Ti-MISO：基于TiLamb的藏文多模态生成式文本摘要

巩鑫[1,2,3], 闫晓东[1,2,3,*] , 常浩远[1,2,3], 田金超[1,2,3]
[1]中央民族大学　信息工程学院，北京　100081
[2]国家语言资源监测与研究民族语言中心
[3]民族语言智能分析与安全治理教育部重点实验室
*通信作者：闫晓东
{23302166,yanxiaodong,23302133}@muc.edu.cn,18904477381@163.com

摘要

为了解决现有单一文本特征生成的藏文摘要质量较低的问题，提出了一种基于TiLamb的多模态生成式文本摘要模型——Ti-MISO。该模型采用ViT（Vision Transformer）模型从图像中提取视觉特征，同时利用预训练微调的TiLamb（Tibetan Large Language Model Base）模型提取藏文文本特征，再通过跨模态交叉注意力机制实现图文特征深层次融合，最终将融合的特征送入模型，借助束搜索算法平衡生成质量更高的摘要。为验证方法有效性，与基于相同语料的其他四种模型进行了对比实验。实验结果表明，Ti-MISO在ROUGE-1、ROUGE-2、ROUGE-L和BLEU四项评价指标上均取得最佳成绩，显示出模型在融合视觉与语言信息、提升摘要质量方面的显著优势。此外，通过一系列消融实验进一步验证了采用ViT模型进行图像特征提取及交叉注意力融合策略的重要性。加入图像信息后采用交叉注意力机制进行特征融合，使融合后的特征保留更多关键信息，帮助模型更加精确地捕捉重点，从而生成的摘要在概括性和可读性上都有明显提升。

关键词： 藏文多模态摘要 ； TiLamb ； 交叉注意力机制 ； 特征融合 ； 束搜索

Ti-MISO: Multimodal Abstractive Summarization of Tibetan Based on TiLamb

Xin Gong[1,2,3] Xiaodong Yan[1,2,3,*],Haoyuan Chang[1,2,3],Jinchao Tian[1,2,3]
[1]School of Information Engineering,Minzu University of China,Beijing 100081
[2]National Language Resources Monitoring and Research Center on Minority Languages
[3]Key Laboratory of Ethnic Language Intelligence Analysis and Security Governance of MOE
[*]Corresponding Author: Xiaodong Yan
{23302166,yanxiaodong,23302133}@muc.edu.cn,18904477381@163.com

Abstract

To address the issue of low-quality Tibetan summaries generated solely based on textual features, a multimodal generative text summarization model based on TiLamb—named Ti-MISO—is proposed. The model employs a ViT (Vision Transformer) to extract visual features from images, while a pre-trained and fine-tuned TiLamb (Tibetan Large Language Model Base) extracts Tibetan text features. It then employs a cross-modal cross-attention mechanism to deeply fuse these visual and textual features. Finally, the fused features are fed into the model, and a beam search algorithm is utilized to balance and generate higher-quality summaries.To validate the effectiveness of the method, comparative experiments were conducted against four other models based on the same corpus. Experimental results demonstrate that Ti-MISO achieved the best

performance on the ROUGE-1, ROUGE-2, ROUGE-L, and BLEU evaluation metrics, highlighting the model's significant advantage in fusing visual and linguistic information to enhance summary quality. Furthermore, a series of ablation studies further confirmed the importance of employing the ViT model for image feature extraction and the cross-attention fusion strategy. Incorporating image information with a cross-attention mechanism for feature fusion allows the fused features to retain more key information, assisting the model in more precisely capturing the main points, thereby significantly improving both the conciseness and readability of the generated summaries.

Keywords: Tibetan Multimodal Summarization , TiLamb , Cross-Attention Mechanism , Feature Fusion , Beam Search

1 引言

在信息时代的快速发展背景下，数据量呈现指数级增长，如何高效提取和压缩信息成为亟待解决的问题。文本摘要生成技术作为自然语言处理领域的重要研究方向，能够帮助用户快速获取关键信息，提高信息处理效率。目前有很多文本摘要生成技术能够依据文档内容生成简洁摘要，但这些技术大多仅聚焦于文本本身，较少涉及其他模态信息的处理。因此，在面对包含视觉信息的文档时，这些方法往往存在一定的局限性，生成的摘要中未能充分利用和整合文档中的视觉信息。已有实验证明，相较于仅使用单一文本特征生成的摘要，融合文本与图像特征生成的摘要可以利用视觉信息，从而来提升生成摘要的质量。随着深度学习技术的不断发展，多模态生成摘要逐渐成为研究热点。

藏文作为中国少数民族语言之一，在我国及周边地区总计约有800万人使用藏语 (王羿钦, 2023)。但藏文信息化起步较晚，目前还没有较为有效的藏文多模态文本摘要系统。相较于中英文多模态文本摘要的相关研究，藏文多模态文本摘要面临着更为复杂的挑战与困难。首先，藏语词形变化复杂且缺乏明确的词边界，增加了自动分词和语义理解的难度，影响摘要生成效果 (李芬芳, 2023)。其次，藏语的语义表达丰富且高度依赖上下文，导致现有针对中英文设计的多模态模型难以直接适用。最后，缺乏大规模、高质量的藏文图文数据集，限制了模型的训练和泛化能力。这些因素使藏文多模态摘要任务相比其他语言更具挑战性。如何克服上述瓶颈，充分发挥藏语丰富的语言特点并有效利用有限的数据资源，是推动藏文多模态文本摘要研究向前发展的关键问题，也是未来相关领域亟待解决的重大挑战。

本文提出一种基于TiLamb的藏文多模态文本摘要生成方法，利用ViT模型和TiLamb模型提取图像与文本特征，再通过跨模态交叉注意力机制实现信息融合，最终结合束搜索算法生成文本摘要。本研究通过引入视觉特征提取、跨模态交互机制和束搜索生成策略，旨在探索高效的多模态融合方法，进而为实现藏文多模态文本摘要生成提供全新的技术解决方案。

本文的贡献如下：

1) 提出Ti-MISO多模态摘要生成模型框架。构建了一个融合图像与藏文文本的多模态文本摘要生成模型，整体架构包括特征提取、特征融合和摘要生成三个核心模块，专为藏文任务设计，填补了该语言在多模态生成领域的空白。

2) 引入视觉提示策略以增强多模态输入表达。通过将图像特征向量转换为可读的"视觉提示字符串"，并嵌入到藏文提示模板中，使得模型能在训练与推理中同时感知图像语义和文本语境，实现有效的视觉-语言对齐。

3) 利用跨模态交叉注意力机制实现特征融合。创新性地采用图像特征作为查询向量（Query），文本特征作为键值对（Key/Value）进行多头注意力计算，有效引导语言模型感知图像中的关键信息，提升生成内容与图像语义的一致性。

4) 引入束搜索解码策略优化摘要生成质量。在文本生成阶段应用束搜索算法，以平衡生成质量与计算效率，输出更具连贯性和上下文一致性的摘要结果，显著优于贪心搜索等简单策略。

2 相关工作

2.1 单模态文本摘要生成方法

文本摘要是自然语言处理领域的重要研究方向，其核心目标是利用计算机技术对文本进行分析、归纳，并自动生成摘要。根据摘要的生成方式，可将文本摘要方法分为抽取式摘要和生成式摘要。抽取式摘要通过从原文中挑选出能代表文档核心内容的关键句或词，并加以排序和组合形成摘要，而生成式摘要则依托深度学习技术，从文本中提取语义信息，理解文本内容，并以流畅、凝练的方式生成能够准确表达文档主旨的摘要 (全安坤et al., 2024)。

经典的抽取式摘要方法包括Lead3和TextRank (Mihalcea and Tarau, 2004)。其中，Lead3方法通过提取文档的前三句话作为摘要，虽然简单直接，但通常能够保证摘要涵盖文章的核心内容。TextRank 是一种基于图的排序算法，借鉴了PageRank (程齐凯et al., 2019)网页排序算法的思想，计算每个句子的得分，并选择得分较高的句子进行组合生成摘要。尽管抽取式摘要方法在语法正确性方面表现较好，但由于缺少必要的连接词，摘要在语义连贯性、逻辑性和一致性方面仍存在一定的改进空间。序列到序列模型 (Sutskever et al., 2014)为生成式摘要奠定了基础。然而，该模型在生成摘要时可能会出现重复生成和未登录词的问题，并且在处理较长文本时可能会造成信息丢失。为了解决这些问题，引入注意力机制的序列到序列模型能够帮助解码器更有效地关注输入文本中与当前输出最相关的部分，从而减少信息损失 (Shi et al., 2021)。束搜索算法通过在每个时间步保留概率最高的前K个候选，然后基于这些候选继续向前搜索，直到生成完整的序列。这样可以保证不会错过全局最优解，同时控制搜索空间的大小 (邵景晨et al., 2024)。

藏文文本摘要生成研究相对较晚，2010年以后零星出现了一些报道，其主要研究工作包括：安见才让(2010)提出了一种基于句子抽取的文本摘要算法，该算法将每个句子的权重拆分为特征词权重和句子结构权重，并根据权重挑选候选句子，随后通过平滑处理提取出质量较高的摘要。南奎娘若等人(2016)采用权重度量和不同特征的加权方法来抽取敏感的藏文文本摘要。李维(2020)提出了两种藏文文本摘要生成方法，其中一种对TextRank算法进行了改进，通过将外部语料库的信息以词向量形式融入算法，对句子中每个词语进行高维映射形成句向量，再通过迭代评分选取得分最高的句子并重排序，从而生成高质量的文本摘要。另一种结合抽取式和生成式方法的藏文文本摘要统一模型 (Yan et al., 2020)。该模型首先利用双向Bi-GRU神经网络从藏文新闻中提取关键信息句子，然后将指针网络融入基于注意力机制的Seq2Seq模型中，以生成高质量的摘要。李亮(2020)预训练一个藏文ALBERT模型，将藏文抽取式文本摘要任务转化为句子分类问题，从而验证了预训练语言模型在该任务中的显著有效性。黄硕等人(Huang et al., 2023)提出了一种基于端到端预训练模型（CMPT）的藏文生成式文本摘要方法，通过去噪和对比学习的预训练策略，以及编码器和解码器的联合训练。

2.2 多模态文本摘要生成方法

目前，虽然仅依靠文本信息的生成式摘要方法已经取得了较好的效果，但随着社会的发展，信息载体往往呈现出多模态的特点，研究者开始探索融合文本、图像等多模态信息的生成式摘要方法。刘泽宇等人(2017)提出了一种基于图像的中文摘要生成方法，通过融合单标签视觉特征和多标签关键词实现图像摘要，但仅侧重视觉模态。陈祥(2020)采用预训练的视觉问答模型抽取图片中的关键信息，结合文本内容生成摘要，该方法主要侧重于对文本信息的理解。何丽(2021)提出了一种图文摘要生成方法，该方法采用BART模型对文本进行处理，利用VGG网络作为图片编码器，并使用多层感知机（MLP）作为图像解码器。该方法并未将文本特征与图片特征映射至同一语义空间，而是分别基于各自的特征生成对应的文本摘要和图片。全安坤等人(全安坤et al., 2024)提出了一种融合文本和图片特征的中文摘要生成方法，使用BERT提取文本特征、ResNet提取图片特征，并通过注意力机制进行跨模态特征融合，最终将融合特征输入指针生成网络生成高质量摘要。随着CLIP(Radford et al., 2021)、BLIP(Li et al., 2022)和LLaVA(Liu et al., 2023)等大规模视觉-语言预训练模型的发展，多模态摘要研究迎来突破。这些模型通过联合编码视觉与语言信息，结合跨模态对比学习等方法，大幅提升了语义理解与生成能力，为多模态摘要提供了更有效的技术路径。

目前，多模态摘要的研究主要集中于中英文语料，由于藏文语义和结构的复杂性，给多模态摘要带来了一定的挑战，因而利用多模态特征生成藏文摘要的相关研究相对较少。在已有工作的基础上，本文提出了一种基于TiLamb的藏文多模态文本摘要生成方法。该方法通过交叉注

意力机制实现文本特征与对应图片特征的跨模态融合，从而保留更多关键语义信息，最后结合束搜索算法生成摘要。实验结果表明，该方法生成的摘要质量显著优于仅依赖文本模态生成的摘要。

3 模型架构

本文使用图片特征与文本特征进行融合，并利用融合后的特征进行文本摘要生成。该方法是由特征提取、特征融合与文本摘要生成三部分组成。第一部分中，针对文本内容，使用微调的TiLamb模型 (Z et al., 2024)进行特征提取；针对图片内容，使用ViT模型 (Dosovitskiy et al., 2020)进行特征提取。第二部分对提取出的文本特征和图片特征进行特征融合。通过跨模态交叉注意力模块将文本模态和图片模态之间的信息进行交互，计算出交叉注意力输出，使这两种模态特征进行有效融合。第三部分是将融合后的特征结合束搜索算法进行在藏文文本摘要的生成。基于TiLamb的藏文多模态文本摘要生成模型框架如图1所示。

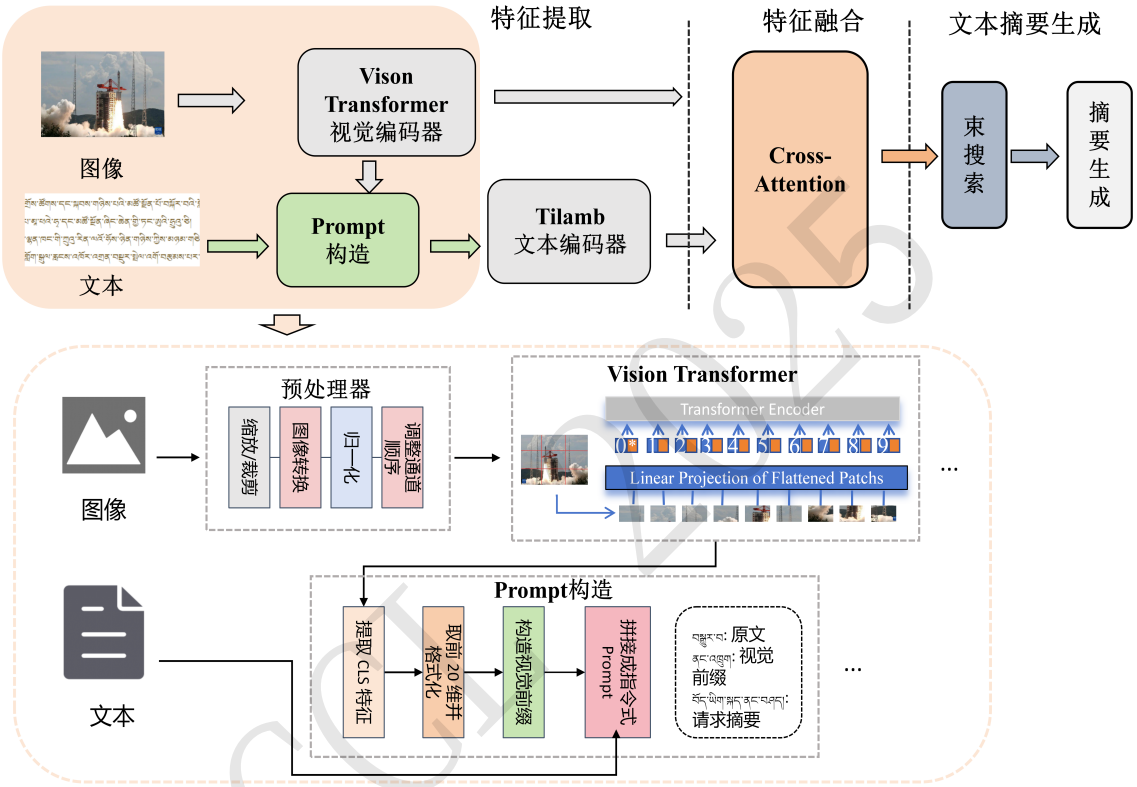

图 1: Ti-MISO模型框架图

3.1 特征提取

3.1.1 图像特征提取

在图像领域中，通常使用卷积神经网络 (Y, 2015)提取图像特征，但随着网络深度的加深以获取更高层次的图像表征，常常会面临梯度回传消失的问题。为了缓解这一问题，ResNet模型引入了残差单元，通过在网络中增加跨层连接，使非相邻层之间建立起更短的路径，从而在反向传播时能够有效传递梯度，减轻梯度消失的影响 (Philipp et al., 2017)。ViT 模型的提出，旨在引入自注意力机制以建模图像中长距离的依赖关系，从而弥补CNN 在捕捉全局上下文信息方面的不足。利用Transformer架构对图像块进行建模，增强了模型对图像全局信息的捕捉能力，提升了图像特征提取的效果。因此，本文采用ViT模型对数据集中的图片进行特征提取。

图像特征的提取流程采用端到端处理逻辑，通过PIL库加载图像并转换为RGB三通道格式，这样可以消除原始图像色彩模式差异对模型输入的干扰。在预训练阶段使用ViT模型对图像进行标准化和分块编码，生成符合模型要求的张量格式。在模型推理环节，采用无梯度计算模式调用预训练ViT模型进行前向传播，提取最后一层隐藏状态中的语义特征。最终通过沿图

像分块序列执行全局平均池化，将分散的局部视觉特征聚合成768维的全局表征向量，该向量既保留了图像的整体语义信息，又可与文本特征进行高效对齐，为多模态融合奠定了视觉语义基础，用于后续摘要生成任务。

3.1.2 文本特征提取

文本特征提取主要分为两部分：预处理阶段构造文本输入，以及在模型内部的文本特征提取。

预处理阶段构造文本输入：从每个样本中提取三个关键信息："Content"（文本指令）、"Title"（目标文本）和"Image Paths"（图像路径列表）。

图像路径部分用于提取图像特征，生成"视觉前缀"。所有提取到的图像特征会被收集到列表中。从ViT的最后一层隐藏状态中取出CLS token向量以代表整张图的全局语义，对多张图依次提取CLS向量后进行拼接然后求均值，得到统一的768维图像特征将；若没有提取到有效特征，则生成一个全零的图像特征向量。然后根据图像特征向量是否全为零，进行接下来的构造视觉前缀。若全为零，说明没有有效的图像信息，则将视觉前缀设为固定占位符字符串；若不全为零，则将图像特征向量的前20个维转换为由逗号分隔的字符串，并构造视觉前缀。

从样本中读取文本指令（instruction）和前面生成的视觉前缀（vision_prompt），形成多模态的输入提示。具体格式如下表1：

提示文本格式
prompt = f"བསྒྱུར་བ: {instruction}\nནང་འཇུག: {vision_prompt}\nབོད་ཡིག་སྐད་ནང་བཤད།:"

表 1: 输入提示文本格式

在预处理阶段，将构造的提示文本进行编码。同时对目标文本进行编码，将提示文本的input_ids与目标文本的input_ids进行拼接，形成一个完整的输入序列。原始文本数据（包括文本指令与目标文本）在预处理阶段被规范化为固定长度的Token序列与相应标签，并与提取的图像特征相融合，构建出完整的语言提示。该过程为后续多模态模型的前向传播与高效训练奠定了坚实基础。

模型内部的文本特征提取：在模型的前向传播中，将预处理后得到的文本序列送入TiLamb模型，利用模型内部多层Transformer编码器生成深层语义表示，即本文提取的文本特征，维度为4096。为了降低计算复杂度和后续融合的需要，将4096维的文本特征经过两个线性投影层，将维度降为768，为后续多模态融合奠定基础。

3.2 特征融合

在进行多模态特征融合时，常见的方法是将来自不同模态的特征直接进行拼接或相加，以实现融合效果。然而，由于各模态特征通常分布在不同的特征空间中，简单的拼接或相加难以充分捕捉模态间的关联性，从而可能导致信息损失，影响融合效果的有效性。相较于简单的拼接或相加等传统特征融合方式，交叉注意力机制 (Lin et al., 2022)通过对不同模态特征向量之间执行张量积运算，实现不同模态间更深层次的信息交互与语义对齐。该机制能够有效挖掘不同模态间的关联性，从而获得更具表征能力的融合特征，保留更多关键信息。因此本文采用交叉注意力机制对文本特征与图像特征进行融合。交叉注意力机制模型如图2所示。

首先进行特征投影对齐，文本特征$T \in R^{B\times S\times 1024}$通过线性层$W_t \in R^{1024\times 768}$投影到图像维度：

$$T' = ReLU\left(T \cdot W_t\right) \quad \left(T' \in R^{B\times S\times 768}\right) \tag{1}$$

图像特征$I \in R^{B\times 1024}$扩展为查询向量：

$$\mathrm{Q} = \mathrm{I}.unsqueeze(0) \quad \left(Q \in R^{1\times B\times 768}\right) \tag{2}$$

然后将投影后的文本特征转置为键值对K，$\mathrm{V} \in R^{S\times B\times 768}$，计算多头注意力：

$$Attn(\mathrm{Q},\ \mathrm{K},\ \mathrm{V}) = Softmax\left(\frac{\mathrm{QK^T}}{\sqrt{\mathrm{d_k}}}\right)\mathrm{V} \quad \left(\mathrm{d_k} = 768/heads\right) \tag{3}$$

第二十四届中国计算语言学大会论文集，第551页-第562页，济南，中国，2025年8月11日至14日。

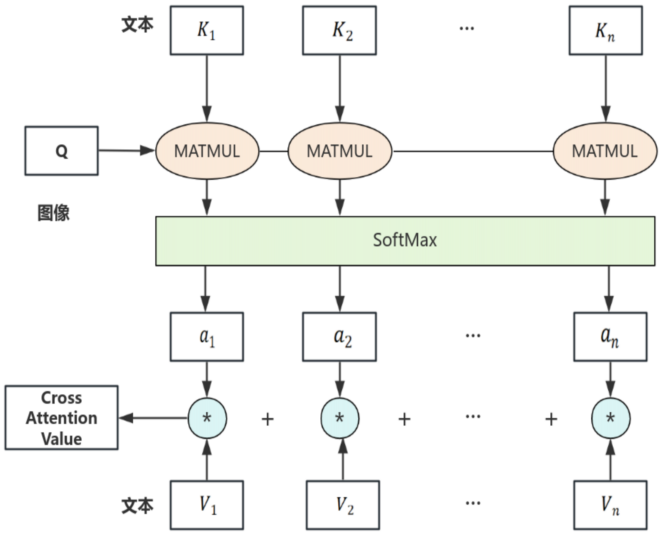

图 2: 交叉注意力机制模型

输出注意力特征$O \in R^{1\times B\times 1024}$,将注意力输出通过$W_o \in R^{768\times 1024}$投影回文本维度并转置:

$$O' = (O \cdot W_o) \cdot transpose(0,1) \quad \left(O' \in R^{1\times B\times 1024}\right) \tag{4}$$

与原始文本特征进行残差相加:

$$T_{fused} = T + O' \quad \left(T_{fused} \in R^{B\times S\times 1024}\right) \tag{5}$$

最终，融合后的特征T_{fused}通过语言模型头生成预测文本。

3.3 文本摘要生成

自回归生成算法是按顺序逐个生成序列中的元素，每个元素的生成都依赖于之前已生成的部分。本文模型输出是一个基于自回归语言建模的文本序列。为获得质量更高、上下文一致性更强的生成结果，在推理阶段引入了束搜索作为解码算法。束搜索是一种启发式搜索算法，通过在解码过程中动态维护k个最优候选序列，平衡生成质量与计算效率。

设第t步的候选集为$H_t = \left\{\left(s^{(i)}, \log P^{(i)}\right)\right\}_{i=1}^{k}$，其中$s^{(i)}$表示候选序列，$\log P^{(i)}$为累积对数概率。解码过程的数学形式化如下:

$$H_{t+1} = top - k_{s'\in Extend(H_t),|s'\leq t+1|}\left[\log P\left(s'\right)\right] \tag{6}$$

通过最大路径的累积对数概率选择$top-k$候选:

$$\log P\left(s'\right) = \sum_{\tau=1}^{|s'|} \log P\left(y_\tau \mid y_{<\tau}, x\right) \tag{7}$$

其中扩展操作:

$$Extend(\mathrm{s}) = \{\mathrm{s} \oplus [\mathrm{w}] \mid \mathrm{w} \in Top - \mathrm{M}(\mathrm{P}(\mathrm{w} \mid \mathrm{s}))\} \tag{8}$$

其中$\oplus$表示序列拼接，$top-M$选取当前预测概率最高的M个候选词。
束搜索算法伪代码实现如下表2:

束搜索算法

输入：模型；初始输入序列；候选序列数量；允许生成的新 token 的最大数量

输出：从起始序列经过多步扩展后生成的、累计对数概率最高的候选序列

```
初始候选集 = [([BOS], 0.0)]
for t in 1...max_new_tokens:
    新候选集 = []
    for 候选序列, 累积概率 in 当前候选集:
        if 序列以</s>结尾:
            加入最终结果集
            continue
        生成下一词概率分布 → p = model(候选序列)
        取 Top-M 个候选词（M=2k）
            for 每个候选词 w:
                新序列 = 候选序列 + [w]
                新概率 = 累积概率 + log(p[w])
                加入新候选集
    当前候选集 = 新候选集中概率最高的 k 个
    当所有候选都生成</s>时提前终止
返回在概率最高的候选序列作为最终输出
```

表 2: 束搜索算法伪代码

4 实验

4.1 数据集

目前，大多数融合不同模态特征的文本摘要研究主要集中在中英文语料上，而针对藏文这一低资源语言的研究较为稀缺，现有的研究多聚焦于单模态的文本摘要语料，缺少包含多模态信息的数据集。本文使用Python爬虫工具从西藏日报藏文媒体、藏族朗诵平台、中央广播电视总台藏语频道以及中国藏族网通等多个藏文微信公众号和藏文网站上进行爬取,内容涵盖新闻资讯、民俗文化、政策宣传、人物纪实等多个领域。最终收集到了38256篇藏文原始图文数据，其中包括文章的标题、正文内容以及与文章匹配的图片。采用文章标题作为参考摘要，同时为提高数据质量，剔除过长或过短的文章正文内容以及过短的文章。对于图片内容，去除了文章中仅作为装饰用途的图片以及公众号自身的二维码等无关信息的图片等操作，最终保留了10000条语料作为实验的数据集，按照8:2比例划分为训练集（8000条）和测试集（2000条）。

4.2 参数设置

表3和表4分别展示了预处理阶段的训练参数设置情况，以及采用Lora方法微调TiLamb模型的参数设置情况。

参数	值	参数	值
per_device_train_batch_size	4	num_train_epochs	20
gradient_accumulation_steps	2	lr_scheduler_type	cosine
max_grad_norm	0.5	warmup_ratio	0.05
learning_rate	2e-4	fp16	True
save_strategy	steps	optim	adamw_torch
save_steps	2000	weigh_decay	0.01

表 3: 训练参数设置

参数	值
r	16
lora_alpha	32
target_modules	q_proj, v_proj
lora_dropout	0.05
task_type	TaskType.CAUSAL_LM

表 4: Lora微调时参数设置

4.3 评测方法

文本摘要的评价方法分为人工评测与自动评测两类。其中，人工评测依赖于领域专家对摘要样本进行主观质量判定，该方法虽能深度捕捉语义准确性和逻辑连贯性等维度，但其评测过程需消耗大量人力资源，导致实施成本显著增加，尤其难以适配大规模语料库的评测需求。相比之下，自动评测方法构建了基于算法模型的评估框架，其核心是通过量化计算生成摘要与参考摘要之间的语义相似度来实现质量评估。该方法采用诸如BLEU、ROUGE等文本匹配算法，能够高效处理海量文本数据，显著提升评测效率。本文采取BLEU、ROUGE系列评价指标进行评价生成藏文摘要的质量。计算公式如下:

$$ROUGE-\mathrm{N}=\frac{\sum_{\mathrm{S}\in\{Refsummaries\}}\sum_{\mathrm{n}-gram\in\mathrm{S}}Count_{match}(\mathrm{n}-gram)}{\sum_{\mathrm{S}\in\{Refsummaries\}}\sum_{\mathrm{n}-gram\in\mathrm{S}}Count(\mathrm{n}-gram)} \tag{9}$$

$$BLEU=BP\cdot\exp\left(\sum_{n=1}^{N}w_n\log p_n\right) \tag{10}$$

$$BP=\begin{cases}1 & if|\mathrm{Gen}|>|Ref|\\ \mathrm{e}^{1-|Ref|/|\mathrm{Gen}|} & otherwise\end{cases} \tag{11}$$

$$p_n=\frac{\sum_{c\in Gen}\sum_{gram_n\in C}\min\left(Count\left(gram_n\right),\max_{Ref}Count\left(gram_n\right)\right)}{\sum_{c\in Gen}\sum_{gram_n\in C}Count\left(gram_n\right)} \tag{12}$$

其中$Refsummaries$表示引用摘要，$Count(\mathrm{n}-gram)$表示引用摘要中的个数，$Count_{match}(\mathrm{n}-gram)$表示生成的摘要和引用摘要中公用个数。$N=4$，$w_n=1/N$。

4.4 实验结果与分析

本节使用构建的Ti-MISO模型在多模态藏文生成式文本摘要任务上进行推理，在构建的数据集的基础上进行了实验。因为目前除Ti-MISO模型以外，还没有其他的模型可以做藏语的多模态生成式摘要任务，为了验证基于TiLamb的藏文多模态文本摘要生成方法的的有效性，将本文方法与以下四种方法进行实验对比:

(1) **Tibetan+ViT**：采用与Ti-MISO模型相同的训练语料、相同的模型架构以及相同的训练生成策略，将微调的TiLamb-7B模型替换成Tibetan-Alpaca-7B大模型。

(2) **Yak-Llama2+ViT**：采用与Ti-MISO模型相同的训练语料、相同的模型架构以及相同的训练生成策略，将微调的TiLamb-7B模型替换成Yak-Llama2-7B大模型。

(3) **Deepseek+ViT**：采用与Ti-MISO模型相同的训练语料、相同的模型架构以及相同的训练生成策略，将微调的TiLamb-7B模型替换成DeepSeek-R1-Distill-Llama-8B大模型。

(4) **CMPT+ViT**：采用与Ti-MISO模型相同的训练语料以及相同的图像编码器，采用预训练的文本编码器CMPT来提取语义特征，然后通过交叉注意力层进行特征融合，最后利用带有指针生成机制的LSTM解码器生成摘要。

模型	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
Tibetan+ViT	31.96	13.57	30.74	21.66
Yak-Llama2+ViT	33.83	14.75	32.76	31.76
Deepseek+ViT	25.92	9.23	24.56	18.61
CMPT+ViT	24.29	10.23	23.85	17.39
Ti-MISO	**34.58**	**15.81**	**33.17**	**32.51**

表 5: 不同模型对比实验结果

对比结果如表5所示，可以看出本文提出的Ti-MISO模型在四项指标上均取得了最佳成绩，其中ROUGE-1、ROUGE-2 和ROUGE-L 分别达到34.58、15.81 和33.17，BLEU 得分为32.51，明显优于其余对比模型。与表现较为接近的Yak-Llama2+ViT 模型相比，Ti-MISO 在ROUGE 系列指标上平均提升约0.75 个百分点，BLEU 得分亦提升了0.75，表明本模型在生成摘要的覆盖性、准确性及语言流畅性方面具备显著优势。

相较之下，Deepseek+ViT和CMPT+ViT等模型在各项指标上表现较弱，ROUGE-1 分别仅为25.92 和24.29，BLEU 得分亦明显偏低，表明这类模型在处理藏文及其图文对齐关系时存在较大局限，主要原因可能在于Deepseek模型未经过藏文任务大规模数据的微调训练，对藏文理解能力较弱。CMPT模型相较于TiLamb模型训练语料较少，不能充分的理解藏文，导致各项指标较低。

图3为生成摘要实例，该新闻报道来自青海省海南藏族自治州湖区乡，2024年6月28日举办的“习近平总书记视察三周年纪念活动暨文化月”活动。原文中明确提及了活动的时间、地点以及参与人数，但对于活动氛围、群众情感表达及组织形式等方面的描写较为简略。相较而言，图像中呈现了更加丰富的现场信息，例如大型横幅标语、参与者身着民族服饰组成方阵，以及背景中的文艺展示场景等。这些视觉要素为模型提供了事件类型、组织规模与情绪色彩等方面的重要语义补充，有助于其更全面地理解语境，从而生成内容更贴切、表述更生动的摘要。

Ti-MISO 在模型架构上引入了多层次的图文交叉注意力机制，使得视觉特征能够更有效地指导文本生成，提升了生成摘要与图像内容的一致性。同时，借助于LoRA 微调策略以及束搜索生成策略，Ti-MISO在低资源环境下依然展现出良好的泛化能力和生成质量。

4.5 消融实验

为了验证基于TiLamb的藏文多模态文本摘要生成方法中prompt构造中采用图像特征维度以及使用ViT模型提取图片特征及使用交叉注意力机制进行跨模态特征融合的有效性，进行如下消融实验。

4.5.1 prompt模板的消融实验

在上述数据集上进行消融实验，以评估所提方法的有效性，结果如表6所示。消融实验说明如下：

消融实验A1：在prompt构造过程中，不加入图像特征维度，只包含文本指令，其他采用上述模型的框架内容。

消融实验A2：在prompt构造过程中，加入图像特征维度前40维，构造视觉前缀，其他采用上述模型的框架内容。

消融实验A3：在prompt构造过程中，加入全部768维图像特征维度，构造视觉前缀，其他采用上述模型的框架内容。

模型	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
A1	32.36	13.96	29.29	28.43
A2	34.49	14.89	32.83	32.14
A3	**34.83**	15.63	33.06	**32.57**
Ti-MISO	34.58	**15.81**	**33.17**	32.51

表 6: 视觉前缀维度的消融实验结果

原文：

2024ལོའི་ཟླ་6ཡིའི་ཚེས་28ཉིན། མཚོ་ཕྱོགས་ཞིང་ཚན་མཚོ་ཐང་པོའི་རྫོང་མི་སེར་ལུང་ཚང་ཚ་ཕྱོང་མཚོའི་འཁྲིལ་ཤང་སེ་གཡུས་དགོན་སྡེའི་གཞུང་གིས་“བདུན་གཅིག་སྨོན་བས་བསུ། འབྲང་བས་སྐུར་ཉེན་འཚེལ་ཞུ། སྤྱི་ལྷག་དྲུག་ཅི་ཞི་ཅིན་རིང་གིས་ཤར་ཚོར་གཙིགས་ཞིབ་མཚོ་ནས་གཤུམ་འགོར་བར་གྱས་པས་བསྡོད་” ཅེས་པ་དང2024ལོའི་མཚོའི་འཁྲིལ་ཤང་གི་“འཚར་ཡན་ཐྱི་མཚོ་འཁྲིལ། རིག་གནས་ཐྱི་པ་ས་” ཞེས་པའི་རིག་གནས་ཟླ་བའི་ཐེང་བཞུར་ཅིད་ཕྱོགས་བཀྲུད་པ་ཤར་ཚ་ཚེ་ཕྱོང་མཚོའི་ཤང་གི་བ་ཚོག་ཐྱེ་བ་ནས་ཐྱེས་མཚོ་བཅན་ཁྲབ་ཀྱལ་ཏང་། ཤར་ཚ་ཕྱོང་གི་འཚེལ་ཡོད་མཚོ་ཁྲིད་དང་མཚོ་འཁྲིལ་ཤང་གི་ལས་ཅིད་མང་ཚོགས་སོགས་ཀི500ཕྱག་ཞུགས་ཡོད། ཅིད་ཕྱོའི་དཀིགས་ཡུལ་ཉི། ཐྱི་ལྷག་དྲུའ་ཅི་ཞི་ཅིན་རིང་གིས་ཤར་ཚོར་གཙིགས་ཞིའ་མཚང་ནས་ཁོ་ཅོག་ཐྱེས་ཁྲིད་ཁང་ཚོགས་ཀྱི་ཐ་ཚུན་ཚོགས་པར་གཏོང་བ། རྒྱ་ཆེ་བའི་ལས་ཅིད་མང་ཚོགས་ཀྱིས་ཐྱ་མཐུན་དུ་འབད་འཚབ་སྐོད་མཚང་ཚོགས་ཀྱི་བཀག་ཅིན་པར་ཐྱེ་ཁྲིད་ཐྱ་བཅས། ཅེས་ཚོགས་ཀྱི་བསྡུས་པའི་རིག་གནས་འཚོ་བ་ཐུན་སྐྱེས་ཚོགས་པར་གཏོང་བ། རྒྱ་ཆེ་བའི་ལས་ཅིད་མང་ཚོགས་ཀྱིས་ཐྱ་མཐུན་དུ་འབད་འཚབ་སྐོད་མཚང་ཚོགས་ཀྱི་བཀག་ཅིན་པར་ཐྱེ་ཁྲིད་ཐྱ་བཅས།

2024 年 6 月 28 日，在青海省海北藏族自治州刚察县的湖区乡，当地政府举办了“七天欢庆迎接、致以生日祝福”活动，庆祝习近平主席亲自考察以来的三周年纪念。在活动期间，以“生态湖区，文化之源”为主题的文化月也在湖区乡正式启动。活动由湖区乡各村庄共同组织，吸引了刚察县和乡政府的负责人，以及 500 多名当地群众参加。活动的核心目标是庆祝习近平主席的考察三周年，并表达对党的感恩和忠诚。此次文化月旨在促进当地的生态保护与经济发展，进一步鼓励各界团体的积极参与。广大群众将通过团结合作和持续努力，为地区繁荣贡献力量。

图片：

参考摘要：

འཕྲལ་ལོ་2ཟླ་མཚོ་འཁྲིལ་ཤང་གི་རིག་གནས་ཟླ་བ་དང་དགའ་སྐོན་སྤྱི་ཐྱེལ། མི།500ལྷག་ཅང་གི་དྲིན་ཚོར་དང་ཤོས་ལྷེས་འཚང་སྐྱོན།

2024 年 6 月 28 日青海湖区乡启动文化月暨庆祝活动，500 人共感党恩、坚定跟党走。

生成摘要：

འཕྲལ་ལོའི་ཟླ་ཤ་ཡིའི་ཚེས་2ཉིན་ཤོང་ཚ་ཕྱོང་མཚོའི་འཁྲིལ་ཤང་དུ་དྲུའ་ཅི་ཞི་ཅིན་གཙིགས་ཞིབ་ཚོ་གསུམ་དང་རིག་གནས་ཟླ་བའི་ཐེང་བཞུར་ཐྱེལ། མི།500ཕྱག་ཞུ།

2024 年 6 月 28 日，湖区乡举办习近平总书记视察三周年纪念暨文化月活动，500 余人参与。

图 3: 生成摘要实例

实验结果表明，采用不同长度的视觉前缀都能显著优于不使用图像特征的基线，其中20维的设置在模型性能与计算成本之间达到了最佳折中。具体来说，利用ViT的CLS token作为视觉特征，它能够高效地聚合图像的全局语义；而截取前20维不仅保留了关键信息，还显著减少了prompt长度和内存占用，便于在资源受限的场景中高效部署。对比不使用视觉提示（A1）及使用更多维度（A2、A3）的消融实验，验证了该策略在提升多模态藏文摘要生成质量方面的有效性。

4.5.2 融合机制的消融实验

在上述数据集和小试实验的基础上进行消融实验，评估所提方法的有效性，结果如表7所示。消融实验说明如下：

消融实验B1：仅利用文本特征进行摘要生成任务，利用微调的TiLamb模型提取文本特征，将提取的文本特征融入语言模型prompt中，采用束搜索策略进行文本摘要生成。

消融实验B2：利用Resnet模型进行图片特征提取，并利用交叉注意力机制进行特征融合，将融合后的特征输入多模态模型中，采用束搜索策略中进行摘要的生成。

消融实验B3：利用ViT模型进行图片特征提取，并利用简单的拼接方式进行特征融合，将融合后的特征输入多模态模型中，采用束搜索策略中进行摘要的生成。

消融实验B4：利用ViT模型进行图片特征提取，并利用自注意力机制进行特征融合，将融

合后的特征输入多模态模型中，采用束搜索策略中进行摘要的生成。

消**融实验B5**：不采用prompt构造，直接利用ViT模型和TiLamb模型分别对图片特征和文本特征进行提取，并利用交叉注意力机制进行特征融合，将融合后的特征输入多模态模型中，采用束搜索策略中进行摘要的生成。

模型	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
B1	25.67	10.03	23.38	26.28
B2	34.49	14.17	32.87	31.30
B3	28.03	14.30	27.06	27.94
B4	30.94	15.17	29.94	29.96
B5	33.48	13.09	30.49	30.58
Ti-MISO	**34.58**	**15.81**	**33.17**	**32.51**

表 7: 融合机制的消融实验结果

消融实验B1仅使用微调的TiLamb模型提取文本特征，直接融入prompt生成摘要，结果表明单靠文本信息可能无法捕获全部关键信息，从而导致生成的摘要在表达上存在模糊或偏颇的情况；消融实验B2采用Resnet模型提取图像特征，并结合交叉注意力进行融合，结果显示Resnet模型可能在细粒度语义表达上存在不足；消融实验B3利用ViT模型提取图片特征，并通过简单拼接与文本特征融合，结果证明ViT模型能为摘要生成提供有力补充，但简单拼接容易忽略各模态之间固有的差异和特性，没有专门的对齐或归一化步骤，可能导致部分模态的信息被稀释或误解；消融实验B4同样基于ViT模型，结果表明单一的全局自注意力可能无法有效区分局部和全局层次的信息交互。虽然自注意力机制在某些细粒度方面具有优势，但在跨模态信息全面融合方面，其效果不如更加针对跨模态交互设计的交叉注意力机制。消融实验B5结果表明prompt构造有效引导模型聚焦核心信息，移除后性能下降。

综合上述消融实验结果，本文提出的采用基于ViT模型提取图片特征，将图像特征的前20维构造prompt，并采用交叉注意力机制进行跨模态特征融合，在绝大多数指标上均优于各消融方案。这充分验证了所提出的prompt构造和交叉注意力融合策略对于提升藏文多模态文本摘要生成质量的有效性和必要性。

5 总结

本文针对藏文自动摘要生成任务中多模态信息利用不足的问题，提出了一种基于TiLamb的藏文多模态生成式文本摘要模型Ti-MISO。该模型充分融合了文本与图像信息，通过采用ViT模型提取图像特征构造prompt，以及利用跨模态交叉注意力机制实现深层次特征融合，进而采用束搜索策略生成高质量的摘要。实验结果表明，Ti-MISO在ROUGE-1、ROUGE-2、ROUGE-L和BLEU指标上均显著优于传统单模态文本摘要方法及其他多模态对比模型，验证了本方法在捕捉图像语义和文本信息交互方面的有效性和必要性。此外，通过详尽的消融实验分析，我们进一步证明了采用prompt模块和交叉注意力机制的重要性，为藏文多模态自动摘要研究提供了新的思路与解决方案。未来的工作将在扩大数据规模、数据集类型、优化跨模态特征融合策略及推广至其他低资源语言的自动摘要任务等方向展开探索。

6 致谢

本论文由国家社科基金重点项目(22&ZD035)，中央民族大学项目(GRSCP202316, 2023QNYL22, 2024GJYY43)资助。

参考文献

Rada Mihalcea and Paul Tarau. 2004. *TextRank: Bringing Order into Text.* In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing,pages 404–411.

Sutskever I, Vinyals O, Le Q V. 2014. *Sequence to sequence learning with neural networks.* Advances in neural information processing systems,27.

Shi T, Keneshloo Y, Ramakrishnan N, et al. 2021. *Neural Abstractive Text Summarization with Sequence-to-Sequence Models.* ACM Transactions on Data Science, 2(1): 1-37.

Xiaodong Yan, Xiaoqing Xie, Yu Zou, and Wei Li. 2020. 基于统一模型的藏文新闻摘要*(abstractive summarization of tibetan news based on hybrid model).* In Proceedings of the 19th Chinese National Conference on Computational Linguistics, pages 479–490.

Huang S, Yan X, OuYang X, et al. 2023. 基于端到端预训练模型的藏文生成式文本摘要*(abstractive summarization of Tibetan based on end-to-end pre-trained model).* Proceedings of the 22nd Chinese National Conference on Computational Linguistics, 2023: 113-123.

Chen Y. 2015. *Convolutional neural network for sentence classification.*

Liu H, Li C, Wu Q, et al. 2023. *Visual instruction tuning.* Advances in neural information processing systems, 2023, 36: 34892-34916.

Li J, Li D, Xiong C, et al. 2022. *Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.* International conference on machine learning. PMLR, 2022: 12888-12900.

Radford A, Kim J W, Hallacy C, et al. 2021. *Learning transferable visual models from natural language supervision.* International conference on machine learning. PmLR, 2021: 8748-8763.

Philipp G, Song D, Carbonell J G. 2017. *The exploding gradient problem demystified-definition, prevalence, impact, origin, tradeoffs, and solutions.* arXiv preprint arXiv:1712.05577.

Dosovitskiy A, Beyer L, Kolesnikov A, et al. 2020. *An image is worth 16x16 words: Transformers for image recognition at scale.* arXiv preprint arXiv:2010.11929.

Wenhao Z, Yuan S, Xiaobing Z. 2024. *TiLamb:* 基于增量预训练的藏文大语言模型*(TiLamb: A Tibetan Large Language Model Based on Incremental Pre-training).* Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference), 2024: 254-267.

Lin H, Cheng X, Wu X, et al. 2022. *Cat: Cross attention in vision transformer.* 2022 IEEE international conference on multimedia and expo (ICME). IEEE, 2022: 1-6.

刘泽宇,马龙龙,吴健,孙乐. 2017. 基于多模态神经网络的图像中文摘要生成方法. 中文信息学报, 31(06):162-171.

程齐凯,王佳敏,陆伟. 2019. 基于引用共词网络的领域基础词汇发现研究. 数据分析与知识发现,3(06):57-65.

全安坤,李红莲,张乐,吕学强. 2024. 融合内容和图片特征的中文摘要生成方法研究. 数据分析与知识发现,8(03):110-119.

邵景晨,柴玉梅,王黎明. 2024. 基于语义加权的双层*LSTM*图像描述生成方法研究. 计算机应用与软件, 41(10):155-162.

南奎娘若and 安见才让. 2016. 基于敏感信息的藏文文本摘要提取的研究. 网络安全技术与应用, (04):58-59.

李维, 闫晓东, and 解晓庆. 2020. 基于改进*textrank* 的藏文抽取式摘要生成. 中文信息学报, 34(9):36–43.

安见才让. 2010. 藏文搜索引擎系统中网页自动摘要的研究. 微处理机, 31(05):77-80.

刘何丽. 2021. 基于多模态神经网络的图文摘要生成方法研究. 北京邮电大学.

王羿钦. 2023. 藏文文本摘要方法的研究. 中央民族大学.

李芬芳. 2023. 藏文摘要生成关键技术研究. 兰州大学.

陈祥. 2020. 基于多模态数据的文本摘要生成研究. 电子科技大学.

李亮. 2020. 基于*ALBERT*的藏文预训练模型及其应用. 兰州大学.