arXiv:2303.03131v2 [cs.CV] 8 Mar 2023

视频问答　　基于 CLIP 引导的视觉-文本注意力进行回答

Shuhong Ye[⋆$]、*Weikai Kong*[⋆$]、*Chenglin Yao*[⋆]、*Jianfeng Ren*[⋆†]、*Xudong Jiang*[‡]

[⋆]中国宁波诺丁汉大学计算机科学学院

[†]中国宁波诺丁汉大学诺丁汉宁波中国卓越研究与创新中心，中国宁波诺丁汉大学，中国

[‡]南洋理工大学电气与电子工程学院

摘要

视频与文本的跨模态学习在视频问答（VideoQA）中发挥着关键作用。本文提出一种视觉文本注意力机制，利用在大量通用域图像-语言对上训练的对比式语言-图像预训练（CLIP），引导视频问答的跨模态学习。具体而言，我们首先使用 TimeSformer 提取目标应用领域的视频特征，并使用 BERT 提取文本特征；随后，通过特定领域学习，利用 CLIP 从通用知识域提取一组视觉文本特征。接着，我们提出一种跨域学习方法，用于提取目标域与通用域中视觉特征和语言特征之间的注意力信息。最后，将 CLIP 引导的视觉文本特征集合进行融合，以预测答案。在 MSVD-QA 和 MSRVTT-QA 数据集上的评估结果表明，所提出的方法优于当前最先进的方法。

索引词— 视频问答、CLIP、跨模态学习、跨域学习

1. 引言

视频问答（VideoQA）在视觉-语言导航领域 [1], 多媒体推荐 [2], 和通信系统中日益流行 [3]。其目标是正确回答有关视频的问题，这要求深入理解视频场景、问题语义以及细粒度的视觉-语言对齐。研究人员已提出许多跨模态学习方法 [4–7]。然而，由于训练数据有限，答案空间中的一些语言概念没有对应的视频样本，导致缺乏语言监督。这一问题阻碍了语言特征与相应视觉特征之间的准确配对，因而限制了现有模型的跨模态学习能力 [8–14]。

改进视频问答模型主要有两个方向。1）利用循环神经网络 [9, 10], 图神经网络 [11, 14, 15], 条件关系网络 [13],或基于注意力的模型 [8, 12]，从经过标注的视频-文本对中挖掘更深层的相关性。2）从大规模预训练视觉-语言模型中引入通用域知识，以克服语言监督不足的挑战，例如例如，HERO [16],ClipBert [6], 和 JustAsk[17]。预训练模型借助额外的通用域知识，能够描述目标应用域中未见过的答案，从而显著提升下游视频问答任务的性能，并在许多视频问答数据集上取得最先进的结果 [4, 8, 18]。

需要指出的是，目标域与通用域之间可能存在知识差异。未能弥合这种差异可能会导致跨模态知识表示中的冲突。为解决这一问题，我们在 CLIP 模型 [19]的指导下，提出了一种两阶段跨域跨模态学习框架。CLIP 通过大规模图文对进行预训练，因此能够描述更多未见过的答案。它可以对视频帧和文本描述中的显著内容进行特征双向转换。在第一阶段，以领域特定的方式提取视觉语言特征。在目标域中，通过 TimeSformer [20] 提取视频特征，并使用变换器[21, 22]对问题和答案文本进行编码。为引入大规模预训练 CLIP 中的通用领域知识，提取关键视频帧作为显著内容，并将其输入 CLIP 以生成 CLIP引导的视觉特征；同时，将问题和答案输入 CLIP 以生成 CLIP引导的语言特征。在第二阶段的跨域学习中，为弥合知识差异，设计了四个 CLIP 引导的视觉-文本编码器，以利用跨模态跨域注意力信息。最后，融合四组视觉文本特征来预测答案。

我们的贡献主要有三点：1）所提出的方法利用预训练 CLIP从通用领域知识中有效提取视觉和语言特征，并利用 TimeSformer和Transformer从目标域中提取这些特征。

$ 作者贡献均等。本研究部分lǎšu中国国家自然科学基金（项目编号：72071116）资助，部分lǎšu宁波市科学技术局（项目编号：2019B10026和2022Z173）资助。

2）所提出的CLIP引导的视觉-文本注意力机制，有效地将通用领域知识融入目标域，实现视频问答中的跨模态跨领域学习。3）在两个大型基准数据集上的实验结果表明，所提出的方法显著优于最先进的视频问答模型。

2. 提出的方法

2.1. 所提方法概述

所提出的 CLIP 引导跨域视频问答（CCVQA）模型的框图如图 1 所示。该模型由两个主要模块组成。1）域特定学习。首先以域特定的方式提取视觉和语言特征。具体而言，采用 TimeSformer [20] 从目标视频问答域的视频序列中提取视觉特征。为融入通用领域知识，首先选取视频的关键帧，并将其输入 CLIP [19]的图像流，以提取与语言监督兼容的视觉特征。同样，Transformer[21, 22] 和 CLIP [19] 的文本流分别以问题句子和提示词作为输入，从而提取目标域和通用域中的文本特征。2）跨域学习。在分别提取目标域或通用域中的视觉和语言特征后，通过跨域学习整合这四组特征。我们设计了四个 CLIP 引导的视觉-文本编码器，通过注意力机制实现跨领域跨模态学习。最后，通过多层感知机（MLP）融合提取的四组特征，并采用答案解码器 [10, 13] 得出正确答案。

2.2. 特定领域学习

给定一个需要回答问题的视频，视觉特征提取、语言特征提取和视觉-语言对齐是回答问题的三个关键步骤[8–14], , 而通用领域知识的缺乏往往使开放式视频问答极具挑战性 [6, 23, 24]。为应对这一挑战，所提出的方法中集成了 CLIP [19]，以融入通用领域知识。所提出的 CCVQA 模型由两个视觉编码器和两个语言编码器组成。

在目标域中，使用 12 层的 TimeSformer [20]对嵌入每个视频帧中的时空信息进行编码。从视频中随机采样 16 帧，以尽可能保留更多信息，同时降低计算负载。TimeSformer 将每帧划分为 N_v 个图像块，并通过线性投影层生成图像块标记。添加可学习的位置嵌入后，将这些词元输入注意力块，在时间和空间维度上执行自注意力。一个时序融合层沿时间维度聚合注意力块的输出，以获得序列特征 $\boldsymbol{H}_v = [\boldsymbol{v}_{cls}, \boldsymbol{v}_1, ..., \boldsymbol{v}_{N_v}]$，其中 $\boldsymbol{v}_i \in \mathbb{R}^d$。 d 是视频特征的维度。 $\boldsymbol{v}_{cls}$ 是分类标记。

在目标域中，使用一个 6 层 BERT-base 模型 [25] 对语言特征进行编码 $\boldsymbol{H}_q$。给定由 N_q 个词元组成的输入问题，BERT 会依次对输入执行自注意力计算，并输出一个嵌入序列 $\boldsymbol{H}_q = \left[\boldsymbol{q}_{cls}, \boldsymbol{q}_1, ..., \boldsymbol{q}_{N_q}\right] \in \mathbb{R}^{N_q \times d}$， $\boldsymbol{q}_i \in \mathbb{R}^d$，其中 $\boldsymbol{q}_{cls}$ 是[CLS] 词元。 d 是问题特征的维度，与视频特征的维度相同。与视频编码器类似，会将可学习的位置嵌入添加到文本词元中。

为了融入通用领域知识，我们设计了一个 CLIP 引导的视觉编码器。由于使用 CLIP 对整个视频序列进行编码非常耗时，因此仅根据颜色直方图对比度选择关键帧[1]，并使用 CLIP ViT-B/32 模型进行编码 [22]。CLIP 在从互联网收集的 4 亿个图文对上进行预训练，以将视觉文本特征映射到联合嵌入空间，从而捕捉编码后的图像特征与语言特征之间的相似性，并作为额外的语言监督，引导后续的文本-图像对齐。CLIP 首先通过线性投影将关键帧转换为 N_k 图像块。添加位置嵌入后，对这些图像块以及[CLS] 标记执行自注意力，输出关键帧的序列特征$\boldsymbol{H}_k = [\boldsymbol{k}_{cls}, \boldsymbol{k}_1, ..., \boldsymbol{k}_{N_k}]$， $\boldsymbol{k}_i \in \mathbb{R}^d$。

CLIP 引导的文本编码器旨在利用通用领域知识提取语言特征。使用给定的问题生成提示，从问题中提取关键信息，并从给定的答案空间中选取可能的答案。请注意，我们处理的是开放式问答，因此没有答案选项，只有一个非常庞大的 C 可能答案空间。与 CLIP 引导的视觉编码器类似，CLIP 语言编码器会对提示句进行分词，将其线性投影到嵌入空间，并添加位置嵌入，同时执行层归一化操作。随后，对句子嵌入执行自注意力，并生成一批特征向量 $\boldsymbol{H}_t \in \mathbb{R}^{C \times N_t \times w}$，其中 N_t 表示每个编码后提示特征 $[\boldsymbol{t}_1, ..., \boldsymbol{t}_{N_t}]$的序列长度， $\boldsymbol{t}_i \in \mathbb{R}^w$而 w 表示提示特征的维度。

2.3. 跨域学习

以特定领域的方式提取视觉特征 $\boldsymbol{H}_v$, $\boldsymbol{H}_k$ 和语言特征 $\boldsymbol{H}_q$, $\boldsymbol{H}_t$ 后，我们设计了四个跨域跨模态视觉-文本编码器，用于捕捉这些特征之间的注意力信息。我们设计了一个 6 层共享权重Transformer $\mathcal{T}$，用于对 $\boldsymbol{H}_q$ 与 $\boldsymbol{H}_v$之间以及 $\boldsymbol{H}_q$ 与$\boldsymbol{H}_k$之间的注意力信息进行建模。它直接将拼接后的多模态特征作为输入，对配对视觉特征执行自注意力

[1]代码可在以下地址获取https://github.com/keplerlab/katna

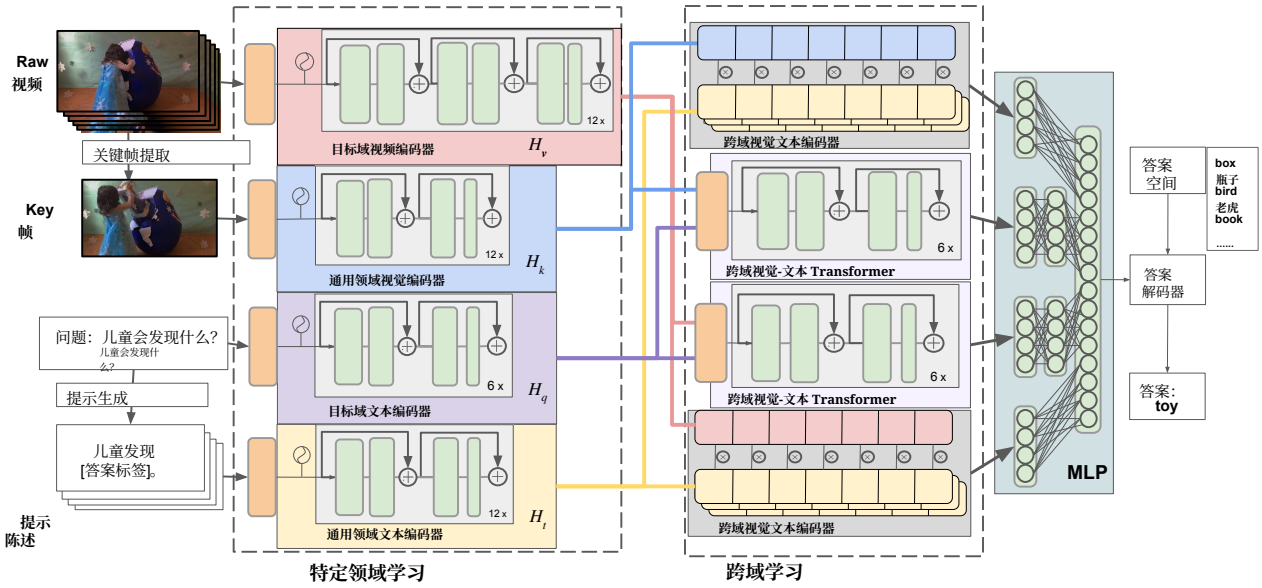

图 1。所提出的 CCVQA 概述。它由两个主要模块组成。1）特定领域学习，包括一个用于提取时空特征 $\boldsymbol{H}_v$的视频编码器、一个用于编码来自目标域的问题描述 $\boldsymbol{H}_q$ 的文本编码器、一个在语言监督下编码对象表示 $\boldsymbol{H}_k$ 的 CLIP 引导的视频关键帧编码器，以及一个在通用域中通过视觉监督生成语言特征 $\boldsymbol{H}_t$ 的 CLIP 引导的候选答案编码器。2）跨域学习，包括四组 CLIP 引导的视觉-文本编码器，通过注意力机制建模跨域跨模态特征交互。最后，使用多层感知机融合这些特征，并将其输入解码器以预测答案。

和语言特征，并使用多层感知机生成类别标记，格式如公式(1)和(2)所示。

$$\boldsymbol{H}_{qv} = \boldsymbol{W}_{qv}\mathcal{T}(\boldsymbol{H}_q;\boldsymbol{H}_v) + \boldsymbol{b}_{qv}, \quad (1)$$

$$\boldsymbol{H}_{qk} = \boldsymbol{W}_{qk}\mathcal{T}(\boldsymbol{H}_q;\boldsymbol{H}_k) + \boldsymbol{b}_{qk}, \quad (2)$$

其中，类别标记 $\boldsymbol{H}_{qv}$，$\boldsymbol{H}_{qk} \in \mathbb{R}^{\boldsymbol{a}}$。$\boldsymbol{W}_{qv}$、$\boldsymbol{b}_{qv}$、$\boldsymbol{W}_{qk}$ 和$\boldsymbol{b}_{qk}$ 为可训练参数。

CLIP 编码的提示的大小 $\boldsymbol{H}_t \in \mathbb{R}^{\boldsymbol{a}\times \boldsymbol{N}_k \times w}$ 远大于问题特征的大小 $\boldsymbol{H}_q \in \mathbb{R}^{\boldsymbol{N}_q \times d}$。将 $\boldsymbol{H}_t$ 与 $\boldsymbol{H}_v$ 或 $\boldsymbol{H}_k$ 拼接会产生长序列，而在 Transformer 中计算此类序列可能耗费数个 GPU 年 [19]。因此，为了通过 Transformer 中的自注意力简化注意力信息的提取，我们对提示文本特征 $\boldsymbol{H}_t$ 与视觉特征 $\boldsymbol{H}_v$或 $\boldsymbol{H}_k$之间的交互执行点积。 [CLS] 词元 $\boldsymbol{H}_t$、$\boldsymbol{H}_v$ 和 $\boldsymbol{H}_k$ 被线性投影为大小分别为 $\mathbb{R}^{\boldsymbol{a}\times w}$、$\mathbb{R}^{w}$ 和 $\mathbb{R}^{w}$的向量。然后，这两个向量被投影并相乘，以生成另外两个类别标记，

$$\boldsymbol{H}_{tv} = P(\boldsymbol{H}_t) \odot P(\boldsymbol{H}_v), \quad (3)$$

$$\boldsymbol{H}_{tk} = P(\boldsymbol{H}_t) \odot P(\boldsymbol{H}_k), \quad (4)$$

其中 $\boldsymbol{H}_{tv}$，$\boldsymbol{H}_{tk} \in \mathbb{R}^{a}$、$\odot$表示点积，而 $P(\cdot)$ 表示投影。

随后，利用加权多头融合整合四个词元 $\boldsymbol{H}_{qv}$、$\boldsymbol{H}_{qk}$、$\boldsymbol{H}_{tv}$ 和 $\boldsymbol{H}_{qk}$。对于每个词元，利用一个线性层进行特征对齐，使得词元中的每一项表示答案空间中某一类别的置信度。最终融合的跨域跨模态特征 $\boldsymbol{H}$ 如公式（5）所示导出。

$$\boldsymbol{H} = \boldsymbol{W}_1\boldsymbol{H}_{qv} + \boldsymbol{W}_2\boldsymbol{H}_{qk} + \boldsymbol{W}_3\boldsymbol{H}_{tv} + \boldsymbol{W}_4\boldsymbol{H}_{tk} + \boldsymbol{b} \quad (5)$$

其中 $\boldsymbol{W}_1$、$\boldsymbol{W}_2$、$\boldsymbol{W}_3$、$\boldsymbol{W}_4$ 和 $\boldsymbol{b}$ 是可训练的融合权重。

3. 实验结果

3.1. 实验设置

本文提出的 CCVQA 在两个用于开放式视频问答的公开数据集上进行评估：MSVD-QA [8]和 MSRVTT-QA [8]，这两个数据集分别由原始数据集 MSVD [26] 和 MSRVTT [27], 中的视频生成，并结合自动生成的问题—答案标注对构成。MSVD-QA 共包含 1.9K 个视频和 50k 组问答对，MSRVTT-QA 共包含 10K 个视频和 243k 组问答对。根据问题的开头词，问题分为五种类型：“what”“who”“when”“where”和“how”。答案均为单词形式的名词或动词，用于描述视频中的具体物体或抽象概念。本文遵循标准的数据集划分方式 [8]。在 MSVD-QA 数据集上，61% 的数据用于训练，13% 用于验证，26% 用于测试。在 MSRVTT-QA 数据集上，训练集、验证集和测试集的划分比例

分别为 65%、5% 和 30%。对于开放式问题的答案，MSVD-QA 使用 2, 423 个候选答案，MSRVTT-QA 使用 1, 500 个。本文提出的方法与最先进模型进行了比较。AMU [8], Co-mem [9], HME [10],HGA [11], SSML [28], QUEST [12], HCRN [13], STN [29],DualVGR [15], 和 HQGA [14] 被选为具有代表性的跨模态学习方法。ClipBERT [6], CoMVT [23], SSRea [24], 和 VQA-T [17]被选为典型的预训练模型增强方法。

带有预训练权重的 CLIP ViT-B/32 架构取自 OpenAI 的官方发布版本[2]。所有视频帧均调整为 224×224。在语言特征编码方面，问题被转换为陈述句形式，MSRVTT-QA 数据集包含 1,500 个句子，MSVD-QA 数据集包含 2, 423 个句子，分别与答案候选项的数量相同。所有实验均在单个 NVIDIA V100 GPU 上进行。初始学习率设置为5×10^{-5}，并在后续训练轮数中线性衰减。两个数据集训练过程的批量大小均为 24 。采用带有权重衰减 1×10^{-3} 的 AdamW 优化器。模型在 MSVD-QA 上经过 15 个训练轮数、在 MSRVTT-QA 上经过 10 个训练轮数后收敛，分别消耗120 个 GPU 小时和 400个 GPU 小时。

3.2. 与最先进方法的比较 ls

方法	MSVD-QA	MSRVTT-QA
AMU [8]	32.0	32.5
Co-mem [9]	31.7	32.0
HME [10]	33.7	33.0
HGA [11]	34.7	35.5
SSML [28]	35.13	35.06
QUEST [12]	36.1	34.6
HCRN [13]	36.1	35.6
TSN [29]	36.7	35.4
DualVGR [15]	39.03	35.52
HQGA [14]	41.2	38.6
ClipBERT [6]	-	37.4
CoMVT [23]	42.6	39.5
SSRea [24]	45.5	41.6
Ours	**46.6**	**42.4**

表 1。MSRVTT-QA 和 MSVD-QA 数据集上与最先进方法在 Top-1准确率（%）方面的比较。

与最先进方法的比较结果列于表1中。结果证实，我们的模型以显著优势胜过所有对比方法。所提出的方法将 Top-1准确率从 SSRea [24] 在 MSVD-QA 数据集上取得的45。5% 提升至 46.6%，并将 MSRVTT-QA 数据集上的准确率从41。6% 提升至 42。4% ，分别为。与 HQGA [14] 和 DualVGR [15]等未使用预训练模型的视频问答方法相比，所提出的模型在 MSVD-QA 数据集上的准确率高出近5。4% ，在 MSRVTT-QA 数据集上高出3。8% ，这验证了利用额外通用领域知识的有效性。与 SSRea [24]和 CoMVT [23], 等其他预训练模型相比，所提出的模型在 MSVD-QA 数据集上的准确率进一步提高约1。1% ，在 MSRVTT-QA 数据集上进一步提高0。8%，这表明我们的视觉和语言特征编码器以及CLIP引导的视觉-文本注意力机制是有效的。

3.3. 消融研究

为展示各项贡献带来的性能提升，我们进行了消融研究。我们将提出的 CCVQA 与两个基线进行比较：1）**不使用 CLIP 的 CCVQA**，即不使用 CLIP 提供的语言和视觉线索；2）**不使用跨域学习的 CCVQA**，即仅在目标域／通用域内进行视觉-文本学习。如表2所示，与**不使用 CLIP 的 CCVQA**相比，CCVQA 在 MSVD-QA 和 MSVRTT-QA 数据集上分别取得了 1.9% 和 1.1% 的显著性能提升，证明了我们采用 CLIP 引导设计的有效性。通过提出的跨域学习方法，CCVQA 在 MSVD-QA 和 MSVRTT-QA 数据集上也分别取得了 0.7% 和 0.3% 的性能提升。

方法	MSVD-QA	MSVRTT-QA
CCVQA（不使用 CLIP）	44.7	41.3
CCVQA（不使用跨域学习）	45.9	42.1
CCVQA	**46.6**	**42.4**

表2。在 MSRVTT-QA 和 MSVD-QA 数据集上以 Top-1准确率(%) 衡量的消融研究。

4. 结论

为应对视频问答中语言监督不足这一挑战，本文提出一种由 CLIP 引导的跨域视频文本编码器，将 CLIP 的通用领域知识转换到应用领域。具体而言，预训练 CLIP 利用通用领域知识对关键视频帧和提示文本进行编码。结合由 TimeSformer 提取的视频特征，以及在应用领域中由 BERT 编码的问题文本特征，通过跨域学习，利用点积和共享权重 Transformer，从视觉和文本特征中提取注意力信息。最后，从答案空间中解码出答案。在 MSVD-QA 和 MSRVTT-QA 这两个开放式视频问答数据集上对所提出的 CCVQA 进行了评估，结果表明，与最先进模型相比，该方法的性能持续且显著提升。

[2]CLIP 模型可在以下地址获取：https://github.com/openai/CLIP

参考文献

[1] P. Anderson、Q. Wu、D. Teney、J. Bruce、M. Johnson、N. S¨underhauf、I. Reid、S. Gould 和 A. Van Den Hen- gel，《视觉与语言导航：在真实环境中解读基于视觉的导航指令》，发表于 *CVPR*，2018年，页码 3674–3683。1

[2] J. Chen、H. Zhang、X. He、L. Nie、W. Liu 和 T. Chua，“注意力协同过滤：采用物品级和组件级注意力的多媒体推荐”，载于*SIGIR*，2017年，第 335–344 页。1

[3] M. Bica、K. Huang、K. Byrne 和 J. Fuhrmann，《基于互信息的联合雷达与蜂窝通信系统雷达波形设计》，载于*ICASSP*，2016年，第3671–3675页。1

[4] J. Lei、L. Yu、M. Bansal 和 T. L. Berg，《Tvqa：局部化、组合式视频问答》，载于*EMNLP*，2018年，第1369–1379页。1

[5] J. Lei、T. L. Berg 和 M. Bansal，“Tvqa+：面向视频问答的时空定位”，载于*ACL*，2020，第 8211–8225 页。

[6] J. Lei、L. Li、L. Zhou、Z. Gan、T. Berg、M. Bansal 和 J. Liu，“少即是多：通过稀疏采样进行视频与语言学习”*CVPR*，2021，第 7331–7341 页。1，2，4

[7] A. Guzhov、F. Raue、J. Hees 和 A. Dengel，“Audioclip：将 CLIP 扩展到图像、文本和音频”，载于 *ICASSP*，2022 年，第 976–980 页。1

[8] D. Xu、Z. Zhao、J. Xiao、F. Wu、H. Zhang、X. He 和 Y. Zhuang，“通过逐步细化外观和运动上的注意力实现视频问答”，载于 *ACM MM*，2017年，第 1645–1653 页。1，2，3，4

[9] J. Gao、R. Ge、K. Chen 和 R. Nevatia，《用于视频问答的运动-外观协同记忆网络》，载于*CVPR*，2018年，第6576–6585页。1，4

[10] C. Fan、X. Zhang、S. Zhang、W. Wang、C. Zhang 和 H. Huang，《用于视频问答的异构记忆增强多模态注意力模型》，载于*CVPR*，2019年，第1999–2007页。1，2，4

[11] P. Jiang 和 Y. Han，“利用异构图对齐进行视频问答推理”，发表于*AAAI*，2020，卷 34，第 11109–11116 页。1，4

[12] J. Gao、Z. Ge、X. Sun 和 H. Wang，“为而治之：问题引导的时空上下文注意力用于视频问答”，发表于*AAAI*，2020，第34卷，第11101–11108页。1，4

[13] T. Le、V. Le、S. Venkatesh 和 T. Tran，“用于视频问答的分层条件关系网络”，发表于*CVPR*，2020，第9972–9981页。1，2，4

[14] J. Xiao、X. Shang、A. Yao 和 T. Chua，《仅为条件图层次结构的视频：用于多粒度问答》，发表于*AAAI*，2022 年，第 2804–2812 页。1，2，4

[15] J. Wang、B. Bao 和 C. Xu，“Dualvgr：一种用于视频问答的双视觉图推理单元，”《*IEEE* 多媒体汇刊》，第 24 卷，第 3369–3380 页，2021 年。1，4[16] L. Li、Y. Chen、Y. Cheng、Z. Gan、L. Yu 和 J. Liu，“Hero：用于视频的层次编码器+ 语言全方位表征预训练”，载于 *EMNLP*，2020，第 2046–2065 页。1[17] A. Yang、A. Miech、J. Sivic、I. Laptev 和 C. Schmid，“Just ask：从数百万段旁白视频中学习回答问题”，载于*CVPR*，2021，第 1686–1697 页。1，4

[18] Y. Kim、C. Kim、一种用于视频问答和检索的联合序列融合模型”，载于 *ECCV*，2018年，第 471–487 页。1

[19] A. Radford、J. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark 等人，“从自然语言监督中学习可迁移视觉模型”，发表于 *ICML*，2021年，第 8748–8763 页。1，2，3

[20] C. Bertasius、H. Wang 和 L. Torresani，“时空理解，时空注意力是否就是你所需要的一切？”，发表于 *ICML*，813–824，第 139 卷，第 4 页。1，2

[21] A. Vaswani、N. Shazeer、N.and Parmar、J. Uszkoreit、L. Jones、A. Gomez、Lu. Kaiser 和 I. Polosukhin，“注意力就是你所需要的一切”，*NeurIPS*，第 30 卷，2017年。1，2

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, 和 N. Houlsby，“一张图像值 16×16 个词：用于大规模图像识别的变换器”，发表于 *ICLR*，2021 年。1，2

[23] P. Sennrich、C. Schmid 和 D. Lagun，《逆转前行：视觉语境化话语》，载于*CVPR*，2021年，第16877–16887页。2，4

[24] W. Yu、H. Zheng、M. Li、L. Ji、L. Wu、N. Xiao 和 N. Duan，《从内部学习：用于视频问答的自驱动孪生采样与推理》，*NeurIPS*，第34卷，第26462–26474页，2021年。2，4

[25] J. Devlin、M. Chang、K. Lee 和 K. Toutanova，《BERT：用于语言理解的深度双向变换器预训练》，*arXiv*预印本 *arXiv:1810.04805*，2018年。2

[26] D. Chen 和 W. Dolan，《收集高度平行的数据用于释义评估》，载于*ACL*，2011年，第190–200页。3

[27] J. 徐、梅、Msr-vtt 和 Y. 鲁，“用于连接视频与语言的大型视频描述数据集”，载于*CVPR*，2016年，第5288–5296页。3

[28] E. Amid、P. Bronstein、R. Pereira 和 J. Anderson，“利用梯度上升进行自监督多模态学习的噪声估计”，发表于*AAAI*，2021，第 35 卷，第 6644–6652 页。4

[29] T. Yang、Z. Zha、H. Xie、M. Wang 和 H. Zhang，“用于视频问答的时空感知管状切换网络”，发表于*ACMMM*，2019年，第 1184–1192 页。4