QMAVIS：融合大型多模态模型的长视频音频理解

Zixing Lin[1], JialeWang[1],GeeWah Ng[1],Lee Onn Mak[1],
Chan Zhi Yang Jeriel[2]*, Jun Yang Lee[3]*, Yaohao Li[3]*

[1]Q Team，新加坡内政团队科技局（HTX），新加坡[2]新加坡国立大学[3]
南洋理工大学，新加坡

arXiv:2601.06573v1 [cs.AI] 10 Jan 2026

摘要——**传统上，视频音频理解领域的大型多模态模型（LMM）仅在时长几分钟的较短视频上进行评估。本文介绍了QMAVIS（Q Team 多模态音视频智能研判系统），这是一种通过融合大型多模态模型、大型语言模型和语音识别模型的后期融合构建而成的新型长视频音频理解流程。QMAVIS 填补了长视频分析领域的空白，尤其适用于时长从几分钟到一小时以上的长视频，为情报研判、视频内容分析、具身人工智能等应用开辟了新的可能性。使用 QMAVIS 进行的定量实验表明，在包含音频信息的长视频数据集 VideoMME（含字幕）上，其表现比 VideoLLaMA2 和 InternVL2 等最先进的视频音频大型多模态模型提升了 38.75%。在 PerceptionTest 和 EgoSchema 等其他具有挑战性的视频理解数据集上进行的评估显示，性能最高提升了 2%，表明其具备竞争力。定性实验还表明，QMAVIS 能够在理解整体叙事的同时，提取长视频中不同场景的细微差异。此外，还进行了消融研究，以确定融合流程中各组成部分的影响。**

索引词—**视频音频模型、长视频分析、大型多模态模型、多模态融合、大语言模型。**

一、引言

早期的大型语言模型（LLM），如 GPT-3 [1]和 LLaMA [2]，凭借处理自然语言提示并生成连贯回复的卓越能力，掀起了人工智能创新浪潮。乘着这股浪潮，大型多模态模型（LMM）应运而生，将视觉等文本之外的其他模态融入其中。GPT-4[3]和 LLaVA [4]等模型融合了语言理解与视觉理解，使人工智能能够通过图像和文本更全面地与世界互动。随后，面向视频理解的 LMM 取得了进一步进展。Video-LLaVA [5], 、Video-ChatGPT [6],和 QCaption [7]等模型与流水线通过引入时间推理和连续帧处理，进一步拓展了人工智能的能力，为视频字幕生成和问答开辟了新的可能，使人工智能向人类认知迈进了一步。然而，人类感知中一个至关重要的方面在以往研究中却在很大程度上遭到忽视：此前的研究在很大程度上忽视了一种感知方式：音频。

能够联合理解视频和音频的多模态模型，即视频音频大语言模型，是一个新兴研究领域，对于人工智能系统理解世界变得日益重要。正如人类观看视频或探索周围世界时会同时接收音频和视觉信息并作出回应一样，这些人工智能模型的目标是在通常的文本信息基础上，将视觉事件与听觉线索相结合，从而更全面地呈现场景。具有开创性的代表模型包括 Video-LLaMA [8] 和 PandaGPT [9]。通过融合文本、视觉和音频线索，视频音频大语言模型能够更全面地感知周围世界，并有望在多媒体内容分析（如电影）、公共安全情报研判、自动驾驶汽车和具身人工智能等领域带来广泛应用。

尽管视频音频大语言模型近期取得了进展，**在处理更长的视频（例如几分钟到超过一小时）并全面理解视觉和音频信息方面仍存在不足**，这主要是由于架构和计算限制 [10]。在架构方面，现有的视频大语言模型无论是否将音频作为额外模态，通常都会对视频进行大幅度下采样，仅从完整视频中提取少量（静态）帧和音频片段。这会增加遗漏时间上下文以及对长内容进行片段化理解的可能性，尤其是对于包含多次场景变化的视频而言 [7]。在计算方面，长视频中的数千个视觉标记超出了典型大语言模型的上下文窗口，导致模型无法一次性对其进行全面分析 [10]。这一尚未解决的不足促成了 QMAVIS 的诞生。
QMAVIS 专门针对长视频音频理解，通过对模型输出进行后期融合，以可管理的片段处理视频；与静态帧快照相比，这种方式能够保留时间信息，同时规避上下文窗口限制。

* 研究工作是在 Q Team、HTX 实习期间开展的。

A. 主要贡献

我们提出了 **QMAVIS**（**Q** 团队-**M**多模态**A**音频**V**视频**I**智能**S**理解）流程。该流程采用后期融合策略，通过将输入分块，分别在大语言模型、视频大语言模型和音频文本模型中进行分析，然后融合它们的输出，以形成基于输入提示的连贯响应，从而为长视频保留丰富的多模态上下文。QMAVIS 的设计也是模块化的：未来可以替换为能力更强的模型，也可以换用针对特定应用场景微调的模型。整个 QMAVIS流程均可完全在本地部署环境中运行，并使用开源模型，无需依赖外部 API；对于要求数据隐私的应用场景而言，这是一个重要要求。

通过在标准化的视频理解基准测试上进行实验，我们证明了 QMAVIS 在处理长视频音频信息方面的卓越能力。**值得注意的是，QMAVIS 的表现优于 VideoLlama 等领先的基线视频大语言模型2 [8] 以及 InternVideo2 [11]，在 VideoMME 数据集上的表现提升了 38.75%；该数据集中超过一半的视频为时长从 4 分钟至 1 小时以上的长视频**。在 PerceptionTest 上，该基准包含丰富而复杂的视频分析任务，QMAVIS 的表现较基线最高提升了 2%。

我们还进行了消融研究，以评估不同融合组件（例如响应聚合大语言模型）对整体结果的影响。

II. 相关工作与融合技术

A. 视频大语言模型

Video-LLaVA [5] 和 Video-ChatGPT [6] 等视频大语言模型能够处理复杂的视频内容、理解提示并生成连贯的响应。这使其能够执行视频字幕生成和基于自然语言的视频问答等任务。然而，这些模型在几秒至 1–2 分钟的短视频片段上表现最佳，在处理多场景视频和捕捉时间关系方面则存在困难 [5]。QCaption 采用后期融合方法，整合来自大语言模型和大语言视觉模型的分块输出，以分析更长的视频；与 Video-LLaVA 等模型相比，在更长视频的视频字幕生成和问答任务中，性能最高分别提升了 44.2% 和 48.9% [7]。然而，这些工作仅限于分析视频，无法将音频作为额外模态进行处理。

B. 视频-音频大多模态模型

近期的视频音频大语言模型探索了不同的融合策略，以结合视频、音频和文本上下文，实现全面的场景分析。**VideoLLaMA** [8] 采用中期至后期融合方法，通过学习到的查询变换器，在语言域中对齐视觉和音频特征。具体而言，系统分别使用视频和音频 Q-Former 对视频帧和音频信号进行编码，然后将其投影到大语言模型的嵌入空间中，以生成基于输入提示的最终答案。**Intern-Video 2** [11], 通常被认为是当前最先进的视频音频模型之一，它采用后期融合方法来整合视频和文本嵌入，以供大语言模型生成响应。该论文提出了一种强调时空一致性的新方法——在训练过程中生成视频-音频-语音字幕，以确保视觉和听觉信息在文本中得到联合 grounded，从而实现更好的对齐。类似地，**PandaGPT** [9] 也提出了一种后期融合方法，借助 ImageBind 多模态嵌入模型为图像和音频生成通用嵌入，并将其输入大语言模型。值得注意的是，尽管该方法仅使用图文对进行训练（即未明确使用音频输入），却使模型展现出了涌现的视听理解能力。早期融合技术，例如在输入时合并音频和视觉输入，在近期的视频音频大语言模型中并不常见。相反，中期至后期融合的变体占据主导地位，主要通过共享嵌入空间或通用模态（例如文本）实现后期阶段的融合。无论采用何种融合方法，既有工作都集中于训练模型处理时长约为 10 至 30 秒的短视频片段 [5], [8], [11], ，并未明确关注从几分钟到超过一小时的更长视频。

III. 方法

A. 融合架构

为实现结合音频上下文的高效长视频字幕生成，QMAVIS 在单一流程中融合了三个大型模型的输出：视频字幕生成大多模态模型、音频转录模型和大语言模型（图 1）。

B. 视频分块

首先，将包含音轨的输入视频切分为 N 较短的视频块，每块约 30 秒，C_i，适用于 $i \in \{1, 2, \ldots, N\}$；视频越长，视频块就越多。视频片段中的每一帧都会被纳入其中一个视频块，以确保分析过程中不会遗漏任何信息。随后，将每个视频块中的视觉信息（视频片段），V_i，适用于 $i \in \{1, 2, \ldots, N\}$；以及音频信息（音频文件），$A_i$，适用于 $i \in \{1, 2, \ldots, N\}$，分别提取并编入索引，以供进一步分析。

C. 视频字幕生成与音频转录

在此阶段，会分别处理每个片段 C_i 的视觉 V_i 和音频 A_i 信息。

视频片段 V_i 与输入提示 X_{in} 一同输入视频大语言模型 f_{video}，以生成受提示 $X_{V,\ i}$约束的字幕或回答，如公式 1 所示。每项基准测试实验中的输入提示 X_{in} 均保持一致，但会根据所用数据集在不同实验间有所变化。例如，在定性实验中，X_{in} 只是“详细描述这段视频”；而对于 PerceptionTest，X_{in} 则是“请结合以下问题：{问题}，详细描述这段视频”，因为该数据集提出的输入提示各不相同。QMAVIS 实验及消融实验所用的视频大语言模型统一为 Qwen2-VL-72B [12]。

$$X_{v,i} = f_{\text{video}}(X_{\text{in}}, V_i), \quad \text{for } i \in \{1, 2, \ldots, N\} \tag{1}$$

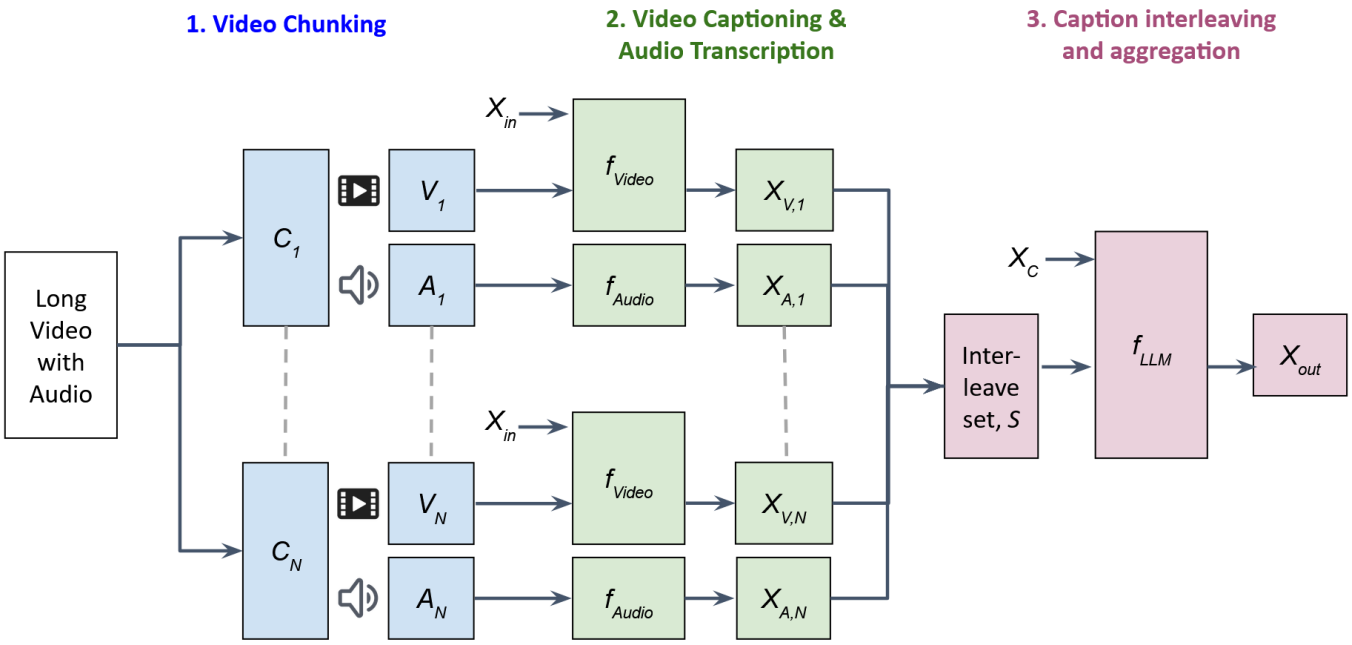

图 1. QMAVISFusion 架构。

使用 Whisper Large v3 [13]，语音翻译／自动语音识别（ASR）模型 f_{audio}，将音频片段 A_i 转录为文本。与视频字幕生成任务不同，音频转录输出 $X_{A,\,i}$不受输入提示约束，而是将音频简单地逐一转换为文本，如公式 2 所示。在此过程中，首先将分块的音频波形转换为频谱图，然后输入一个序列到序列模型。该模型使用涵盖多种任务的大规模多语言语音语料库数据集进行训练，任务包括多说话人分离，从而生成准确的音频片段转录结果，以供进一步分析。

$$X_{A,i} = f_{\text{audio}}(A_i), \quad \text{for } i \in \{1, 2, \ldots, N\} \tag{2}$$

D. 字幕交错与聚合

对每个独立片段完成描述和转录后，我们将其组装成一个集合 $\mathcal{S}$，其中包含每个片段的视频描述 $X_{V,\,i}$，并与其对应的音频转录交错排列，$X_{V,\,i}$。这样，我们确保每个片段都结合视觉和音频信息进行整体分析。公式 3 展示了交错操作。

$$\mathcal{S} = \bigcup_{i=1}^{N} \{X_{V,i}, X_{A,i}\} \tag{3}$$

随后将整个集合 $\mathcal{S}$ 输入大语言模型，以生成一份汇总文本回复，其中参考了从各个片段中提取的信息。该回复由第二个输入提示 X_C引导，要求大语言模型考量各个信息片段并将其汇总为一份完整、连贯的报告，同时关注长视频中所呈现场景的细微差异。

无论数据集如何，X_C 在所有实验中都保持一致。如果集合中的令牌数量超出大语言模型的上下文窗口，则会递归地汇总这些字幕。

E. 消融研究

为研究融合架构中各组件的影响，我们通过系统地禁用选定组件进行了消融实验。

一项实验禁用了字幕聚合大语言模型，并将整个集合 $\mathcal{S}$ 作为文本输出的简单拼接直接呈现为最终答案，不进行进一步总结或处理。该实验旨在研究字幕聚合阶段可能损失细节与获得更简洁连贯的回答之间的权衡。

另一项实验禁用了音频转录模型，以评估音频信息损失的影响。

IV. 结果

A. 评估指标

评估多模态模型需要一种定量指标，用于衡量其跨视频、音频和文本等不同模态进行解读、推理和理解的能力。对于多项选择问答（QA）任务，最广泛使用的评估方法之一是 Top-1准确率。

该指标在 Perception Test mc-VQA、VideoMME 和 EgoSchema 等基准测试上对多模态模型进行零样本评估时表现稳健。这些基准测试考察多模态理解的各个方面，包括一般理解、上下文识别以及特定任务

推理。通过评估模型预测的正确性，Top-1 准确率为模型评估提供了一项有效的比较指标。

在这项评估中，模型需要回答根据真实世界视频和音频数据编制的多项选择题。任务旨在确定模型处理复杂多模态输入并从一组选项中选出正确答案的准确程度。

Top-1准确率按以下公式计算：

$$\text{Top-1 Accuracy} = \left(\frac{\text{Number of Correct Predictions}}{\text{Total Number of Questions}} \right) \times 100 \tag{4}$$

为进一步分析模型性能，可以计算置信区间，以量化报告的准确率在统计上的显著性

$$CI = \hat{p} \pm Z \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \tag{5}$$

其中，$\hat{p}$ 是观察到的Top-1准确率，N 是问题总数，Z 是与给定置信水平相对应的临界值（例如，95%置信水平对应1.96）。

B. 基准数据集

我们深入探讨所用的三个基准数据集：*PerceptionTestmc-VQA*、*VideoMME*和*EgoSchema*。

1) 感知测试：

2) VideoMME：*VideoMME*，由 Chaoyou Fu 等人提出。 [**?**],重点评估多模态模型在基于视频的问答（QA）和视频描述任务中的表现。该基准涵盖六个领域和30个子领域的多样化视频。VideoMME 的一大特点是重点关注时长各异的视频，从短至11秒到长达1小时以上的视频均有涉及。在我们的实验中使用的 VideoMME 数据集中，**52%的视频为时长在4分钟至1小时以上的长视频**，其余48%的视频则较短，时长从11秒到4分钟不等。

VideoMME 使用多模态输入，包括视频、字幕和音频，确保评估全面。高质量的标注可提供精准评估。该评估旨在衡量模型保留短期上下文知识、识别中等时长视频中的趋势和模式，以及从长期序列中推断结论的能力。除了关注视频时长外，VideoMME 还使用多种数据模态，除了视频帧外，还包括字幕和音频。这使得评估模型的多模态推理能力更加全面。

3）*EgoSchema*：EgoSchema 基准由 Karttikeya Mangalam 等人创建[14], ，旨在测试多模态模型理解以自我为中心的视频数据中的长期事件并进行推理的能力。与侧重于短期视觉感知的其他数据集（如感知测试）不同 [15] ，EgoSchema 提供了更长的视频片段，时长中位数约为 100 秒。该数据集涵盖广泛的真实场景，从日常活动到高度特定任务的环境，挑战模型长时间保持上下文意识。该基准的一项关键特点是能够评估不同层次的时间推理能力。测试内容包括识别短期活动（如演奏乐器或做笔记），以及更复杂的中期任务（如整理工作空间或跟踪一段对话的进展）。模型还需要具备长期推理能力，从延长的视频序列中抽象出总体行为模式，并识别随时间逐渐显现的主题和目标。该基准要求模型找出并非紧接发生的活动之间的联系，从而对多模态模型的长期上下文推理能力进行严苛评估。

C. 定性实验

我们还针对时长超过一小时的视频进行了定性实验。这些视频通常包含多个场景，并围绕一个贯穿始终的叙事或故事情节展开。实验旨在比较 QMAVIS 与基线模型在把握每个场景细节的同时，跟踪整体叙事的能力。

D. 基线

我们选取了三种知名的视频音频大语言模型作为基线，以验证 QMAVIS 融合流程。

InternVideo2 [11]是一种先进的视频音频基础模型，采用后期融合方法，确保视觉信息和听觉信息共同以文本为基础，从而实现更好的对齐。**PandaGPT** [9] 是一种多模态指令遵循模型，能够处理图像和音频。它通过为图像和音频输入生成通用嵌入来实现后期融合，再通过投影层将这些嵌入输入大语言模型。**VideoLLaMA2**[8] 是另一种知名的视频音频大语言模型，它采用后期融合，先分别编码视觉和音频特征，再在共享嵌入空间中对齐，然后将其输入大语言模型进行推理和生成。QMAVIS 与每种视频音频大语言模型均在上述三个基准数据集 Video MME、PerceptionTest 和 EgoSchema 上进行了评估。三个数据集均报告了 Top-1准确率。

E. QMAVIS

参见表 I，QMAVIS 在 VideoMME（含字幕）上取得 66.46 分，较表现次佳的模型 VideoLlama2 提高了 **38.75%** 。VideoMME 中有 52% 的视频时长在 4 分钟至 1 小时以上，其中约 15% 是时长超过 15 分钟的长视频。所有视频都配有对应的音轨。这表明，采用融合方法分析长视频具有优势：将长视频拆分为当前视频大语言模型擅长处理的较短视频块，并利用大语言模型的聚合能力生成连贯的最终答复。我们针对不同的视频时长进行了实验，



Video: 6 Days Jungle Survival Camping with NO Food - US Marine Corps Jungle Survival School
Duration: 1:13:01
Link: https://www.youtube.com/watch?v=2OdYv8YKSaI
Prompt: Describe this video in detail.

QMavis

The video depicts a group of men participating in a survival training exercise in a forest. The training is part of a larger role-play war game scenario…
Key Elements of The Video:

- Classroom Instruction: Participants receive instruction on survival techniques, including fire starting and using natural materials… They are taught the Universal Edibility Test (UET) …
- Outdoor Activities: At a demonstration survival site, they practice setting up camp with fire, hammocks, and tarps… They learn to create a torch by wrapping a ball of material on a stick…
- Survival Techniques: Training on how to find and purify water using a grail (a water filtration device)... identifying and eating cicadas, described as tasting like a plant-meat hybrid… Guidance on cleaning and cooking a chicken provided by an instructor...
- Environmental Conditions: Set in a forest with dense vegetation and trees… Exposure to rain, mud, and food scarcity; some participants even sleep in the rain and mud…
- Group Dynamics: Emphasis on teamwork, with members offering tips and advice… participants are seen joking around and playing with cicadas...

Summary: A detailed, immersive training exercise focusing on fire starting, water purification, and foraging for food... Participants work together as a team under challenging conditions, including rain and food scarcity... The exercise culminates with a march through the jungle seeking food—specifically a deadly viper called a habu—and cooking ferns for dinner...

InternVideo2

Starts with a man wearing a hat speaking to the camera, then running down a path in the woods…

- Shows a man holding a snake and another reacting in surprise…
- A group discussion around a jar of water…
- A man with a clear bag containing a frog shown to a group of women…
- Scenes of a couple conversing in a car/house, a man jumping on a vehicle, and various shots of soldiers and civilians…
- Includes military elements: soldiers with backpacks, discussing experiences and engaging around a campfire, setting up tents, and building a fire…

PandaGPT

The video shows a person wearing camouflage clothing and a hat, walking through a forest with their back towards the camera. They appear to be looking at something in the distance. The trees are tall and leafy, and there is a stream running through the area

VideoLlama2

The video starts with a man holding up a jar filled with something while talking to others inside a room. Then, we see soldiers gathered around a fire in the woods, cooking food and enjoying their time together.

Next, we see soldiers sitting around fires at night, presumably sharing stories and bonding over their shared experiences. The scene then shifts to soldiers eating food outside their base, possibly during a break or after completing a mission.

Later on, soldiers are seen sitting around fires again, this time accompanied by music playing in the background. They seem to be enjoying themselves and unwinding after a long day.

Finally, we see soldiers gathered indoors, possibly inside a training facility or barracks. They appear to be engaged in conversation and possibly discussing tactics or planning for upcoming missions.

Overall, the video seems to depict soldiers enjoying their downtime and bonding with each other amidst their rigorous training and demanding schedules.

图2. QMAVIS、InternVideo2、PandaGPT和VideoLLaMA2之间的定性结果

并根据视频时长优化参数。对于短视频和中等长度视频，我们使用时长为 60 秒的视频块，采样率为每秒 1 帧（帧每秒）。对于长视频，则采用 600 秒、0.5 帧每秒，以便在最后一步进行更全面、上下文依赖更少的聚合。我们观察到，分块方法能够

表 I QMAVIS 与其他视频理解模型在多个基准测试中的性能比较。

S/N		方法	所使用的视频大语言模型	音频	视频理解基准		
					VideoMME（含字幕）	感知测试	EgoSchema
1	基线	VideoLLaMA2	VideoLLaMA2	已包含 in VideoLLaMA2	47.9	57.5	63.9
2		InternVideo2（基础版）	InternVideo2	包含在 InternVideo2	41.9	52.16	55.0
3		PandaGPT	ImageBind	包含 in PandaGPT	22.58	31.63	24.0
4	QMAVIS	QMAVIS（完整版）	Qwen2VL	Whisper-LargeV3	**66.46***	57.72	**65.0**
5	消融研究	QMAVIS（消融实验 1 - 无大语言模型）	Qwen2VL	Whisper-LargeV3	63.00*	53.14	62.8
6		QMAVIS（消融实验 2 - 无语音转文本）	Qwen2VL	None	48.18*	**58.37**	-

* 部分数据集（占总数据集的 75%），实验仍在进行中。

实现强大的长视频理解，并保留直接输入整段视频时可能丢失的细节。

同样，在侧重于稳健性和复杂情境的 PerceptionTest 上，不含音频组件的 QMAVIS 比 VideoLLaMA2 提升了**1.51%** 。尽管差异较小，这种方法仍展现出 QMAVIS 在具有挑战性的视频理解任务中超越最先进模型的能力。分块方法让视频大语言模型能够聚焦于视频中的小片段，以较高的采样率分析复杂场景的细节。随后，这些较短的分块描述会传递给融合大语言模型；它在记忆相关任务中发挥重要作用，使模型能够在文本空间中更充分地理解上下文。

我们还观察到，QMAVIS 在 EgoSchema 上也有所提升。该数据集用于衡量模型在现实世界场景以及短期和长期上下文推理中的表现。QMAVIS 的表现比 VideoLLaMA 高**1.72%**，这表明聚合大语言模型能够在各个片段的视频描述之间建立时间关联，同时也体现了视频大语言模型对视频片段所生成的简洁描述的作用。

在这三个基准测试中，与 PerceptionTest 和 EgoSchema 相比，我们观察到 QMAVIS 在 VideoMME 上的性能提升最大。这可能归因于 QMAVIS 在长视频理解方面的优势，相较于处理较短视频的能力而言。尽管 PerceptionTest 和 EgoSchema 的视频长度都超过了视频理解数据集的平均水平，但它们仍主要由时长分别约为 22 秒和 100 秒的视频构成；而 VideoMME 则包含更多时长显著更长的视频，每段从 4 分钟到 1 小时以上不等。视频越长，分块与聚合流程的优势就越突出，因此与其他模型相比，整体结果更优。

F. 消融实验结果

我们还通过禁用流程中的某些组件，对 QMAVIS 进行了消融研究。此举旨在探究字幕生成与聚合阶段可能损失的细节，与回答更加简洁连贯之间的权衡。

对于 VideoMME（含字幕），我们首先尝试使用视频大语言模型进行聚合，而不使用专用大语言模型。我们观察到，与完整版 QMAVIS 相比，性能下降了**3.72%** ，但仍领先于 VideoLLaMA2**38.75%**。这证明了视频大语言模型在结合额外音频信息处理大规模文本上下文方面具有强大的推理能力。尽管其性能不及专用聚合大语言模型，但这为更轻量级的部署提供了一种可靠的替代实现方案：通过减少所使用的模型数量，可以在性能可接受的下降范围内完成部署。我们还尝试移除音频转录模型，这使完整模型的性能显著下降，准确率仅为**48.18%** 。失去对音频模态的支持后，QMAVIS 退化为纯视觉实现，导致其在所有与音频相关的任务上表现不佳。然而，该方法仍以**0.28%**的幅度略微领先于 VideoLLaMA2。

我们也对 Perception- Test 和 EgoSchema 进行了相同的消融研究。对于这两项基准测试，我们观察到，在不使用大语言模型进行聚合时，结果表现出与 VideoMME 类似的趋势。然而，在第二项消融实验中，PerceptionTest 的结果显示，相较于完整版 QMAVIS流程，性能有所提升，准确率提高了**0.65%** 。由于 PerceptionTest 的许多音频任务涉及非语音音频线索，而非纯语音，移除音频转录组件可能减少了对非语音音频信号的潜在误解所造成的不准确性。对于 PerceptionTest，音频转录模型可能会对整个流程产生负面影响。

G. 定性结果

图2比较了 QMAVIS 与基准模型在带有音频的长视频（时长超过一小时）上的回答。这些视频因包含多个各具特色的场景和贯穿始终的叙事而被选中。

InternVideo2：能够对人物进行非常细致的描述——细致到他们的穿着和动作——但无法识别多个场景，也无法传达整体故事情节。

PandaGPT：提供简洁的单帧描述，能够很好地捕捉一个场景，但缺少理解长视频所需的时间和叙事上下文。

VideoLlama2：提供侧重叙事的摘要，涵盖群体互动和连续事件；然而，它无法充分细致地呈现场景级细节。

QMAVIS：擅长提取视频中的关键场景并逐一进行详细描述，随后全面概述事件的整体发展脉络。这种方法使 QMAVIS 能够同时捕捉细粒度细节和更宏观的叙事上下文。

总体而言，QMAVIS 能够有效平衡细粒度细节与对长篇复杂视频叙事结构的理解，因而优于基准模型。

V. 结论

本文介绍了 QMAVIS，这是一种新颖的音频-视频字幕生成与问答流程，采用后期融合方法来有效处理长视频和音频内容（最长可超过一小时），在保留视频中各个场景细节的同时，捕捉整体叙事。这弥补了视频音频大语言模型的不足：这类模型通常只能处理时长约为 10–30 秒的短片段。在 VideoMME、PerceptionTest 和 EgoSchema 等视频理解基准测试中，QMAVIS 的表现优于 VideoLLaMA2 和 InternVideo2 等领先基线模型，分别高出**38.75%**、1.51% 和 1.72%。值得注意的是，QMAVIS 在长视频音频理解方面的能力，从其在 VideoMME 数据集上的显著提升中可见一斑；该数据集一半以上的视频时长为 4 分钟至一小时以上。

未来工作。研究范围有望扩展到音频模型，这类模型不仅能够转录语音，还能理解语气以及背景噪声、音乐和非语言线索（例如拍手声、脚步声）等非语音信号，从而对视频进行更丰富的分析。此外，也可以探索字幕交错阶段的其他可能性：不再按每个片段交错音频和视频信息，而是将交错扩展到多个相邻片段，以期改善对时间关系的理解。还可以调整融合流程中的某些超参数，例如片段长度、片段数量、视频字幕生成中的帧采样率，或聚合大语言模型的上下文长度，以获得计算效率更高的结果。我们也有望开发专门的长视频音频评估数据集，因为大多数现有数据集（包括本文使用的数据集）仍缺乏更丰富的长视频资源，目前视频时长大多只有几分钟。最后，与大多数基于生成式人工智能的流水线一样，输入提示始终有进一步优化的空间，从而生成更精确的答案，或更贴合特定应用场景的答案。

总而言之，我们希望 QMAVIS 能为采用融合方法的稳健音视频大语言多模态模型流水线的进一步研究奠定基础。

致谢

本研究由新加坡内政部（MHA）下属法定机构内政部科技局（HTX）资助。

参考文献

[1] Tom B. Brown、Benjamin Mann、Nick Ryder 等。语言模型是少样本学习器。神经信息处理系统进展，33:1877–1901，2020。[2] Hugo Touvron、Louis Martin、Kevin Stone 等。Llama 2：开放基础模型与微调聊天模型。*arXiv* 预印本 *arXiv:2307 09288*，2023。。[3] OpenAI。GPT-4 技术报告。*arXiv* 预印本 *arXiv:2303.08774*，2023。[4] Haotian Liu、Chunyuan Li、Qingyang Wu 和 Yong Jae Lee。视觉指令微调。*arXiv* 预印本 *arXiv:2304.08485*，2023。[5] Bin Lin、Yang Ye、Bin Zhu、Jiaxi Cui、Munang Ning、Peng Jin 和 Li Yuan。Video-LLaVA：通过先对齐后投影学习统一视觉表征。*arXiv* 预印本 *arXiv:2311.10122*，2023。[6] Muhammad Maaz、Hanoona Rasheed、Salman Khan 和 Fahad Shahbaz Khan。Video-ChatGPT：借助大型视觉语言模型实现细致的视频理解，2023。[7] Jiale Wang、Gee Wah Ng、Lee Onn Mak、Randall Cher、Ng Ding Hei Ryan 和 Davis Wang。QCaption：融合大型多模态模型，实现视频字幕生成与问答。*2024*年第*27*届国际信息融合会议（*FUSION*），第1–8页，2024年7月。[8] Hang Zhang、Xin Li 和 Lidong Bing。Video-LLaMA：用于视频理解的指令微调音视觉语言模型。*2023*年自然语言处理实证方法会议论文集：系统演示，第543–553页，2023。[9] Yixuan Su、Tian Lan、Huayang Li、Jialu Xu、Yan Wang 和 Deng Cai。PandaGPT：一个模型遵循所有指令。载于第*1*届驯服大型语言模型研讨会：交互式助手时代的可控性！（*TLLM 2023*）论文集，第11–23页。计算语言学协会，2023。[10] Yan Shu、Zheng Liu、Peitian Zhang、Minghao Qin、Junjie Zhou、Zhengyang Liang、Tiejun Huang 和 Bo Zhao。Video-XL：用于小时级视频理解的超长视觉语言模型。*arXiv* 预印本 *arXiv:2409.14485*，2024。[11] Yi Wang、Kunchang Li、Xinhao Li、Jiashuo Yu、Yinan He、Chenting Wang、Guo Chen、Baoqi Pei、Ziang Yan、Rongkun Zheng、Jilan Xu、Zun Wang、Yansong Shi、Tianxiang Jiang、Songze Li、Hongjie Zhang、Yifei Huang、Yu Qiao、Yali Wang 和 Limin Wang。InternVideo2：扩展基础模型以实现多模态视频理解。载于欧洲计算机视觉会议（*ECCV*），2024。[12] Shuai Bai、Keqin Chen、Xuejing Liu 等。Qwen2.5-VL 技术报告。*arXiv* 预印本 *arXiv:2502.13923*，2025年2月。[13] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine Mcleavey 和 Ilya Sutskever。通过大规模弱监督实现稳健的语音识别。技术报告，OpenAI，2023。[14] Karttikeya Mangalam、Raiymbek Akshkulakov 和 Jitendra Malik。EgoSchema：用于超长视频语言理解的诊断基准。*arXiv* 预印本 *arXiv:2308.09126*，2023。[15] Viorica Pˇatrˇaucean、Lucas Smaira、Ankush Gupta 等。感知测试：评估人工智能模型中的多模态感知与理解。*arXiv* 预印本 *arXiv:2305.13786*，2023。