

Vidi2.5: 面向视频理解与创作的大型多模态模型

Large Multimodal Models for Video Understanding and Creation

智能编辑团队*, 智能创作 (Intelligent Creation), 字节跳动

美国圣何塞 / 西雅图

<https://bytedance.github.io/vidi-website/>

摘要

视频已成为互联网上的主要通信和创意表达媒介, 带动了高质量视频生产的迫切需求。Vidi 模型不断朝下一代视频创作方向演进, 在多模态时序检索 (TR) 中实现了最先进的性能。Vidi2 是其第二代版本, 通过细粒度时空定位 (STG) 深化视频理解, 并在视频问答 (VQA) 中扩展了能力, 实现真正的多模态推理。给定文本查询, Vidi2 既能定位对应的时间戳, 也能在该时间范围内识别目标对象的边界框。这种端到端的时空定位能力打开了复杂剪辑场景的应用空间——如情节理解、自动镜头切换等。为评测 STG 在真实场景中的表现, 我们推出新基准 VUE-STG, 在视频时长、查询格式、标注质量和评测指标上都超越了现有 STG 数据集。同时, 我们将 VUE-TR 升级为 VUE-TR-V2, 提升了视频长度和查询分布的均衡性。Vidi2 在 VUE-TR-V2 和 VUE-STG 上超越 Gemini 3 Pro Preview 和 GPT-5 等领先专有系统, 在视频问答上与相近规模的开源模型相当。最新的 **Vidi2.5** 采用强化学习训练, 大幅增强了 STG 能力, TR 和视频问答性能也优于前代。本次更新还推出 **Vidi2.5-Think** 用于复杂情节推理。我们提出新基准 VUE-PLOT 来评测情节理解, 包含「人物」和「推理」两个赛道。**Vidi2.5-Think** 在细粒度人物识别上超越 Gemini 3 Pro Preview, 在复杂情节推理上相当。此外, 我们在一个具有挑战性的真实应用——视频剪辑规划上展示了 Vidi2.5 的实力。后训练变体 **Vidi-Edit** 生成结构化剪辑计划, 指定叙事结构、音频属性和视觉剪辑意图, 证明了模型在复杂实际任务中的价值。

*完整贡献者名单见第 7 节。

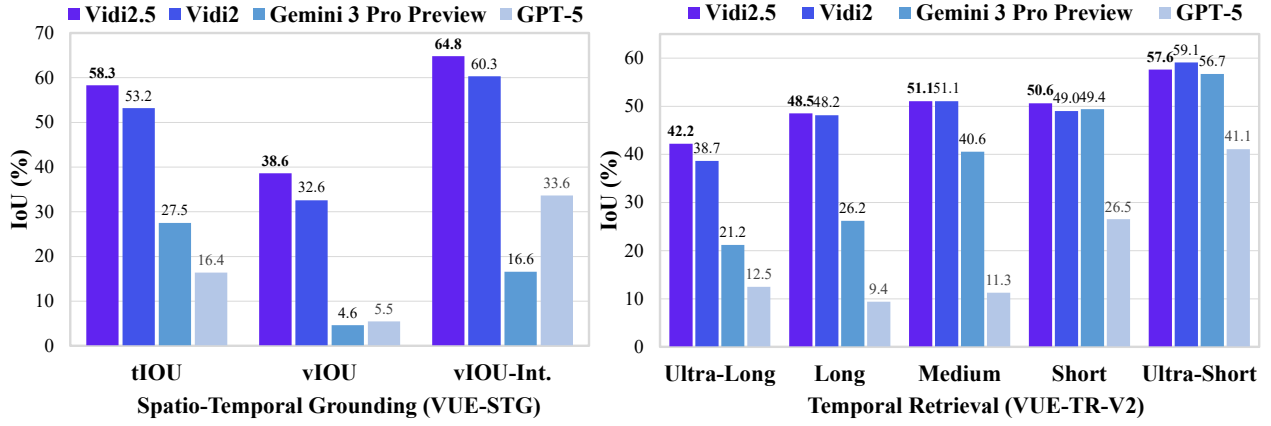


图 1: 所提基准上的时空定位和时序检索结果。

1 引言

视频作为多模态媒介，已是在线通信和信息共享的主要渠道。但对多数用户来说，尤其在移动设备上上进行修剪和编辑时，生成高质量视频仍然困难。**Vidi** [17] 首版本在文本、视觉和音频模态中展示强大的时序理解，显示了自动视频修剪和理解的潜力。为构建能处理复杂编辑的下一代视频创作系统，Vidi 不断演进，增强多模态感知和推理。

第二代版本——**Vidi2** [18]，首次实现了端到端时空定位（STG）——既识别文本查询对应的时间段，也定位这些帧内相关对象的空间位置。现有学术模型 [5, 6, 9, 10, 13, 20] 和最先进的工业系统（如 Gemini 3 Pro Preview [1, 3] 和 GPT-5 [11]）都不能用统一文本格式生成这样的细粒度结果（见图 1）。如图 2 所示，查询「a man standing up from a kneeling position（一个人从跪姿站起来）」是个困难情景——黑暗中有多个个人物。Vidi2 准确定位了时间范围，用精确边界框在端到端框架中将目标人物与他人区分。这要求深入理解所述动作的视觉和时序动态。这种细粒度时空感知展现了 Vidi2 在高级编辑工作流中的潜力——如情节级理解、自动多视角切换、构图感知重帧等。

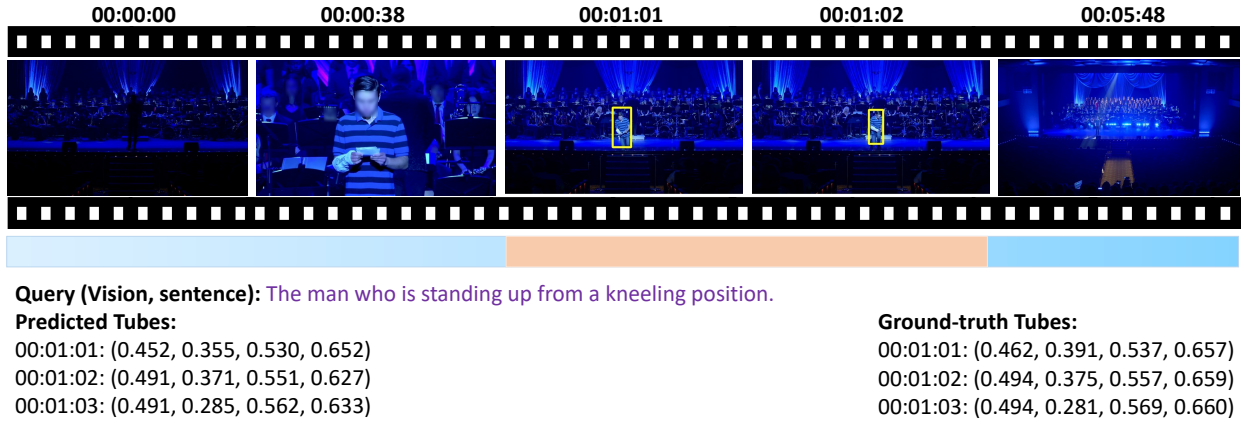


图 2: 时空定位查询示例及其对应的时间范围和目标轨迹（带边界框的时间戳，以黄色显示）。边界框采用百分比坐标表示。示例视频总时长为 387 秒（即 06:27），查询已转换为名词式格式。面部区域已模糊处理以保护隐私。

为评测真实场景中的时空定位（STG），我们推出新基准 VUE-STG，视频时长从 10 秒到 30 分钟，远

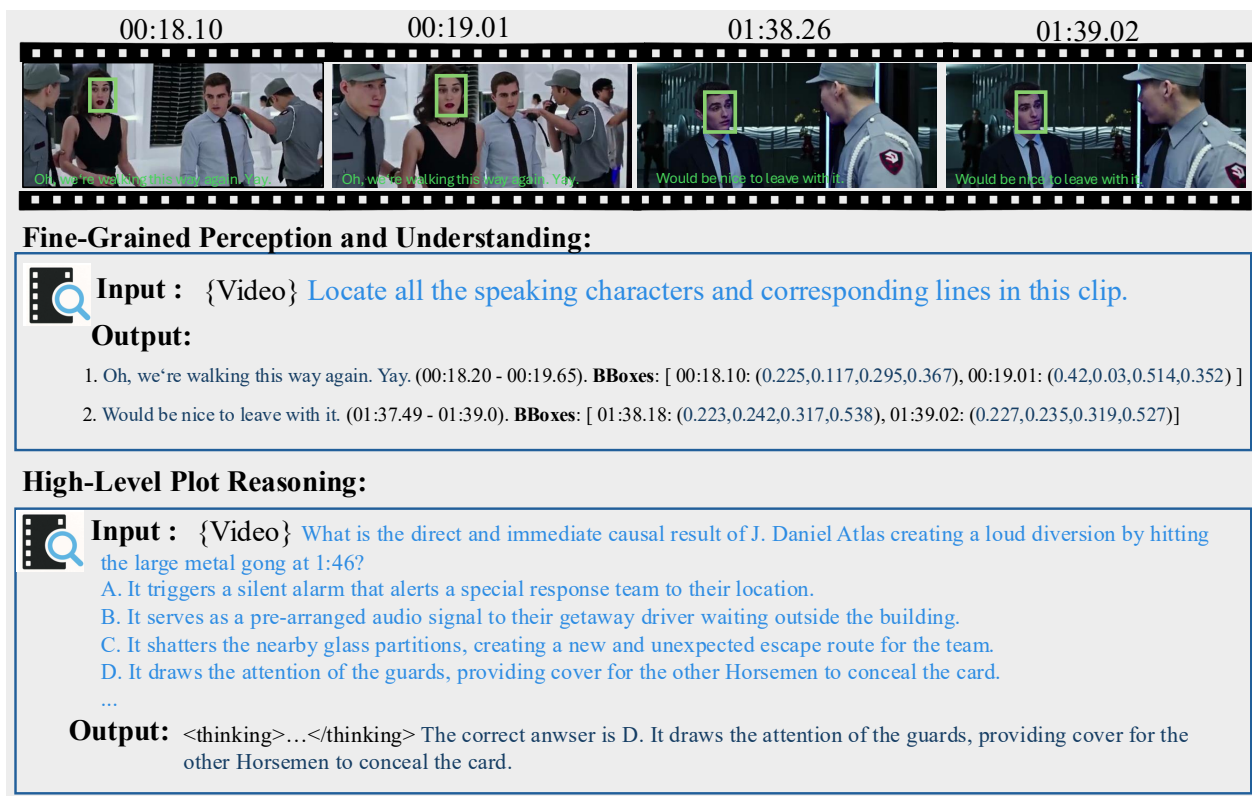


图 3: VUE-PLOT 中的情节理解和推理任务示例。「人物」赛道需要密集的说话者定位和语音识别,而「推理」赛道通过多选视觉问答来评测叙事理解、人物关系和编辑技术。说话者的脸部和说话内容在视频中展示以便更好地可视化。

超现有 STG 数据集 [2, 14, 23, 24]。VUE-STG 包含更长的视频和轨迹标注、更复杂的多模态查询（常含音频线索辅助定位）。所有时间戳和边界框都精确手工标注,确保时空对齐可靠。同时,我们升级 VUE-TR 为 VUE-TR-V2,增加了人工编写的自由形式查询,视频长度分布更加均衡（长视频和超长视频比例提升）,使任务更贴近实际且有挑战。

Vidi2 在 VUE-STG 上超越了 Gemini 3 Pro Preview 和 GPT-5,展现大型多模态模型中最先进的时空定位。相比前代,Vidi2 在时序检索中也大幅进步,尤其在长视频上。在 VUE-TR-V2 上,它在「中等」到「超长」视频类别上超越 Gemini 3 Pro Preview*和 GPT-5,印证了其时序推理的领先性。此外,Vidi2 拓展到通用视频问答,迈向具有强大时空感知的全面视频理解基础模型。

最新的 Vidi2.5 在第 1、3.3、4.4、5 节引入重大更新,介绍新模型、基准和应用。补充结果见第 2、4 节。基础 Vidi2.5 模型采用多模态强化学习 (RL) 训练和可验证奖励,提升了 STG、TR 和视频问答的性能。为在复杂场景（如情节理解）中增强多模态推理,我们推出推理模型 Vidi2.5-Think。受 DeepSeek-R1 [7] 和 OpenAI o1 [8] 启发,Vidi2.5-Think 支持推理时缩放以生成更高质量响应。与此不同,Vidi2.5-Think 用可验证奖励和生成式奖励预测的组合优化,特别针对视频中的细粒度细节捕捉和复杂情节推理。

我们推出新基准 VUE-PLOT 全面评测细粒度感知和高级情节理解。VUE-PLOT 包含「人物」和「推理」两个赛道。「人物」赛道基于真实世界视频,要求模型密集定位说话人并识别说话内容,评测细粒度人物

*请注意, Gemini 3 Pro 在时序检索上相比 Gemini-2.5-Pro-0325 有显著改进。

感知。「推理」赛道针对高级情节理解，通过多选视觉问答挑战模型对叙事动态、人物关系和编辑技术的推理。Vidi2.5-Think 在人物定位和语音识别上超越 Gemini 3 Pro，在复杂情节推理上相当。

除了模型和基准外，Vidi2.5 还应用于实际视频创作工作流。我们将视频剪辑规划作为代表应用，通过后训练变体 Vidi-Edit 实现——它将原始视频资产和可选用户意图转换为结构化剪辑计划。计划明确指定叙事结构、旁白内容、音频属性和视觉剪辑意图，作为下游创作的高级规范。这展示了 Vidi2.5 如何支持标准视频理解之外的复杂真实视频编辑工作流。

2 模型概览

Vidi2 基于 Vidi [17] 进行重大升级——在架构、训练数据和任务能力上都有提升，最重要的是加入了时空定位（STG）和视频问答（VQA）。这些改进共同实现了跨文本、视觉和音频模态的细粒度多模态理解。

架构： Vidi2 采用 Vidi [17] 的多模态架构，用于联合处理文本、视觉和音频，在编码器和 LLM 主干（如 Gemma-3 [15]）中加入关键增强。本报告所有结果基于 12B 参数配置。为在不同长度视频上实现稳健性能，我们重设自适应 token 压缩策略，平衡短视频和长视频的表示效率。单个图像视为 1 秒的无声视频，实现了图像和视频输入的统一编码。

训练数据： Vidi2 采用与前代 Vidi 类似的训练管道，但扩展了数据量和多样性。合成数据对覆盖和稳定性至关重要，但增加真实视频比例显著提升了所有视频任务的性能。时序检索监督微调（SFT）数据集在质量和数量上都得到扩展。同时，我们在 SFT 中引入通用问答数据，强化了模型的开放式推理能力，提升了在高光检测、密集描述和视频摘要等下游应用上的多功能性。为了在可验证任务上进行 Vidi2.5 的 RL 训练，我们还收集了额外的人工标注数据。

时空定位： 为支持新增的时空定位（STG）能力，我们在所有训练阶段引入了任务特定数据。我们利用现有图像级空间定位数据集合成大规模时空视频定位对，有效连接了空间和时序对齐。同时，我们策划了大量真实世界视频 STG 标注，这对模型以高精度定位时间范围和边界框至关重要。

视频问答： 前代 Vidi [17] 验证了时序感知多模态对齐训练对时序检索的价值（见 [17] 第 4 节）。Vidi2 进一步验证了这种对齐策略对通用视频问答的推广性。通过在后训练阶段加入视频问答数据，Vidi2 能以高准确度回答视觉和听觉问题。我们观察到，扩展多模态对齐数据持续改善了整体视频问答性能。

推理模型： Vidi2.5 引入推理模型 Vidi2.5-Think，通过推理时缩放增强视频中的情节推理。Vidi2.5-Think 使用基于 VERL [12] 的自定义强化学习框架训练。训练数据包含多样化多模态指令示例，涵盖视频感知、音频理解、叙事理解和摄影编辑技术——既有多选视觉问答，也有开放式推理任务。模型采用可验证奖励和语言模型评审相结合的方式优化，评估推理预测、生成响应及其一致性的质量。

3 评测基准

为全面评估 Vidi2 的性能和泛化能力，我们考虑三类代表性的视频理解任务：时空定位（STG）、时序检索（TR）和视频问答（VQA）。前两类任务采用我们新提出的基准进行评测（VUE-STG 和 VUE-TR-V2），用于评估真实视频剪辑和理解场景中的细粒度时序和时空推理能力。对于视频问答任务，我们采用广泛使用的公开基准，包括 LVBench [19]、LongVideoBench [21] 和 VideoMME [4]。所有视频问答数据集均采用多选题格式，保证多模态推理性能评测的标准化、客观性和可复现性。这些基准构成统一且全面的评测框架，涵盖检索、定位和视频推理。我们进一步引入 VUE-PLOT 与 Vidi2.5 配套，用于真实视频剪辑场景中的复杂剧情理解。

3.1 VUE-STG

我们引入 **VUE-STG**，一个新的 Video Understanding Evaluation（视频理解评测）基准，专门设计用于推进真实场景中的 **Spatio-Temporal Grounding**（时空定位）。VUE-STG 克服了以往学术基准的四个关键不足 [2, 14, 23, 24]：视频时长、查询格式、标注质量和指标设计。我们对所有真值标签进行人工精心标注，相比现有基准实现更准确、更一致的标签质量。

Duration Category	Ultra-short	Short	Medium	Total
	< 1 min	1 – 10 mins	10 – 30 mins	
# Videos	126	294	562	982
# Queries (Tubes)	206	461	933	1,600
# Boxes	1,063	3,741	7,343	12,147
Video Hours	0.82	26.28	177.69	204.79

表 1: 所提 VUE-STG 基准的视频时长分布。该基准涵盖宽泛的视频长度范围，从极短（< 1 分钟）片段到中等（10 – 30 分钟）视频，使得能够在多样化的真实场景中全面评估模型性能。

3.1.1 数据分布

VUE-STG 基于 VUE-TR [17] 构建，采用公开视频，标注数据向开放研究社区提供。如表 1 所示，该基准包含 982 个视频、1,600 个查询，总时长超过 204.79 小时。它支持属性分层以实现细粒度性能分析，具体包括以下几个维度。

- **视频时长。**与短视频为主的现有数据集不同，VUE-STG 的视频长度从 10 秒到 30 分钟，覆盖真实场景的绝大部分时长范围。我们按 VUE-TR [17] 将视频分为三个均衡的桶：极短（< 1 分钟）、短（1 – 10 分钟）和中等（10 – 30 分钟）。
- **查询。**VUE-STG 的每个查询都经过人工仔细验证，清晰指向视频中的目标对象，消除定位中的语义歧义。当描述涉及多个对象时，我们调整措辞明确指向目标对象。例如，面对”一个球员被装上救护车”这样的模糊表述，我们改为”球员被装入的救护车”或”被装入救护车的球员”，具体取决于指代对象。这种标注策略消除语义歧义，确保查询与对象指代的可靠性。

Object Size	Small < 10%	Medium 10% – 30%	Large > 30%
# Queries (Tubes)	746	511	343

表 2: VUE-STG 基准中标注时空管的对象大小分布。对象大小定义为每个时空管内的平均边界框面积，按视频帧面积归一化。

Tube Duration	Micro-Short < 3s	Ultra-Short 3s – 10s	Short 10s – 60s
# Queries (Tubes)	179	1013	408

表 3: VUE-STG 基准中标注时空管的时序时长分布。时空管涵盖从短暂时刻（< 3 秒）到较长活动（10–60 秒）的时序范围，能够对时序定位性能进行细粒度分析。

- **时空管**。每个目标对象按每秒 1 帧的频率标注为时空管，确保长视频中定位的一致性。对象大小统计基于每个时空管的平均边界框面积计算，呈现均衡分布（表 2）。时空管时序时长同样呈现多样性分布（表 3）：大部分跨度为 3-10 秒，也包含短暂事件和长时间活动，支持全面评测。

值得特别说明：VUE-STG 中不是所有时空管都在时序上严格连续。实际场景中，目标对象因镜头切换、遮挡或场景转换而消失。我们刻意保留这种时序分割的时空管，反映真实的视频动态。这使模型能在更逼真的长视频场景中接受评测，其中对象间歇性地进入和离开视野。

3.1.2 评测指标

我们采用 时序定位准确率 和 时空定位准确率 两项指标评估模型性能。时序指标衡量预测时间跨度与真值标注的对齐程度；时空指标评估预测边界框随时间的定位精度。所有结果均报告为 VUE-STG 测试集的样本级平均值。其中，vIoU 为主排名指标，综合捕捉时序和空间对齐准确率。

时空管表示。时空管表示为随时间变化的边界框：

$$B(t) = (x_0(t), y_0(t), x_1(t), y_1(t)), \quad (1)$$

其中 $(x_0(t), y_0(t))$ 和 $(x_1(t), y_1(t))$ 分别为左上和右下坐标。设 $B^{\text{pred}}(t)$ 和 $B^{\text{gt}}(t)$ 分别为预测和真值时空管，定义在时序支撑集 $T_{\text{pred}}, T_{\text{gt}} \subset \mathbb{R}$ 上。时序轴通过均匀采样离散化，采样率统一为每秒 1 帧。

边界框交并比 (bIoU)。我们将下式定义为两个边界框 B_1 和 B_2 之间的空间重叠指标

$$\text{bIoU}(B_1, B_2) = \frac{\text{Area}(B_1 \cap B_2)}{\text{Area}(B_1 \cup B_2)} \quad (2)$$

时序指标： 时序定位指标评估预测时间区间与真值的对齐程度，包括 tIoU、tP 和 tR。设 $T_{\cap} = T_{\text{pred}} \cap T_{\text{gt}}$ 和 $T_{\cup} = T_{\text{pred}} \cup T_{\text{gt}}$ 。时序精确率（tP）衡量预测时间跨度与真值的重叠程度，召回率（tR）衡量真值被预测覆盖的程度：

$$\text{tP} = \frac{|T_{\cap}|}{|T_{\text{pred}}|}, \quad \text{tR} = \frac{|T_{\cap}|}{|T_{\text{gt}}|}.$$

时序交并比（tIoU）以交并比的形式衡量预测和真值时间区间的对齐程度：

$$\text{tIoU} = \frac{|T_{\cap}|}{|T_{\cup}|}.$$

时空指标： 时空指标评估模型在整个视频中定位目标对象的能力，包括 vIoU、vIoU-Int.、vP 和 vR。

帧级交并比。 对于任意 $t \in T_{\text{pred}} \cup T_{\text{gt}}$ ，我们定义帧级交并比为

$$\text{IoU}_t = \begin{cases} \text{bIoU}(B^{\text{pred}}(t), B^{\text{gt}}(t)), & t \in T_{\cap}, \\ 0, & \text{其他}, \end{cases}$$

其中 $T_{\cap} = T_{\text{pred}} \cap T_{\text{gt}}$ 是时序交集。

时序交集上的累积交并比。 在 $T_{\cap}$ 外， $\text{IoU}_t = 0$ ，故累积交并比简化为在 $T_{\cap}$ 上的求和：

$$S = \sum_{t \in T_{\cap}} \text{IoU}_t.$$

时空精确率和召回率（vP 和 vR）。 分别衡量预测和真值时间跨度上的平均帧级交并比：

$$\text{vP} = \frac{S}{|T_{\text{pred}}|}, \quad \text{vR} = \frac{S}{|T_{\text{gt}}|}.$$

时序并集上的空间交并比（vIoU）。 对时序并集 $T_{\cup} = T_{\text{pred}} \cup T_{\text{gt}}$ 上的帧级交并比求平均，评估任一时空管存在时间范围内的时空准确率：

$$\text{vIoU} = \frac{S}{|T_{\cup}|}.$$

时序交集上的空间交并比（vIoU-Int.）。 为完整起见，我们还计算两个时空管都定义的时间区间上的平均交并比：

$$\text{vIoU-Int.} = \frac{S}{|T_{\cap}|}.$$

3.2 VUE-TR-V2

基于 VUE-TR [17] 的框架，我们推出改进版 VUE-TR-V2，提供更均衡的视频长度分布和更优质的查询标注。

数据分布： 如表 4 所示，VUE-TR-V2 保持与 VUE-TR 相当的视频数量，但总时长从 107.87 小时增至 311.11 小时，翻倍扩展。模态和格式分布（图 5）与原始基准保持一致。VUE-TR-V2 增加了长视频和超长视频的比例，形成更均衡的长度分布，覆盖电影混编、长视频等多样化真实场景，打破短片段主导的局面。

评测指标： 沿用 VUE-TR [17] 基准，评测指标采用时序维度精确率、召回率和交并比（IoU）的 AUC 分数，分别记为第 4 节的 $\bar{P}$ 、 $\bar{R}$ 和 IoU 。其中， IoU 为主排名指标。

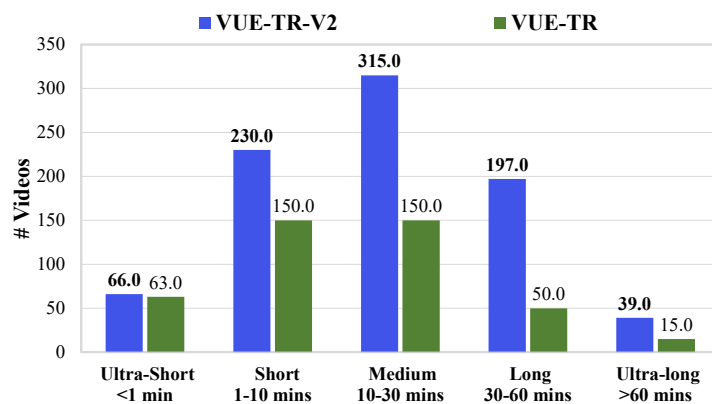


图 4: VUE-TR-V2 和 VUE-TR 基准之间的视频分布对比。

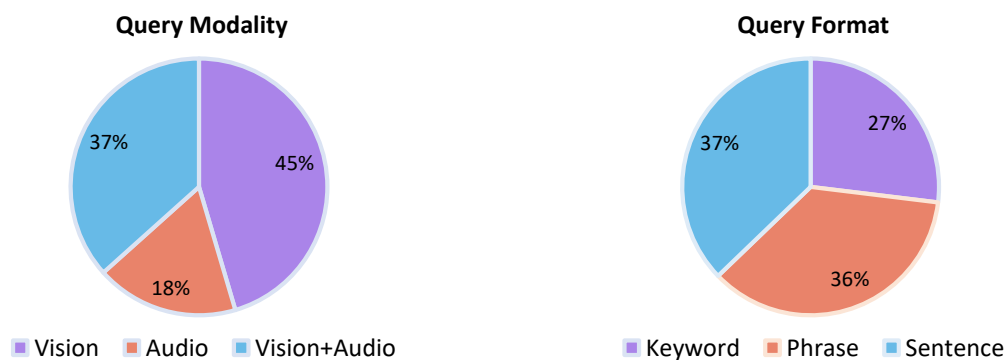


图 5: VUE-TR-V2 基准中查询模态和格式的分布。

Duration Category	Ultra-short	Short	Medium	Long	Ultra-long	Total
	< 1 min	1 – 10 mins	10 – 30 mins	30 – 60 mins	> 60 mins	
# Videos	66	230	315	197	39	847
# Queries	134	390	430	426	220	1,600
Video Hours	0.74	19.85	97.08	145.95	47.50	310.72

表 4: 所提 VUE-TR-V2 评测基准中视频的时长分布。数据集涵盖宽泛的视频长度范围，从极短片段 (< 1 分钟) 到超长视频 (> 1 小时)，能够在多样化的真实场景中全面评估时序检索模型。VUE-TR-V2 的总视频时长 (310.72 小时) 是之前 VUE-TR 基准视频时长 (107.87 小时) 的 2.8 倍以上。

Duration	Character			Reasoning	
	# Videos	# Segments	# Bboxes	# Videos	# QAs
<90s	34	547	1276	7	65
90-150s	248	5630	13422	56	521
150-210s	235	6517	15944	58	488
>210s	29	860	2441	16	140
Total	546	13554	33083	137	1214

表 5: VUE-PLOT 基准的数据分布统计。该基准涵盖了从短片段 (< 90s) 到较长序列 (> 210s) 的多样化视频时长范围，「人物」赛道的总视频时长为 22.45 小时，「推理」赛道为 6.11 小时。这种全面覆盖使得能够在各种时序尺度上进行稳健的模型评测。「人物」赛道提供细粒度的定位标注 (段落和边界框)，而「推理」赛道通过多个领域的复杂推理任务挑战模型。

3.3 VUE-PLOT

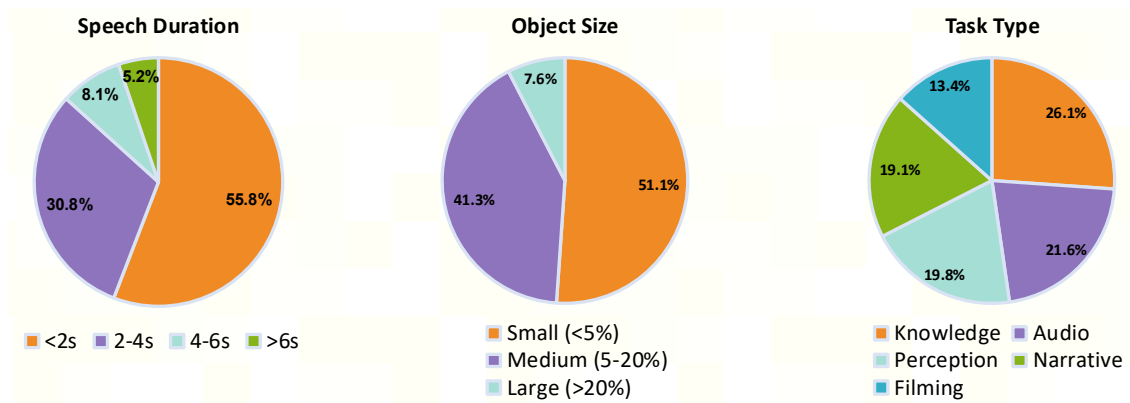


图 6: VUE-PLOT 基准 *Character* (角色) 赛道的言语时长分布 (左) 和边界框面积分布 (中)，以及 *Reasoning* (推理) 赛道的任务类型分布 (右)。

VUE-PLOT 的 *Character* (角色) 赛道是细粒度视频理解的基础, 包含 546 个视频, 总时长 22.45 小时。视频分为四个时长区间确保时序多样性 (表 5): < 90 秒、90-150 秒、150-210 秒和 > 210 秒, 中间区间分布均衡。该基准进行密集标注, 包含 13,554 个言语片段和 33,083 个边界框。片段时长主要集中在短长度, 需要精确时序定位 (图 6 左)。边界框涵盖多种尺度: 小 51.1% ($< 5\%$ 面积)、中等 41.3% (5%-20% 面积)、大 7.6% ($> 20\%$ 面积) (图 6 中), 对模型尺度鲁棒性构成挑战。

Reasoning (推理) 赛道为高级推理场景设计, 从 137 个视频 (总时长 6.11 小时) 的多样性集派生 1,214 个问答对。问题分布于五种任务类型 (图 6 右): *Narrative and Structural Understanding* (叙事和结构理解)、*Perception and Understanding* (感知和理解)、*Professional Filming and Editing Techniques* (专业拍摄和剪辑技术)、*Social Cognition and Knowledge Integration* (社交认知和知识整合)、*Speech, Audio, and Sound Effect Reasoning* (言语、音频和音效推理)。这种多样性覆盖多个认知维度。文本复杂度高: 平均问题 24.8 字、平均选项 24.5 字, 需要强大的多模态推理能力。

评测指标: 我们采用不同指标评估模型在剧情理解和推理上的性能。

时序定位交并比 (tIoU) 用于评估预测和真值转写文本片段间的时序对齐准确率。基于时序交并比 (IoU), 每个包含相关说话角色的真值片段与至多一个预测片段匹配。时序定位分数为所有匹配片段对的平均交并比:

$$\text{tIoU} = \frac{1}{N_{\text{match}}} \sum_{i=1}^{N_{\text{match}}} \text{IoU}(s_i^{\text{gt}}, s_i^{\text{pred}}).$$

词错误率 (WER) 衡量在正确时序定位下的语音转写质量。连接所有匹配的真值片段和对应预测文本, 计算其间的词错误率:

$$\text{WER} = \frac{S + D + I}{N},$$

其中 S 、 D 、 I 分别为替换、删除、插入次数, N 为真值转写文本字数。

边界框交并比。对于包含空间标注的匹配转写片段, 使用边界框交并比评估视觉定位准确率。真值和预测框基于最小时间戳差异进行时序对齐 (设置为 20 毫秒), 为每对匹配框计算交并比。最终分数为所有匹配框的平均边界框交并比:

$$\text{sIoU} = \frac{1}{N_{\text{box}}} \sum_{j=1}^{N_{\text{box}}} \text{IoU}(b_j^{\text{gt}}, b_j^{\text{pred}}).$$

多选视频问答准确率。在多选设置中评估高级剧情理解, 衡量模型的答题准确率。模型从固定选项集中选择一个答案, 准确率定义为预测答案与真值完全一致的问题所占比例:

$$\text{Acc}_{\text{MC}} = \frac{1}{N_{\text{q}}} \sum_{i=1}^{N_{\text{q}}} \mathbb{I}(a_i^{\text{pred}} = a_i^{\text{gt}}),$$

其中 N_{q} 为多选问题总数, a_i^{pred} 和 a_i^{gt} 分别为第 i 题的预测和真值答案, $\mathbb{I}(\cdot)$ 为指示函数。

4 实验

为了验证 Vidi2 的有效性, 我们在三个代表性任务上评测了模型: 时序检索、时空定位和视频问答。前两个任务在自建基准 VUE-TR-V2 和 VUE-STG 上测试, 后者采用公开的视频问答数据集 [19, 21, 4] 评估通用多模态理解能力。VUE-PLOT 用于评测 Vidi2 及竞争模型的复杂情节理解能力。

4.1 时空定位

我们在 VUE-STG 基准上对比 Vidi2 与领先的大型多模态模型 (LMM), 包括 Gemini 3、GPT-5 和 Qwen3-VL。系统预测完整的时空管, 因此我们从时间和空间两个维度分析其表现。指标定义见第 3.1.2 节。

Vidi2 在所有时间和时空指标上均达到最优性能, 全面超越 Gemini 3、GPT-5 和 Qwen3-VL。这表明该模型不仅时序定位更准确, 空间定位在整个视频序列上也显著更优。

4.1.1 LMM 推理设置

为确保公平比较, 我们设计了任务对齐的提示词和标准化的输出格式, 使所有模型都在一致的设置下执行时空定位。Vidi2 将时空管输出为离散、归一化的边界框序列, 每个时间戳一帧; 其他 LMM 则根据各自 API 以不同形式生成时间预测。下面详述各模型的输入、提示词和输出配置, 以及如何统一转换预测为标准化表示以实现一致评测。

Gemini 3 输入/输出格式: 我们评测 Gemini 3 Pro Preview 模型, 该模型通过 URL 接受视频输入。设计的提示词指示模型输出包含时间戳和边界框的 JSON 数组:

```
Answer with timestamps and bounding boxes and do not output explanation.
Output only a JSON array.
Timestamp format: MM:SS with zero-padding (00-59 for MM and SS).
Bounding box format: [x0, y0, x1, y1] with values in [0, 1000] normalized to the video frame.
```

Example:

```
[{"timestamp": "00:30", "box_2d": [100, 200, 300, 400]},
 {"timestamp": "05:00", "box_2d": [150, 250, 350, 450]}
```

```
What are all the timestamps and positions corresponding to the text query:
"yellow school bus which parked on street near the tree with yellow leaves"?
```

提示词遵循 Gemini API 官方指南: MM:SS 时间戳格式基于视频理解规范[†], [0, 1000] 坐标系则按图像理解文档定义[‡]。

GPT-5 输入/输出格式: GPT-5 API 不支持直接视频输入, 改为支持每次最多 120 帧的图像序列。采样规则如下: 1) 视频短于两分钟时, 以 1 FPS 提取帧, 与 Vidi2 输入配置一致; 2) 长视频统一二次采样以满足 120 帧限制。GPT-5 提示词如下:

```
Given the frames of video, please find all the objects corresponding to the text query:
"yellow school bus which parked on street near the tree with yellow leaves".
```

```
Output only a JSON array.
```

[†]<https://ai.google.dev/gemini-api/docs/video-understanding>

[‡]<https://ai.google.dev/gemini-api/docs/image-understanding>

Frame index format: 0-based integer.

Bounding box format: [x0, y0, x1, y1] with normalized value in 0.000 ~ 1.000

Example:

```
[{"frame": 3, "box": [0.051, 0.252, 0.323, 0.954]},  
{"frame": 5, "box": [0.372, 0.353, 0.634, 0.955]}].
```

模型输出采样帧的边界框。

Qwen3-VL 输入/输出格式： Qwen3-VL 的测试参考官方博客[§]的时空定位示例和手册[¶]的交错时间戳-图像模式，对 Qwen3-VL-32B-Instruct 进行评测。

按照手册约定，将视频转换为采样帧序列，提示词交错显式时间戳指示符（如 “<0.0 seconds>”）和相应的帧图像。所有时间戳-图像对后，追加文本定位查询：

```
Given the query "yellow school bus which parked on street near the tree with  
yellow leaves", for each frame, detect and localize the visual content described  
by the given textual query in JSON format. If the visual content does not exist in  
a frame, skip that frame.
```

```
Output Format: [{"time": 1.0, "bbox_2d": [x_min, y_min, x_max, y_max]},  
{"time": 2.0, "bbox_2d": [x_min, y_min, x_max, y_max]}, ...].
```

Qwen3-VL 输出单位为秒的时间戳，边界框坐标在 [0, 1000] 整数范围内，与 Gemini API 一致。

时间和空间归一化： 四个模型的时间定位信号表示互不兼容。为实现公平比较，将所有预测统一转换为秒：1) Gemini 3 的 MM:SS 时间戳解析转换为秒；2) GPT-5 的帧索引按采样策略映射为秒；3) Qwen3-VL 直接以秒报告，无需转换。转换后，所有时间戳四舍五入到整数秒，得到一致的每秒表示。

空间归一化将所有边界框映射到 [0, 1] 共享坐标系。Gemini 和 Qwen3-VL 的 [0, 1000] 坐标除以 1000 归一化，GPT-5 已生成归一化坐标。最终得到统一的空间表示，用于计算所有空间和时空指标。

4.1.2 性能评测

时间性能： 表 6 汇总了不同视频长度、真值时间时长和对象大小下的时序定位结果。Vidi2 在所有条件下均优势明显，达到最高的时间精确率（73.00）、召回率（59.80）和 IoU（53.19）。差距在中等长度视频上最为显著，Gemini 3 Pro Preview、GPT-5 和 Qwen3-VL-32B 均显著下滑。这表明 Vidi2 即使在长视频或细粒度时间条件下也能稳健地进行时间对齐。

时空性能： 表 7 展示了时空定位的准确性。Vidi2 在所有类别上依然最优，vP 为 44.56、vR 为 36.32、vIoU 为 32.57，在交集指标 vIoU-Int 上更具优势（60.30）。这表明该模型不仅识别准确的时间范围，空间定

[§]<https://qwen.ai/blog?id=qwen3-vl>

[¶]https://github.com/QwenLM/Qwen3-VL/blob/main/cookbooks/video_understanding.ipynb

Major Type	Category	Metric(%)	Vidi2.5	Vidi2	Gemini 3 Pro Prev.	GPT-5	Qwen3-VL -32B
Overall	Overall	tIoU	58.34	53.19	27.50	16.40	25.91
		tP	72.14	73.00	51.91	38.29	45.29
		tR	68.00	59.80	35.26	19.53	39.19
Video Length	Ultra-Short ($< 1m$)	tIoU	73.13	61.43	38.68	64.66	54.33
		tP	85.99	85.04	58.91	78.34	76.54
		tR	79.89	66.80	52.29	75.80	69.22
	Short ($1 - 10m$)	tIoU	65.31	61.48	35.39	19.73	38.97
		tP	79.40	81.83	66.49	43.86	58.66
		tR	74.04	68.36	43.59	24.27	57.27
	Medium ($10 - 30m$)	tIoU	51.63	47.27	21.13	4.10	13.19
		tP	65.48	65.98	43.15	23.86	30.45
		tR	62.38	54.01	27.37	4.77	23.64
Tube Duration	Micro-Short ($< 3s$)	tIoU	53.40	47.24	25.92	16.70	24.44
		tP	59.48	56.49	37.06	30.61	44.74
		tR	68.44	59.78	39.66	20.95	35.75
	Ultra-Short ($3s - 10s$)	tIoU	58.60	54.66	29.51	17.63	25.80
		tP	71.01	73.13	51.77	38.17	46.23
		tR	68.25	60.85	38.03	21.22	40.21
	Short ($10s - 60s$)	tIoU	59.84	52.14	23.20	13.23	26.85
		tP	80.49	79.93	58.71	41.25	43.23
		tR	67.17	57.18	26.43	14.72	38.18
Object Size	Small ($< 10\%$)	tIoU	56.35	51.46	28.35	17.57	26.01
		tP	71.05	70.79	52.00	40.94	46.46
		tR	65.55	57.94	37.16	20.46	38.99
	Medium ($10\% - 30\%$)	tIoU	60.98	54.51	27.23	16.66	26.85
		tP	73.98	73.85	51.58	36.91	45.04
		tR	71.15	61.56	34.41	20.04	40.76
	Large ($> 30\%$)	tIoU	58.72	54.97	26.06	13.49	24.32
		tP	71.78	76.55	52.23	34.77	43.18
		tR	68.62	61.21	32.38	16.76	37.29

表 6: VUE-STG 基准上的时序定位结果。报告时间精确率 (tP)、召回率 (tR) 和交并比 (tIoU)。tR 和 tIoU 在真值管上取平均, tP 在预测管上取平均。结果按视频长度、标注管时长和对象大小进一步分组。

Major Type	Category	Metric(%)	Vidi2.5	Vidi2	Gemini 3 Pro Prev.	GPT-5	Qwen3-VL -32B
Overall	Overall	vIoU	38.64	32.57	4.61	5.47	5.12
		vIoU-Int.	64.84	60.30	16.59	33.64	18.47
		vP	47.26	44.56	8.95	13.01	8.61
		vR	44.71	36.32	5.71	6.50	7.49
Video Length	Ultra-Short ($< 1m$)	vIoU	48.25	36.70	7.77	20.60	11.30
		vIoU-Int.	64.42	57.86	19.44	31.36	20.56
		vP	56.17	50.06	12.44	25.19	16.11
		vR	52.05	39.46	9.96	23.98	13.95
	Short ($1 - 10m$)	vIoU	45.23	39.62	6.36	6.86	8.63
		vIoU-Int.	67.78	62.97	18.45	34.45	21.01
		vP	54.68	52.41	12.36	14.97	12.37
		vR	50.60	43.45	7.61	8.43	12.56
	Medium ($10 - 30m$)	vIoU	33.27	28.18	3.05	1.44	2.02
		vIoU-Int.	63.29	59.37	14.33	34.52	15.08
		vP	41.63	39.46	6.49	8.47	4.75
		vR	40.18	32.10	3.84	1.68	3.55
Tube Duration	Micro-Short ($< 3s$)	vIoU	33.35	27.96	3.93	5.60	4.31
		vIoU-Int.	59.05	57.28	12.98	32.95	16.06
		vP	36.98	33.25	5.85	10.52	8.10
		vR	41.31	34.74	5.20	6.74	5.34
	Ultra-Short ($3s - 10s$)	vIoU	39.22	34.01	4.79	5.74	4.57
		vIoU-Int.	65.73	60.98	16.30	33.40	17.32
		vP	47.10	45.17	8.66	12.90	8.07
		vR	45.32	37.52	6.08	6.94	7.15
	Short ($10s - 60s$)	vIoU	39.54	31.03	4.48	4.76	6.85
		vIoU-Int.	64.99	59.83	18.43	34.27	22.45
		vP	52.19	48.00	11.03	14.14	10.16
		vR	44.67	34.01	5.01	5.29	9.27
Object Size	Small ($< 10\%$)	vIoU	29.08	23.31	2.33	3.66	1.49
		vIoU-Int.	49.86	44.13	7.90	21.38	5.88
		vP	36.20	32.08	4.36	9.06	2.73
		vR	33.25	25.71	2.85	4.18	2.17
	Medium ($10\% - 30\%$)	vIoU	46.08	39.42	5.83	7.34	7.08
		vIoU-Int.	73.99	71.21	20.41	42.19	23.65
		vP	55.33	52.97	10.92	15.93	11.29
		vR	53.37	44.44	7.16	8.69	10.20
	Large ($> 30\%$)	vIoU	48.38	42.50	7.77	6.61	10.09
		vIoU-Int.	82.82	77.22	30.22	48.08	38.02
		vP	64.33	58.88	18.88	33.33	18.88
		vR	61.33	55.55	17.77	31.11	17.77

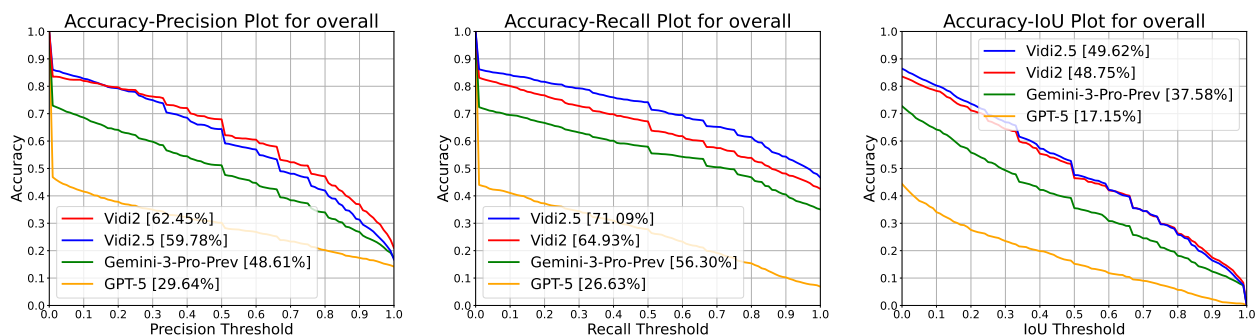


图 7: VUE-TR-V2 基准上时序检索的整体性能曲线。我们报告了不同模型在各种阈值下的准确性。

位也在整个管内保持一致。Vidi2.5 通过强化学习训练进一步提升 vIoU 至 6.07%。相对而言，Gemini 3 Pro Preview、GPT-5 和 Qwen3-VL-32B 在长视频或小对象上严重退化，空间跟踪在扩展时间范围上不稳定。

4.2 时序检索

与 Gemini 3 Pro Preview[3] 和 GPT-5[11] 等三个最先进的专有模型对比。这些模型因性能强、模式支持广泛，成为实际应用中的有力基准。

最新 Gemini 3 Pro Preview 原生支持 '*.mp4' 视频输入，最长超过 2 小时。相比早期 0325 版，新版本稳定遵循时序检索任务，输出 HH:MM:SS 格式的时间戳。示例提示词如下：

Answer with time ranges and do not output explanation. What are all the time ranges corresponding to the text query: "QUERY"?.

最新 GPT-5 API 不支持音频，改为以 1 FPS 提取帧作为输入。参考 Vidi[17] 设计包含上下文示例的提示词，确保输出仅为帧索引范围。示例指令如下：

The input images are frames from a video. Output the frame indexes that correspond to the text query: "QUERY". Only output the index range, for example, 2-4, 6-8.

GPT-5 稳定遵循该提示词，生成干净可解析的帧范围，随后转换为时间间隔进行评测。

表 8 按长度、查询格式和模式子集呈现详细结果。Vidi2 在总体 IoU 上显著领先 Gemini 3 Pro Preview 和 GPT-5 (> 10%)，在大多数子集上的 IoU、精确率和召回率也均出色。更高的总体 IoU 得分确保自动视频编辑和修剪应用中有更好的用户体验。

图 7 展示了不同指标的 ROC 曲线，Vidi2 在全阈值范围内的 IoU、精确率和召回率均超过所有竞争对手。这体现了该模型在细粒度时间理解上的优势，是实际视频编辑任务的核心能力。Vidi2.5 通过强化学习进一步提升总体 IoU。

Category	Metric(%)	Vidi2.5	Vidi2	Gemini 3 Pro Prev.	GPT-5
Overall	IoU	49.62	48.75	37.58	17.15
	$\bar{P}$	59.78	62.45	48.61	29.64
	$\bar{R}$	71.09	64.93	56.30	26.63
Ultra-Short ($< 1m$)	IoU	57.60	59.09	56.72	41.08
	$\bar{P}$	66.81	69.76	68.56	63.84
	$\bar{R}$	80.66	82.48	75.22	48.19
Short ($1 - 10m$)	IoU	50.63	49.01	49.40	26.52
	$\bar{P}$	63.61	65.38	61.83	41.64
	$\bar{R}$	73.31	66.05	74.93	38.91
Medium ($10 - 30m$)	IoU	51.06	51.05	40.56	11.28
	$\bar{P}$	59.94	63.42	50.48	20.51
	$\bar{R}$	73.55	67.91	63.69	21.50
Long ($30 - 60m$)	IoU	48.54	48.15	26.20	9.39
	$\bar{P}$	56.96	60.54	36.73	20.21
	$\bar{R}$	68.90	62.56	39.36	16.05
Ultra-Long ($> 60m$)	IoU	42.22	38.65	21.19	12.49
	$\bar{P}$	53.86	54.55	32.18	21.49
	$\bar{R}$	60.73	50.98	30.11	22.27
Keyword	IoU	48.56	47.73	38.41	23.17
	$\bar{P}$	59.64	63.21	52.43	41.07
	$\bar{R}$	70.63	63.72	56.00	32.59
Phrase	IoU	51.22	50.11	37.84	17.02
	$\bar{P}$	61.66	63.80	48.94	29.18
	$\bar{R}$	73.10	67.07	57.40	25.88
Sentence	IoU	48.83	48.17	36.73	12.92
	$\bar{P}$	58.07	60.59	45.51	22.26
	$\bar{R}$	69.47	63.75	55.46	23.04
Audio	IoU	45.64	46.90	33.27	6.93
	$\bar{P}$	54.26	57.51	39.42	12.16
	$\bar{R}$	71.94	67.09	58.26	23.21
Vision	IoU	52.77	51.04	39.54	20.62
	$\bar{P}$	61.37	63.12	52.16	35.78
	$\bar{R}$	73.13	66.94	54.76	26.85
Vision+Audio	IoU	47.65	46.81	37.26	17.85
	$\bar{P}$	1660.51	64.03	48.66	30.17
	$\bar{R}$	68.13	61.38	57.26	28.04

4.3 通用视频问答

Vidi2 支持视频问答,这是相比前代 Vidi 模型的重大进展。在三个代表性基准 LVBench [19]、LongVideoBench [21] 和 VideoMME [4] 上评测了模型和主流多模态模型。Vidi2 在通用视频问答上落后于领先的专有模型 Gemini-2.5-Pro [3], 但与规模相当的开源模型 Qwen2.5-VL-7B [16] 相比具有竞争力。这表明 Vidi2 有扎实的多模态推理能力, 尽管主要针对检索和定位任务优化。

	LVBench [19]	LongVideoBench [21]	VideoMME [4]
Gemini-2.5-Pro [3]	78.7	84.3	-
Qwen2.5-VL-7B [16]	45.3	54.7	65.1
Vidi2	45.8	57.1	63.5
Vidi2.5	45.2	58.9	63.6

表 9: 在流行基准上的视频问答性能。

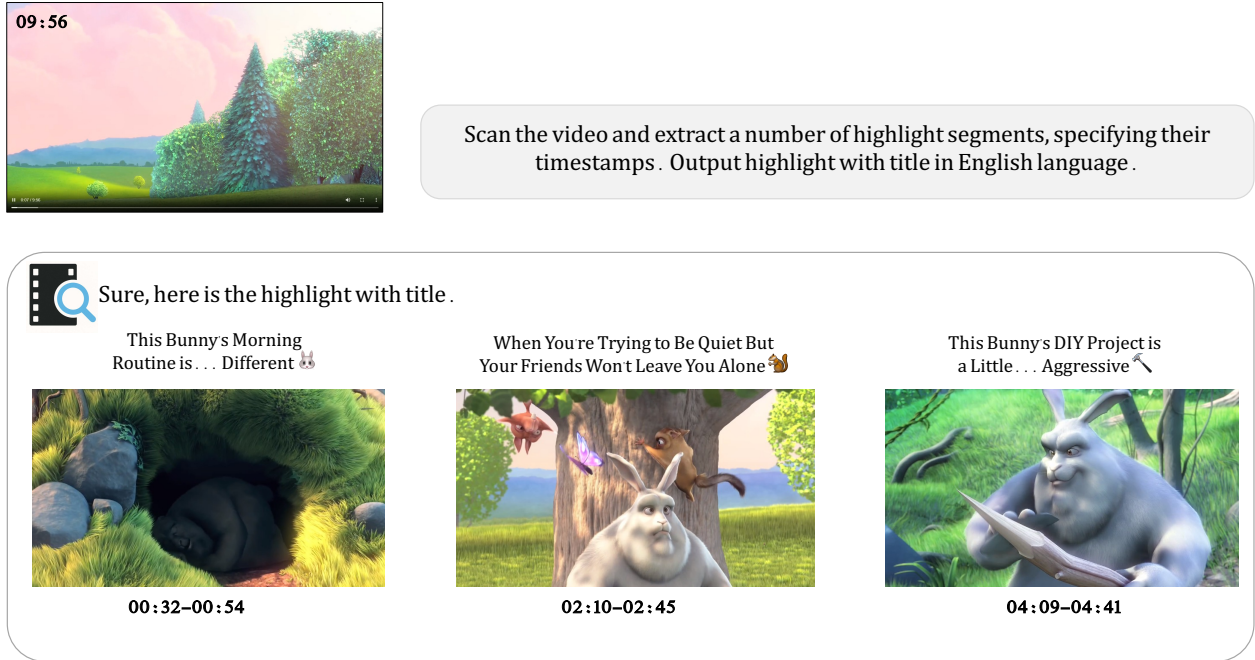


图 8: 精彩片段提取应用示例。

强大的时间建模能力使 Vidi2 能适配精彩片段预测等下游任务。给定长视频, 模型自动生成多个精彩片段及简洁标题, 用户可直接发布短片。如图 8 所示, 生成的标题与视频内容紧密对齐, 展示了时间精准性和语义理解。类似地, 如图 9 所示, 模型也能为长视频生成带总结和标题的章节。

4.4 情节理解与推理

LMM 推理设置。针对 VUE-PLOT 的角色赛道, 设计了任务对齐的提示词和结构化的输出格式, 促使多语言大模型 (MLLM) 在执行主动说话人转写和人脸跟踪时生成稳定可解析的输出。这确保了与 Vidi2.5 的公



Segment this video into logical chapters with timestamps and titles.

Sure, here is the chapter description and title .

Introduction to the Big Buck Bunny	The Quest for the Butterfly	Preparing for the Hunt
00:00–01:24 The film opens with a serene view of a meadow . A large, gentle rabbit emerges from its burrow and is immediately smitten with a beautiful butterfly, setting up the central conflict of the story .	01:24–03:53 Determined to win the butterfly's affection, the Big Buck Bunny embarks on a journey . He encounters a variety of forest creatures, including a squirrel, a mouse, and a porcupine, who all seem to be mocking or hindering his efforts .	03:53–05:21 Feeling frustrated, the Big Buck Bunny decides to take a more aggressive approach . He trains himself, crafting a spear and practicing his hunting skills, preparing to catch the butterfly by force .

图 9: 章节总结应用示例。

平比较。下面详述各比较模型的输入和输出配置。

Gemini 3 和 Qwen3-Omni 输入/输出格式: 评测 Gemini 3 Pro Preview[3] 和 Qwen3-Omni-Thinking[22], 两者均处理视频和音频输入。视频统一以 1 FPS 采样。为减少提示词变化的影响, 两个模型应用相同的结构化提示词, API 级别通过 JSON 模式统一输出格式。Qwen3-Omni 的结构化输出已在最新 vLLM 版本支持^{||}。在此约束下, 两个模型需生成包含精确时间戳和归一化坐标的 JSON 对象:

```

Transcribe the active speaker's speech and track their face
using bounding boxes. Output only a single JSON object and
do not output explanation.
Timestamp format: Total seconds (float).
Bounding box format: [x0, y0, x1, y1] with values in [0, 1] normalized
to the video frame, tightly enclosing the speaker's face.
Sampling strategy: Sample at the first visible frame and every subsequent
integer second.
Example:
{
  "text": "Hello everyone.",
  "start": 62.4,

```

^{||}<https://github.com/vllm-project/vllm>


```

"end": 65.0,
"boxes": [
  {"timestamp": 62.4, "box_2d": [0.400, 0.150, 0.600, 0.350]},
  {"timestamp": 63.0, "box_2d": [0.405, 0.155, 0.605, 0.355]}
]
}

```

赛道	类别	指标 (%)	Models				
			Vidi 2.5	Gemini 3 Pro	GPT 5	Qwen3 VL	Qwen3 Omni
人物	-	时序交并比	71.63	66.04	-	-	50.68
		空间交并比	55.89	13.24	-	-	6.80
		词错误率 ↓	23.20	29.00	-	-	58.12
推理	感知与理解	准确率	66.25	80.42	55.83	35.00	40.83
	语音、音频和音效推理	准确率	74.43	66.03	55.34	45.42	27.10
	社交认知和知识整合	准确率	62.78	62.15	53.94	31.23	23.03
	叙事和结构理解	准确率	55.17	59.48	51.72	25.86	25.00
	专业摄影和剪辑技术	准确率	61.35	50.92	55.21	30.67	24.54
	整体	准确率	64.33	64.58	54.37	33.94	28.01

表 10: VUE-CHAR 基准上的评测结果，报告「人物」（人物）赛道中的时序交并比（tIoU）、空间交并比（sIoU）和 ASR 错误，以及「推理」（推理）赛道中的问题准确率。

性能评测。 表 10 详述了定位精度和推理准确率结果。Vidi2.5-Think 在 VUE-CHAR 的空间 IoU 上明显超越 Gemini 3 Pro Preview[3] (> 42%)，词错误率也最低。针对角色赛道的详细分析：虽然 Gemini 3 Pro Preview 因感知能力强保持微弱的总体准确率领先，但 Vidi2.5-Think 在语音、音频和音效推理（74.43% vs 66.03%）和专业摄影和编辑技术（61.35% vs 50.92%）上表现更优。优越的音视频对齐能力为理解复杂电影叙事和微妙角色互动提供更好的体验。图 10 和图 11 展示了 Vidi2.5-Think 的案例。Vidi2.5-Think 能密集定位活跃说话人并识别其内容（见图 10）。同时提供逐步推理推导正确答案，将复杂情节理解过程分解为一系列可解释的中间推断，体现了对视频内容时间和因果推理的类人理解（见图 11）。

5 应用

除了标准的多模态理解基准，Vidi2.5 还通过后训练支持多种下游应用，这些应用涉及结构化推理、长范围规划和多模态决策。本工作以 ** 视频剪辑规划 ** 为代表性应用案例，在实际场景中展示这些能力。



图 10: Vidi2.5-Think 在视频中密集说话人定位和语音识别的示例案例。真值标注和模型预测分别用绿色和红色在输入视频中标记以便更好地可视化。



Question: In the exchange from 0:17 to 0:20, how does the vocal delivery of Leonard and Willie establish the dysfunctional nature of their partnership?

- A. They speak in perfect unison, indicating they are a highly practiced and professional team with a shared script.
- B. Leonard speaks calmly and authoritatively, while Willie repeatedly interrupts with rapid, corrective statements, showing his insubordination and need for validation.**
- C. Willie whispers his lines to Leonard, showing he is afraid to speak directly to the family and is reliant on Leonard for instructions.
- D. Leonard's voice cracks with uncertainty, while Willie's is steady, suggesting Willie is the secret leader of the pair.
- E. Both men speak in a slow, menacing drawl, using identical intonation to project a unified and intimidating front to the family.
- F. Leonard shouts all of his lines while Willie speaks softly, establishing a classic 'loud cop/quiet cop' dynamic to confuse their targets.

Ground-Truth: <think> Leonard attempts to deliver a monologue ("I worked for Mr. Big."), but Willie immediately cuts in to amend it ("He used to work for us... We worked for Mr. Big."). Willie's delivery is faster and more insistent than Leonard's. This isn't collaboration; **it's a comedic power struggle where the subordinate is constantly trying to co-opt the leader's authority**, establishing the core friction of their relationship seen throughout the clip. </think> <answer>**B**</answer>

VIDI2.5-Think: <think> Leonard attempts to speak for the both of them ('We work for Mr. B'), but Willie immediately and repeatedly corrects him ('I work for Mr. B', 'We work for Mr. B'), highlighting a fundamental disagreement and Willie's inability to let Leonard take the lead.</think> <answer>**B**</answer>

Question: From the moment the thugs enter to the moment Ma Bell is struck (0:01 - 1:28), how many distinct characters are visible on screen inside the house?

- A. Two: Leonard and Willie. While other characters are visible in the background, the scene's primary action and ...
- B. Three: Leonard, Willie, and Ma Bell. A fourth character, Cheryl, is partially visible in a doorway but is obscured ...
- C. Four: Leonard, Willie, Ma Bell, and Cheryl.**
- D. Five: Leonard, Willie, Ma Bell, Cheryl, and a fifth figure, likely another woman, who is briefly visible deep in the ...
- E. Five: Leonard, Willie, Ma Bell, Cheryl, and a young child who can be glimpsed for a single frame peeking from behind ...
- F. The number is intentionally ambiguous, but at least four. While Leonard, Willie, Ma Bell, and Cheryl are clearly visible, shadows and movements in other rooms seen through doorways

Ground-Truth : <think> This question requires careful observation and tracking of all individuals throughout the specified time period. At the start of the scene, Leonard and Willie are present. Ma Bell and Cheryl are also clearly visible in the room and react to the thugs. A careful scan of the entire sequence from 0:01 to 1:28 confirms that **only these four characters are present inside the house**. No other family members or individuals appear in the background or enter the scene. </think> <answer>**C**</answer>

VIDI2.5-Think: <think>Throughout the scene from 0:01 to 1:28, four characters are clearly visible and play significant roles: the two thugs (Leonard and Willie), and two residents of the house (Ma Bell and Cheryl).</think> <answer>**C**</answer>

图 11: Vidi2.5-Think 在视频中情节推理的示例案例。Vidi2.5-Think 提供逐步推理以得出正确答案。正确选项用绿色标记以便更好地可视化。

5.1 视频剪辑规划

视频剪辑规划制定一组相互依赖的高级剪辑决策，确定如何从原始资产生成最终视频。给定原始资产集合（图像、视频）和可选用户提示，目标是生成剪辑计划，包括叙事结构、旁白内容、音频属性和视觉意图。

剪辑计划采用高级文本或结构化表示，统一叙事、音频和视觉设计意图。这种设计使下游系统能协调执行音频合成、视觉效果参数化和视频渲染。

5.2 剪辑计划规范

剪辑计划作为语义蓝图，定义了一组高级决策，共同指定最终视频的结构、呈现和意图。这些决策充当下游执行系统的统一指令（图 12）。

叙事结构 定义视频的分层和时序编排，涉及片段选择、段落提取和排序。确定编辑的时序和语义骨架，决定视觉内容的排列方式，实现预期的叙事呈现。

旁白内容 指定与叙事结构对齐的叙述材料，包含语义内容、表达风格和时序同步。决定叙述如何支持故事情节，为下游语音合成提供指导。

音频属性 定义音频元素：音乐风格、情绪、节奏特征、说话人属性等，为视觉叙事提供补充。构成视频的听觉框架，确保音效设计与视觉进展协调一致。

视觉剪辑意图 用语义术语描述风格和节奏指令，包括过渡设计、强调提示和视觉指令。规定如何应用视觉效果以支持叙事结构和节奏。

综上，这些决策类别定义了剪辑计划的规范空间，捕捉视频的预期结构和呈现，同时对执行细节保持灵活性。

5.3 用于剪辑计划生成的后训练

为支持视频剪辑规划，我们对 Vidi2.5 进行后训练，使其多模态推理能力适配剪辑计划生成的结构化要求。我们称这个后训练配置为 Vidi-Edit。该阶段不改变模型架构，而是调整预训练表示，使其生成符合规范的剪辑计划。

后训练采用额外的任务特定监督，形式为与文本或结构化剪辑计划配对的多模态输入。监督信号强调格式遵守和跨叙事、旁白、音频、视觉维度的语义一致性，使模型生成连贯、规划导向的输出。

5.4 剪辑计划实现

Vidi-Edit 生成的剪辑计划通过结构化执行管道实现，将高级规划意图转化为具体操作。生成的计划（自然语言或结构化文本表示）由翻译阶段处理，将语义规范映射到可执行组件（图 12）。

在执行阶段，剪辑计划的各方面通过专门模块实现：音乐属性从精心策划的音乐库检索；叙述内容由文本转语音（TTS）生成器根据说话者属性合成；视觉意图通过效果检索和参数化模块映射到视觉效果原语。

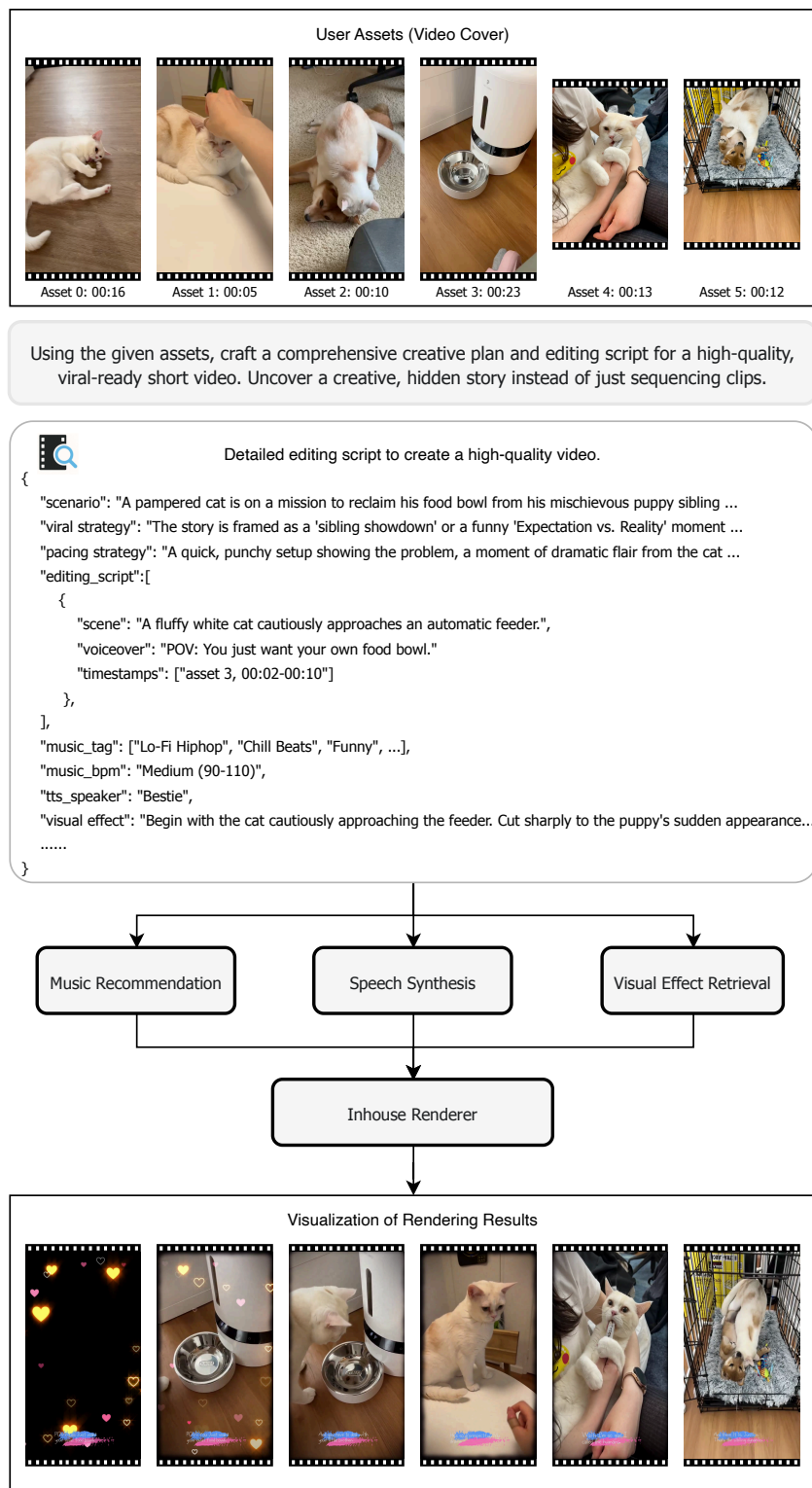


图 12: Vidi-Edit 执行管道的概览。生成的剪辑计划指定了高级语义意图，包括叙事结构、旁白内容、音频属性和视觉剪辑指令，随后转化为具体操作，如音乐推荐、语音合成、视觉效果检索和视频渲染。这种设计明确地将高级规划与执行分离开来，允许模型专注于语义推理，同时下游系统处理实现。

生成的音频资产和效果参数送入渲染系统，集成视频片段、合成音频、文本覆盖和视觉效果，输出最终视频。这种模块化设计将高级决策与实现解耦，让规划模型专注于语义推理，执行组件处理资产选择、合成和渲染。该架构可进一步扩展为编辑代理系统，Vidi-Edit 作为高级决策模块，执行管道将决策转化为具体操作。

5.5 定性示例

我们展示定性示例，检查 Vidi-Edit 生成的剪辑计划，说明高级意图如何在输出中体现。图 13 14 15 16 展示代表性渲染帧，可推断叙事结构、片段安排和旁白进展。

在这些示例中，Vidi-Edit 生成连贯的剪辑计划，将原始片段组织成明确的时序结构。片段按故事情节排列，包括背景设置（例如开场钩子）、内容发展和结尾总结。旁白与时序结构对齐，构成贯穿视频的连贯叙事。

静态帧无法直接显示音频设计和视觉效果规范等属性。项目网站提供额外定性结果、示例视频和交互式界面，可深入探索模型能力。

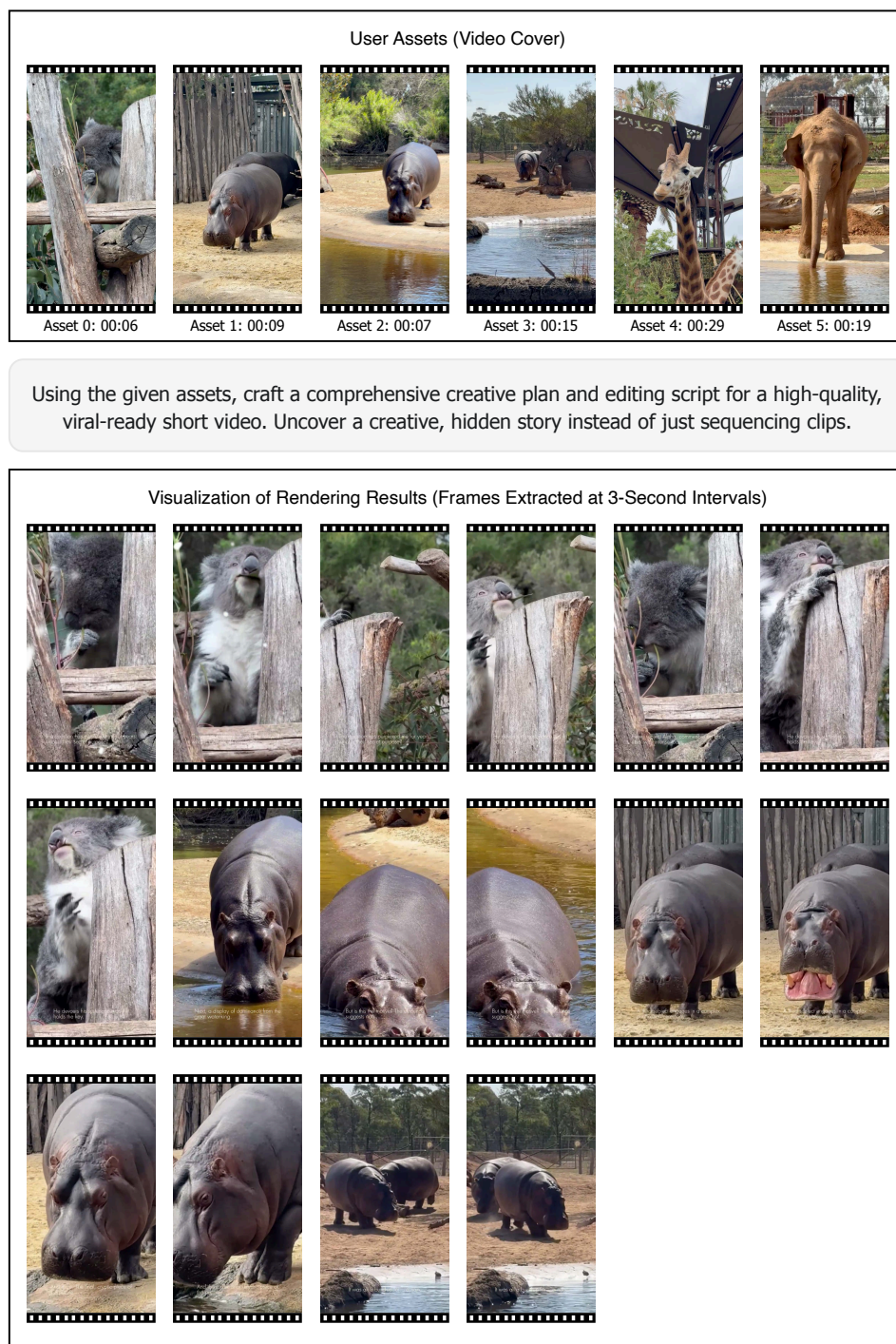
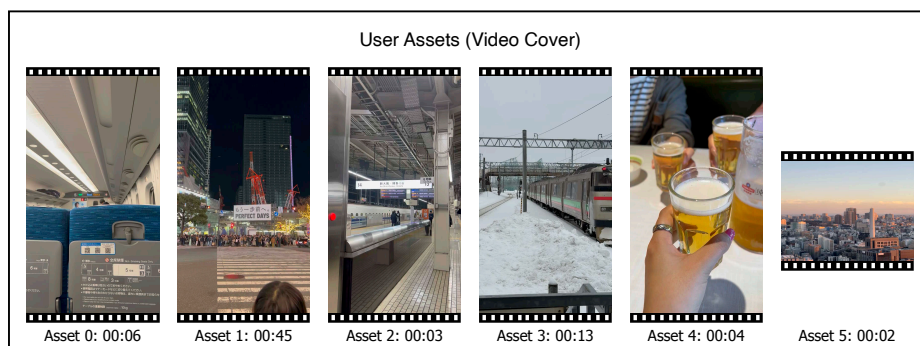


图 13: Vidi-Edit 为单主题场景（动物园）生成的剪辑计划定性示例。所有输入资产均来自同一动物园环境。模型从多个动物片段中识别显著时刻，组织成具隐含时序结构的连贯叙事，而非简单视觉相似度连接。



Using the given assets, craft a comprehensive creative plan and editing script for a high-quality, viral-ready short video. Uncover a creative, hidden story instead of just sequencing clips.



图 14: Vidi-Edit 为单主题城市旅游场景生成的剪辑计划定性示例。给定来自城市游历的异构资产 (交通、街道、夜生活、社交时刻)，模型构建统一故事线，捕捉跨场景的时序进展和体验流。

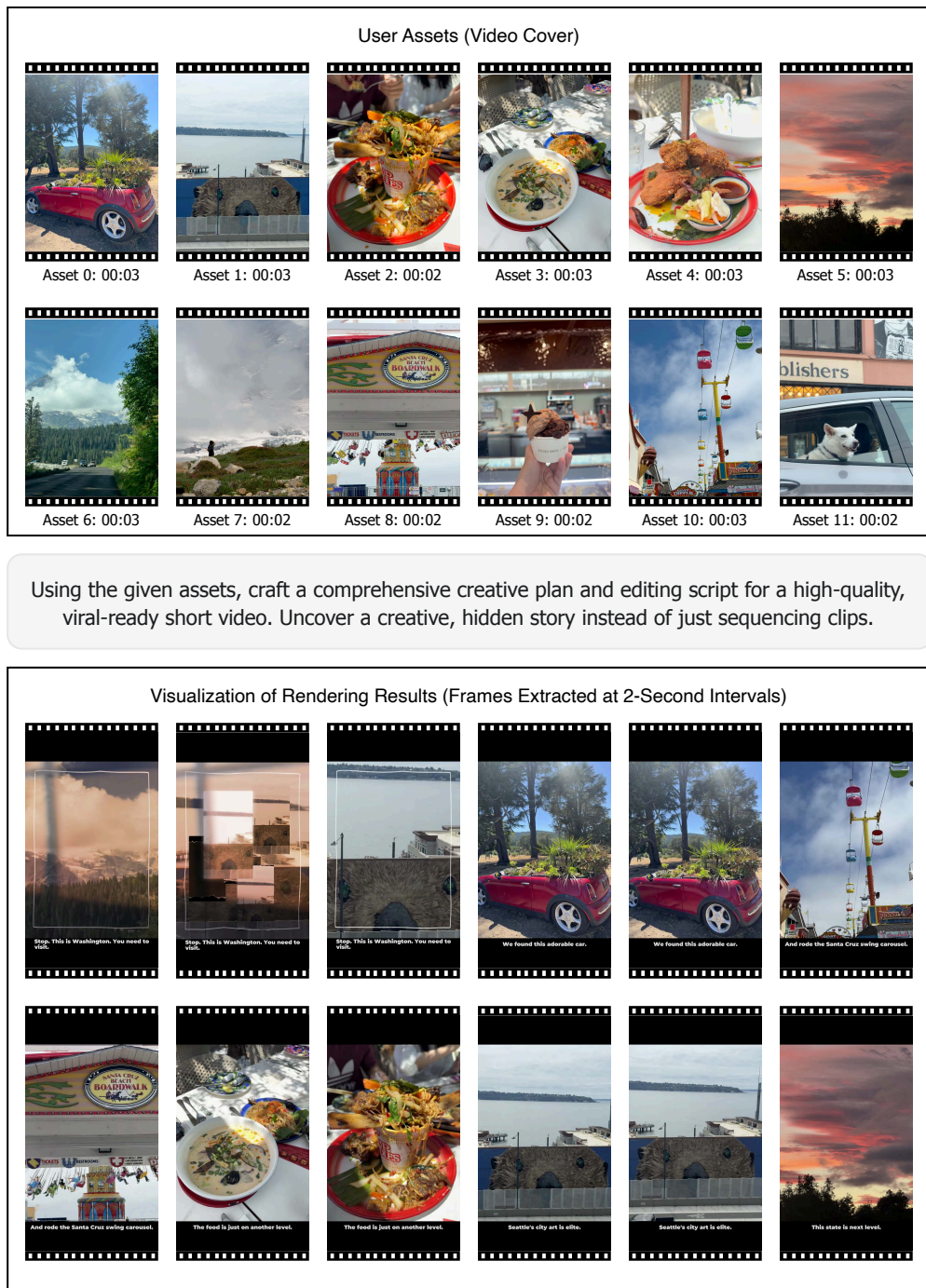


图 15: Vidi-Edit 为西雅图单主题城市市场景生成的剪辑计划定性示例。每个输入仅包含几秒内容，孤立状态下时序上下文有限。尽管如此，模型推断片段间的高级语义关系，组织为连贯叙事，展示对稀疏分割输入的鲁棒性。

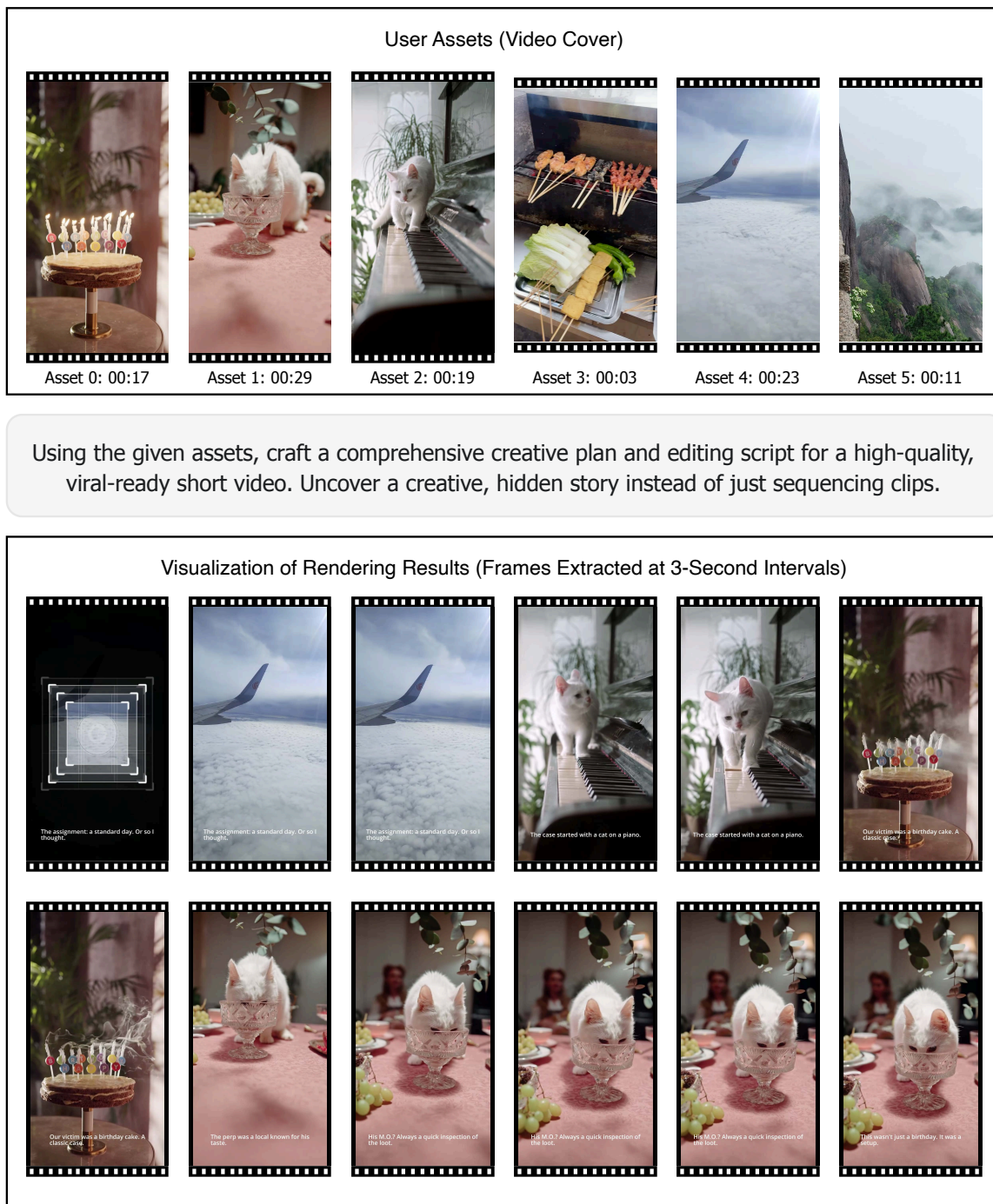


图 16: Vidi-Edit 为混合内容场景生成的剪辑计划定性示例。输入资产语义多样化，缺乏明显的时序或主题顺序。Vidi-Edit 选择性识别和排列片段构建连贯故事线，展示超越低级视觉相似性和朴素连接的高级叙事规划能力。

6 结论

本报告介绍了 Vidi2，它在保持强大视频问答能力的同时实现了前所未有的端到端时空定位。这是下一代视频创作和理解的基础模型。为评测模型在真实视频编辑背景下的表现，我们推出综合基准 VUE-STG，在视频时长、查询格式、标注质量和评测指标上都进步了现有时空定位数据集。同时，我们升级时序检索基准为 VUE-TR-V2，视频长度分布更加均衡（增加长视频和超长视频），查询更贴近用户风格。在公开视频问答数据集上，Vidi2 展现了强大的多模态推理。Vidi2 在 VUE-STG 和 VUE-TR-V2 上超越 Gemini 3 Pro Preview 和 GPT-5，展现最先进的时空理解。在视频问答上与相近规模的领先开源模型（如 Qwen2.5-VL-7B）相当。最后，我们展示了 Vidi2 可能支持的三个前景真实应用。总体而言，这些结果证明了该模型的时空理解和多模态推理能力，为智能、类人的视频编辑和创作系统奠定基础。

最新的 Vidi2.5 采用 RL 训练，在 TR、STG 和视频问答上相比基础 Vidi2 都有进步。我们还推出推理模型 Vidi2.5-Think，支持推理时缩放以处理视频创作中的复杂推理。本报告重点关注情节理解，提出 VUE-PLOT 基准来全面评测细粒度感知和复杂推理。Vidi2.5-Think 在细粒度人物理解上超越 Gemini 3 Pro Preview，在复杂情节推理上相当。

除了基准评测外，我们还展示了 Vidi2.5 可通过后训练调整以支持实际视频创作工作流。以视频剪辑规划为代表，后训练变体 Vidi-Edit 从原始视频资产生成结构化剪辑计划，指定叙事结构、音频属性和视觉剪辑意图。这展现了 Vidi2.5 的时空理解和多模态推理如何转化为真实视频编辑系统的可执行规划表示。

7 贡献者

核心贡献者——研究——基础（按字母顺序）

Chia-Wen Kuo, Dawei Du, Fan Chen, Guang Chen, Haoji Zhang, Sijie Zhu, Xin Gu, Zhenfang Chen.

核心贡献者——研究——应用（按字母顺序）

Feng Gao, Lu Xu, Wen Zhong, Ye Yuan, Yiming Cui.

核心贡献者——基础设施（按字母顺序）

Fanding Lei, Haojun Zhao, Tong Jin.

团队负责人

Longyin Wen*, Xiaohui Shen.

贡献者（按字母顺序）

Chuang Huang, Jin Liu, Jingjing Zhuge, Lili Fang, Lingxi Zhang, Lu Guo, Lusha Li, Qihang Fan, Rachel Deng, Shaobo Fang, Shu Zhang, Stuart Siew, Weiyan Tao, Yicheng He, Zhihua Wu, Zuhua Lin.

参考文献

- [1] <https://deepmind.google/models/gemini/>.
- [2] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee Kenneth Wong. Weakly-supervised spatio-temporally grounding natural sentence in video. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1884–1894, 2019.
- [3] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- [4] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025.
- [5] Xin Gu, Heng Fan, Yan Huang, Tiejian Luo, and Libo Zhang. Context-guided spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18330–18339, 2024.

*Corresponding author.

- [6] Xin Gu, Yaojie Shen, Chenxi Luo, Tiejian Luo, Yan Huang, Yuewei Lin, Heng Fan, and Libo Zhang. Knowing your target: Target-aware transformer makes better spatio-temporal video grounding. *arXiv preprint arXiv:2502.11168*, 2025.
- [7] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv*, 2025.
- [8] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv*, 2024.
- [9] Yang Jin, Zehuan Yuan, Yadong Mu, et al. Embracing consistency: A one-stage approach for spatio-temporal video grounding. *Advances in Neural Information Processing Systems*, 35:29192–29204, 2022.
- [10] Zihang Lin, Chaolei Tan, Jian-Fang Hu, Zhi Jin, Tiancai Ye, and Wei-Shi Zheng. Collaborative static and dynamic vision-language streams for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23100–23109, 2023.
- [11] OpenAI. GPT-5 System Card. Technical report, OpenAI, August 2025. Version published August 7 2025. Available at: <https://openai.com/index/gpt-5-system-card/>.
- [12] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv*, 2024.
- [13] Rui Su, Qian Yu, and Dong Xu. Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1533–1542, 2021.
- [14] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8238–8249, 2021.
- [15] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.
- [16] Qwen Team. Qwen2.5-vl, January 2025.
- [17] Vidi Team, Celong Liu, Chia-Wen Kuo, Dawei Du, Fan Chen, Guang Chen, Jiamin Yuan, Lingxi Zhang, Lu Guo, Lusha Li, et al. Vidi: Large multimodal models for video understanding and editing. *arXiv preprint arXiv:2504.15681*, 2025.
- [18] Vidi Team, Celong Liu, Chia-Wen Kuo, Chuang Huang, Dawei Du, Fan Chen, Guang Chen, Haoji Zhang, Haojun Zhao, Lingxi Zhang, Lu Guo, Lusha Li, Longyin Wen, Qihang Fan, Qingyu Chen, Rachel Deng, Sijie Zhu, Stuart Siew, Tong Jin, Weiyan Tao, Wen Zhong, Xiaohui Shen, Xin Gu, Zhenfang Chen, and Zuhua Lin. Vidi2: Large multimodal models for video understanding and creation. *arXiv preprint arXiv:2511.19529*, 2025.

- [19] Weihang Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967, 2025.
- [20] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Videogrounding-dino: Towards open-vocabulary spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18909–18918, 2024.
- [21] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.
- [22] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025.
- [23] Masataka Yamaguchi, Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Spatio-temporal person retrieval via natural language queries. In *Proceedings of the IEEE international conference on computer vision*, pages 1453–1462, 2017.
- [24] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10668–10677, 2020.