实用的描述，我们计划建立更全面的质量定义，并改进评估方法，以更高效地开展评估。

另一个重要方向是情感化语音。证据表明，盲人和低视力受众始终更偏好人工旁白，因为人工旁白清晰且富有情感深度，而合成语音往往难以做到这一点（Fernández-Torné，2015年；Walczak 和 Fryer，2018年；RNI（[n.d.]））。我们计划在 GenAD 和 AdaptAD 中集成表现力更强的文本转语音语音，这有助于超越功能性无障碍，带来更丰富、更引人入胜的体验。

8 结论

本文介绍了ADx3，一种协作工作流，它整合了人工智能生成的基线、人在回路的优化与自适应用户控制，以大规模支持高质量的无障碍音频描述。我们的评估与

无障碍专家表明，尽管视觉语言模型在以无障碍为重点的提示引导下能够生成“好”描述，但在优先级排序、时机和呈现方式方面仍存在挑战。RefineAD和AdaptAD通过实现编辑细节，并根据多样化的盲人及低视力人士的偏好对旁白进行个性化，弥补了这些不足。

我们强调，ADx3 是一项探索性贡献。该框架展示了如何将自动化、人工编辑和交互式查询整合到工作流程中，以支持有意义且更具适应性的音频描述，但仍需通过盲人和低视力受众进行全面验证。未来的工作将通过扩展数据集和开展大规模评估来推进这项研究，并让无障碍专家、专业描述员、新手描述者，尤其是盲人和低视力用户参与持续的纵向研究。通过将可扩展的自动化与包容性协作相结合，ADx3 将音频描述定位为动态过程，而非固定产品，在以盲人和低视力受众的自主权与体验为中心的同时，拓展他们获取数字媒体的渠道。

GenAD 提示模板

本节包含我们系统中使用的完整提示模板，包括初始指南设置、生成、优化和润色提示。

A.1 指南提示

```
    指南  ="""
    音频描述指南：
    - 简洁、客观地描述你所看到的内容。
    - 始终按照屏幕文字显示的内容逐字朗读。
    - 描述应客观、准确、精确。
    - 尽可能使用上下文中的规范术语和名称。
    - 使语气和情绪与视频相符。
    - 不要过度描述——少即是多。
    - 不要解读或评论你所看到的内容。
    - 不要提前透露尚未发生的惊喜。

    重要角色识别：
    - 根据上下文认出角色时，务必使用其具体姓名。
    - 每个场景开始前，仔细查看上下文，识别所有有名有姓的角色。
    - 根据上下文信息，尽可能具体地识别。
    """

    指南_提示词 = """
    你是一名专业的音频描述员，请遵循以下指南：

    {guidelines}

    你理解这些指南了吗？请回复“YES”并简短确认。
"""
```

图 15：用于视觉语言模型初始化的指南提示

A.2 视频生成提示

```
提示词 = """
场景时长：{scene_duration:.2f} 秒

上下文：
{context}

你正在分析一个视频场景。请识别上下文中提到的具体人物、地点以及任何重要元素。


首先，生成一个屏幕文字事件的 JSON 数组。
文字事件（"type": "屏幕文字"）：
- 捕捉所有可见的屏幕文字。
- 不要包含文字转录或对话。
- 重要：对于每个文字事件，，都要包含准确的 `开始 _文本出现的时间` （以秒为单位）


包括：
- 标题、章节标题、姓名 - 信息



不包括：
- 品牌徽标和水印 - 网络徽标 - 社




第二，生成一个视觉事件 JSON 数组。
- 使用简洁的措辞，对场景提供丰富的视觉描述 - 详细描述每个动作 - 始终使用上下文中的具体角色姓名







规则：
- 将输出格式设为 JSON 数组。每个对象应包含：- `start_time`（秒数） - `
```

图16：用于生成场景级描述和文字事件的提示词

A.3 内联优化提示词

```
提示词 = """
你正在优化一段视频的一组视觉描述。
原始描述："{combined_text}"

可用时长：{available_duration:.2f} 秒

任务：
合并并精简这些描述，使其适合在 {available_duration:.2f} 秒内完成。

指导原则：
- 编写连贯流畅的描述
- 保持动作顺序，并使用简洁自然的语言
- 确保最终描述必须在时间限制内读完

输出格式：
仅提供优化后的描述文本，不要附加解释。
"""
```

图17：内联描述优化提示

A.4 重试优化提示

```
重试_提示词 ="""
你正在为视频优化多条视觉描述。
上一次尝试："{optimized_text}"

这段描述需要 {tts_duration:.2f} 秒才能读完，但你只有 {
    available_duration:.2f} 秒可用。
你需要将其缩短 {tts_duration - available_duration:.2f} 秒。

任务：
请创建一个更短的版本，必须在时间限制内完成。

指南：
- 保留最关键的视觉元素，删除冗余细节
- 使用简洁自然的语言

输出格式：
仅提供缩短后的描述，不要添加其他内容。
"""
```

图18：用于重试失败的行内优化的提示词

A.5 扩展描述过滤提示

```
提示词 = """
你是一位无障碍专家，需要为每个场景选择一个视觉描述，将其转换为音频
    描述，供盲人和低视力用户使用。

上下文：
- 音频描述应简洁，仅在必要时打断音频
- 口述转录文本是主要信息来源
- 仅当音频中缺少关键视觉信息时使用音频描述

输入：
当前场景转录文本：
{transcript_text}

累计转录文本：
{cumulative_文字转录}

累计描述：
{previous_描述_文本}

待评估的视觉描述：
{clip['文本']}

评估标准：
如果符合以下任一条件，请加入描述（必要时为 = true）：
- 传达音频中未提及的视觉细节
- 描述重要的无声动作或关键事件
- 确定其他情况下含义不明确的人物或地点
- 引入新颖或独特的视觉元素
- 标注场景变化或未明确说明的时间跳跃

输出格式：
返回一个 JSON 数组，其中每个条目都包含 "necessary": true：
- "id"：描述的索引
- "necessary"：true 或 false
- "reason"：简短说明
"""
```

图 19：用于筛选扩展的音频描述候选项的提示词

A.6 教程类 SceneMergingPrompt

```
提示词 = """
任务：
将屏幕文字和视觉元素合并为对场景的一个连贯描述，然后
    缩短描述，使其符合时间限制。

重要：
- 保留并包含所有测量值，且必须完全一致
- 最终描述必须控制在 {available_scene_time:.2f} 秒以内
"""
```

图20：用于合并和压缩说明性描述的提示词

B AdaptAD 提示模板

B.1 D 键

```
查询 == “描述场景”:
提示词 = f"""视频场景上下文:
    {场景_信息_文本}
    用户查询：{query}
    重要提示：您正在查看视频中的两个帧。
    - 第一帧是接近时间戳 {keyframe_time:.2f}s 时捕获的关键帧。
    - 第二帧是在时间戳 {exact_time:.2f}s 时捕获的确切帧。
    """
```

图 21：用户按下 *D* 键以请求场景的更多描述时显示的提示词

B.2 Q 键

```
提示词 = f"""视频场景上下文:
    {场景_信息_文本}
    用户查询：{query}
    重要提示：你正在查看视频中的两帧。
    - 第一帧是接近时间戳 {keyframe_time:.2f}s 时捕获的关键帧。
    - 第二帧是在时间戳 {exact_time:.2f}s 时捕获的确切画面。
    利用上下文，以一句非常简洁且信息丰富的回答回应{query}

    当查询是“为什么”问题时，您的回答必须包含具体的因果细节
        细节
    （明确涉及人员、事件发生时间及所发生的事情），并直接
        回答该查询。
    如果问题询问地点，请具体说明场景。
    请勿添加超出范围的外部细节。
    请勿提及画面编号或时间戳。
    """
```

图22：用户按下按键 *D* 以请求场景的更多描述时显示的提示词

C 音频描述多维评估模型

本附录部分重现了提供给专家评分者的完整文档。

概述：此模型从两个主要维度评估 YouTube 视频的人类及人工智能生成的音频描述：**内容**（五项标准）和**格式**（两项标准）。

一、内容

（*1*）准确——无错误内容

定义：描述 提供准确无误的视觉信息，正确识别实际发生的事情 .
没有事实性错误，也没有 误导性信息。

评估标准 (1–5)：

5：所有视觉元素均符合事实。对实际发生情况的描述没有错误。事实准确性完美。4：基本符合事实，仅有不会造成误导的轻微错误。描述总体准确。3：总体符合事实，但存在一些明显错误。大体准确，但有一些疏漏。2：存在多处事实错误，导致对实际发生情况产生误导。描述的准确性较差。

1：重大事实事实错误，或信息完全不正确。未能准确描述正在发生的事情。

（2）优先排序——语境与推断

定义： 描述通过根据细节的语境重要性和推断价值来选择内容，实现了最佳优先级排序。例如，描述会优先选取语境丰富的细节，而非“哈利跑进森林”与“一个男孩跑进森林”这类泛泛的描述；也会做出合理推断，例如描述为“一个穿着足球队服的男孩”，而不是“一个穿红色球衣、穿着高筒袜的男孩”。

评估标准（1–5）：

5：平衡恰到好处——对理解最重要的元素进行了完美的优先级排序。选择与语境相关的细节和恰当的空间信息。4：优先级排序良好但不够完美——内容略显泛泛或略显冗长。对纳入哪些内容的选择总体不错。3：优先级排序尚可，但存在明显失衡——要么遗漏了一些重要细节，要么纳入了一些不必要的信息。2：优先级排序较差——要么重要信息不完整，要么纳入了过多不重要的细节。对重要内容的取舍不当。1：存在重大问题——要么重要信息存在重大缺漏，要么事无巨细地描述所有内容，包括不重要的元素。未能清晰地优先呈现重要内容。

(3) 恰当——受众与目的契合度

定义： 描述所用的语言、细节程度和风格应符合内容类型，并契合目标受众的体验需求。例如，娱乐视频的描述应增强观赏乐趣，教育视频的描述应帮助受众理解，而教学视频则应让观众能够跟随或复现视频中展示的步骤。

评估标准（1–5）：

5：完全契合——语言和细节程度精准匹配受众的理解能力与内容目的，描述充分支持预期体验。4：契合度良好，仅有细微偏差——总体上适合受众和目的，但在语气、复杂程度或重点方面偶有不妥。3：契合度尚可，但存在明显脱节——部分满足受众需求和内容目的，但语言水平或实际需求方面的匹配不够一致。2：契合度较差——经常使用不适合受众的语言，或未能支持内容目的。描述往往与预期目标背道而驰。1：完全不契合——语言和表达方式完全不适合受众，和／或有损内容目的。显然未考虑谁会使用这些内容，也未考虑使用目的。

(4) 一致性 — 一致性与连贯性

定义： 描述在整个视频中保持术语、风格和语气一致，从而支持连贯统一的叙事。

评估标准（1–5）：

5：术语和风格完全一致。叙事流畅连贯。4：总体一致，仅有细微差异。叙事整体上仍然连贯。3：一致性尚可，但术语或风格有一些明显变化。2：措辞或语气经常不一致，叙事变得难以理解。1：完全无法保持一致。叙事支离破碎或缺乏连贯性。

(5) 平等性 — 客观性与非诠释性

定义： 该描述通过保持客观、不含个人解读、偏见或不必要的评论，确保平等获取。

评估标准（1–5）：

5：完全客观。无个人解读。使用恰当的描述性语言，不含编辑性评论。4：总体客观，仅有极少轻微的解读性片段。3：大体客观，但存在一些不必要的解读。2：解读性语言频繁。描述中可见一定偏见。1：高度解读且带有偏见。大量个人评论妨碍平等获取。

II. 格式

（*1*）策略性地使用描述方法（内嵌式与扩展式）

定义： 根据内容特点，在内嵌式和扩展式描述方法之间做出恰当选择。

内嵌式描述是以下情况下的标准且首选方法：

- 原始内容时序中存在足够的自然停顿 国家无障碍媒体中心（2017年）
- 可在现有音频间隙中充分描述视觉内容3Play Media（2020年）

扩展式描述适用于以下情况：

- 文字密集视频，例如幻灯片放映或讲座录制视频 3Play Media（2020年）
- 对白密集视频，因为音频描述不应盖过人们的谈话声 3Play Media（2020年）
- 包含重要音乐或声音的嘈杂视频，因为音频描述可能会削弱这些元素的效果 3Play Media（2020年）
- 包含短镜头和/或细节极其丰富的画面的视频，标准描述在下一个镜头切换前可能无法完整呈现内容 3Play Media（2020年）
- 核心视觉信息无法充分融入现有的自然停顿中：“如果没有这样的停顿，必须在该处插入扩展描述 国家无障碍媒体中心（2017年）

评估标准（1–5）：

5：方法选择完美——根据专业标准，始终为有足够停顿的内容选择内嵌式，仅在绝对必要时选择扩展式。4：方法选择良好，偶有轻微错误——选择总体恰当，极少不必要地使用扩展描述。3：方法选择尚可，但存在一些不佳选择——有时会不必要地使用扩展式，或在需要扩展描述时错失机会。2：方法选择较差——经常选错方法，要么过度使用扩展描述，要么在需要时未使用。1：方法选择存在严重问题——不了解根据专业标准何时应使用内嵌式或扩展式。

（*2*）时机与位置

定义： 恰当 描述相对于视觉内容和音频元素的恰当放置时机，依据 n
既定的无障碍 性标准。两种描述方法的时机标准如下：

- 不得打断重要对话或关键音效 DCMP（2024年）。
- 应在时间轴上的自然节点插入描述——不要在说话者话音未落时打断，但要利用短暂停顿。只要描述不脱离语境，即使是词语或句子之间的停顿也足够 国家无障碍媒体中心（2017年）。
- 应尽可能将描述安排在靠近视觉动作的位置 3Play Media（2020年）。
- 允许预先描述：若有助于澄清情境，描述可插入“略早于屏幕上动作发生之时”国家无障碍媒体中心（2017年）。

评估标准（1–5）：

5：最佳时机——描述安排在靠近视觉动作的自然停顿中，且不打断关键音频。4：偶尔时机不佳——总体位置良好，但有时描述过早、过晚，或与重要音频略有重叠。3：时机问题明显——描述相对于视觉内容的时机不佳，对对话有一定干扰。2：时机较差——描述经常时机不当，频繁打断对话，或与相关动作相距过远。1：严重时机问题——时机持续不佳，扰乱内容流程并干扰关键音频。

参考文献

[n. d.]. Fundamentals of Audio Description. ([n. d.])

.[无日期.]. RNIB研究报告：用于音频描述的合成语音 | RNIB. https: // www.rnib.org.uk/professionals/research-and-data/reports-and-insight/ rnib-study-report-synthetic-speech-for-audio-description/?utm_source=chatgpt.com[无日期]. WAI-ARIA概述 | Web无障碍倡议（WAI） | W3C. https://www.w3.org/WAI/ standards-guidelines/aria/ 2008. 万维网联盟（W3C）。2008。《理解成功标准1.2.3：音频描述或媒体替代方案（预录内容）》。https://www.w3.org/TR/UNDERSTANDING-WCAG20/ media-equiv-audio-desc-only.html. 3Play Media. 2020. 音频描述（AD）指南。 https://www.3playmedia.com/popular-topics/ audio-description/ Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang、Humen Zhong、Yuanzhi Zhu、Mingkun Yang、Zhaohai Li、Jianqiang Wan、Pengfei Wang、Wei Ding、Zheren Fu、Yiheng Xu、Jiabo Ye、Xi Zhang、Tianbao Xie、Zesen Cheng、Hang Zhang、Zhibo Yang、Haiyang Xu 和 Junyang Lin。2025。《Qwen2.5-VL技术报告》。（2025年2月）。 http://arxiv.org/abs/2502.13923 Daniel Bergin 和 Brett Oppegaard。2025。《媒体无障碍自动化：一种分析生成式人工智能算法中音频描述的方法》。《技术传播季刊》34, 2 (2025), 169–184. doi:10.1080/10572252.2024.2372771 Aditya Bodi、Pooyan Fazli、Shasta Ihorn、Yue Ting Siu、Andrew T. Scott、Lothar Narins、Yash Kant、Abhishek Das 和 Ilmi Yoon。2021。《面向盲人及低视力用户的自动视频描述》。计算系统中的人为因素会议论文集(2021年5月)。 doi:10.1145/3411763.3451810 Carmen J Branje 和 Deborah I Fels《Structured》。 [无日期].《*LiveDescribe*：业余描述撰写者能否创作高质量音频描述？》技术报告。Maryam Cheema、Hasti Seifi 和 Pooyan Fazli。2024。《Describe Now：由用户驱动的盲人及低视力人士音频描述》。（2024年11月）。 http://arxiv.org/abs/2411.11835 Peng Chu、Jiang Wang 和 Andre Abrantes。2024。《LLM-AD：基于大语言模型的音频描述系统》。（2024年5月）。https://arxiv.org/abs/2405.00983v1 DCMP. 2024. 专业指南。https://dcmp.org. Anna Fernández-Torné. 2015.《专业翻译杂志》文本转语音与真人配音音频描述：一项针对加泰罗尼亚语配音电影的接受度研究。（2015）。Chaoyou Fu、Yuhan Dai、Yongdong Luo、Lei Li、Shuhuai Ren、Renrui Zhang、Zihan Wang、Chenyu Zhou、Yunhang Shen、Mengdan Zhang、Peixian Chen、Yanwei Li、Shaohui Lin、Sirui Zhao、Ke Li、Tong Xu、Xiawu Zheng、Enhong Chen、Rongrong Ji、Xing Sun、项目负责人及通讯作者。2024。《Video-MME：首个用于视频分析的多模态大语言模型综合评测基准》。（2024年5月）。 https://arxiv.org/ pdf/2405.21075 Google DeepMind. 2024.《*Gemini 1.5*技术报告》。技术报告。 https://deepmind.google/ technologies/gemini/gemini-1-5/ Spence Green、Jeffrey Heer 和 Christopher D. Manning。2013。人工译后编辑对语言翻译的效用。计算系统中的人为因素会议论文集 (2013), 439–448. doi:10.1145/ 2470654.2470718

Lucy Jiang Crescentia Jung Mahika Phutane、Abigale Stangl Shiri Azenkot、Lucy Jiang、Crescentia Jung、Mahika Phutane 和 Abigale Stangl。 [无日期]. “这在某种程度上取决于情境”：了解盲人与低视力人群在不同观看场景下的视频无障碍偏好。24 ([无日期]). doi:10.1145/3613904.3642238 Tom Kocmi 和 Christian Federmann。 [无日期]. 大语言模型是翻译质量评估的最先进工具。（[无日期]）。 https://github.com/MicrosoftTranslator/GEMBA Samuel Läubli、Mark Fishel、Gary Massey、Maureen Ehrensberger-Dow 和 Martin Volk。 [无日期]. 在真实翻译环境中评估译后编辑效率。（[无日期]）。 http://www.my-across.net/en/ Junnan Li、Dongxu Li、S. Savarese 和 Steven C. H. Hoi。2023。BLIP-2：利用冻结图像编码器和大语言模型启动语言-图像预训练。国际机器学习会议(2023)。 doi:10.48550/ARXIV.2301.12597 Mala D. Naraine、Deborah I. Fels 和 Margot Whitfield。2018。对质量的影响：盲人及低视力观众娱乐体验评分中的愉悦因素：一项定性研究。*PLoS ONE* 13, 12 (2018年12月)。 doi:10.1371/journal. pone.0208165 Rosiana Natalie、Ebrima Jarjue、Hernisa Kacorri 和 Kotaro Hara。2020。ViScene：用于视频场景描述的协作创作工具。*ASSETS*。*ACM*辅助技术会议2020 (2020年10月), 87. doi:10.1145/3373625.3418030 国家无障碍媒体中心。2017。《无障碍数字媒体指南》。 https://ncam.wgbh.org Zheng Ning、Brianna L. Wimer、Kaiwen Jiang、Keyi Chen、Jerrick Ban、Yapeng Tian、Yuhang Zhao 和 Toby Jia-Jun Li。2024。SPICA：通过增强音频描述，为盲人或低视力观众提供交互式视频内容探索。*CHI*计算系统中的人为因素会议论文集（*CHI ’24*），*2024*年*5*月*11*日*â•fi16*，*2024*年，檀香山，美国夏威夷州1 (2024年2月), 18. doi:10.1145/3613904.3642632 OpenAI. 2024. GPT-4o系统卡。（2024）。Amy Pavel、Gabriel Reyes 和 Jeffrey P. Bigham。2020。Rescribe：音频描述的创作与自动编辑。*UIST 2020*——第*33*届*ACM*用户界面软件与技术年度研讨会论文集(2020年10月), 747–759. doi:10.1145/3379337.3415864/SUPPL{_}FILE/3379337.3415864.MP4 Anna Rohrbach、Atousa Torabi、Marcus Rohrbach、Niket Tandon、Christopher Pal、Hugo Larochelle、Aaron Courville 和 Bernt Schiele。2017。《电影描述》。《国际计算机视觉杂志》123, 1 (2017年5月), 94–120. doi:10.1007/S11263-016-0987-1 Habiba Sarhan 和 Simon Hegelich。2023。理解并评估人工智能生成的图像说明在政治图像中造成的危害。《政治科学前沿》5 (2023年9月), 1245684. doi:10.3389/FPOS.2023.1245684/BIBTEX Emilie Schmeidler、Corinne Kirchner、Katharine Bond、Laurie Everett、Jaclyn Packer、Lawrence Scadden、Joel Snyder 和 Karen Wolffe。2000。添加音频描述：会产生不同效果吗？技术报告。Christoph Schuhmann、Romain Beaumont、Richard Vencu、Cade Gordon、Ross Wightman、Mehdi Cherti、Theo Coombes、Aarush Katta、Clayton Mullis、Mitchell Wortsman、Patrick Schramowski、Srivatsa Kundurthy、Katherine Crowson、Ludwig Schmidt、Robert Kaczmarczyk 和 Jenia Jitsev。2022。LAION-5B：用于训练下一代图文模型的开放式大规模数据集。（2022年10月）。http://arxiv.org/abs/2210.08402 Abigale Stangl、Shasta Ihorn、Yue Ting Siu、Aditya Bodi、Mar Castanon、Lothar D. Narins 和 Ilmi Yoon。2023。视频自动音频描述系统中视觉对话智能体的潜力。*ASSETS 2023*——第*25*届国际*ACM SIGACCESS*计算机与无障碍会议论文集(2023年10月), 17. doi:10.1145/3597638.3608402;PAGE:STRING:ARTICLE/CHAPTER TomarSuramya. 2006. 使用FFmpeg转换视频格式。《*Linux Journal*》(2006年6月)。doi:10.5555/1134782. 1134792 Tess Van Daele、Akhil Iyer、Yuning Zhang、Jalyn C. Derry、Mina Huh 和 Amy Pavel。2024。利用分层视频摘要提升短视频的无障碍性。计算系统中的人为因素会议论文集1 (2024年2月), 17. doi:10.1145/3613904.3642839 Agnieszka Walczak 和 Louise Fryer。2018。不同类型音频描述的语音呈现方式：衡量用户的临场感。《翻译学研究》26, 1 (2018年1月), 69–83. doi:10.1080/0907676X.2017.1298634 Yujia Wang 和 Wei Liang。2021。面向无障碍视频的自动音频描述生成。计算系统中的人为因素会议论文集(2021年5月)。doi:10.1145/3411764.3445347/SUPPL{_}FILE/3411764.3445347{_}VIDEOPREVIEW.MP4 Xiaojun Ye、Junhao Chen、Xiang Li、Haidong Xin、Chao Li、Sheng Zhou 和 Jiajun Bu。2024。*MMAD*：多模态电影音频描述。技术报告。11415页。 https://github.com/Daria8976/MMAD.

YouDescribe。 [n. d.]。YouDescribe。访问日期：2025年3月8日。https://www.youdescribe.org/。Beste F Yuksel、Pooyan Fazli、Umang Mathur、Vaishali Bisht、Soo Jung Kim、Joshua Junhee Lee、Seung Jung Jin、Yue-Ting Siu、Joshua A Miele 和 Ilmi Yoon。2020年。通过人在回路机器学习提升视障人士和盲人用户的视频无障碍体验。（2020年）。doi:10.1145/3357236.3395433 Lianmin Zheng、Wei-Lin Chiang、Ying Sheng、Siyuan Zhuang、Zhanghao Wu、Yonghao Zhuang、Zi Lin、Zhuohan Li、Dacheng Li、Eric P Xing、Hao Zhang 和 Joseph E Gonzalez。 [n. d.]。使用 MT-Bench 和聊天机器人竞技场评判 LLM即评委。（[n. d.]）。