Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2025, 14(4), 893-905
Published Online July 2025 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2025.144085

Hans汉斯

面向医学视觉问答定位任务的视觉定位与文本－视觉交互注意力机制

郑少义

温州大学计算机与人工智能学院，浙江 温州

收稿日期：2025年6月9日；录用日期：2025年7月11日；发布日期：2025年7月24日

摘　要

医学领域的视觉问答(VQA)任务旨在针对医学图像中的临床问题生成准确答案。尽管现有医学VQA系统已取得显著进展，但在临床外科手术中，准确识别手术区域位置仍至关重要。因此，将视觉问答定位任务(Visual Question Localized-Answering, VQLA)引入临床手术场景，有助于更有效地辅助医生完成对精确定位要求较高的操作。然而，现有VQLA方法多依赖简单注意力机制进行模态融合，缺乏对同一模态内及跨模态特征的深度交互，导致答案区域定位不准确及问题理解不足。为解决上述问题，本文提出一种融合视觉定位与文本－视觉交互的注意力机制(VLTVI Attention)，从通道维度与空间维度对视觉模态特征进行更全面建模，从而实现对答案区域的精准定位。同时，引入分层结构的文本－视觉交互注意力，以加深模型对问题语义的理解，并增强其推理能力。我们在基于MICCAI EndoVis-2017与EndoVis-2018手术视频构建的两个公共VQLA数据集上开展了大量实证研究，验证了所提方法在医学VQLA任务中的性能优越性，并取得新的最先进性能(state-of-the-art)。此外，本文还提供了详尽的消融实验与可视化分析，以验证关键注意力模块的有效性。

关键词

视觉问答定位(VQLA)，注意力机制

Visual Localization and Text-Visual Interaction Attention for Medical Visual Question Localized-Answering Tasks

Shaoyi Zheng

College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou Zhejiang

Received: Jun. 9th, 2025; accepted: Jul. 11th, 2025; published: Jul. 24th, 2025

文章引用: 郑少义. 面向医学视觉问答定位任务的视觉定位与文本-视觉交互注意力机制[J]. 人工智能与机器人研究, 2025, 14(4): 893-905. DOI: 10.12677/airr.2025.144085

Abstract

The VQA in the medical domain aims to predict answers to clinical questions related to medical images. While existing medical VQA systems have made rapid progress, accurate identification of the surgical site's location is crucial in clinical surgical procedures. Therefore, introducing Visual Question Localized-Answering (VQLA) in clinical surgery can better assist healthcare professionals in addressing issues involving precise location operations. However, existing VQLA methods only use simple attention mechanisms to fuse different modality features, lacking sufficient interaction between individual or different modality features, resulting in inadequate localization of answer regions and understanding of questions. To address this issue, we designed a Visual Localization and Text-Visual Interaction (VLTVI) Attention aimed at more comprehensive modeling of visual modality features from channel and spatial dimensions to accurately locate answer regions. Additionally, hierarchical text-visual interaction attention is designed to deepen the model's understanding of questions and strengthen reasoning of answers. To validate our VLTVI, extensive experiments were conducted on two public VQLA datasets based on surgical videos from MICCAI EndoVis-17 and 18, achieving a new state-of-the-art performance. Furthermore, comprehensive ablation studies and visualizations are provided to validate the essential attention modules of our method.

Keywords

Visual Question Localized-Answering, Attention Mechanism

1. 引言

医学视觉问答(Medical Visual Question Answering, Medical VQA)是视觉问答(VQA)技术在医疗领域中的典型应用，旨在基于医学图像和临床相关问题生成准确答案。一个安全、可靠的医学 VQA 系统不仅可有效辅助医生实现快速诊断，还可缩短医学从业人员的培训周期，或帮助患者进行高效地自我诊断。目前，医学 VQA 已在放射学、病理学以及面向视障人士的辅助技术等领域得到广泛应用。

尽管医学 VQA 具有广阔的应用前景，但其功能仍局限于生成自然语言形式的答案。在对目标定位精度要求极高的场景中(如外科手术)，VQA 系统往往受限于自然语言表达的模糊性，仅能给出低精度、泛化性的相对位置描述，难以满足精确定位的实际需求。为弥补这一不足，研究者在 VQA 任务的基础上提出了视觉问答定位任务(Visual Question Localized-Answering, VQLA)，该任务不仅需要回答临床问题，还需提供图像中与问题相关的空间区域信息，从而辅助医生完成对关键区域的精准识别。

现有的医学 VQLA 方法多沿用传统 VQA 的技术范式，即将基于通用 VQA 数据集预训练的模型微调至医学领域的特定任务上。然而，医学图像高度专业化，常包含复杂的解剖结构、病灶及器官信息，与通用 VQA 图像存在显著差异；同时，医学文本中蕴含的大量专业术语也与通用语料存在语义鸿沟。受限于这些模态间的异质性，通用 VQA 方法中常见的特征融合策略(如加权求和、平均、点积或简单注意力机制)在此场景下难以提取高质量的跨模态表示，导致模型性能提升有限。为了解决这个问题，Bai 等人[1]提出了一种新颖的 CAT-ViL 嵌入方法，通过利用协同注意力和门控机制生成改进的跨模态表示。这类方法虽然改善了文本模态的融合效果，但是只将重心放在了两个模态的融合上，忽略了模型对答案位

置的关注和对问题的深入理解。

为进一步突破上述瓶颈，我们将医学 VQLA 任务的核心挑战归纳为两点：(1) 使模型能够自适应地聚焦于答案区域，以提升定位精度；(2) 使模型能够深度理解问题语义，从而增强推理能力。为此，本文提出一种视觉定位与文本–视觉交互注意力机制(Visual Localization and Text-Visual Interaction Attention, VLTVI)，以协同实现上述两方面目标。

综上，本文的主要贡献如下：

- 系统分析了医学 VQLA 任务中存在的关键问题，指出模型亟需具备对答案区域的自适应关注能力与对问题的深度语义理解能力。
- 提出一种新颖的 VLTVI 注意力模块，由视觉定位注意力与文本–视觉交互注意力两部分组成。前者能够精准定位医学图像中与问题相关的区域，实现自适应注意力聚焦；后者通过层次化语义建模与双向交互机制，逐步提升模型对问题的理解能力与答案推理能力。
- 在 EndoVis-2018 VQLA 与 EndoVis-2017 VQLA 两个数据集上对所提方法进行验证，实验结果表明本方法在两个基准任务上均实现了当前最优性能。

2. 方法

2.1. 网络结构

网络的整体架构如图 1 所示。该模型主要由特征提取器、视觉–文本嵌入模块(GVLE)、我们的 VLTVI 模块以及用于空间区域定位和问题回答的定位与分类头组成。具体而言，VLTVI 模块包括自注意力模块、空间注意力模块、通道级注意力模块和分层视觉–文本交互注意力模块。

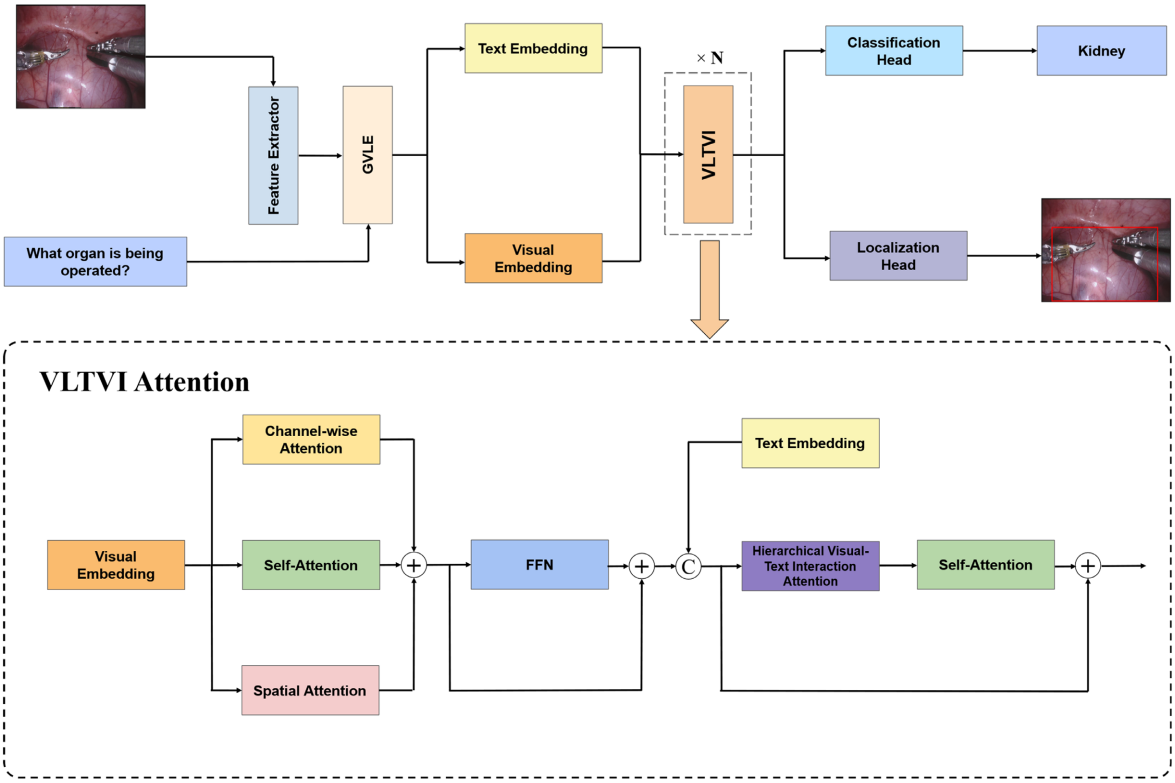

Figure 1. Overview of the model architecture
图 1. 模型整体结构图

特征提取器主要用于提取对应图象和文本的特征用于进一步的处理，首先，给定一幅图像 I 和一个问题 Q，我们首先利用在 ImageNe 上预训练的 ResNet18 作为特征提取器，从 I 中提取图像特征，得到 I'。随后，我们采用视觉－文本嵌入模块模块分别为图像和文本生成嵌入表示：

$$V,T = f_{GVLE}\left(I',Q\right) \tag{2-1}$$

其中 V 和 T 分别表示通过 GVLE 模块获得的视觉和文本嵌入表示，作为后续模块的输入。此处，f_{GVLE} 表示 GVLE 模块中的嵌入操作，能够有效地从输入的文本和图像中提取出文本和视觉嵌入，并通过填充操作确保输出的嵌入具有相同维度的特征。

接下来，我们将视觉和文本嵌入送入 VLTVI 模块，以获得具有更强语义信息的输出特征 X_{out}，并将 X_{out} 输入到定位头和分类头中，最终得到答案和位置信息。

在答案预测任务中，我们采用交叉熵损失函数：

$$L_{CE} = -\sum_{i=1}^{N} y_i \log\left(p_i\right) \tag{2-2}$$

其中，N 表示答案类别的数量，y_i 是真实标签，p_i 代表模型对相应类别的预测概率，对于目标定位任务，我们采用 L_1 损失和 $GIoU$ 损失：

$$\begin{aligned} L_{GIoU} &= 1 - \frac{\left|A \cap B\right|}{\left|A \cup B\right|} + \frac{\left|C \setminus \left(A \cup B\right)\right|}{\left|C\right|} \\ L_1 &= \left|A - B\right| \end{aligned} \tag{2-3}$$

其中，A 和 B 分别表示真实边界框和预测边界框，$\left|\cdot\right|$ 代表面积，C 是最小包围 A 和 B 的矩形区域。

最终损失函数定义如下：

$$L = L_{CE} + L_1 + L_{GIoU} \tag{2-4}$$

2.2. 空间注意力模块

与传统的 VQLA 任务相比，医学领域中的 VQLA 任务会面临更大的挑战，这些挑战源于医学知识的高度专业化和领域的独特性。医学图像通常包含复杂的解剖结构、病变器官或组织，且这些结构在不同的医疗影像技术中，如 X 光片，B 超，CT 或是内窥镜中的形态差异巨大。与一般领域的 VQLA 任务不同，医学 VQLA 不仅要求模型生成正确的答案，还需要根据输入的视觉信息对答案进行精确地定位，这使得问题的复杂度和难度大大增加。上述问题的解决都建立在模型能够准确地聚焦于答案所在的区域。因此，我们提出了一种空间注意力机制，利用平均池化(Average Pooling)和最大池化(Max Pooling)操作来获得空间注意力图，从而使网络能够自适应地关注包含答案的区域，从而提高 VQLA 任务的准确性和可靠性。

空间注意力模块的整体架构如图 2 所示。具体来说，空间注意力机制利用特征图中的空间关系生成空间注意力，使得模型能够集中关注特征图中的显著信息。首先，V 表示通过 GVLE 模块获得的视觉嵌入，并将其重塑为 X_{2D}。然后，沿着 X_{2D} 的通道维度进行平均池化和最大池化操作，并将结果的特征图拼接得到 X'_{2D}，上述的过程可以表示如下：

$$\begin{aligned} X_{avg} &= AvgPool\left(X_{2D}\right), \\ X_{max} &= MaxPool\left(X_{2D}\right), \\ X'_{2D} &= \left[X_{avg} : X_{max}\right] \end{aligned} \tag{2-5}$$

其中 $\left[:\right]$ 表示拼接操作，$AvgPool\left(\cdot\right)$ 和 $MaxPool\left(\cdot\right)$ 分别表示平均池化和最大池化操作。接下来，对 X'_{2D} 应

用一次卷积核大小为 7 × 7 的卷积操作，再经过 Sigmoid 操作，从而获得空间注意力的权重 A_s，这个过程可以表示如下：

$$A_s = \sigma\left(g_{2,1}^{7\times7}\left(X'_{2D}\right)\right) \tag{2-6}$$

其中，$\sigma(\cdot)$ 表示 Sigmoid 函数，$g_{2,1}^{7\times7}(\cdot)$ 表示一个卷积核大小为 7×7，并将通道数从 2 减少到 1 的卷积操作。

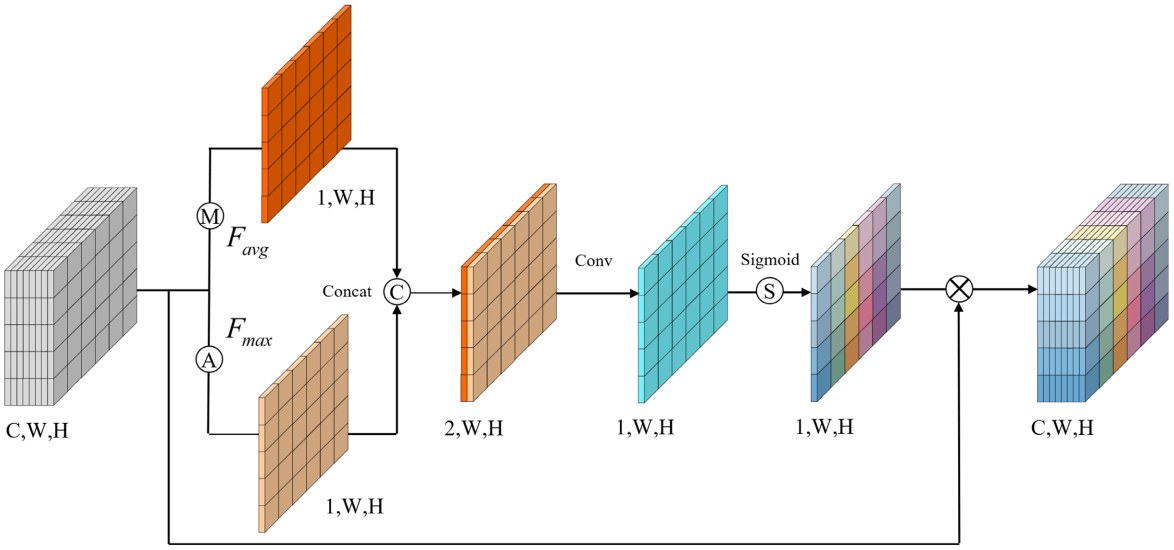

Figure 2. Architecture of the spatial attention module
图 2. 空间注意力模块结构图

最后，将获得的注意力权重 A_s 应用于原始特征 X_{2D}，以得到经过空间注意力加权后的特征 $X'_{spatial}$：

$$X'_{spatial} = A_s \odot X_{2D} \tag{2-8}$$

其中 $\odot$ 表示哈达玛积(Hadamard 积)，即将空间注意力权重应用到每个像素的特征上，增强模型对目标区域的关注。最终，将 $X'_{spatial}$ 重塑为 $X_{spatial}$ 以保持与输入相同的形状，以确保其形状与输入特征一致，从而便于后续的处理。

2.3. 自注意模块

传统的基于卷积的注意力机制受限于卷积的感受野，无法捕获空间长距离依赖，只能捕获局部信息。尤其当答案所对应的物体在图像中占据较大比例时，网络所产生的预测框有可能无法覆盖整个物体，从而出现定位偏差。为此，我们利用自注意力模块来捕获长距离依赖，学习图像中各个位置之间的关联性，即使是图像不同区域的相关信息，也能被有效地捕获和整合。具体而言，给定通过 GVLE 模块获得的视觉嵌入 V 作为输入，我们将自注意力机制分为不同的 h 个自注意力头，每个自注意力头能够独立地学习不同的信息表示，帮助模型捕捉更加复杂的依赖关系，从而增强模型捕捉多种模式和表达能力的能力。对于第 i 个头，我们通过投影操作来获得矩阵 Q_i，K_i 和 V_i，具体的过程如下：

$$\begin{aligned} Q_i &= VW_i^Q, \\ K_i &= VW_i^k, \\ V_i &= VW_i^V \end{aligned} \tag{2-8}$$

W_i^Q，W_i^k 和 W_i^V 都是可学习的参数。第 i 个自注意力头 H_i 的计算过程为：

$$H_i = \text{Softmax}\left(\frac{Q_i K_i^{\mathrm{T}}}{\sqrt{\{d_k\}}}\right)V_i \tag{2-9}$$

其中，d_k 为缩放因子，引入有助于缓解梯度消失的问题，稳定模型的训练过程，提高模型性能。最后，通过融合所有 h 个头的计算结果，我们得到自注意力特征

$$X_{sa} = [H_1 : \cdots : H_h]W^o \tag{2-10}$$

W^o 为可学习参数，h 是自注意力头的数量，在我们的实验中将它的值设置为 12。

通过自注意力机制，我们能够有效地捕捉到图像中长距离的依赖关系，并确保即便是远离答案区域的信息，也能被模型准确理解和利用，从而提升最终的定位精度和答案推理能力。

2.4. 通道注意模块

空间注意力机制与自注意力机制在定位答题对象的空间位置上表现出较高的效能，但这些方法通常缺乏对答题对象本身激活程度的有效增强能力。为了解决这一问题，我们提出了一种新的通道级注意力模块，通过在通道层面生成精细化的注意力图，从而对各个通道进行加权操作，进一步提升目标对象的激活强度。这一模块的引入不仅增强了答题对象本身的显著性，同时还提高了模型对关键特征的关注能力。

通道注意力模块的整体架构如图 3 所示。具体而言，经过 GVLE 模块获得的视觉嵌入记作 V，首先我们将 V 重塑为 X_{2D}。为了降低通道级注意力的计算复杂度，我们将 X_{2D} 的通道数压缩至原始通道数的 $\frac{1}{r}$，获得压缩通道后的特征 X'，具体的过程如下：

$$X' = F_{red}(X_{2D}) = \delta\left(g_{C,\frac{C}{r}}^{1\times1}(X_{2D})\right) \tag{2-11}$$

其中，$\delta(\cdot)$ 表示 ReLU 激活函数，$g(\cdot)$ 表示一个卷积核大小为1×1的卷积操作，用于将输入特征通道数压缩至原始通道数的 $\frac{1}{r}$，并且在实验中我们将 r 设置为 16。

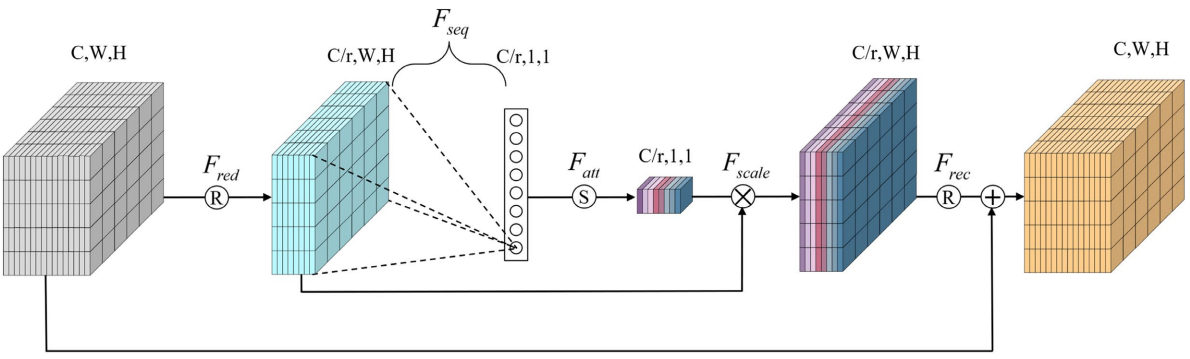

Figure 3. Architecture of the channel attention module
图 3. 通道注意力模块结构图

由于获得的全局描述表示了每个通道的特征，我们需要将全局空间信息压缩到一个通道中。因此，我们采用全局平均池化(GAP)对空间特征进行编码，生成全局特征，编码过程如下：

$$Z = \mathrm{GAP}(X') \tag{2-12}$$

我们使用简单的门控机制，并结合 Sigmoid 激活函数，为每个通道生成权重，从而促进聚合通道信息的利用：

$$Z' = F_{att}(Z) = \sigma\left(\delta\left(ZW_1\right)W_2\right) \tag{2-13}$$

其中，$\delta(\cdot)$ 和 $\sigma(\cdot)$ 分别表示 ReLU 和 Sigmoid 激活函数，W_1 与 W_2 都是可学习参数。

最终，将获得的权重应用于原始特征，从而在通道维度上得到加权结果：

$$\hat{X} = F_{scale}(X', Z') = X' \odot Z' \tag{2-11}$$

为了确保通道级操作前后特征维度的一致性，需将通道数恢复至原始大小。为此，我们通过相同的卷积核大小为1×1卷积操作来重建通道：

$$X_{rec} = F_{rec}\left(\hat{X}\right) = \delta\left(g_{\frac{C}{r},C}^{1\times1}\left(\hat{X}\right)\right) \tag{2-15}$$

其中，$g(\cdot)$ 表示一个卷积核大小为1×1的卷积操作，经过此操作后，通道数从 $\frac{C}{r}$ 恢复至 C。最终，X_{rec} 被重塑为 $X_{channel}$ 作为通道注意力操作后的输出。

2.5. 分层视觉－文本交互注意力模块

除了使模型能够访问包含充分语义信息的图像特征外，解决 VQLA 问题的另一个关键方面是确保模型能够准确理解自然语言问题，并在给定的图像上下文中准确推断答案。因此，我们提出了分层视觉－文本交互注意力机制。该注意力机制模拟了人类语言习得的渐进性特征，通过依次理解单词、短语和句子三个层级，增强了模型对自然语言问题的理解能力。随后，在每个层级，我们采用文本和视觉模态之间的双向交互注意力机制，分别提升模型对问题的理解能力以及其推理答案的能力。

分层视觉－文本交互注意力模块的整体架构如图 4 所示。具体而言，首先，我们将空间注意力、自注意力和通道级注意力的结果融合，以获得融合的视觉嵌入：

$$V_f' = X_{spatial} + X_{sa} + X_{channel} \tag{2-16}$$

由于视觉和文本嵌入属于不同模态，它们的特征空间并不重合。为了有效应用后续的交互注意力，我们首先将这两种模态线性映射到相同的特征空间：

$$\begin{aligned} T' &= \tanh\left(TW_T\right), \\ V' &= \tanh\left(V_f'W_V\right) \end{aligned} \tag{2-17}$$

T' 和 V' 分别表示在同一特征空间中的文本嵌入和视觉嵌入。W_T 和 W_V 是可学习的权重矩阵，$\tanh(\cdot)$ 为双曲正切激活函数。

接下来，我们将问题划分为三个层级：单词层级、短语层级和问题层级。在每个层级上，我们应用视觉－文本交互注意力来增强模型对问题的理解能力以及推理答案的能力。在单词层级，我们定义了一个原子注意力操作 A：

$$A = f_{atten}(X) = \mathrm{Softmax}\left(\tanh\left(X\right)W\right) \tag{2-18}$$

其中，X 和 A 分别表示该注意力操作的输入和输出注意力图。W 为可学习的权重矩阵。进一步地，考虑到问题中的每个单词在重要性上并不相同，我们将原子注意力操作应用于 T'，以放大重要单词的权重，以让模型更关注这些更加重要的单词：

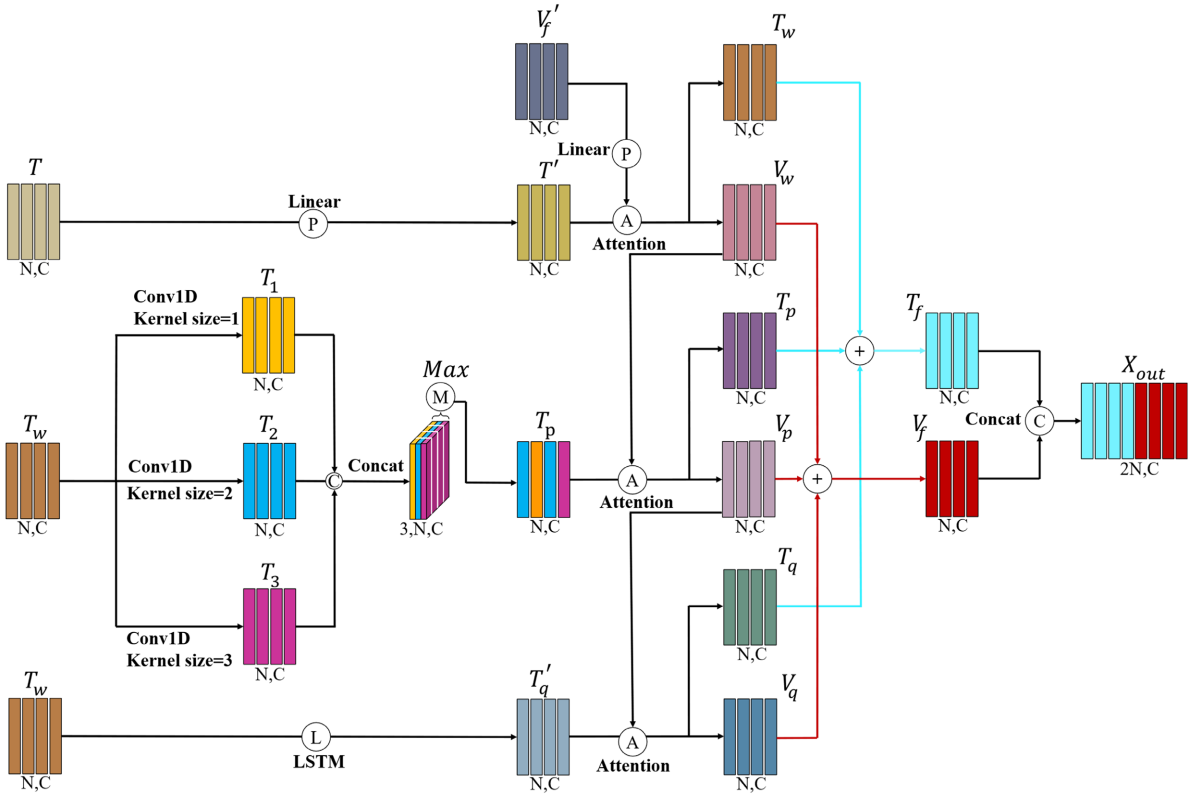

Figure 4. Architecture of the hierarchical visual-text interaction attention module
图 4. 分层视觉－文本交互注意力结构图

$$T_{atten}^{w} = f_{atten}\left(T'W_a\right) \odot T' \tag{2-19}$$

上述式子的 T_{atten}^{w} 表示经过原子注意力操作后的文本嵌入，W_a 为可学习的参数。随后，我们利用 T_{atten}^{w} 和 V' 进行相互引导，完成视觉－文本交互注意力的过程。该过程旨在通过 V' 引导 T_{atten}^{w} 来增强模型对问题的理解，通过 T_{atten}^{w} 引导 V' 来提升模型对答案的推理能力，具体的引导过程如下：

$$\begin{aligned} T_w &= f_{atten}\left(T_{atten}^{w}W_{wt} + V'W_{wv}\right) \odot T_{atten}^{w}, \\ V_w &= f_{atten}\left(V'W_{wv}' + T_{atten}^{w}W_{wt}'\right) \odot V' \end{aligned} \tag{2-20}$$

W_{wt}，W_{wv}，W_{wv}' 和 W_{wt}' 是独立的可学习的参数，f_{atten} 之间是不共享参数的。T_w 和 V_w 分别表示经过视觉-文本交互注意力后的文本和视觉嵌入。

在短语层级，我们考虑由变长单词组成的短语。因此，我们通过利用不同窗口大小的一维卷积来模拟这种情况，生成不同长度短语的特征，表达式为：

$$T_n = g_{C,C}^{n}\left(T'\right),\ n \in \left(1,2,3\right) \tag{2-21}$$

其中，$g_{C,C}^{n}\left(\cdot\right)$ 表示具有窗口大小 n 和输出维度为 C 的一维卷积操作，T_n 表示短语的特征，n 表示短语的长短。随后，我们将这些短语特征融合，得到融合后的短语特征：

$$T_p' = \mathrm{Max}\left(T_1, T_2, T_3\right) \tag{2-22}$$

$\mathrm{Max}\left(\cdot\right)$ 表示沿通道维度取最大值，T_p' 表示融合后的短语特征。类似于单词层级，我们通过原子注意力操作获得增强权重的短语特征 T_{atten}^{p}，并利用 T_{atten}^{p} 和 V_w 进行相互引导，获得短语层级的文本嵌入 T_p 和视觉嵌

入V_p。值得注意的是，在短语层级，V_w替代了V'，z 这样有利于将单词和短语的信息融入到V_p中。此外，借助单词层级的引导，能够缓解训练过程中的复杂性，从而确保模型的稳定性。

在问题层级，我们首先采用 LSTM 从T'中提取问题层级特征T_q'。随后，类似于单词和短语层级，我们利用原子注意力操作和交互注意力操作，通过T_q'和V_p得到问题层级的文本嵌入T_q和视觉嵌入V_q。

在获得了三个层级的视觉和文本嵌入后，我们用线性融合不同层级同一模态的特征，得到融合后的文本嵌入T_f和视觉嵌入V_f：

$$\begin{aligned} T_f &= T_w + T_p + T_q, \\ V_f &= V_w + V_p + V_q \end{aligned} \tag{2-23}$$

然后，我们利用自注意力操作来融合不同模态间的特征，并获得输出特征X_{out}：

$$X_{out} = f_{sa}\left(\left[T_f : V_f\right]\right) \tag{2-24}$$

其中，$f_{sa}(\cdot)$表示自注意力操作，$[:]$表示通道维度上的拼接。

3. 实验与分析

本节将详细阐述实验的环境配置及具体实验操作流程，并与当前最先进的模型进行对比分析，以突出本文所提出模型的优势。此外，实验设计涵盖了定量分析与定性评估两个方面；同时，借助消融实验，将验证所提出模块的有效性。

3.1. 数据库及评价标准介绍

EndoVis-18-VQLA 是一个新近发布的公共数据集，基于 MICCAI 2018 年内窥镜视觉挑战赛数据集构建而成。该数据集包含 14 个机器人手术视频序列。每一帧图像都配有相应的问答对以及边界框注释。数据集涵盖了 18 个类别，包括一个器官、13 种工具交互和四种工具位置。EndoVis-17-VQLA 是从 MICCAI 2017 年内窥镜视觉挑战赛数据集派生而来。该数据集包含 10 个机器人手术视频序列，每个视频序列中的每一帧图像都配有相应的问答对和边界框注释。数据集总计包括 97 帧图像和 472 个问答对。我们将 EndoVis-17-VQLA 用作额外的验证集，以评估模型的泛化能力。

对于这两个数据集，我们采用准确率(Accuracy, Acc)、F 分数(F-Score)和均值交并比(mean Intersection over Union, mIoU)作为评估指标。准确率和 F 值用于衡量模型的分类能力，而均值交并比则评估预测的边界框与真实值之间的相似性，进而衡量模型的定位性能。

3.2. 实施细节

在我们的实验中，所有模型均使用在 ImageNet 上预训练的 ResNet18 作为视觉特征提取器，并以 5 × 5 的图像块提取作为输入特征。此外，还使用了一个经过特定数据集预训练的定制化文本分词器，该分词器涵盖了专门的外科术语。训练过程中，学习率设定为 1×10^{-5}，训练周期为 80 个 epoch，批处理大小为 64，优化器采用了 Adam。所有模型均在 NVIDIA A100 Tensor Core GPU 上使用 PyTorch 框架进行训练。为了确保测试的公平性，所有模型在相同条件下进行测试，使用相同的预测头和损失函数。

3.3. 与最先进模型的结果比较和分析

我们将所提模型与其他最先进(SOTA)方法在 EndoVis-18-VQLA 和 EndoVis-17-VQLA 数据集上进行了比较。比较结果如表 1 所示，经过在相同环境下的严格测试，结果表明我们的模型在两个数据集上显著超越了其他最先进的方法。例如，与之前的最先进模型 GVLE-LViT 相比，我们的方法在 EndoVis-18-VQLA 数据集上取得了 4.48 个百分点的准确率提升和 8.78 个百分点的 F-Score 提升。类似地，在 EndoVis-

17-VQLA 数据集上，我们的方法在 ACC 上提升了 8.05 个百分点，F-Score 上提升了 9.08 个百分点，并在均值交并比(mIoU)上提高了 2.57 个百分点。这些结果表明，我们引入的四种注意力机制使得模型能够在通道和空间维度上更加全面地建模视觉模态特征，从而实现图像区域的精确定位。此外，层次化的视觉－文本交互注意力机制加深了模型对问题的理解，并通过视觉与文本的交互增强了模型对答案的推理能力。

Table 1. Comparison and analysis of results with State-of-the-Art models
表 1. 与最先进模型的结果比较和分析

Method	EndoVis-18-VQLA			EndoVis-17-VQLA		
	ACC	F-Score	mIoU	ACC	F-Score	mIoU
VisualBERT [2]	0.6268	0.3329	0.7391	0.4005	0.3381	0.7073
VisualBERT R [3]	0.6301	0.3390	0.7352	0.4190	0.3370	0.7173
MCAN [4]	0.6825	0.3338	0.7526	0.4137	0.2932	0.7029
VQA-DeiT [5]	0.6104	0.3156	0.7341	0.3797	0.2858	0.6909
MUTAN [6]	0.6283	0.3395	0.7639	0.4242	0.3482	0.7218
MFH [7]	0.6283	0.3254	0.7592	0.4103	0.3500	0.7216
BlockTucker [8]	0.6201	0.3286	0.7653	0.4221	0.3515	0.7288
GVLE-LViT [9]	0.6367	0.3454	0.7624	0.4470	0.4130	0.7125
Ours	**0.6815**	**0.4332**	**0.7698**	**0.5275**	**0.5038**	**0.7382**

3.4. 消融分析

这一节我们主要进行消融实验，以验证我们提出的模型各个模块在 EndoVis-18-VQLA 和 EndoVis-17-VQLA 两个数据集上的效果：

(1) 关键注意力的效果

我们提出的 VLTVI 方法主要由四个基本的注意力机制组成：通道注意力(Channel-wise Attention)、空间注意力(Spatial Attention)、自注意力(Self-Attention)和分层视觉－文本交互注意力(Hierarchical Visual-Text Interaction Attention)。为验证这些注意力机制的有效性，我们设计了一系列消融实验，分析它们对模型性能的影响。表 2 展示了单独使用每种注意力机制及其组合的实验结果。

从实验结果可以明显看出，每种注意力机制的引入均对模型性能产生了积极影响。例如，相较于基础模型(即不使用任何注意力机制)，单独使用空间注意力(SA)后，EndoVis-18-VQLA 数据集的 ACC 从 0.6367 提升至 0.6533，F-Score 从 0.3454 提升至 0.4194，而 EndoVis-17-VQLA 数据集的 ACC 则从 0.4470 上升至 0.4746。类似地，单独使用自注意力(Self)、通道注意力(CA)或分层视觉－文本交互注意力(HA)均能带来不同程度的性能提升。这表明，在通道维度和空间维度进行建模、处理长程依赖关系或帮助模型提取语义信息，都对性能的提升起到了积极作用。当组合使用三种注意力机制时，模型的性能普遍优于单独使用某一注意力模块。最终，当所有四种注意力机制联合使用时，模型在两个数据集上均取得了最优结果。显示出四种注意力机制的互补性及其在复杂医学视觉问答场景中的重要性。这表明，空间注意力(SA)能够有效突出关键区域，自注意力(Self)有助于理解上下文依赖关系，通道注意力(CA)强化了特征提取能力，而分层视觉－文本交互注意力(HA)则增强了跨模态信息的整合。四者协同作用，使得 VLTVI 方法在医学图像分析和问答任务中展现出更强的泛化能力和更精准的预测效果。

Table 2. Effectiveness of different attention mechanisms. SA represents spatial attention, Self represents self-attention, CA represents channel-wise attention, and HA represents hierarchical visual-text interaction attention
表 2. 各注意力的效果，SA 代表空间注意力，Self 代表自注意力，CA 代表通道注意力，HA 代表分层视觉－文本交互注意力

SA	Self	CA	HA	EndoVis-18-VQLA			EndoVis-17-VQLA		
				ACC	F-Score	mIoU	ACC	F-Score	mIoU
				0.6367	0.3454	0.7624	0.4470	0.4130	0.7125
√				0.6533	0.4194	0.7650	0.4746	0.4118	0.7213
	√			0.6504	0.4187	0.7647	0.4788	0.4251	0.7307
		√		0.6407	0.4099	0.7653	0.4852	0.4188	0.7222
			√	0.6515	0.3964	0.7631	0.4725	0.4203	0.7300
	√	√	√	0.6627	0.4138	0.7695	0.4936	0.4403	0.7304
√		√	√	0.6700	0.4262	0.7658	0.5085	0.4520	0.7278
√	√		√	0.6696	0.4199	0.7655	0.4958	0.4305	0.7226
√	√	√		0.6609	0.4319	0.7660	0.4979	0.4312	0.7274
√	√	√	√	**0.6815**	**0.4332**	**0.7698**	**0.5275**	**0.5038**	**0.7382**

(2) 分层视觉－文本交互注意力中不同层次特征的效果

在我们提出的分层文本－视觉注意力(Hierarchical Text-Visual Attention)机制中，文本和视觉特征被分为不同的层次——词汇层次、短语层次和问题层次。每个层次的特征在理解和推理任务中起到了不同的作用，且它们的交互对于模型的整体性能至关重要。为了进一步探讨在层次化文本－视觉交互注意力中逐级处理不同层次特征的重要性，我们在两个数据集上进行了不同层次特征处理的实验。表 3 展示了详细的实验结果，结果表明，仅在词汇层次进行特征交互时，虽然 EndoVis-17-VQLA 数据集上的 ACC 和 F-Score 有所提升，但总体增益有限，而 EndoVis-18-VQLA 数据集上的 mIoU 甚至出现轻微下降。这表明，单纯依赖词汇级特征的建模方式不足以有效提升模型的理解和推理能力，可能是因为词汇级别的信息粒度较低，无法充分表达医学问答任务中的复杂语义。

Table 3. Effectiveness of different feature levels in hierarchical visual-text interaction attention
表 3. 分层视觉-文本交互注意力中不同层次特征的效果

Word-level	Phrase-level	Question-level	EndoVis-18-VQLA			EndoVis-17-VQLA		
			ACC	F-Score	mIoU	ACC	F-Score	mIoU
			0.6367	0.3454	0.7624	0.4470	0.4130	0.7125
√			0.6392	0.3805	0.7623	0.4597	0.4131	0.7131
√	√		0.6497	0.3831	0.7626	0.4682	0.4153	0.7188
√	√	√	**0.6515**	**0.3964**	**0.7631**	**0.4725**	**0.4203**	**0.7300**

当进一步引入短语层次信息时，我们观察到所有指标均有较为稳定的提升。这说明短语级特征的引入能够捕捉更丰富的语境信息，使模型对问题的理解更加准确。最终，当将问题层次信息加入后，模型的性能达到了最佳水平。实验结果充分验证了分层文本－视觉交互注意力的有效性。通过逐层引入不同语义层次的文本信息，使得模型在处理复杂医学问答任务时能够获得更深层次的语义理解和更精准的视

觉定位能力。

(3) 不同融合方法的效果

为了验证在层次化文本–视觉交互注意力之后，整合来自三个不同层次的视觉和文本特征的必要性，我们在两个数据集上实验了多种融合方法。表 4 展示了各个融合方法的详细结果，其中“None”表示未应用任何额外的融合方法，仅依赖初步的特征对齐机制。结果显示，在未使用额外融合方法的情况下，模型的 ACC、F-Score 和 mIoU 均处于最低水平，这表明简单的特征交互不足以充分对齐视觉和文本模态信息，难以有效提升模型的推理能力。

在基于注意力的融合方法中，采用线性注意力(Linear Attention)的方法相较于“None”方法有一定程度的提升，说明显式的特征加权能够增强跨模态特征对齐的效果。然而，与自注意力(Self-Attention)方法相比，线性注意力的提升幅度较小。这可能是因为线性注意力方法通常仅捕获局部或低阶的特征交互，而无法充分建模跨模态特征之间的长程依赖关系。相比之下，采用自注意力机制的融合方法在两个数据集上的所有指标均显著优于其他方法，ACC、F-Score 和 mIoU 均达到了最佳水平。这表明，自注意力机制不仅能够更有效地对齐不同层次的文本和视觉特征，还能够强化模态间的全局信息交互，使得模型在医学视觉问答任务中的推理和定位能力得到提升，能够更准确地刻画文本和视觉之间的复杂关系，提高模型的回答精准度。

这些结果表明，跨模态特征融合的效果大大依赖于所采用的融合方法。基于注意力的融合，特别是自注意力机制，能够更好地捕捉模态之间的长程依赖关系，使得模型能够在多模态学习任务中实现更高效的信息整合和更精确地推理。

Table 4. Effectiveness of different fusion methods
表 4. 不同融合方法的效果

Combination Method	EndoVis-18-VQLA			EndoVis-17-VQLA		
	ACC	F-Score	mIoU	ACC	F-Score	mIoU
None	0.6540	0.4175	0.7612	0.5042	0.4237	0.7209
Linear Attention	0.6584	0.4277	0.7634	0.5106	0.4303	0.7226
Self Attention	**0.6815**	**0.4332**	**0.7698**	**0.5275**	**0.5038**	**0.7382**

4. 总结

本章提出了一种名为 VLTVI 的方法，专门用于医学视觉定位回答(VQLA)任务，该任务旨在根据医学图像和关联的临床问题，预测答案及其在图像中的位置。我们设计并引入了多种创新的注意力机制，旨在通过增强模态内和模态间的交互，进一步提升模型在图像定位和文本问题语义理解方面的能力。在模型的设计与实现过程中，我们设计的方法充分考虑到多模态信息的交互与融合，尤其是在处理医学图像和临床问题时，能够通过细粒度的特征学习，增强模型对局部与全局信息的感知能力。这一策略使得 VLTVI 在复杂的医学情境下，能够提供精确且符合临床需求的答案与定位未来，我们计划进一步提升模型的推理能力，并在更多医学 VQLA 任务场景中开展深入评估。

参考文献

[1] Bai, L., Islam, M. and Ren, H. (2023) Co-Attention Gated Vision-Language Embedding for Visual Question Localized-Answering in Robotic Surgery. In: *Lecture Notes in Computer Science*, Springer, 397-407. https://doi.org/10.1007/978-3-031-43996-4_38

[2] Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J. and Chang, K.W. (2019) VisualBERT: A Simple and Performant Baseline for Vision and Language.

[3] Seenivasan, L., Islam, M., Krishna, A.K. and Ren, H.L. (2022) Surgical-VQA: Visual Question Answering in Surgical Scenes Using Transformer. In: *Lecture Notes in Computer Science*, Springer, 33-43. https://doi.org/10.1007/978-3-031-16449-1_4

[4] Yu, Z., Yu, J., Cui, Y.H., Tao, D.C. and Tian, Q. (2019) Deep Modular Co-Attention Networks for Visual Question Answering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 15-20 June 2019, 6281-6290.

[5] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. and Jegou, H. (2021) Training Data-Efficient Image Transformers Amp; Distillation through Attention. *Proceedings of the* 38*th International Conference on Machine Learning* (*Proceedings of Machine Learning Research*, *Vol*. 139), Location, 18-24 July 2021, 10347-10357.

[6] Ben-Younes, H., Cadene, R., Cord, M. and Thome, N. (2017) MUTAN: Multimodal Tucker Fusion for Visual Question Answering. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 22-29 October 2017, 2631-2639. https://doi.org/10.1109/iccv.2017.285

[7] Yu, Z., Yu, J., Xiang, C.C., Fan, J.P. and Tao, D.C. (2018) Beyond Bilinear: Generalized Multimodal Factorized High-Order Pooling for Visual Question Answering. *IEEE Transactions on Neural Networks and Learning Systems*, **29**, 5947-5959. https://doi.org/10.1109/tnnls.2018.2817340

[8] Ben-Younes, H., Cadene, R., Thome, N. and Cord, M. (2019) BLOCK: Bilinear Superdiagonal Fusion for Visual Question Answering and Visual Relationship Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 8102-8109. https://doi.org/10.1609/aaai.v33i01.33018102

[9] Bai, L., Islam, M., Seenivasan, L. and Ren, H.L. (2023) Surgical-VQLA: Transformer with Gated Vision-Language Embedding for Visual Question Localized-Answering in Robotic Surgery. 2023 *IEEE International Conference on Robotics and Automation* (*ICRA*), London, 29 May-2 June 2023, 6859-6865. https://doi.org/10.1109/icra48891.2023.10160403