E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP183;TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2025)12-3760-22

论文引用格式:Xie J L, Zhang R F and Li G B. 2025. Video question answering with large language models: a survey. Journal of Image and Graphics, 30(12):3760-3781(谢君琳, 张锐斐, 李冠彬. 2025. 大语言模型下的视频问答方法综述. 中国图象图形学报, 30(12):3760-3781)[DOI: 10.11834/jig.240535]

大语言模型下的视频问答方法综述

谢君琳[1],张锐斐[1],李冠彬[2*]

1. 香港中文大学(深圳)理工学院,深圳　518116; 2. 中山大学计算机学院,广州　510006

摘　要:大语言模型在自然语言处理领域取得显著进展,展现出卓越的语言理解和生成能力。然而,尽管这些模型在文本处理方面表现出色,但在应对复杂多模态任务时,尤其在视频问答领域局限性逐渐显现。视频作为一种动态的视觉模态,具有显著的时序依赖性和跨模态信息融合的复杂性,对模型的时序信息处理能力和计算效率提出更高的要求。本文系统回顾基于大语言模型的视频问答模型的研究进展,详细分析非实时视频问答模型与实时视频问答模型的技术特点、优势及其在不同应用场景中的表现。同时,探讨了现有研究中常用的数据集及其评测标准,并总结了当前技术面临的挑战与瓶颈。在此基础上,对未来视频问答模型的发展方向进行前瞻性展望,旨在推动多模态人工智能的进一步发展与应用。

关键词:大语言模型(LLMs);视频问答(Video QA);多模态信息融合;时序信息处理;视频理解

Video question answering with large language models: a survey

Xie Junlin[1], Zhang Ruifei[1], Li Guanbin[2*]

1. School of Science and Engineer, The Chinese University of Hong Kong (Shenzhen), Shenzhen　518116, China;

2. School of Computer Science and Engineer, Sun Yat-sen University, Guangzhou　510006, China

Abstract: In recent years, large language models (LLMs) have achieved remarkable progress in natural language processing (NLP), demonstrating exceptional capabilities in language understanding and generation. These advancements have driven widespread applications in tasks such as text generation, machine translation, question answering, text summarization, and text classification. However, despite their impressive performance in handling and generating text, LLMs face notable limitations when handling highly complex multimodal tasks, particularly in the domain of video question answering (Video QA). Video QA is a particularly challenging task that requires models to comprehend and generate responses based on dynamic visual content, which often includes temporal and auditory information. Unlike static images or purely textual contents, video data contains inherent temporal dependencies, where the meaning of events and actions unfolds over time. This temporal dimension adds substantial complexity to the understanding process because models must not only interpret individual frames but also maintain coherent understanding across sequences of frames within the broader video context. Consequently, effective Video QA demands advanced temporal information processing capabilities that many LLMs, primarily designed for static text, often struggle to handle adequately. Moreover, the multimodal nature of video, which often involves the integration of visual, auditory, and occasionally textual cues, further complicates the task. Effective Video QA requires the model to seamlessly fuse information across these different modalities, ensuring accurate interpretation and

收稿日期:2024-09-06;**修回日期**:2025-04-20;**预印本日期**:2025-04-27

*通信作者:李冠彬　liguanbin@mail.sysu.edu.cn

基金项目:国家自然科学基金项目(62322608)

Supported by: National Natural Science Foundation of China (62322608)

response to questions regarding video content. This process involves understanding visual scenes, recognizing speech or background sounds, and correlating them with the corresponding textual information. The challenge lies not only in processing each modality independently but also in establishing meaningful connections between them to generate coherent and contextually appropriate responses. This paper presents a comprehensive review of the current state of research on Video QA models based on large language models. The technical characteristics, strengths, and weaknesses of non-real-time and real-time Video QA models are also investigated. Non-real-time Video QA models typically operate on pre-recorded video content, allowing them to access and analyze the entire video sequence before generating responses. These models can leverage global contextual information, making such models particularly effective for tasks that require video content analysis, such as video summarization or detailed scene interpretation. However, they may struggle with efficiency and scalability, particularly when handling long videos or large datasets. In contrast, real-time Video QA models are designed to process video streams as they are received, increasing their suitability for applications requiring immediate responses, such as live video monitoring or interactive video systems. However, these models must maintain a balance between processing speed and accuracy due to their frequently limited access to the full temporal context of the video. The paper discusses the challenges encountered by these models in maintaining performance under real-time constraints, including efficient computation and prediction capability based on partial information. Additionally, the paper explores the commonly used datasets in Video QA research, highlighting their features, limitations, and the types of tasks they are designed to address. The evaluation of Video QA models is also examined, focusing on the metrics and benchmarks used to assess their performance. Understanding the strengths and weaknesses of different datasets is crucial for advancing the field, helping in the identification of gaps in current research and guiding the development of robust and versatile models. Finally, the paper addresses the extensive challenges and bottlenecks in the field of Video QA, including the difficulties in scaling models to handle large and diverse video datasets, the need for efficient multimodal fusion techniques, and the computational demands associated with video data processing in real-time. The discussion is further extended to consider the potential future research directions in Video QA, with particular emphasis on improving the temporal reasoning capabilities of LLMs, enhancing their multimodal integration, and developing efficient model architectures that can operate effectively under resource constraints. Overall, while large language models have presented new possibilities in the field of video interpretation, considerable challenges remain in adapting these models to the specific demands of Video QA. Through the systematic review of the current advancements and the presentation of the key obstacles and future directions, this paper aims to contribute to the ongoing efforts to develop highly capable and intelligent multimodal AI systems. The field must continue innovations in the following areas: temporal modeling, where novel architectures that can effectively capture long-range dependencies in video sequences are needed; multimodal representation learning, where sophisticated approaches for integrating visual, auditory, and textual features could yield substantial improvements. Furthermore, the development of highly efficient training paradigms that can address the computational intensity of video processing while retaining model performance is essential for practical applications. Another critical area for future work focuses on the creation of highly comprehensive and challenging benchmark datasets that effectively reflect real-world scenarios, pushing the boundaries of what current models can achieve. As research in this area progresses, addressing these challenges will be crucial for realizing the full potential of LLMs in video interpretation applications. Achieving this goal will require AI systems that can interpret and reason about dynamic visual content with a level of proficiency comparable to human cognition. The integration of advanced techniques from computer vision, speech processing, and natural language understanding will be pivotal in developing truly multimodal systems capable of managing the complexity and variability in real-world video data. Through continued innovation and interdisciplinary collaboration, the field can overcome current limitations and drive the development of next-generation video understanding technologies with broad applicability across domains such as education, entertainment, surveillance, and human-computer interaction.

Key words: large language models(LLMs); video question answering(Video QA); multimodal information fusion; temporal information processing; video understanding

0 引 言

近年来，大语言模型（large language models，LLMs），如GPT（generative pre- trained Transformer）（Achiam等，2024）、PaLM（pathways language model）（Anil等，2023）和LLaMA（large language model meta AI）（Touvron等，2023）在自然语言处理领域取得显著进展。通过大规模的预训练和复杂的架构设计，这些模型展现出卓越的语言理解和生成能力，广泛应用于文本生成（Mo等，2024）、机器翻译（Feng等，2025）、问答系统（Pan等，2023）、文本摘要（Feng等，2025）以及文本分类（Wei等，2023）等任务。这些成就标志着通用人工智能迈出了重要一步。然而，尽管这些模型在处理和生成纯文本方面表现出色，但在应对复杂的多模态任务时，其局限性也逐渐显现，特别是在需要结合视觉、听觉等多感官信息的情境中，仅依赖文本的大语言模型难以充分理解和处理复杂的任务（Tang等，2023）。

人类与世界的互动不仅依赖于语言，还通过视觉、听觉和触觉等多感官的信息输入和处理（Zhang等，2019），形成多维度的感知体验。因此，随着研究的深入，研究者开始探索视觉语言大模型（vision-language model，VLM）（Zhang等，2024a），这类模型不仅能够处理语言信息，还能够理解和生成图像内容，在多模态任务中展现出强大的潜力。然而，视频模态的理解和生成（Lavee等，2009）任务由于其独特的特征和复杂性，给基于大语言模型的视频理解带来了巨大挑战。

视频作为一种动态的视觉模态，具有以下独特的特征和挑战：1）动态性与时序依赖性。视频不仅展示视觉内容，还包含物体、场景及事件随时间演化的过程。大语言模型在捕捉静态图像时表现优秀，但要理解视频中的帧间关联性和时序动态性，尤其是长时间跨度的复杂变化，需要模型具备更高的时序信息处理能力。这意味着模型不仅需要能够理解单帧的图像信息，还需要通过捕捉帧间变化来把握整体情境。如何有效地将时间维度的信息融入到大语言模型的理解过程中，是当前研究的一大难题。2）跨模态信息融合。视频通常伴随着音频信息，在理解视频内容时，需要有效地融合视觉与听觉模态的信息。这对大语言模型提出更高的要求，模型不仅要能处理单一模态的数据，还必须在不同模态之间建立有效的联系，以便更准确地解读视频中的复杂情境。例如，在一段对话视频中，语言和视觉的结合能够帮助模型更准确地理解说话者的意图，而单独依赖语言或视觉信息可能无法捕捉到对话的全部语义。如何在保持大语言模型语言生成能力（You等，2025）的同时，增强其对多模态信息的处理能力，是实现真正智能视频理解的关键。3）数据量与计算复杂性。视频数据量通常比图像数据大得多，处理视频需要更多的计算资源和存储空间。视频不仅包含大量的帧，还引入时间维度，导致模型的计算复杂性显著增加。尤其是在需要实时处理的应用场景中，模型的计算效率成为一个关键问题。处理如此庞大的数据量，不仅要求模型架构的高效性，还需要在训练和推理过程中采用优化策略，以便在有限的资源下实现高性能。当前的研究正试图在模型复杂性与实际应用需求之间找到一个平衡点。

在大语言模型出现之前，传统的处理方法（Fan等，2021）往往依赖于精心设计的特征提取算法和任务特定的优化技术。这些方法通常需要通过反复训练和微调，针对每一个任务进行定制化设计，无法很好地应对多样化的场景需求，使得构建一个通用的视频多模态大模型变得异常困难。然而，尽管大语言模型在处理纯文本方面具有显著优势（Trummer，2024），但要让它们有效地理解视频内容并构建强大的视频问答模型并不简单。视频的动态性、跨模态信息融合的复杂性以及庞大的数据量和计算要求（Singh和Cuzzolin，2016），使得研究者必须在模型架构和训练方法上做出创新。

面对上述挑战，研究者提出多种方法试图将大语言模型的强大语言处理能力应用于视频问答中。其中，最具代表性的两个方向是基于大语言模型的非实时视频问答模型和实时视频问答模型，如图1所示。这两类模型各有其独特的应用场景和技术特点，并在处理视频内容时展现出不同优势：1）非实时视频问答模型。这类模型主要针对预录视频，模型可以利用整个视频的全局信息进行理解和生成。在这种设置下，模型能够访问视频的全部帧，使其可在捕捉视频的全局上下文时展现出较强能力，并能够根据视频内容的整体结构和主题进行更准确的问答。由于模型可以使用视频的全局信息，可以分析整个视频的时间线、场景转换以及关键情节，从而更好地理解视频的核心信息，使得模型不仅能回答与

具体帧相关的问题，还能回答涉及视频整体情节、人物发展以及复杂因果关系的问题。此外，这类模型还能够识别和理解视频中的隐喻、象征以及其他需要深度语义分析的内容，从而提供更加丰富和精确的回答。2)实时视频理解模型。实时视频理解模型是指能够在视频流输入的过程中，实时进行内容分析和理解的人工智能模型。与传统的离线视频理解模型不同，实时视频理解模型必须具备在接收视频数据的同时，立即进行分析和生成输出的能力。这类模型需要在计算效率和时序信息捕捉上达到高度平衡，以确保在处理视频的同时，不会出现明显的延迟。

本文系统回顾了当前基于大语言模型的视频问答模型方法，梳理现有技术在非实时和实时场景中的应用与挑战。总结来看，尽管大语言模型为视频问答带来了新的可能性，但要充分利用其潜力，还需要克服诸多技术障碍。在未来研究中，如何进一步提升模型的时序信息处理能力、实现更有效的多模态信息融合以及提高模型在实时场景中的应用效果是需要重点关注的领域。此外，随着多模态预训练模型的发展，将探索更大规模、更复杂的数据集，以进一步提升模型的泛化能力和鲁棒性。低资源条件下的视频问答任务也将成为未来的一个重要研究方向，特别是在移动设备和嵌入式系统中，如何在有限的计算资源下保持模型的高效性和准确性，将对技术的应用提出新的要求。总之，大语言模型的出现和发展为视频问答乃至视频理解领域都带来新的机遇和挑战。随着研究的深入和技术的不断创新，未来有望看到更加智能和高效的视频问答系统，为多模态人工智能的发展开辟新的道路。本文将在后续内容中，全面系统地阐述当前主流大语言模型下的视频问答方法的研究进展，以便研究人员能够更好地了解该领域的研究现状以及发展与应用。

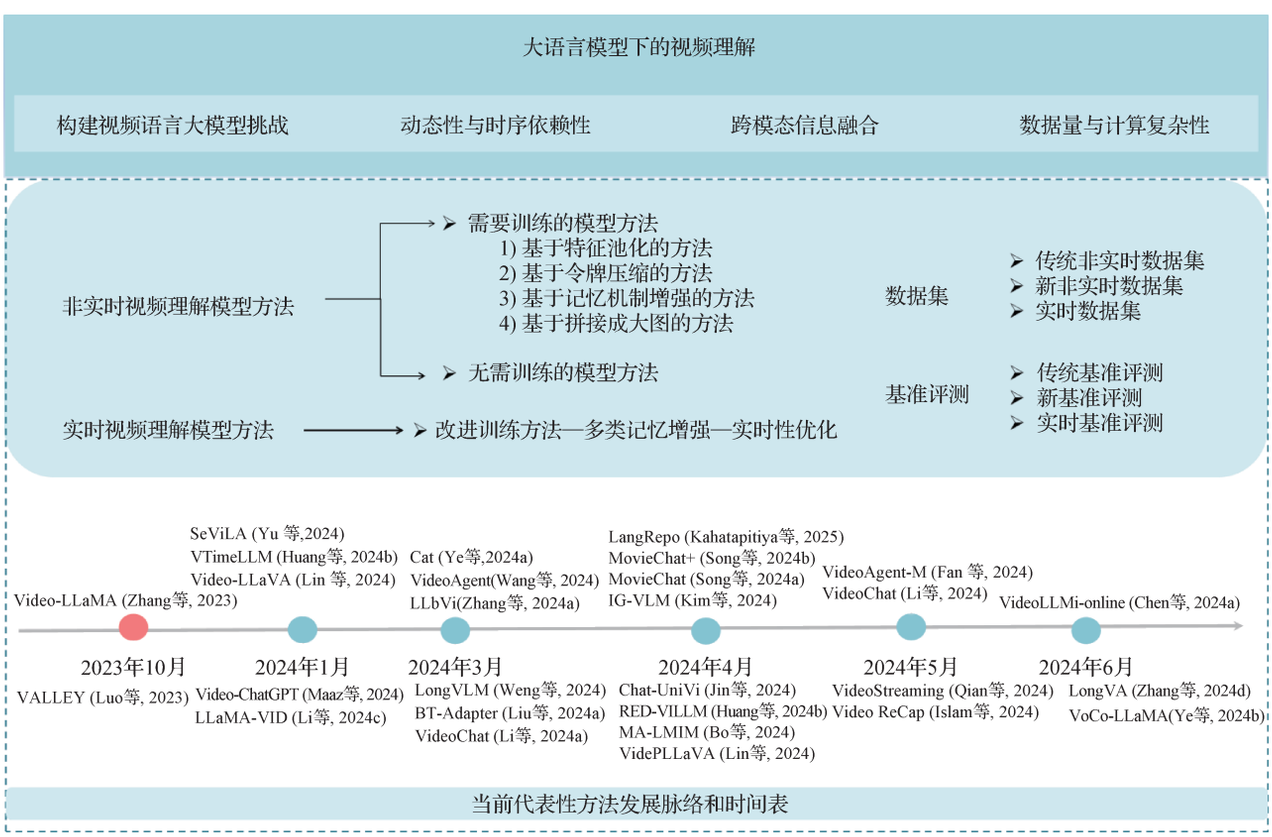


图 1　本文结构框架

Fig. 1　Overall framework of this paper

1　经典的视频问答模型回顾

1.1　基于注意力机制的早期模型

早期模型中非常重要的一类方法为注意力机制。通过注意力机制，在视频问答的任务中，模型能够有选择性地关注视频中的关键帧和重要特征，并将其与问题信息进行动态关联，从而实现跨模态的信息融合和推理。这种基于注意力的方法不仅能够捕捉视频中的时序依赖关系，还能建立问题与视觉

内容之间的语义对应，为准确的视频问答提供重要支撑。

HCRN（hierarchical conditional relation network）方法（Le等，2020）中，基于条件关系网络（conditional relationship network，CRN）的方法核心在于注意力机制设计，通过问题作为条件特征来引导对视频内容的注意力分配，CRN单元的堆叠形成层次化注意力结构使不同分支关注不同视频片段，同时实现了跨时序的关系性注意力建模，构建多层次的视觉—语言关系图，从而在视觉和语言模态间建立有效的注意力引导机制，最终在视频问答任务中取得优异性能。类似地，r-STAN（spatio-temporal attention network）方法（Zhao等，2017）提出一个层次化时空注意力网络，通过时空注意力编码—解码学习框架来处理视频问答任务。该方法的注意力机制在空间和时间两个维度上起作用：空间上关注视频帧中的关键区域，时间上捕捉视频内容的动态变化，并通过多步推理过程将问题信息与视频的时空特征进行动态关联，从而有效学习视频与问题的联合表示，克服了仅关注静态图像的局限性。

1.2 基于构建记忆机制的早期模型

基于记忆机制的早期视频问答模型主要通过设计外部记忆模块来增强多模态信息的存储和检索能力。这类模型通常采用记忆网络架构，将视频的视觉特征和问题的语言特征存储在记忆单元中，通过读写操作实现信息的动态更新和检索。从而实现对视频内容的长期记忆和细粒度理解。模型通过记忆模块存储关键信息，在问答过程中可以根据问题内容有选择地检索和利用相关特征，增强了模型处理长视频序列的能力。同时，记忆机制也为多模态特征的交互提供了新的范式，通过记忆的读写操作实现特征的动态融合，提升了模型的推理能力。这类方法为后续视频问答任务中的特征存储和检索机制提供了重要的技术思路。

Fan等人（2019）提出一个基于异构记忆机制（heterogeneous memory enhancement，HME）的视频问答框架，核心是设计了3个记忆模块：1）异构记忆模块存储和学习视频的表观以及运动特征中的全局上下文信息；2）问题记忆模块专门用于理解复杂问题语义并突出查询主体；3）推理记忆模块通过自更新的方式存储多步推理的中间状态。在问答过程中，模型先通过输入与记忆内容的交互来获取全局语义，然后基于存储的记忆信息进行多轮推理，不断更新和优化记忆状态直至得出最终答案。这种层次化的记忆机制设计在4个VideoQA基准数据集上都展现出了优越性能。

相似地，Gao等人（2019）提出一个运动—表观共记忆网络（motion-appearance co-memory network，MACM）。该网络构建运动和表观特征的共同记忆空间，使两种模态信息能够相互提供注意力线索。同时通过时序卷积—反卷积网络生成多层次的上下文记忆表示，并设计了动态记忆组合机制来根据不同问题的需求灵活检索相关的时序信息。

1.3 基于优化多模态融合的早期模型

在视频问答任务中，多模态融合模型通过设计特征对齐和跨模态交互机制来处理视频和问题文本之间的关联。这类方法通过时空特征提取器捕获视频的动态信息，并利用双线性池化、图网络或变压器等架构将视觉特征与问题表示进行动态关联，从而实现视频内容与问题信息的深度融合。

MCB（multimodal compact bilinear）方法（Fukui等，2016）提出多模态紧凑双线性池化作为一种新的特征融合方法，通过近似计算视觉和文本向量的外积，相比传统的元素级加乘或简单拼接方法具有更强的特征表达能力。在模型架构上，它被用于两个融合阶段：首先融合特征以生成空间注意力，然后将注意力加权的视觉特征与问题特征进行深层融合，这种层次化的融合策略在Visual7W和VQA（visual question answering）数据集上都展现出优越的性能。

类似地，DF（dynamic fusion）方法（Gao等，2019）从多模态信息流的角度创新性地解决了特征融合问题，提出一种动态融合框架，通过模态内和模态间的信息流交互来实现特征的深度融合。其中，模态内注意力流可以在另一模态的条件下动态调节目标模态的注意力分布，而模态间的交替信息传递则能够有效捕获视觉和语言领域之间的高层交互关系。这种双向动态的信息流动机制显著提升了特征融合的效果，使模型在VQA 2.0数据集上达到了当时最优的性能。

在大模型兴起之前，视频问答领域的研究主要集中在上述3类经典方法，这些早期方法为视频内容理解奠定了重要基础，但在处理长视频序列、复杂场景理解以及跨模态对齐等方面仍存在局限性。随着大规模预训练模型的出现，视频问答领域迎来了

新的发展机遇和挑战。在后续内容中，将介绍大模型时代下的视频问答模型。

2　非实时视频问答模型研究现状

非实时视频问答模型旨在处理和理解包含复杂时空信息的视频数据，并在此基础上生成自然语言的回答。与实时视频处理范式不同，非实时视频问答模型主要针对非实时场景，其核心思想是将所有视频帧及用户的查询或指令一次性输入到大模型中，通过模型的强大推理能力生成答案。为了更好地理解这一研究领域，首先对这一问题进行定义，并概述当前的总体框架。接着，详细介绍该框架下的主要研究方法、相关的数据集与任务，以及模型的评测方式。

2.1　问题定义与总体框架

非实时视频问答模型主要应用于无时间限制的场景中，允许系统将整个视频序列和用户的查询或指令一次性输入模型中进行处理。如图 2 模型所示，给定整个视频序列 $\boldsymbol{V}=\{\boldsymbol{v}_0,\boldsymbol{v}_1,\cdots,\boldsymbol{v}_n\}$，其中，$\boldsymbol{v}_i$ 表示视频的第 i 帧，同时用户的查询或指令表示为 Q，目标是得到整个视频中 Q 相关内容的答案，即

$$A = LLM\left(Connector\left(Encoder_v(\boldsymbol{V})\right), Encoder_q(Q)\right) \quad (1)$$

式中，$Encoder_v$ 是视频处理器，$Encoder_q$ 是问题处理器，$Connector$ 是模态转化器，将视频模态转换到与问题模态同一个特征空间，更便于大语言模型理解。

当前的研究方法通常采用如下总体框架：首先通过视频编码器对视频特征进行编码，将视频的视觉信息转化为可计算的特征表示。接着使用一个连接器将这些视频特征映射到与查询文本特征共享的空间中，从而实现多模态特征的对齐。在该空间中，视频特征与用户查询或指令的文本特征能够进行有效匹配。最终，这些对齐后的特征会被送入大语言模型中，以生成符合用户需求的答案或执行相应的任务。下一小节将详细介绍当前基于大语言模型的非实时视频问答模型研究方法。

2.2　研究方法

当前的研究方法为了更好地让大语言模型理解视频内容，主要聚焦于总体框架中的视频编码器和连接器部分的优化。这种优化旨在解决两个主要挑战：一是如何构建有效的视频时序理解方法，二是如何在不增加过多计算资源和存储空间开销的前提下处理视频数据。在现有研究中，如图 1 所示，本文将这些优化方法分为两大类：需要进行数据训练的方式和不需要进行训练的方式。在需要训练的模型类别中，进一步细分出多个子类，主要依据处理视频模态特征的不同方式。这些子类的区分主要基于对视频处理部分的不同改进方法，包括视频编码器和连接器的优化方式，如图 2 所示。

对比两个类型的方法，需要进行数据训练的方式通常涉及对总体框架中基础组件的优化，包括大语言模型，通过大规模数据集上的直接训练或微调，模型能够学习到更加细致和复杂的视频时序特征表示。这种方法的优势在于，经过训练的模型通常能够更好地理解视频内容的复杂性，具备较强的泛化能力，尤其是在处理多样化的场景和任务时表现优异。然而，这种方式的劣势也很明显，即需要大量的计算资源和时间来完成模型的训练过程，尤其是面对大型视频数据集时，训练成本和时间开销都会显著增加。此外，训练过程还可能对存储空间产生较高的需求，以保存中间模型参数和训练数据。而不需要进行训练的方式则通过预设的算法或规则直接处理视频数据，这些方法通常基于现有的特征提取技术，利用已有的模型或特征表示来实现视频内容的理解。其主要优势在于不需要额外的数据训练，因而可以显著减少计算资源和时间的消耗，适合在计算资源有限或时间要求紧迫的场景中使用。这种方式的另一个优势是其灵活性高，可以快速部署并应用于不同的任务和场景。然而，其劣势在于，由于缺乏针对性的训练，模型在处理复杂和变化多样的视频内容时可能表现不足，难以捕捉到视频中的细微时序特征，导致理解和输出的准确性可能不如经过训练的模型。

接下来，将按照分类结构逐一介绍每个子类的方法和技术，相关方法总结如表 1 所示。在各个小节中，将详细讨论各子类的特点、优势以及其在实际应用中的适用性，并分别对需要训练的方法和不需要训练的模型进行介绍。通过这种分类和分析，希望为读者提供一个清晰框架，以理解当前实时视频问答大模型的研究进展及面临的挑战。

2.2.1　基于特征池化视频处理方法

在最初的视频处理模块优化中，主要采用一种较为直接简单的处理方式。首先使用图像编码器

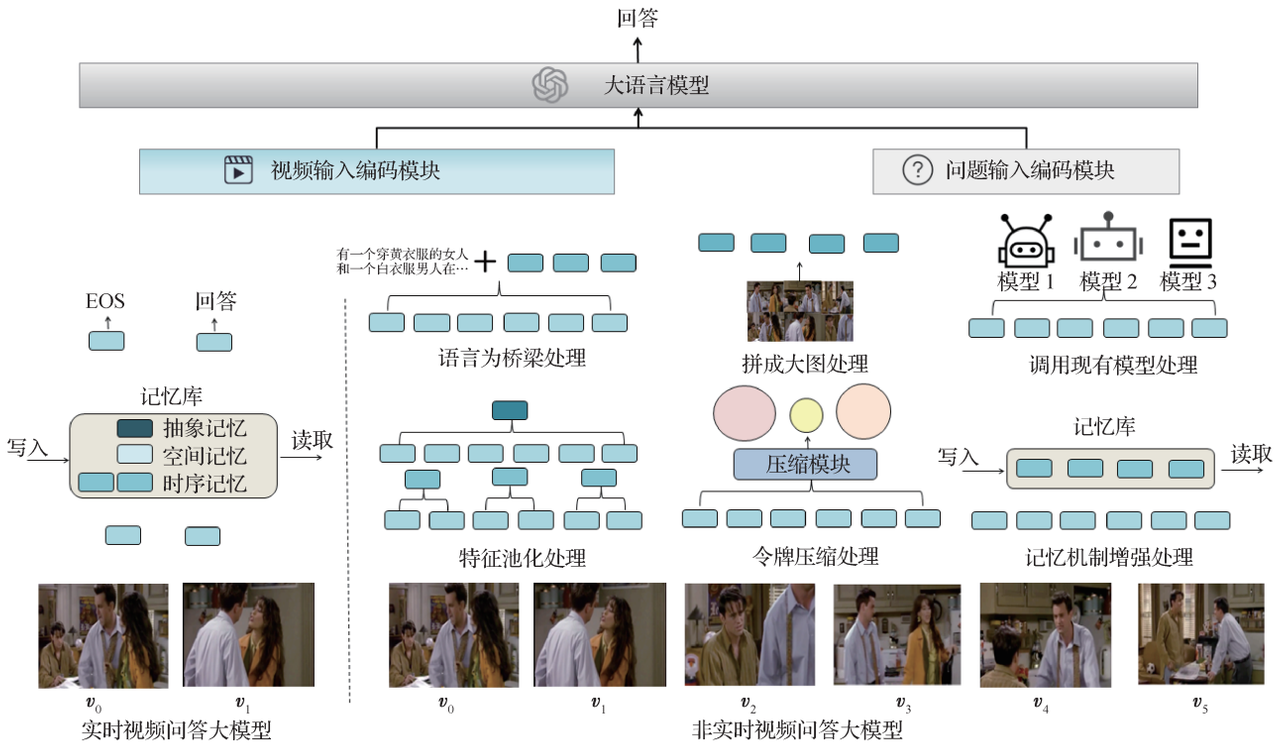


图2　当前方法模型可视化展示

Fig. 2　Visualization of the current methodology model

(通常是clip模型)对视频中的每一帧进行编码得到每一帧的高维特征表示,接着简化对整个视频的处理,将所有帧的特征向量进行平均池化操作,从而生成一个统一的特征表示来概括整个视频的内容。具体而言,Lin等人(2024)提出Video-LLaVA(large language and vision assistant)方法,通过巧妙利用空间和时间池化技术,有效地将视频处理为一系列时序密集的图像,并从中提取关键的空间和时间特征。空间池化技术用于整合每一帧图像中的空间信息,将复杂的图像特征压缩为更为简洁的表示,而时间池化则通过处理多个时间帧之间的关联,提取出反映视频动态变化的时序特征。这两种池化方法相结合,不仅显著降低了模型的计算复杂度,还能保留视频的关键时序信息,使得视频表示更加高效和精准。然而,这种"暴力"的编码方式存在明显的不足。首先,它忽略了视频中帧与帧之间的时序关系和动态变化,仅依赖于静态帧的平均特征来代表整个视频,无法捕捉视频中时间维度的复杂信息。这在处理需要时序理解的任务时表现尤为不足,如动作识别或事件检测等。其次,平均池化的操作在压缩特征表示的同时,可能会丢失一些重要的时序细节和动态信息,导致模型在理解视频的全局内容时不够准确。

基于此,Luo等人(2024)提出两种改进的平均池化类模型方法的结构,以增强模型对视频时序信息的理解。第1种结构引入了可学习的线性层,通过学习每一帧的时间重要性得分来进行加权平均,从而更好地保留每个时间帧的相对重要性;第2种结构在此基础上进一步优化,通过将所有视频帧图像的空间编码特征输入到单层变压器编码器中,提取时间变化特征,并将其与平均池化特征相结合,使模型能够更加精准地捕捉视频中的动态信息。这两种结构分别在不同层次上改进了平均池化方法,提升了模型对时间维度的理解和处理能力。

除此之外,Xu等人(2024)提出PLLaVA(pooling LLaVA)方法,该方法的核心在于设计了一种高效的时序特征池化机制,通过对视频帧序列的特征分布进行自适应平滑,有效降低了异常帧特征的过度影响,显著提升了现有图像—语言预训练模型在视频理解任务中的表现。

2.2.2　基于令牌压缩视频处理方法

基于令牌压缩的视频处理方法相比于基于平均池化类方法,具备更强的时序信息保留和动态变化捕捉能力。平均池化类方法在处理视频时,通过对不同时间戳的特征进行简单平均,容易导致时间维

表 1　大语言模型下视频问答的代表性算法归纳总结和总览

Table 1　Comprehensive overview and summary of representative algorithms for video question answering in large language models

方法类型		代表性方法	核心特点/代码链接
无需训练		VideoAgent	将视频理解过程定义为状态、动作和观察的序列，并使用大语言模型作为控制这一过程的核心代理。公开代码链接：https://github.com/wxh1996/VideoAgent
		VideoAgentM	将视频中信息整理成一个结构化的记忆库，帮助大语言模型像人类一样理解和推理视频内容，同时利用现有工具模型回答问题。公开代码链接：https://github.com/YueFan1014/VideoAgent
		Video ReCap	引入递归的视频语言架构，利用语言层次结构逐步生成视频的多层次描述。公开代码链接：https://github.com/md-mohaiminul/VideoRecap
需要训练	特征池化	Video-LLaVA	利用空间和时间池化技术将视频处理为一系列时序密集图像，从中提取关键的空间和时间特征。
		Valley	提出两种改进的平均池化类模型方法的结构，以增强模型对视频时序信息的理解。
	令牌压缩	LLaMA-VID	将每个视频帧图像表示为两种不同的令牌：上下文令牌和内容令牌。公开代码链接：https://github.com/dvlab-research/LLaMA-VID
		VideoChat	基于动态令牌选择和聚合的机制，能够在视频处理时动态地选择视频帧特征中最具代表性的特征令牌聚合，通过引入一种简单有效的池化策略来平滑沿时间维度的特征分布，从而减少极端特征的主导影响。公开代码链接：https://github.com/magic-research/PLLaVA
		Chat-UniVi	通过引入基于 K 近邻的密度峰值聚类算法的令牌合并方法，逐步合并具有相似语义意义的视觉令牌，从而获得动态视觉令牌。公开代码链接：https://github.com/PKU-YuanGroup/Chat-UniVi
		VoCo-LLaMA	首个利用大语言模型内在功能进行视觉压缩的方法。公开代码链接：https://github.com/Yxxxb/VoCo-LLaMA
	记忆机制	MA-LMM	引入一种长时记忆库，通过顺序处理视频帧并将提取的特征存储在记忆库中，实现长时间视频的有效建模。公开代码链接：https://github.com/boheumd/MA-LMM
		MovieChat	将短期记忆作为视频信息的快速处理和存储单元，长期记忆用于整合和保持关键信息。公开代码链接：https://github.com/rese1f/MovieChat
	拼成大图	LongVA	将视频表示为扩展的图像形式，通过这种编码方式，大语言模型可以将整个视频视做一个整体。

度上信息的丢失或混淆，无法区分和保留每一帧的相对重要性。虽然一些改进方法已经在一定程度上缓解了这些问题，但平均池化类方法仍可能在理解视频中的复杂动态变化和关键事件时表现不足，从而对模型的时序理解能力造成负面影响。

与此不同，基于令牌压缩的方法通过对视频帧的特征进行更细致的处理，能够在压缩特征表示的同时保留更多的时序信息。令牌压缩方法通常通过选择性地保留和聚合最具信息量的视频特征令牌，避免了平均池化中的信息平滑问题。这种方法不仅能够有效减少计算和存储的开销，还能更好地捕捉视频中的关键帧和动态变化，使得模型对时序关系的理解更加精确。因此，令牌压缩方法在处理复杂时序关系的任务中表现出更高的效率和准确性。

具体而言，Li 等人（2024c）提出 LLaMA-VID，这是一种处理长视频令牌生成问题的创新方法，核心思想是将每个视频帧图像表示为两种不同的令牌：上下文令牌和内容令牌。上下文令牌旨在根据用户输入编码图像的整体上下文，将更广泛的图像内容有效地压缩成一个令牌。同时，内容令牌捕捉每个帧的更细节的方面。与上述将视频每一帧只用两类令牌表示不同的是，Li 等人（2024c）提出一种基于动态令牌选择和聚合的机制，能够在视频处理时动态地选择视频帧特征中最具代表性的特征令牌，并对其进行聚合，在减少计算量的同时保持时序信息的完整性。这种方法在视频理解任务中的表现优异，

特别是在需要精确捕捉时序关系的场景下，如动作识别、事件检测等。Zhang等人（2023）提出一种结合Q-Former架构的令牌压缩策略，在视频处理过程中，通过这种架构生成的视觉查询令牌来选择性地保留关键帧特征，在减少数据冗余的同时保留了重要的时序信息。此外，为了增强对多模态数据的理解，该方法融合了音频和视觉特征，使得模型能够在压缩后的特征表示中同时捕捉到视频和音频信号的动态变化。这种方法在视频对话任务中表现出了较好的性能，有效提升了模型的推理效率和时序信息的捕捉能力。Weng等人（2024）提出LongVLM方法，这是一种简单而有效的模型，用于高效的长视频理解，通过将长视频划分为多个短期片段，并为每个片段提取局部特征，以保留其时序顺序。此外，LongVLM通过层级令牌合并模块对视觉令牌进行聚合，压缩每个片段的特征表示，在保持计算和存储开销可控的情况下，实现对视频细节的精细捕捉。为了增强上下文理解，LongVLM将每个视频帧的全局语义信息集成到局部特征中。最终，这些特征序列被输入到大语言模型中，在因果注意力机制的帮助下，模型同时实现了短期片段序列的时序结构建模以及全局语义信息的注入。这一创新方法能够在精细化的层面上理解长视频内容，并保持了计算效率和准确性。

更有新意的是，Jin等人（2024）提出的Chat-UniVi方法，通过引入独特的令牌合并方法逐步合并具有相似语义意义的视觉令牌，从而获得动态视觉令牌。首先使用视觉变换器初始化视觉令牌，然后通过应用基于K近邻的密度峰值聚类算法（density peaks clustering based on K-nearest neighbors，DPC-KNN）逐步对令牌特征进行分组和合并。在处理视频时，利用DPC-KNN对帧特征进行事件提取，每个合并步骤中，分配到同一聚类的视觉令牌通过平均其令牌特征进行合并，最终为大模型提供多尺度的表示形式。高层表示包含高层次语义概念，而低层表示则强调视觉细节。这个方法具有两个显著优势：首先，其统一的图像和视频建模方法允许在混合数据集上进行训练，使其能够直接应用于图像和视频任务；其次，多尺度表示形式有助于对图像和视频的全面理解，使此模型方法能够适应多种任务。在图像和视频理解任务上的评估显示，此模型方法在理解图像和视频方面表现出色，且在多模态大语言模型的图像和视频联合训练中展现了明显的优势。

除了上述几种方法，还有一些方法使用语言模型进行视频特征令牌压缩。VideoStreaming方法（Qian等，2024）是一种新颖的记忆传播流式编码架构，结合自适应记忆选择来顺序编码长视频为简化记忆，并生成参考相关时间戳的回答。方法的核心思想是通过保留代表性空间线索和时序动态，同时减少视频中的时间冗余，以实现高效的视频编码。具体而言，将长视频分割成多个短片段，并顺序地对每个片段进行编码。在编码每个片段时，首先引用前一片段的编码结果作为历史记忆，然后将其与当前片段的特征连接，并输入到一个小型的解码器语言模型中。由于其自回归性质，序列的信息自然会积累到最后几个令牌中，因此方法将这些最后的几个令牌作为更新后的记忆，涵盖了当前时间戳之前的视频信息。通过这种流式编码，VideoStreaming能够显式地考虑长期时序关系，并保持一个固定长度的记忆来表示任意长度的视频。除此之外，方法会存储所有片段的历史记忆，并选择与特定问题密切相关的固定数量的记忆子集。在对每个片段进行流式编码时，方法还会在序列末尾添加一个总结令牌，作为片段指示符，用一个令牌总结片段内容。然后，在给定特定问题的情况下，其将最终迭代中获得的简化记忆与问题进行连接，并通过相同的小型语言模型，最终通过计算问题指示符与所有历史片段指示符的相似性，选择出与问题最相关的记忆。最后，将这些自适应选择的记忆输入到大语言模型中以进行详细的问答。类似地，Ye等人（2024b）提出首个利用大语言模型内在功能进行视觉压缩的创新方法VoCo-LLaMA。该方法引入VoCo（vision compression）令牌的设计理念，通过精心设计的注意力机制重组，实现视觉和文本令牌之间的有效隔离与交互，使大语言模型能够自然地承担起视觉压缩功能。该方法突破性地利用大语言模型的变压器层级结构进行视觉信息压缩，通过VoCo令牌作为信息桥梁从顶层变压器激活状态中提取关键视觉信息，实现了高效的模态对齐。在推理阶段，该方法创新性地引入缓存机制来存储和重用压缩后的变压器激活状态，在保持83.7%性能的同时实现了576倍的压缩率，并在计算效率上取得显著突破，包括缓存存储减少99.8%、FLOPs（floating point operations per second）减少94.8%和推理时间减少69.6%。此外，VoCo-

LLaMA 成功扩展到视频处理领域，通过生成时间序列 VoCo 令牌序列来建模视频时序信息，为多模态大模型的发展开创了新的研究方向。

令牌压缩类方法虽然在短时间内提高视频处理的效率，但也存在一些明显的劣势。首先，它在处理长时间依赖的信息时表现不足，可能无法充分捕捉和利用视频中涉及长时间跨度的重要特征。此外，压缩过程可能导致信息丢失，特别是在多样性高或复杂的视频内容中，这种丢失可能对后续处理产生负面影响。由于其压缩决策通常是静态的，缺乏动态调整的能力，这可能导致对时序信息理解的不灵活和不精准。

2.2.3 基于记忆机制增强视频处理方法

与上述两类方法不同的是，这类方法更侧重于通过引入动态的存储单元或记忆库来维护长时间跨度的视频特征。记忆机制能够在处理长视频时，保留和追踪重要的时序信息，使模型能够在不同时间点动态地更新和检索这些特征。这在处理长时间视频或需要对历史信息进行持续跟踪的任务中尤为有效。具体而言，He 等人（2024）提出 MA-LMM（memory-augmented large multimodal model）方法，核心在于引入一种长时记忆库，通过顺序处理视频帧并将提取的特征存储在记忆库中，来实现长时间视频的有效建模。该方法采用一种在线处理策略，逐帧处理视频并将视频特征动态地存储在长时记忆库中。记忆库通过自回归方式聚合和捕捉历史视频信息，使模型在后续的视频处理过程中能够灵活地参考和利用这些累积的特征。此外，为了提高处理效率，研究团队提出一种记忆库压缩方法，通过选择与平均最相似的相邻帧特征，使得记忆库的长度相对于输入视频长度保持恒定，不仅保留了视频的全部时间信息，还显著减少了长视频中的时间冗余，从而优化了模型的整体性能。类似地，Song 等人（2024）提出 MovieChat 方法，通过引入一种结合短期和长期记忆的机制来处理长视频任务。方法的核心在于将短期记忆作为视频信息的快速处理和存储单元，而长期记忆则用于整合和保持关键信息。具体而言，采用滑动窗口方法提取视频特征，并将这些特征表示为令牌形式，逐帧顺序地输入到固定长度的短期记忆中。当短期记忆达到其容量极限时，最早的令牌会被移出并整合到紧凑的长期记忆中，从而保留长时间视频的关键时序信息。这种机制不仅在视频的时序信息管理上表现出色，还通过持续的短期记忆更新和有效的长期记忆管理，显著减少了视频处理过程中的计算复杂度和内存消耗。

2.2.4 基于拼接成大图的视频处理方法

本小节介绍一种通过将视频帧图像直接拼接成一个大图来对视频进行处理的方法。通过这种拼接视频帧生成大图的方法，模型能够在处理视频时，将视频中的多个帧视做一幅扩展的图像，从而避免了传统方法中因逐帧处理导致的上下文割裂问题。这种处理方式不仅提高模型在视觉信息检索和理解中的效率，还使得模型能够更好地处理长时间跨度的视频内容。具体而言，Zhang 等人（2024d）提出一种名为 UniRes 的统一编码方案，该方案将视频表示为扩展的图像形式，通过这种编码方式，大语言模型可以将整个视频视做一个整体，就像处理一篇长文本一样，在较长的上下文内进行推理和理解。为了实现这一点，研究人员首先扩展了大语言模型的上下文长度，使其能够处理更长的文本数据。然后，他们将这种上下文扩展能力直接应用于多模态模型中，通过将视频帧拼接成扩展图像，语言模型得以在视觉领域中进行长上下文的推理。 这种方法利用了大语言模型在理解和生成长文本时的优势，使其能够在处理长视频时，保持对全局信息的掌握和理解。

Kim 等人（2024） 提出一种名为 IG-VLM（image grid VLM）的创新性方法，该方法通过将视频的多个采样帧按照网格布局拼接成一幅复合图像，从而在一幅图像中融入视频的时间信息。这种方法的核心思想是，通过将视频的时间维度转化为空间维度，使得大语言模型能够处理传统上被认为不适合的视频任务。 视觉语言模型通常只能处理文本标记和图像标记，其中图像标记仅包含空间信息，无法传递时间信息。这一差异使得传统观点认为视觉语言模型难以用于视频理解任务。然而，IG-VLM 通过将视频帧拼接成一个复合图像，将时间信息嵌入到单一图像中，使得视觉语言模型可以利用其强大的图像处理能力来分析视频。这一方法带来了多重优势。首先，它使得现有的高性能视觉语言模型能够直接应用于视频分析任务，无需为视频数据设计专门的模型。其次，由于完全依赖于预训练且冻结的 VLM，避免了任何视频数据的训练需求，从而大大简化了视频分析的流程。最后，这种方法通过消除多阶段

基础模型的复杂性，简化了处理流程，使得整个方法更加高效且易于实现。实验结果显示，只要为每个基准选择合适的视觉语言模型，IG-VLM方法在9个基准测试中的表现超越了现有的先进方法，特别是在5个开放式视觉问答基准和5个多选视觉问答基准中的4个中取得了领先。

虽然将视频帧拼接成一幅大图的方法具有一系列优势，但这种方法并不完全符合人类感知视频时的自然方式。在人类感知视频内容时，时间序列的信息是动态且连续的，通过观看一系列帧的顺序变化来理解事件的进展、动作的连贯性和场景的转变。 将视频帧静态地拼接成一幅图像，虽然可以在一定程度上保留时间信息，但却丢失了动态变化和时间进程的自然流动性。这种静态拼接的方式将时间维度压缩进空间维度，虽然能够让视觉语言模型处理视频内容，但却无法捕捉到视频中帧与帧之间的动态关系。 人类的大脑在处理视频时，会利用连续的时间线索来推断动作的因果关系、对象的运动轨迹以及事件的发展过程。而拼接后的图像只是一种静态的表达方式，无法完全体现这些复杂的时间相关信息。此外，这种方法在一定程度上简化了视频的复杂性，但也可能导致对某些细微但关键的时间信息的忽略。 例如，视频中的某些动作可能只有在特定的时间点上才具有意义，而这种拼接方式可能无法充分体现这些时间点的重要性。这样一来，尽管拼接后的图像可能包含了视频的大部分视觉信息，但对于时间敏感的内容，仍然可能无法提供与动态视频相同的理解效果。

2.2.5　无需训练的非实时视频多模态大模型

与需要额外训练的传统方法不同，一类新兴的视频处理模型通过巧妙利用大语言模型的语言理解优势，实现了零训练、高效的视频分析范式。这类方法主要分为两种路线：第1种是将视频信息转换为语言描述，利用大语言模型强大的语言理解能力直接进行分析和推理；第2种则是将大语言模型作为调度器，通过协调和组合已训练好的视频处理模块完成特定任务。这些方法的共同特点是充分利用了大语言模型在语言理解方面的优势，通过语言作为中介桥梁来处理视频信息。这种设计思路不仅避免了耗时的模型训练过程，还大幅降低了计算资源的消耗，使得视频处理系统的开发和部署更加灵活高效。

具体而言第1种方法中，Kahatapitiya等人(2025)提出一种名为“语言仓库”的新型表示方法。该方法通过迭代更新的机制，处理与视频片段相对应的文本描述，生成一种具备高度可解释性的表示形式。语言仓库的核心思想是在文本操作的基础上，进行读写操作：写入操作会自动修剪冗余的文本信息，生成精简且有效的描述，确保大语言模型能够更好地利用上下文；而读取操作则从仓库中提取存储的语言信息，结合多种时间尺度的数据以及其他可选的元数据(例如时间戳)，进行进一步的分析和推理。

Islam等人(2024)提出一种创新方法，该方法通过引入递归的视频语言架构，利用语言层次结构来逐步生成视频的多层次描述。这种方法通过在不同层次上生成描述，从短片段开始，然后逐级将生成的描述和视频特征结合起来，最终得到更高层次的总结和概括。首先，方法采用了递归的视频语言架构，使其能够在不同的层次上生成视频描述。在最低层次上，模型从短视频片段中提取特征，生成几秒钟内的片段描述。随着层次的提升，模型利用从上一层次生成的描述和稀疏采样的视频特征，进一步生成当前层次的视频描述。这种递归的设计有效地利用了不同层次视频描述之间的协同作用，使得模型能够高效处理长达数小时的视频输入。此外，这种架构设计还使得模型能够利用现代大语言模型的强大推理能力，对视频内容进行更深入的分析和总结。其次，该方法还引入了一种层次化的课程学习策略，从训练模型生成短视频片段的描述开始，逐步引入中长度片段的描述数据，最终训练模型生成长视频的总结性描述。通过这种逐步递进的学习策略，模型能够逐渐学习和掌握视频的层次结构，从生成低层次的短描述逐步过渡到生成高层次的长摘要。通过逐步处理越来越长的文本片段，语言仓库能够学习到更高层次的语义信息，特别是长时间依赖关系，从而提升模型在长时间视频推理任务中的表现。更有新意的是，Ye等人(2024a)提出一种利用语言桥接视觉和音频内容的创新方法，这一方法通过将视频中的视觉和音频特征与文本进行深层次的融合，使得模型能够更好地捕捉和理解视频内容中的语义信息。这种方法的核心在于利用大语言模型的语言理解能力来处理多模态数据，通过引入语言桥梁，使得不同模态之间的信息能够更紧密地关联在一起。具体而言，模型首先从视频中提取视觉和音频特征，

然后通过语言将这些特征与对应的文本信息进行整合。在这一过程中，模型能够捕捉到视频中的关键细节，并且通过语言的纽带，将这些细节与视频语义内容紧密结合，从而在语义层面上对视频信息进行更深层次的理解。这一方法克服了传统多模态学习模型在处理复杂视频内容时所遇到的诸多挑战。首先，通过语言作为中介，模型可以更精确地对视频中的细节进行语义理解，这不仅有助于提高视频内容分析的准确性，还能够在处理与视频相关的问答任务时提供更具针对性和意义的答案；其次，该方法能够有效减少多模态之间的语义偏差，使得模型在面对复杂和多样化的视频场景时，能够更好地捕捉和表达视频内容的核心信息，从而提升多模态问答任务的整体性能。

Zhang 等人（2024a）提出一种名为 LLoVi 的创新性框架，该框架专注于长时段视频的问答任务，利用语言作为连接视频内容和长时段推理的桥梁。与以往的长时段视频模型不同，这一方法并不依赖于特殊的长时段视频模块（如内存队列、状态空间层等），而是采用了一种简洁而有效的双阶段策略。首先，框架将长视频分割为多个短片段，并通过预训练的帧/片段级视觉描述器将这些片段转换为简短的文本描述。 这些文本描述以时间顺序连接在一起，并作为输入传递给大语言模型，以执行长时段推理，完成视频问答任务。这种方法充分利用了大语言模型在处理长时段信息方面的优势，通过语言中介，实现了对长视频的有效理解和推理。为了进一步提升这一框架的效果，研究人员还引入了一种新颖的多轮总结提示策略。该策略首先要求大语言模型对短期视觉描述进行总结，然后再基于模型生成的视频摘要回答给定的问题。 由于生成的视觉描述可能包含噪声或冗余信息，这种总结方案能够有效地过滤可能分散注意力的无关信息，并消除冗余句子，从而显著提升大语言模型在长时段视频问答任务中的推理能力。

在第 2 种方法中，Wang 等人（2024）提出 VideoAgent 方法，该方法模拟了人类理解长视频的认知过程。VideoAgent 以大语言模型为核心，通过多步骤循环来理解视频内容。大语言模型首先浏览视频关键帧以获取整体内容，随后判断当前信息是否足够回答问题。若信息不足，它会确定需要获取的额外信息，使用 CLIP（contrastive language-image pre-training）模型来找到包含这些信息的新帧，并通过视觉模型将新帧转换为文本描述，从而更新对视频的理解。该方法将视频理解过程定义为状态、动作和观察的序列，以大语言模型作为控制代理，通过 CLIP 模型和视觉模型作为工具，实现了大语言模型的视觉理解和长上下文检索能力。

类似地，Fan 等人（2024）提出 VideoAgentM 方法，用于理解和处理视频内容。这个方法的核心是将视频中的信息整理成一个结构化的记忆库，帮助大语言模型像人类一样更好地理解和推理视频内容，同时利用现有的工具模型来回答问题。这个方法首先将视频分成多个短小片段（每段大约 2 s），并为每个片段生成一段文字描述，这些描述被存储在一个叫做“时间记忆”的库中。此外，他们还会记录视频中出现的所有物体和人物的信息，这些信息被存储在另一个叫做“对象记忆”的库中。当需要回答一个关于视频的问题时，大语言模型会作为规划器首先将问题分解成多个小任务，并使用不同的工具模型来处理这些任务。例如，它可能会先从时间记忆库中提取与问题相关的片段，或从对象记忆库中查找某个特定物体或人物的相关信息。通过组合这些工具的结果，大语言模型最终生成对问题的完整回答。这个方法的特别之处在于，它利用了大语言模型的强大推理能力，通过灵活调用现有的工具模型来处理视频内容，而不需要额外的训练。这种方法不仅提高效率，还减少了开发过程中的资源消耗。测试结果表明，VideoAgent 在多个视频理解任务上表现优异，显著提升了准确率。

尽管基于语言桥接的视频处理方法展现出显著优势，但仍面临着一些固有的挑战和局限。首要的挑战在于其对中介语言质量的高度依赖，视频内容到语言描述的转换过程中可能出现信息损失、细节遗漏或表达不准确等问题。特别是在处理复杂的时序关系时，语言描述往往难以完整和精确地捕捉视频中的动态变化。此外，将视频信息转换为语言描述并进行处理的过程也带来了额外的计算开销。

相比之下，之前介绍的直接训练视频理解模型的方法虽然需要更多的训练资源，但能够更直接地处理视频数据。基于语言桥接的方法的另一个明显局限在于其性能严重依赖于现有视频模型的能力水平。这种“零训练”的便利性是以牺牲模型的潜在上限为代价的，如果底层调用的视频处理模型存在偏

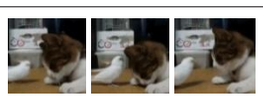
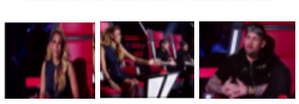
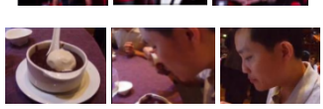
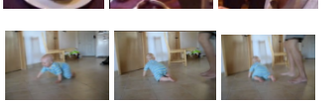
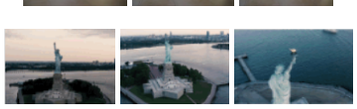
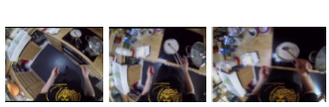
差或能力不足，即使大语言模型的调度和推理再优秀，最终的处理效果也会受到制约。这种对既有模型的依赖性使得系统的整体性能很难突破现有视频模型的能力边界。

2.3 数据集和任务

在大语言模型普及之前，视频问答模型通常依赖于特定的数据集和任务范式，模型的训练和测试都基于专门设计的数据集。然而，随着大语言模型的发展，尤其是那些具备强大跨模态处理能力的模型，视频问答任务的研究范式正在发生显著转变。这一转变促使新型数据集和任务设计的出现，特别是模态对齐数据集和指令微调数据集。模态对齐数据集是为提升大语言模型处理多模态信息（如文本、图像、视频）的能力而开发的，如表2所示。这些数据集通常包括文本与视频的配对，旨在帮助模型在不同模态间建立更精确的关联。这类数据集在大语言模型的预训练阶段被使用，以增强模型的跨模态理解能力，使其能够从视频内容中提取信息并生成准确的文本描述，从而更好地执行视频问答任务。指令微调数据集则旨在增强模型对复杂指令的理解能力，使其能够根据用户的具体需求提供更加准确和相关的回答。这些数据集通常包含各种复杂的指令及其对应的视频内容，帮助模型在实际应用中更好地适应和应对不同的问答需求。接下来，将具体介绍传统数据集和新型数据集的特点，以便读者更好地理解它们的独特性。

表2 视频问答当前数据集总览和归纳

Table 2 Overview and summary of the current video question answering datasets

数据集	视频展示	问题	答案	特点
TGIF-QA		猫做了三次什么？	低下头	约165 K个GIF较短视频片段，超过120 K个问答对
MSRVTT-QA		谁在和评委说话？	女孩	约10 000个视频片段 约243 000个问答对
ActivityNet-QA		这个碗是什么形状的？	圆形	约5 800个长视频， 超过58 000个问答对
Next-QA		婴儿爬行后，人干了什么？	跟着他	约5 440个视频包， 约50 000个问答对
VideoChatGPT-100k		视频拍摄于哪里？	纽约	约100 000个问答对， 问题涉及不同的推理层次
Online-video		接下来我应该干什么？	煎牛排	该数据集通过150个问题模板，利用大语言模型从海量离线视频注释中合成流式问答对，覆盖过去、现在和未来的多时间推理层次

2.3.1 传统数据集

1）TGIF-QA数据集（Jang等，2017）。是一个专注于短视频理解的大规模数据集，2017年发布，包含72 000个精选的GIF（graphics interchange format）动画片段和165 000个问答对。每个GIF片段平均长度为3.1 s，配有多个针对不同方面的问题。问题分为4类：重复计数（如“人挥手几次？”）、动作识别（如“这个人在做什么？”）、状态转变（如“这个人的表情从什么变成了什么？”）和帧问答（如“桌子上有什么东西？”）。这些问答对的设计目的是测试人工智能（artificial intelligence，AI）模型对动态视觉内容的理解能力，特别是在捕捉动作、状态变化和时序关系方面的表现。该数据集已成为评估视频问答系统的重要基准之一。

2）MSRVTT-QA数据集（Xu等，2017）。是一个基于MSR-VTT（Microsoft research video to text）视频数据集构建的大规模视频问答数据集，由微软研究院于2017年发布。该数据集包含10 000个来自不同领域的视频片段，总时长超过41 h，配有243 000个问答对。问题类型可以分为5类：询问“什么”的问

题(如“他们在做什么?”)、询问“谁”的问题(如“视频中的主角是谁?”)、询问“如何”的问题(如“他是怎么打开门的?”)、询问“哪里”的问题(如“这个场景在哪里?”)以及询问“什么时候”的问题(如“这个视频是在白天还是晚上拍摄的?”)。这些视频涵盖了日常生活、体育运动、电影片段、新闻报道等多个领域,每个视频都配有约20句描述性文本。数据集的多样性和规模使其成为评估视频理解和跨模态学习模型的重要基准。

3)ActivityNet-QA数据集(Yu等,2019)。包含58 K个问答对和5 K段长视频,专注于长时视频的理解与问答任务。与短视频数据集不同,ActivityNet-QA的视频长度通常从几分钟到几十分钟不等,问题范围也覆盖视频的全局内容。这些问题需要模型从长时间跨度的视频中提取和理解关键信息,并准确回答与全局视频内容相关的问题,如“整个视频的主题是什么?”、“主要人物在做什么?”等。由于视频时间较长,ActivityNet-QA对模型的记忆力和全局语义理解能力提出更高要求,适用于研究复杂场景的全局语义理解和信息提取。

4)Next-QA数据集(Xiao等,2021)。专注于多步推理和因果关系的视频问答任务,包含50 K个问答对和多样化的视频内容。这个数据集的独特之处在于问题设计上强调对视频中事件顺序和因果关系的理解,要求模型进行多步推理。例如,问题可能涉及视频中的因果链或时间序列,如“在某个动作之前发生了什么?”或“接下来会发生什么?”。Next-QA数据集(Xiao等,2021)广泛应用于评估模型在推理任务中的表现,特别是在复杂场景中理解事件之间的因果关系和时序关系。这种数据集对模型的逻辑推理能力和深层语义理解提出极高的挑战,适合开发和测试具有高级推理能力的AI系统。

2.3.2 新数据集

这类数据集主要指两类:模态对齐数据集和指令微调数据集。通常,训练视频问答大模型,第1阶段使用模态对齐数据集,第2阶段使用指令微调数据集。

当前模态对齐任务使用的数据集主要来自WebVid 2.5M(Bain等,2021)数据集中的视频—字幕对。WebVid 2.5M数据集是一个庞大且多样化的资源库,包含约250万对视频和字幕,是目前规模最大的公开视频—字幕数据集之一。该数据集涵盖广泛的内容类型和场景,包括自然风光、城市生活、社会活动和体育赛事等,不同的视频展现了从日常生活到专业领域的丰富多样的情境,并且数据集中的字幕详细且自然,能够准确描述视频中的视觉内容,这些字幕不仅包括静态的视觉元素,还涵盖动作、事件顺序、人物互动和情感表达等动态信息,使得数据集在多模态学习任务中具有极高的应用价值。通过这些字幕,模型能够学习如何将视觉内容与自然语言描述进行精确匹配,从而在模态对齐任务中表现出色。此外,该数据集的视频内容质量较高,字幕生成经过仔细筛选和优化,确保字幕与视频内容的高度一致性。这种一致性对于训练模态对齐模型至关重要,因为它能够使模型在处理复杂的现实场景时保持准确的理解和表达能力。数据集的规模和多样性也使得它特别适合用于训练和测试不同类型的多模态模型,无论是用于生成自然语言描述的视频字幕模型,还是用于检索相关视频的文本查询模型,都能在这一数据集上得到充分的锻炼和验证。

当前的指令微调数据集有ShareGPT4Video数据集(Chen等,2024b),该数据集是一个庞大且多样化的视频—文本对数据集,专为提升视频理解和视频生成任务而设计。该数据集包含约480万对视频和高质量的文本描述,这些描述是通过一个创新的滑动窗口差分标注方法生成的,确保文本描述的时间一致性和细节准确性。它的数据来源非常广泛,涵盖了多个领域,包括野生动物、烹饪、体育、自然风景、第一人称视角的人类活动和自动驾驶场景。为了保证数据的多样性和高质量,数据集采用语义过滤策略,过滤掉冗余内容,确保每个视频片段都具有独特的主题和情境。此外,该数据集通过GPT-4模型和定制的ShareCaptioner-Video模型生成了详细的时间序列描述,这使得模型能够更好地理解和生成具有复杂时间结构的视频内容。凭借这些特性,ShareGPT4Video成为当前最先进的指令微调数据集之一,在提升大规模视觉语言模型和文本生成视频模型的方面表现出色。

除此之外,VideoChatGPT-100k数据集(Maaz等,2024)涵盖100 000个多模态对话样本。这个数据集独特之处在于它不仅包含文本对话,还集成视频内容,帮助模型在处理视频和文本信息时实现更好的理解和推理能力。每个对话样本都关联有视频片段,这使得模型能够通过视觉信息捕捉上下文,增

强对话的自然性和准确性。通过使用 VideoChat-GPT-100k 数据集进行微调，模型可以更好地应对真实世界中的多模态交互场景，如虚拟助理、视频客服和教育领域的应用。此外，数据集还提供了丰富的场景多样性和对话复杂性，这有助于模型在各种情况下保持稳健性。

除了上述训练数据集外，还有一些新出的专门用于测试视频大模型的数据集。MVBench 数据集（Liu 等，2024b）通过动转静方法构建 20 个视频理解任务，包含 23 000 多个多选题问答对。这些任务主要评估模型在运动方向判断、速度变化检测、时序顺序理解等基础时序感知能力，物体状态变化、场景转换、物体交互等视觉状态转换能力，因果关系推断、意图预测、情感变化等高级认知推理能力，以及时间定位、空间追踪、动作持续时长等时空定位理解能力。每个任务都经过精心设计，确保只有通过对整个视频序列的理解才能正确回答，而无法仅依靠单帧图像作答。这种多层次的任务设计使该数据集成为评估视频大模型时序理解能力的重要基准数据集，其任务涵盖了从基础感知到高级认知的完整能力谱系，为模型的综合评估提供了可靠的测试标准。VideoMME 数据集（Fu 等，2024）是首个全面的视频多模态评估基准数据集，发布于 2024 年，包含 900 个视频，总时长达 254 h，每个视频配有 3 个高质量的多项选择题，共 2 700 个问答对。该数据集具有 4 个主要特点：1）视频类型多样性，覆盖 6 个主要视觉领域和 30 个细分领域，包括知识、影视、体育竞技、艺术表演、生活记录和多语言等；2）时长维度丰富，包含短、中、长期视频，时长范围从 11 s 到 1 h 不等；3）数据模态广泛，除视频帧外，还整合了字幕和音频等多模态输入；4）标注质量严格，由专业标注人员反复观看视频内容进行人工标注。

2.4 基准及评估

本小节介绍视频问答任务相关的评估方法，这些方法可以分为 4 类：封闭集评估、开放集评估、对话连续性评估和人机交互评估。

1）封闭集评估。适用于有明确标准答案的问题类型，评估模型的性能主要通过计算其提供的答案与标准答案的一致性。常用的评价指标包括 CIDEr（consensus-based image description evaluation）、METEOR（metric for evaluation of translation with explicit ordering）、ROUGE（recall-oriented understudy for gisting evaluation）和 SPICE（semantic presence in caption evaluation）等，用以量化模型输出与标准答案的相似性。相关的基准示例包括 2.3 节提到的传统基准数据集，如 MSRVTT-QA（Xu 等，2017）、TGIF-QA（Jang 等，2017）、ActivityNet-QA（Yu 等，2019）以及新基准 MVBench（Liu 等，2024b）。

2）开放集评估。应用于没有固定答案的问答任务，模型生成的答案没有预定义的选项。尽管如此，仍需要参考标准答案进行评价，通常会借助诸如 GPT-4 等高级语言模型，来比较模型生成的回答与参考答案的相似度和相关性，从而给出评分。相关的基准示例包括传统基准，如 NextQABench（Xiao 等，2021）、TVQA+Benc（Lei 等，2020）以及新基准 VideoChatGPT Bench（Maaz 等，2024）。

3）对话连续性评估。专注于评估模型在多轮对话中的连贯性和上下文理解能力。这类评估重点考察模型是否能在对话中保持逻辑一致性、情境适应性以及语言生成的自然流畅性。评价指标可能包括 BLEU（bilingual evaluation understudy）、TER（translation edit rate）以及对话特定的评估方法，如对话连贯性得分。相关的基准示例包括 VideoChatGPT Bench（Maaz 等，2024）和 MVBench（Liu 等，2024b）。

4）人机交互评估。旨在测试模型在实际使用环境中的表现，特别是与真实用户进行交互的情况。通过模拟真实的对话场景，评估模型的应答质量、交互流畅度和用户满意度。这种评估方式尽管费时费力，但能够提供关于模型实际应用价值的重要反馈。相关的基准示例包括传统基准 YouCook2Bench（Zhou 等，2018）和新基准 VideoChatGPT Bench（Maaz 等，2024）。

以上为当前的数据集和评估基准介绍，尽管现有的视频多模态评测数据集做出了重要贡献，但仍存在几个亟待解决的局限性。首先，多数现有基准测试主要局限于简单的提取式问题，即答案可以直接从视频内容中找到，这类评测难以真正反映模型的深度理解和推理能力。其次，大多数现有基准测试主要关注短视频（通常在 30 s 以内），这可能无法充分反映现实场景中所需的长期时序依赖性和上下文理解的复杂性。此外，现有模型在细粒度的时空理解方面也存在明显不足：一些模型擅长空间定位但难以处理细粒度的时序理解任务；另一些模型专注于时序理解却无法确定物体的边界框位置，缺乏

统一的时空细粒度多模态理解能力。这种局限性部分源于多模态坐标对齐的困难性，以及在保持视觉细节的同时进行特征压缩的挑战。另一个重要的问题是，这些数据集普遍采用单轮问答格式，可能无法有效评估模型在维持上下文连贯性和进行动态多轮交互方面的能力。同时，许多评测任务过于依赖合成数据，限制了其在真实场景中的适用性，且采用的评估指标（如F1值和ROUGE）可靠性不足。不同数据集的标注质量和一致性也存在差异，这是因为许多数据集依赖半自动化流程或非专业标注人员。

针对这些局限性，未来的改进方向可以考虑以下几个方面：1）构建更加全面和系统的评估框架，包含多层次的理解任务，从基础的视频内容提取到复杂的推理和因果关系理解；2）开发统一的时空细粒度理解评测标准；3）引入多轮对话式评测机制，更好地评估模型的上下文理解和交互能力；4）建立更严格的数据质量控制标准，确保标注的专业性和一致性；5）增加对长视频理解能力的评估，并设计更贴近真实应用场景的测试案例。

3　实时视频问答模型研究现状

实时视频问答模型（杨铮 等，2022）旨在处理和理解实时流式视频数据，并在接收视频帧的同时，生成与视频内容相关的自然语言回答。这类模型在视频数据的输入与推理过程中与非实时视频多模态大模型有显著的区别，特别是在推理阶段视频帧的处理与时序特征的更新方面。下面本文将定义这一问题，并详细介绍实时视频问答模型相关的研究方法、数据集和评测方式，着重分析与非实时模型的不同之处。

3.1　问题定义和总体框架

在实时视频问答模型的框架下，问题可以定义为一个序列建模问题。整个视频序列和非实时模型输入一样，但是实时是在视频序列的某个时间点 t 时，用户进行提问 Q，即假设当前时间点是2，则当前视频输入序列为 $\boldsymbol{V}_2=\{\boldsymbol{v}_0,\cdots,\boldsymbol{v}_2\}$，目标是得到当前时间点下用户提问 Q 的答案 A，具体为

$$A=LLM\left(Connector\left(Encoder_{v_t}\left(\boldsymbol{V}_t\right)\right),Encoder_q\left(Q\right)\right) \quad (2)$$

式中，$Encoder_{v_t}$ 是到当前时间点 t 下的视频处理器，往往要求有更高的实时性处理；$\boldsymbol{V}_t$ 是指当前时间点下包含的视频序列，而整个视频序列是 $\boldsymbol{V}$。

实时视频多模态大模型与非实时模型相比，面临更加复杂的挑战，主要体现在对时间敏感性和处理效率的要求上。首先，实时模型需要具备强大的时间对齐和事件识别能力。在非实时模型中，模型可以在获取整个视频之后进行全局分析，识别事件和对齐用户查询。然而，实时模型必须在每一帧视频输入时，立即识别和处理关键事件，确保不会错过任何与用户查询相关的时刻，例如用户要求在某个特定时刻提醒时，模型需要实时扫描每一帧，准确识别并做出响应。其次，低延迟响应是实时模型的核心需求。非实时模型可以在处理完成后生成响应，不受时间压力的限制，而实时模型则要求在接收视频帧的同时，几乎同步生成输出，尤其在自动驾驶等场景中，任何延迟都可能带来安全隐患，因此模型架构必须经过优化，以在保证准确性的同时，达到极低的延迟。最后，实时模型在处理过程中还需维持对长时序信息的理解和记忆。与非实时模型可以依赖全局视频信息进行上下文处理不同，实时模型必须在逐帧输入的条件下，设计出高效的记忆机制，确保在回答诸如“我已经完成了哪些步骤”之类的长时序问题时，能够准确整合和回顾整个视频流信息。这种对时间敏感性、处理效率以及长时序信息管理的综合要求，使得实时视频多模态大模型在设计上更加复杂，通常需要在轻量化和高效性之间找到平衡。尽管实时模型的总体框架与非实时模型类似，都包含视频编码器、连接器、问题编码器和大语言模型等基本组件，但为了满足实时性的要求，需进行一系列的优化与改进。下一小节将详细介绍当前研究中用于实现这些改进的方法。

3.2　研究方法

当前关于实时视频问答大模型的研究方法非常有限，仅有两种方法。本文将介绍这两种方法的特点，并重点说明它们相比非实时视频问答模型的改进之处。

具体而言，Chen 等人（2024a）提出一个创新框架，旨在提高模型的时间对齐能力、支持长时间的视频上下文处理，并实现高效推理。为了解决模型实时回答的问题，作者引入一种新的训练目标，称为流式 EOS（end of sequence）预测。这一目标帮助模型在处理视频流时，判断何时应该作出响应，何时应保持沉默。具体来说，当视频帧输入模型时，模型会检

查用户的指令，判断是否需要输出答案；如果不需要回答，模型会输出EOS（即“结束标记”），跳过这一轮对话，从而学会在适当的时候进行回答。与传统的多模态大模型通过预测下一个标记来训练的方式不同，流式EOS标记不会出现在模型的输入或输出序列中，但它可以与常规的自回归损失结合使用。这种训练方法使得模型能够保持持续的开放状态，用户可以随时提问。当问题涉及到过去或当前的视频内容时，模型会实时回答；而对于未来的视频内容，模型会判断是否已经到了合适的时间点，并在到达时即时作答。这种方法不仅减少了GPU内存的使用，还降低了对上下文长度的依赖，实现了更快速的推理和回答。此外，模型的回答时机更加灵活，实时性也显著提升。

与上述通过改进训练方式来构建实时视频多模态大模型不同，Zhang等人（2024c）提出名为Flash-VStream的模型，这是一种能够实时处理超长视频流并响应用户查询的视频语言模型。Flash-VStream模型采用类似人类处理流程的“4步2过程”设计：视频帧编码器如同人眼处理视觉信息，大语言模型如同大脑处理语言信息。其时空抽象检索记忆库可以在线实时压缩并更新必要的视觉信息并进行存储，从而实现高效的视频处理和实时响应用户指令。具体而言，当多个视频帧通过编码后得到视频特征后，模型构造的时空抽象检索记忆库对这些特征进行实时压缩存储。该记忆库主要包括4类记忆信息：当前视频帧的空间记忆信息、时序记忆信息、抽象记忆信息和检索记忆信息。空间记忆中包含最近输入的视频帧的详细空间信息，并使用先进先出队列进行存储；时序记忆中包含时序动态信息，可视为当前所有输入视频帧中的事件信息，通过聚类方式获取；抽象记忆包含更高层次的语义内容，主要通过将空间和时间记忆中的内容综合成抽象的语义概念；检索记忆则通过识别和检索最重要的帧特征，专注于存储精确的空间细节。通过将原视频特征压缩成4类不同粒度的特征，模型可以更高效地处理视频特征，从而更快地完成用户指令的响应。这个设计大大提高了模型的处理效率和实时性，使得它能够在处理超长视频时仍保持高效和精确。

3.3 数据集

本节介绍现有实时视频问答数据集，包括数据来源、数据处理和标注，以及数据集的构建方式。

Chen等人（2024a）构建了一个实时视频对话数据集online-video，主要依靠先进模型（如GPT-4）来生成新数据或优化现有数据集。数据来源主要包括流式视频集（Ego4D）和非流式视频集（comprehensive instructional video dataset，COIN）。流式视频集是通过实时拍摄和传输方式收集的视频数据，而非流式视频集则是通过预先录制和存储视频的方式收集的数据。虽然流式视频集在一定程度上符合训练实时模型的需求，但其先前的标注主要是针对视频短片段的。为了更好地满足实时训练的需求，作者对这两种视频集分别采用了不同的处理方式，从而基于它们构造了实时视频理解数据集。对于流式视频集中的标注，采用了与人工标注者相同的提示来指导先进模型，对5 min视频进行实时生成叙述，这些叙述作为新的标注数据；对于非流式视频集，通过提取时间段注释，将其组织成包含视频活动描述的语言提示，并创建一个包含150个问题的模板库，这些问题涵盖视频的过去、现在和未来事件。通过随机插入查询问题，在视频的关键时间戳（如活动状态变化的时间点），提示先进模型根据当前问题生成响应，从而生成时间上多样化的对话数据。通过上述方法，集合流式视频集新生成的叙述和非流式视频集新生成的对话，构造了一个更丰富和动态的数据集。

Li等人（2024）提出一个新的实时视频问答数据集，通过先进模型优化已有的非实时视频数据集。与之前的数据集不同，该数据集的视频长度更长，平均时长为30～60 min。数据集部分来源于Ego4D数据集，另一部分来自Movienet数据集（Huang等，2020）。问题内容多样化，涵盖以下几类：场景总结、动作描述、事件发生、事件顺序叙述和序列验证。问题形式包括开放性问题和是/否双选题。具体而言，作者通过5个步骤进行数据的采集和标注：先从Ego4D数据集中选择10个（每个视频时长1 h）视频，以及从MovieNet数据集中选择22个（每个视频时长30 min）视频，这些视频涵盖了不同类型的内容。接着，使用先进模型GPT-4V对每个视频片段进行密集描述生成，即先将选择的长视频分成多个30 s的片段，并从每个片段中稀疏采样8帧作为输入进模型。每个模型的输出详细描述了对应30 s视频片段的内容，并标记有具体的时间戳。随后，作者使用先进模型GPT-4对由GPT-4V生成的密集描述进行去重和

总结，生成的摘要通常来源于几分钟视频内容对应的多个密集描述，即为一段简洁的场景级描述，整个总结过程中会仔细保留时间戳。然后，使用GPT-4基于场景摘要生成5种类型的问答对，每个问答对从单个或连续的场景摘要中生成，确保问答对仅与时间戳之前的视觉信息相关。最后，标注志愿者判断生成的问答对与视频内容的相关性，并筛选掉以下类型的问答对：问题与视频无关或含糊不清、需要视频之外的额外知识、可以在没有视频的情况下回答、答案错误或含糊不清、问题重复。通过这些步骤，确保最终生成的问答对与视频内容高度相关。

3.4 评测方式

由于实时视频问答模型存在更多的挑战，评测模型指标当前除了回答准确性外，还增加了实时性类的评测。

具体而言，Chen等人（2024a）提出3个指标来进行评测，分别是语言模型指标、时间差异和流畅度。语言模型指标通过语言困惑度来衡量模型在特定时间点的语言建模能力，困惑度越低，表示回答越准确；时间差异通过计算模型回答响应时间戳与预期回答对应的时间戳之间的差异来评估在线助手的时间对齐能力，并将每次对话的时间差异取平均作为指标；流畅度则综合评估了语言建模和时间对齐，通过计算在对话轮次中连续成功标记预测的比例来反映在线对话中的综合语言建模能力。而Zhang等人（2024）提出两个指标来评估方法的计算效率：从提问到回答的平均响应延迟和模型最大视频随机存取存储器的使用情况。

4 结 语

本文系统梳理、介绍了大语言模型下的视频问答模型的研究进展，并将其分为非实时视频问答模型和实时视频问答模型两大类。针对每一大类模型，不仅深入探讨了各种方法的核心原理与架构设计，还对相关的数据集、任务场景、当前的基准与评测指标进行全面综述。尽管当前的视频问答模型研究已取得显著进展，但本文认为未来仍有许多关键问题和潜在研究方向值得进一步探讨。以下是本文提出的几个具有前瞻性的研究方向：

1）长视频相关的任务挑战。随着用户生成内容和长时间视频数据的激增，处理长视频的能力变得愈发重要。目前，大多数视频问答模型主要聚焦于短视频的处理，尚未能充分应对长视频中的复杂时序依赖和信息密集度问题。长视频往往包含更多的场景转换、人物互动以及跨时间段的信息关联，对模型的时序建模能力提出更高的要求。此外，由于长视频的数据量巨大，处理过程中计算资源的消耗与时效性之间的矛盾也更加突出。未来研究可以探索更高效的长视频问答模型，在保证处理速度的前提下，实现对长时间跨度信息的精准捕捉，同时优化计算资源的使用，从而提高长视频问答的效率和准确性。

2）视频问答模型的“幻觉”问题。在视频问答任务中，多模态大模型经常会因数据分布不均或模型训练不足而产生“幻觉”，即生成与输入内容不符的答案。这一问题在生成类任务中尤为显著，特别是在处理复杂视频内容时，幻觉现象更加常见，可能导致模型输出结果的可信度降低。此外，模态间的信息融合不当也会加剧幻觉问题的发生。尽管本文未深入探讨这一问题，但它是现有视频问答模型中不可忽视的问题。未来研究应进一步分析并缓解幻觉现象，以确保模型输出的信息准确且一致。

3）实时视频问答模型的创新方法探讨。实时处理是视频问答模型面临的一大挑战，尤其是在低延迟与高准确性之间的平衡方面。现有的实时视频问答模型仅在部分方法上进行了优化，但在应对复杂场景和多任务处理时仍显不足。随着实时应用场景的不断扩展，如视频会议、直播监控等，模型需要在极短时间内对多模态信息进行高效处理。未来研究可以探索更多实时处理的创新方法，例如利用边缘计算和分布式处理技术，以提高模型的实时性和准确性，更好地满足实际应用需求。

4）更多模态的融入帮助。当前视频问答模型的研究主要集中在视觉模态的分析与处理上，即通过视频中的图像信息进行场景理解。然而，对于视频所携带的音频模态和文本模态（如字幕）的研究相对较少。这些模态在视频问答中起着重要的补充作用，能够提供视觉信息无法涵盖的关键信息，从而提升模型的整体表现。音频模态可以帮助模型理解情感语调、环境音效等情境信息，而字幕文本模态则能够提供明确的语义线索和情节提示。未来研究应更多地融入多元模态信息，发展更加全面且精确的视频问答模型。

综上所述,通过对以上问题与方向的深入研究,期望能够推动视频问答模型的理论进步与实际应用,以更好地应对复杂多样的视频分析任务。

参考文献(References)

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman F L, Almeida D, Altenschmidt J, Altman S, Anadkat S, Avila R, Babuschkin I, Balaji S, Balcom V, Baltescu P, Bao H M, Bavarian M, Belgum J, Bello I, Berdine J, Bernadett-Shapiro G, Berner C, Bogdonoff L, Boiko O, Boyd M, Brakman A L, Brockman G, Brooks T, Brundage M, Button K, Cai T, Campbell R, Cann A, Carey B, Carlson C, Carmichael R, Chan B, Chang C, Chantzis F, Chen D, Chen S, Chen R, Chen J, Chen M, Chess B, Cho C, Chu C, Chung H W, Cummings D, Currier J, Dai Y X, Decareaux C, Degry T, Deutsch N, Deville D, Dhar A, Dohan D, Dowling S, Dunning S, Ecoffet A, Eleti A, Eloundou T, Farhi D, Fedus L, Felix N, Fishman S P, Forte J, Fulford I, Gao L, Georges E, Gibson C, Goel V, Gogineni T, Goh G, Gontijo-Lopes R, Gordon J, Grafstein M, Gray S, Greene R, Gross J, Gu S S, Guo Y F, Hallacy C, Han J, Harris J, He Y C, Heaton M, Heidecke J, Hesse C, Hickey A, Hickey W, Hoeschele P, Houghton B, Hsu K, Hu S L, Hu X, Huizinga J, Jain S, Jai S, Jang J, Jiang A, Jiang R, Jin H Z, Jin D, Jomoto S, Jonn B, Jun H, Kaftan T, Kaiser Ł, Kamali A, Kanitscheider I, Keskar N S, Khan T, Kilpatrick L, Kim J W, Kim C, Kim Y, Kirchner J H, Kiros J, Knight M, Kokotajlo D, Kondraciuk Ł, Kondrich A, Konstantinidis A, Kosic K, Krueger G, Kuo V, Lampe M, Lan I, Lee T, Leike J, Leung J, Levy D, Li C M, Lim R, Lin M, Lin S, Litwin M, Lopez T, Lowe R, Lue P, Makanju A, Malfacini K, Manning S, Markov T, Markovski Y, Martin B, Mayer K, Mayne A, McGrew B, McKinney S M, McLeavey C, McMillan P, McNeil J, Medina D, Mehta A, Menick J, Metz L, Mishchenko A, Mishkin P, Monaco V, Morikawa E, Mossing D, Mu T, Murati M, Murk O, Mély D, Nair A, Nakano R, Nayak R, Neelakantan A, Ngo R, Noh H, Ouyang L, O'Keefe C, Pachocki J, Paino A, Palermo J, Pantuliano A, Parascandolo G, Parish J, Parparita E, Passos A, Pavlov M, Peng A, Perelman A, de Avila Belbute Peres F, Petrov M, de Oliveira Pinto H P, Pokorny M, Pokrass M, Pong V H, Powell T, Power A, Power B, Proehl E, Puri R, Radford A, Rae J, Ramesh A, Raymond C, Real F, Rimbach K, Ross C, Rotsted B, Roussez H, Ryder N, Saltarelli M, Sanders T, Santurkar S, Sastry G, Schmidt H, Schnurr D, Schulman J, Selsam D, Sheppard K, Sherbakov T, Shieh J, Shoker S, Shyam P, Sidor S, Sigler E, Simens M, Sitkin J, Slama K, Sohl I, Sokolowsky B, Song Y, Staudacher N, Such F P, Summers N, Sutskever I, Tang J, Tezak N, Thompson M B, Tillet P, Tootoonchian A, Tseng E, Tuggle P, Turley N, Tworek J, Uribe J F C, Vallone A, Vijayvergiya A, Voss C, Wainwright C, Wang J J, Wang A, Wang B, Ward J, Wei J, Weinmann C J, Welihinda A, Welinder P, Weng J Y, Weng L L, Wiethoff M, Willner D, Winter C, Wolrich S, Wong H, Workman L, Wu S, Wu J, Wu M, Xiao K, Xu T, Yoo S, Yu K, Yuan Q M, Zaremba W, Zellers R, Zhang C, Zhang M, Zhao S J, Zheng T H, Zhuang J, Zhuk W and Zoph B. 2024. GPT-4 Technical Report [EB/OL]. [2024-09-06]. https://arxiv.org/pdf/2303.08774.pdf

Anil R, Dai A M, Firat O, Johnson M, Lepikhin D, Passos A, Shakeri S, Taropa E, Bailey P, Chen Z F, Chu E, Clark J H, El Shafey L, Huang Y P, Meier-Hellstern K, Mishra G, Moreira E, Omernick M, Robinson K, Ruder S, Tay Y, Xiao K F, Xu Y Z, Zhang Y J, Abrego G H, Ahn J, Austin J, Barham P, Botha J, Bradbury J, Brahma S, Brooks K, Catasta M, Cheng Y, Cherry C, Choquette-Choo C A, Chowdhery A, Crepy C, Dave S, Dehghani M, Dev S, Devlin J, Díaz M, Du N, Dyer E, Feinberg V, Feng F, Fienberg V, Freitag M, Garcia X, Gehrmann S, Gonzalez L, Gur-Ari G, Hand S, Hashemi H, Hou L, Howland J, Hu A, Hui J, Hurwitz J, Isard M, Ittycheriah A, Jagielski M, Jia W H, Kenealy K, Krikun M, Kudugunta S, Lan C, Lee K, Lee B, Li E, Li M, Li W, Li Y G, Li J, Lim H, Lin H Z, Liu Z T, Liu F, Maggioni M, Mahendru A, Maynez J, Misra V, Moussalem M, Nado Z, Nham J, Ni E, Nystrom A, Parrish A, Pellat M, Polacek M, Polozov A, Pope R, Qiao S Y, Reif E, Richter B, Riley P, Ros A C, Roy A, Saeta B, Samuel R, Shelby R, Slone A, Smilkov D, So D R, Sohn D, Tokumine S, Valter D, Vasudevan V, Vodrahalli K, Wang X Z, Wang P D, Wang Z R, Wang T, Wieting J, Wu Y H, Xu K, Xu Y H, Xue L T, Yin P C, Yu J H, Zhang Q, Zheng S, Zheng C, Zhou W K, Zhou D, Petrov S and Wu Y H. 2023. PaLM 2 Technical Report [EB/OL]. [2024-09-06]. https://arxiv.org/pdf/2305.10403.pdf

Bain M, Nagrani A, Varol G and Zisserman A. 2021. Frozen in time: a joint video and image encoder for end-to-end retrieval//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 1708-1718 [DOI: 10.1109/ICCV48922.2021.00175]

Chen J, Lyu Z Y, Wu S W, Lin K Q, Song C N, Gao D F, Liu J W, Gao Z T, Mao D X and Shou M Z. 2024a. VideoLLM-online: online video large language model for streaming video//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 18407-18418 [DOI: 10.1109/CVPR52733.2024.01742]

Chen L, Wei X L, Li J S, Dong X Y, Zhang P, Zang Y H, Chen Z H, Duan H D, Lin B, Tang Z Y, Yuan L, Qiao Y, Lin D H, Zhao F and Wang J Q. 2024b. ShareGPT4Video: improving video understanding and generation with better captions//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #614

Fan C Y, Zhang X F, Zhang S, Wang W S, Zhang C and Huang H.

2019. Heterogeneous memory enhanced multimodal attention model for video question answering//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 1999-2007 [DOI: 10.1109/CVPR.2019.00210]

Fan H Q, Murrell T, Wang H, Alwala K V, Li Y H, Li Y L, Xiong B, Ravi N, Li M, Yang H C, Malik J, Girshick R, Feiszli M, Adcock A, Lo W Y and Feichtenhofer C. 2021. PyTorchVideo: a deep learning library for video understanding//Proceedings of the 29th ACM International Conference on Multimedia. [s.l.]: Association for Computing Machinery: 3783-3786 [DOI: 10.1145/3474085.3478329]

Fan Y, Ma X J, Wu R J, Du Y T, Li J Q, Gao Z and Li Q. 2024. VideoAgent: a memory-augmented multimodal agent for video understanding//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 75-92 [DOI: 10.1007/978-3-031-72670-5_5]

Feng Z P, Zhang Y, Li H, Wu B, Liao J Y, Liu W Q, Lang J, Feng Y, Wu J and Liu Z Z. 2025. TEaR: improving LLM-based machine translation with systematic self-refinement//Findings of the Association for Computational Linguistics. Albuquerque, New Mexico, USA: Association for Computational Linguistics: 3922-3938 [DOI: 10.18653/v1/2025.findings-naacl.218]

Fu C Y, Dai Y H, Luo Y D, Li L, Ren S H, Zhang R R, Zhou C Y, Shen Y H, Zhang M D, Chen P X, Li Y W, Lin S H, Zhao S R, Li K, Xu T, Zheng X W, Chen E H, Shan C F, He R and Sun X. 2024. Video-MME: the first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis [EB/OL]. [2024-09-06]. https://arxiv.org/pdf/2405.21075.pdf

Fu R G, Li B, Gao Y H and Wang P. 2016. Content-based image retrieval based on CNN and SVM//Proceedings of the 2nd IEEE International Conference on Computer and Communications. Chengdu, China: IEEE: 638-642 [DOI: 10.1109/CompComm.2016.7924779]

Fukui A, Park D H, Yang D, Rohrbach A, Darrell T and Rohrbach M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding//Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas, USA: Association for Computational Linguistics: 457-468 [DOI: 10.18653/v1/D16-1044]

Gao P, Jiang Z K, You H X, Lu P, Hoi S C H, Wang X G and Li H S. 2019. Dynamic fusion with intra-and inter-modality attention flow for visual question answering//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 6639-6648 [DOI: 10.1109/CVPR.2019.00680]

He B, Li H D, Jang Y K, Jia M L, Cao X F, Shah A, Shrivastava A and Lim S N. 2024. MA-LMM: memory-augmented large multimodal model for long-term video understanding//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13504-13514 [DOI: 10.1109/CVPR52733.2024.01282]

Huang Q Q, Xiong Y, Rao A Y, Wang J Z and Lin D H. 2020. MovieNet: a holistic dataset for movie understanding//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 709-727 [DOI: 10.1007/978-3-030-58548-8_41]

Islam M M, Ho N, Yang X T, Nagarajan T, Torresani L and Bertasius G. 2024. Video ReCap: recursive captioning of hour-long videos//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 18198-18208 [DOI: 10.1109/CVPR52733.2024.01723]

Jang Y, Song Y L, Yu Y, Kim Y and Kim G. 2017. TGIF-QA: toward spatio-temporal reasoning in visual question answering//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 1359-1367 [DOI: 10.1109/CVPR.2017.149]

Jin P, Takanobu R, Zhang W C, Cao X C and Yuan L. 2024. Chat-UniVi: unified visual representation empowers large language models with image and video understanding//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13700-13710 [DOI: 10.1109/CVPR52733.2024.01300]

Kahatapitiya K, Ranasinghe K, Park J and Ryoo M S. 2025. Language repository for long video understanding//Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: OpenReview.net

Kim W, Choi C, Lee W and Rhee W. 2024. An image grid can be worth a video: zero-shot video question answering using a VLM. IEEE Access, 12: 193057-193075 [DOI: 10.1109/ACCESS.2024.3517625]

Lavee G, Rivlin E and Rudzsky M. 2009. Understanding video events: a survey of methods for automatic interpretation of semantic occurrences in video. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 39(5): 489-504 [DOI: 10.1109/TSMCC.2009.2023380]

Lin B, Ye Y, Zhu B, Cui J, Ning M, Jin P, and Yuan L. 2024. Video-LLaVA: learning united visual representation by alignment before projection//Proceedings of 2024 Conference on Empirical Methods in Natural Language Processing. Florida, USA: ACL: 5971-5984

Le T M, Le V, Venkatesh S and Tran T. 2020. Hierarchical conditional relation networks for video question answering//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9969-9978 [DOI: 10.1109/CVPR42600.2020.00999]

Lei J, Yu L C, Berg T and Bansal M. 2020. TVQA+: spatio-temporal grounding for video question answering//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [s. l.]: Association for Computational Linguistics: 8211-8225 [DOI: 10.18653/v1/2020.acl-main.730]

Li K C, He Y N, Wang Y, Li Y Z, Wang W H, Luo P, Wang Y L,

Wang L M and Qiao Y. 2024a. VideoChat: chat-centric video understanding [EB/OL]. [2024-09-06]. https://arxiv.org/pdf/2305.06355.pdf

Li Y W, Wang C Y and Jia J Y. 2024c. LLaMA-VID: an image is worth 2 tokens in large language models//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 323-340 [DOI: 10.1007/978-3-031-72952-2_19]

Liu R Y, Li C, Tang H R, Ge Y X, Shan Y and Li G. 2024b. ST-LLM: large language models are effective temporal learners//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 1-18 [DOI: 10.1007/978-3-031-72998-0_1]

Liu Y, Duan H D, Zhang Y H, Li B, Zhang S Y, Zhao W B, Yuan Y K, Wang J Q, He C H, Liu Z W, Chen K and Lin D H. 2024c. MMBench: is your multi-modal model an all-around player?//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 216-233 [DOI: 10.1007/978-3-031-72658-3_13]

Luo R P, Zhao Z W, Yang M, Dong J W, Li D, Wang T, Qiu M H, Hu L M and Wei Z Y. 2024. Valley: video assistant with large language model enhanced ability//Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria: OpenReview.net

Maaz M, Rasheed H, Khan S and Khan F S. 2024. Video-ChatGPT: towards detailed video understanding via large vision and language models//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. [s. l.]: Association for Computational Linguistics: 12585-12602

Mo Y H, Qin H, Dong Y S, Zhu Z Y and Li Z L. 2024. Large language model (LLM) AI text generation detection based on transformer deep learning algorithm [EB/OL]. [2024-09-06]. https://arxiv.org/pdf/2405.06652.pdf

Pan J T, Lin Z Y, Ge Y Y, Zhu X T, Zhang R R, Wang Y, Qiao Y and Li H S. 2023. Retrieving-to-answer: zero-shot video question answering with frozen large language models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Paris, France: IEEE: 272-283 [DOI: 10.1109/ICCVW60793.2023.00035].

Qian R, Dong X Y, Zhang P, Zang Y H, Ding S R, Lin D H and Wang J Q. 2024. Streaming long video understanding with large language models//Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #3792

Singh G and Cuzzolin F. 2016. Untrimmed video classification for activity detection: submission to ActivityNet challenge [EB/OL]. [2024-09-06]. https://arxiv.org/pdf/1607.01979.pdf

Song E X, Chai W H, Wang G H, Zhang Y C, Zhou H Y, Wu F Y, Chi H Z, Guo X, Ye T, Zhang Y T, Lu Y, Hwang J N and Wang G A. 2024. MovieChat: from dense token to sparse memory for long video understanding//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 18221-18232 [DOI: 10.1109/CVPR52733.2024.01725]

Tang Y L, Bi J, Xu S T, Song L C, Liang S S, Wang T, Zhang D A, An J, Lin J Y, Zhu R Y, Vosoughi A, Huang C, Zhang Z L, Liu P X, Feng M Q, Zheng F, Zhang J G, Luo P, Luo J B and Xu C L. 2023. Video understanding with large language models: a survey [EB/OL]. [2024-09-06]. https://arxiv.org/pdf/2312.17432.pdf

Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E and Lample G. 2023. LLAMA: open and efficient foundation language models [EB/OL]. [2024-09-06]. https://arxiv.org/pdf/2302.13971.pdf

Trummer I. 2024. Large language models: principles and practice//Proceedings of the 40th IEEE International Conference on Data Engineering (ICDE). Utrecht, the Netherlands: IEEE: 5354-5357 [DOI: 10.1109/ICDE60146.2024.00404]

Wang X, Zhang Y, Zohar O, and Yeung-Levy S. 2024. VideoAgent: long-form video understanding with large language model as agent//Proceedings of 2024 European Conference on Computer Vision. Cham: Springer Nature, Switzerland: 58-76 [DOI: 10.1007/978-3-031-72989-8_4]

Wei F S, Keeling R, Huber-Fliflet N, Zhang J P, Dabrowski A, Yang J C, Mao Q and Qin H. 2023. Empirical study of LLM fine-tuning for text classification in legal document review//Proceedings of 2023 IEEE International Conference on Big Data (BigData). Sorrento, Italy: IEEE: 2786-2792 [DOI: 10.1109/BigData59044.2023.10386911]

Weng Y, Han M, He H, Chang X, and Zhuang B. 2024. LongVLM: efficient long video understanding via large language models//Proceedings of 2024 European Conference on Computer Vision. Milan, Italy: European Computer Vision Association: 453-470 [DOI: 10.1007/978-3-031-70070-1_25]

Xiao J B, Shang X D, Yao A and Chua T S. 2021. NExT-QA: next phase of question-answering to explaining temporal actions//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 9772-9781 [DOI: 10.1109/CVPR46437.2021.00965]

Xu D J, Zhao Z, Xiao J, Wu F, Zhang H W, He X N and Zhuang Y T. 2017. Video question answering via gradually refined attention over appearance and motion//Proceedings of the 25th ACM International Conference on Multimedia. Mountain View, USA: Association for Computing Machinery: 1645-1653 [DOI: 10.1145/3123266.3123427]

Xu L, Zhao Y L, Zhou D Q, Lin Z J, Ng S K and Feng J S. 2024. PLLaVA: parameter-free LLaVA extension from images to videos for video dense captioning [EB/OL]. [2024-09-06]. https://arxiv.org/pdf/2404.16994.pdf

Yang Z, He X W, Wu J H, Wang X and Zhao Y. 2022. Edge computing technologies for streaming video analytics. Scientia Sinica Informa-

tionis，52(1)：1-53（杨铮，贺骁武，吴家行，王需，赵毅. 2022. 面向实时视频流分析的边缘计算技术. 中国科学：信息科学），52(1)：1-53［DOI：10.1360/SSI-2021-0133］

Ye Q L，Yu Z T，Shao R，Xie X Y，Torr P and Cao X C. 2024a. CAT：enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios//Proceedings of the 18th European Conference on Computer Vision. Milan，Italy：Springer：146-164［DOI：10.1007/978-3-031-72684-2_9］

Ye X B，Gan Y K，Huang X K，Ge Y X and Tang Y S. 2024b. VoCo-LLaMA：towards vision compression with large language models［EB/OL］.［2024-09-06］. https://arxiv.org/pdf/2406.12275.pdf

You Z，Wen Z Q，Chen Y F，Li X，Zeng R H，Wang Y W and Tan M K. 2025. Toward long video understanding via fine-detailed video story generation. IEEE Transactions on Circuits and Systems for Video Technology，35(5)：4592-4607［DOI：10.1109/TCSVT.2024.3514820］

Yu Z，Xu D J，Yu J，Yu T，Zhao Z，Zhuang Y T and Tao D C. 2019. ActivityNet-QA：a dataset for understanding complex web videos via question answering//Proceedings of 2019 AAAI Conference on Artificial Intelligence. Honolulu，USA：AAAI Press：9127-9134［DOI：10.1609/aaai.v33i01.33019127］

Zhang C，Lu T X，Islam M M，Wang Z Y，Yu S B，Bansal M and Bertasius G. 2024a. A simple LLM framework for long-range video question-answering//Proceedings of 2024 Conference on Empirical Methods in Natural Language Processing. Miami，USA：Association for Computational Linguistics：21715-21737［DOI：10.18653/v1/2024.emnlp-main.1209］

Zhang D Z，Yu Y H，Dong J H，Li C X，Su D，Chu C H and Yu D. 2024b. MM-LLMs：recent advances in multimodal large language models//Findings of the Association for Computational Linguistics. Bangkok，Thailand：Association for Computational Linguistics：12401-12430［DOI：10.18653/v1/2024.findings-acl.738］

Zhang H，Li X and Bing L D. 2023. Video-LLaMA：an instruction-tuned audio-visual language model for video understanding//Proceedings of 2023 Conference on Empirical Methods in Natural Language Processing：System Demonstrations. Singapore，Singapore：Association for Computational Linguistics：543-553［DOI：10.18653/v1/2023.emnlp-demo.49］

Zhang H J，Wang Y Q，Tang Y S，Liu Y，Feng J S，Dai J F and Jin X J. 2024c. Flash-VStream：memory-based real-time understanding for long video streams［EB/OL］.［2024-09-06］. https://arxiv.org/pdf/2406.08085.pdf

Zhang P Y，Zhang K C，Li B，Zeng G T，Yang J K，Zhang Y H，Wang Z Y，Tan H R，Li C Y and Liu Z W. 2024d. Long context transfer from language to vision//Proceedings of the 13th International Conference on Learning Representations. Singapore，Singapore：OpenReview.net

Zhang S F，Zhai J H，Xie B J，Zhan Y and Wang X. 2019. Multimodal representation learning：advances，trends and challenges//Proceedings of 2019 International Conference on Machine Learning and Cybernetics（ICMLC）. Kobe，Japan：IEEE：1-6［DOI：10.1109/ICMLC48188.2019.8949228］

Zhao Z，Yang Q F，Cai D，He X F and Zhuang Y T. 2017. Video question answering via hierarchical spatio-temporal attention networks//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne，Australia：AAAI Press：3518-3524

Zhou L W，Xu C L and Corso J. 2018. Towards automatic learning of procedures from web instructional videos//Proceedings of 2018 AAAI Conference on Artificial Intelligence. New Orleans，USA：AAAI Press［DOI：10.1609/aaai.v32i1.12342］

作者简介

谢君琳，女，博士研究生，主要研究方向为多模态大模型。
E-mail：223010150@link.cuhk.edu.cn
李冠彬，通信作者，男，教授，主要研究方向为视觉理解与生成、多模态大模型、具身智能。
E-mail：liguanbin@mail.sysu.edu.cn
张锐斐，男，博士研究生，主要研究方向为多模态大模型。
E-mail：223010140@link.cuhk.edu.cn