中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(2025)10-3361-16

论文引用格式:Luo S, Qian W H and Liu P. 2025. Fusion of prompt learning and visual semantic generation for image captioning of Dongba paintings. Journal of Image and Graphics, 30(10):3361-3376(罗霜, 钱文华, 刘朋. 2025. 结合提示学习和视觉语义生成融合的东巴画图像描述. 中国图象图形学报, 30(10):3361-3376)[DOI:10.11834/jig.240601]

结合提示学习和视觉语义生成融合的东巴画图像描述

罗霜,钱文华*,刘朋

云南大学信息学院,昆明　650504

摘　要:**目的**　东巴画是纳西族传统艺术的瑰宝,其画面视觉元素丰富、色彩分明,具有鲜明的地域文化特色和民族特征。针对现有图像描述方法在东巴画描述中存在的领域偏移问题,提出一种结合提示学习和视觉语义—生成融合的东巴画图像描述方法。该方法引入内容提示模块和视觉语义—生成融合损失,旨在引导模型学习东巴画的主题信息,提升描述的准确性和文化表达能力。**方法**　采用编码器—解码器(encoder-decoder)架构实现东巴画图像描述的生成。编码器采用卷积神经网络(convolutional neural network,CNN)捕获图像中关键的语义信息,并将这些特征整合到解码器编码层中的归一化层,控制文本描述的生成过程;解码器采用Transformer结构实现,利用自注意力机制有效地捕捉输入序列中的长距离依赖关系,使模型关注输入序列中的关键信息。此外,本文在解码器之前引入了内容提示模块。该模块通过图像特征向量得到图像的主体、动作等信息,并将其构建成提示信息作为描述文本的后置提示。通过后置提示信息,解码器能有效地关注描述文本中具体的文化场景和细节特征,增强对东巴画特定图案和场景的识别与理解能力。同时,引入了视觉语义—生成融合损失,通过优化该损失,引导模型提取东巴画中的关键信息,从而生成与图像保持高度一致的描述文本。**结果**　实验结果表明,在东巴画测试集上,本文所提方法在BLEU_1(bilingual evaluation understudy)到BLEU_4、METEOR(metric for evaluation with explicit ordering)、ROUGE(recall-oriented understudy for gisting evaluation)和CIDEr(consensus-based image description evaluation)评价指标上分别达到0.603、0.426、0.317、0.246、0.256、0.403和0.599,东巴画图像描述文本在主观质量也得到了更好的效果。**结论**　本文方法显著增强了模型对东巴画图像主题和民族文化特征的捕捉能力,有效提升了生成描述在准确性、语义关联性和表达流畅性方面的表现。

关键词:东巴画;图像描述;提示学习;视觉语义—生成融合;领域偏移

Fusion of prompt learning and visual semantic generation for image captioning of Dongba paintings

Luo Shuang, Qian Wenhua*, Liu Peng

School of Information Science and Engineering, Yunnan University, Kunming　650504, China

收稿日期:2024-10-21;**修回日期**:2025-02-12;**预印本日期**:2025-02-19

*通信作者:钱文华　whqian@ynu.edu.cn

基金项目:国家自然科学基金项目(62162065);云南大学"双一流"建设联合专项(202401BF070001-023);云南省科技厅应用基础研究计划项目(202201AT070167);云南省"视觉与文化计算创新团队"项目(202505AS350009)

Supported by: National Natural Science Foundation of China (62162065); Joint Special Project Research Foundation of Yunnan Province (202401BF070001-023); Yunnan Fundamental Research Projects (202201AT070167); Yunnan Province Visual and Cultural Innovation Team (202505AS350009)

Abstract: Objective Dongba painting is a treasured form of Naxi traditional art. It is rich visual elements, distinct colors, and strong regional cultural and national characteristics. The application of existing image description methods to Dongba paintings is hindered by domain bias, with the model generating text features aligned with those of natural images. The conventional approaches to image description require the use of predefined sentence structures and depend on the quantity and quality of datasets. The descriptions generated by such methods are often repetitive and lack diversity. When these methods are directly applied to Dongba painting descriptions, they present certain limitations. Multi-task pretraining models jointly train multiple vision-language objectives, including image-text contrastive learning, image-text matching, and image-conditioned language modeling. However, these image captioning models tend to generate generic descriptions and cannot capture the distinctive style present in ethnic images. Most controlled image captioning models regulate the generation of image descriptions through analyzing textual data and extracting keywords or entities. However, keyword extraction relies on the explicit entity information present within the sentence. It cannot consider the implicit cultural connotations and semantic depth inherent to Dongba paintings. Consequently, the resulting descriptions are simple and one dimensional. To address these issues, this study proposes an image captioning method based on the fusion of prompt learning and visual semantic generation. This approach guides the model in acquiring the cultural background knowledge associated with Dongba paintings, which mitigates the domain bias challenge. **Method** First, this study uses a codec architecture to realize image captioning for Dongba paintings. In the encoding stage, the encoder employs a convolutional neural network to extract the essential semantic information from the Dongba painting image. This information is then integrated into the normalization layer within the coding layer of the decoder, which enables the control of the process of generating textual descriptions. Therefore, the resulting descriptions are semantically aligned with the image content. In the decoding stage, a Transformer structure is employed to efficiently capture long-distance dependencies in the input sequence through a self-attention mechanism. This utilization facilitates the generation of accurate and coherent text output. To circumvent overfitting resulting from the stacking of decoder layers, the proposed model employs a decoder comprising 10 layers. In addition, pretrained BERT weights are employed for initialization, which enhances the performance and convergence speed of the model. Second, we introduce a visual semantic generative fusion loss, which is optimized to guide the model to extract the key information in the Dongba paintings. As a result, descriptive texts that are highly consistent with the images are generated. Meanwhile, the content prompt module is introduced before the decoder. Through a mapper, the cultural background information such as religious patterns, gods, spirits, and hell ghosts in Dongba paintings can be obtained. By combining prompt information and text information, the decoder can more effectively understand and use the content information of the image. This capability improves the relevance and accuracy of the generated description text and the image theme. Finally, to enhance the capacity of the model for targeted learning and description of painting image features, a bespoke dataset for image captioning of Dongba paintings caption is constructed. This dataset is based on the themes of Dongba paintings and the underlying Dongba culture, and it is used for model training. **Result** Several experiments are conducted on the test dataset for Dongba paintings in this study to verify the effectiveness of the proposed method. The experiment includes contrast and ablation experiments. In the ablation experiments, the architectural configuration of the encoder and decoder is retained, and the ablation model is constructed in accordance with this configuration. Three kinds of comparison experiments are formed. In the first part, the effect of different model encoders on image feature extraction and image description text quality is compared. In the second part, the effect of the number of coding layers of model decoders on the quality of generated description text is discussed. In the third part, some advanced algorithms are selected for comparison to verify the superiority of the proposed algorithms. Experimental results show that the proposed method achieves 0.603, 0.426, 0.317, 0.246, 0.256, 0.403, and 0.599 in the evaluation indexes of BLEU-1 to BLEU-4, METEOR, ROUGE, and CIDEr, respectively. It also achieves better results in the subjective quality of image description text for Dongba paintings. **Conclusion** The image captioning method based on the fusion of prompt learning and visual semantic generation for Dongba paintings proposed in this study effectively enhances the capability of the encoder to capture image features and the capability of the decoder to understand the input text. This improvement results in descriptions that are more relevant to the semantics of the images, as well as more accurate and fluid. Furthermore, the method preserves the national characteristics of Dongba paintings. In the future, the relationships between the elements in Dongba

paintings will be comprehensively explored. Furthermore, the size and diversity of the dataset will be expanded. The model structure will be optimized to enhance its robustness. Other datasets will be also used for training purposes, which will expand the scope of application.

Key words: Dongba painting; image caption; prompt learning; visual semantic-generation fusion; domain shift

0 引　言

图像描述是指将图像输入数学模型和神经网络，模型经学习后，指导计算机输出符合图像内容和中文语法规则的描述文本。描述文本不仅需要包含图像中的对象，还必须表达对象之间的关联关系，以及它们的属性和活动状态（Vinyals 等，2015）。目前，图像描述已经广泛应用于视觉生理功能障碍者辅助、智能人机交互及机器人开发（汤鹏杰 等，2017）、视觉问答和医学图像自动语义标注等领域（Waghmare 和 Shinde，2022）。

东巴画是纳西族传统艺术的瑰宝，是东巴文化的重要载体。如图 1 所示，东巴画的绘画形式以线条为主，画面视觉元素丰富、色彩艳丽，具有鲜明的地域特色和民族特征。其中，图 1（a）为一头全身布满红色花纹的虎，它是纳西族的图腾动物和东巴教祭坛的守护神兽。图 1（b）显示了一位身着传统服饰，正骑着梅花鹿的女性。

(a) 示例 1　　(b) 示例 2

图 1　东巴画实例

Fig. 1　Examples of Dongba paintings

((a) example 1; (b) example 2)

东巴画通过其丰富的内容展现了纳西族悠久的历史和深厚的文化底蕴，每幅画作都蕴含着深刻的宗教寓意和民族情感。目前，针对东巴画的研究主要集中在艺术价值（杨鸿荣，2021）和绘画风格（苏泉，2022）等方面，缺乏对东巴画图像及其文化背景进行图像描述的工作。因此，研究东巴画图像描述，对深入地理解纳西族的民族信仰、生活方式以及艺术追求具有重要意义。

图像描述需要计算机掌握多种视觉和语义识别技术，同时还需要将所有检测结果总结为一个自然语言表述的语句（赵永强 等，2023）。早期的图像描述方法主要是基于传统的机器学习，例如 Li 等人（2011）提出的模板匹配图像描述方法、Devlin 等人（2015）提出的基于最近邻检索的图像描述方法。这类方法简单、便捷，但存在一定的局限性。其依赖于预定义的句子结构，或受限于检索库中图像—描述对的多样性和质量，导致生成描述的表达方式单一且重复，缺乏多样性。

随着深度学习的发展，Vinyals 等人（2015）通过传统的卷积神经网络（convolutional neural network，CNN）提取图像特征，并利用长短期记忆网络（long short term memory，LSTM）生成相应的图像描述。单一的 CNN 编码器难以精确捕捉图像中关键的视觉元素，导致生成的描述缺乏细节和准确性。为此，Anderson 等人（2018）引入机器翻译中的注意力机制，通过自下而上的注意力机制从图像中提取对象特征，再利用自上而下的注意力机制对这些特征进行选择和聚焦，使网络有效识别和定位图像中的显著对象和区域。

随着研究的深入，利用辅助任务联合训练已成为提升模型性能的重要手段。Johnson 等人（2016）提出 DenseCap（dense captioning）模型，通过联合训练目标检测和描述生成任务，实现了对图像中多个区域进行密集描述。Li 等人（2022）提出一种多模态混合编码器—解码器模型 BLIP（bootstrapping language image pretraining），用于统一视觉语言理解和生成。该模型使用字幕生成器生成合成字幕，并通过过滤器去除嘈杂的字幕，降低网络数据中的噪声。上述图像描述模型倾向于生成通用的描述，因此图像描述领域开始研究可控图像描述（controllable image captioning，CIC），用控制信号指导模型生成描述文本，提高模型的泛化能力。模型在描述图像内容的同时，需要满足给定的约束条件。Zheng 等人（2019）为了更全面地描述图像中的内容，提出模型

CGO(image captions with guiding objects),模型选取句子中的一个词作为初始点,通过LSTM预测初始点左右两边的内容。Wang等人(2023)将提示学习嵌入到图像描述模型的框架中,通过提示词生成所需的风格化标题。Fei等人(2023)提出模型ViECap(visual entities captioning),利用语法解析器(grammar parser)从文本中得到实体硬提示(entity-aware hard prompts),将提示与屏蔽策略(masking strategy)相结合引导语言模型关注实体信息。Qiu等人(2024)提出的MacCap(mining fine-grained image text alignment captioning)通过过滤检索模块获得记忆库中与图像相关的关键字,通过关键字控制生成图像描述文本。

然而,上述方法在进行东巴画图像描述时存在以下困难:1)过拟合问题。在图像描述任务中,基于深度学习的模型通常依赖于大规模且高质量的图像样本及其详细描述,以确保模型的有效训练和优化。由于东巴画样本的稀缺性和描述信息的不足,模型在训练过程中会过拟合训练数据的细节和噪声,导致其泛化能力降低。2)领域偏移问题。现有多模态预训练图像描述模型虽然在大规模数据集上表现出色,但描述未见领域的图像时,表现出明显的领域偏移问题。模型的先验知识会主导生成过程,导致模型无法捕捉到任务所需的具体语义和细节,使生成的描述与图像无关,出现幻觉现象。现有可控图像描述模型大多通过分析文本信息并提取关键字或实体来控制生成图像描述,但关键字提取依赖于句子的显性实体信息,无法捕捉东巴画中隐含的民族文化内涵和语义层次,导致生成的描述简单片面。

综上所述,由于东巴画语义丰富、样本较少,已有方法应用于东巴画图像描述时,对图像视觉理解不足,且对其丰富的文化内涵描述不充分。为此,本文针对东巴画图像描述进行研究,主要贡献在于:1)建立了一个新的东巴画图像描述数据集。从东巴画相关文献(钱文华 等,2020)收集真实纳西族东巴画,将图像数据按内容分为神像精灵、地狱鬼怪和宗教图案等,并描述了每一类题材的画面内容及其在东巴文化中的含义。2)为了提升生成的描述文本与主题的相关性和准确性,并缓解幻觉现象,本文构建了内容提示模块。该模块利用图像的主体信息及其动作等信息引导模型理解东巴画描述文本的上下文信息,从而使解码器能更有效利用图像的视觉和文本信息。3)为了增强模型对东巴画中图案和场景信息的识别能力,并生成以内容为中心的图像描述,本文引入了视觉语义—生成融合损失。通过优化损失,模型能更有效地提取东巴画中的关键信息,从而生成与图像保持高度一致的描述文本。大量实验表明,本文方法相较现有方法,在客观描述指标上表现更优,并且显著提升图像描述的主观质量。

1 本文方法

1.1 网络整体结构

本文所提结合提示学习和视觉语义—生成融合的东巴画图像描述方法(prompt and visual semantic-generation fusion-based Dongba painting captioning, PVGF-DPC)的整体结构如图2所示。网络主要包括图像编码器、文本解码器以及内容提示模块。首先,对输入数据进行处理。将东巴画图像数据进行统一规范化,再将其输入编码器,输出得到1 280维图像特征向量,并将对应的图像描述文本进行分词处理,得到长度为128的序列。其次,将特征向量$\boldsymbol{e}^x$输入解码器的各个编码层,以控制文本描述的生成。同时,将$\boldsymbol{e}^x$输入内容提示模块得到图像的主体及其动作等信息,并将其构建成文本提示信息$\boldsymbol{e}^p$,再将$\boldsymbol{e}^p$与图像对应的描述文本连接之后输入解码器。最后,通过联合训练内容提示模块和文本生成任务,共同优化图像描述模型的性能。

1.2 编码器

编码器采用卷积神经网络捕获图像中关键的语义信息。MobileNetV2(mobile network)(Sandler等,2018)具有较高的特征提取效率,其倒残差块(inverted residual block)结构和深度可分离卷积(depthwise separable convolution)技术降低了模型的参数量和计算复杂度,从而显著提升了模型的计算效率。因此,本文采用MobileNetV2作为模型的编码器。

为了确保模型输入的一致性,并保留东巴画的艺术细节和内容特征,本文计算了东巴画数据集中所有图像长宽的平均值,并将它们统一规范化为长宽均为299像素,且保持通道数为RGB3通道。经过这一处理后,将图像送入编码器进行特征提取。

图像进入编码器后,首先经过一个初始卷积层和一个3×3的标准卷积层,输出32个通道。这一过

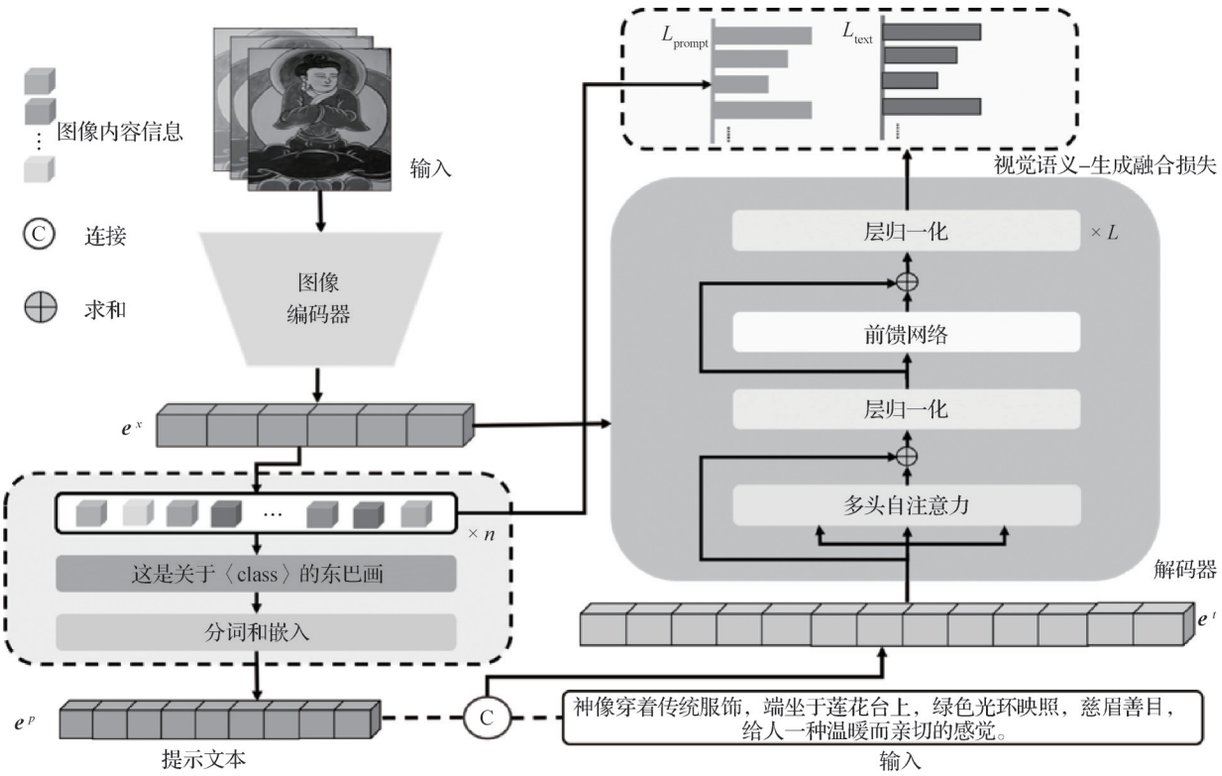


图2　模型结构

Fig. 2　Structure of model

程减小输入图像的尺寸，并提取东巴画的基本视觉特征，如色彩分布和基本轮廓等。接着，图像特征通过16个倒残差块进行处理。通过引入升维和深度可分离卷积技术，倒残差块能够在保持计算效率的同时，提高网络提取细粒度东巴画图像特征的能力。倒残差块由3个主要部分组成：扩展层、深度卷积层和投影层，其结构如图3所示。

扩展层通过增加特征图的维度，帮助网络更好地捕捉东巴画中复杂的文化元素，如服饰细节、纹理图案以及象征性的符号。深度卷积层提取东巴画的空间特征，能够识别东巴画中的人物姿态、面部表情以及细腻的背景层次感。最后，投影层将特征图压缩，减少冗余信息，同时保留东巴画的艺术元素，如人物轮廓、文化符号和色彩搭配等，为后续处理提供简洁有效的视觉特征表示。经过所有倒残差块的处理后，通过1 × 1卷积层将输出通道数增加到1 280，并通过全局平均池化（mean pooling）将特征图转换为一个1 280维的特征向量。

1.3　解码器

为了充分理解东巴画图像描述文本的上下文信息，本文将内容提示模块识别的图像主体及其动作

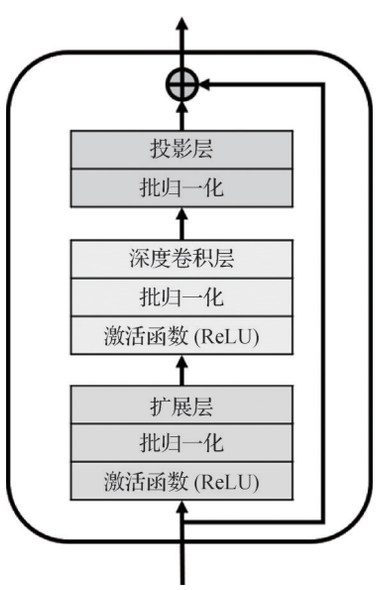


图3　倒残差块结构

Fig. 3　Structure of inverted residual block

等信息作为图像描述文本的后置提示辅助解码器对描述文本的理解。同时，将东巴画图像的特征向量嵌入解码器，帮助模型生成与图像内容紧密相关的描述。由于Transformer（Vaswani等，2017）结构通过自注意力机制能有效地捕捉输入序列中的长距离依赖关系，有助于生成准确且连贯的文本。为此，本文

采用包含Transformer结构的模型作为解码器。BERT（bidirectional encoder representations from Transformers）模型（Devlin等，2019）具备强大的语义理解能力，能够捕捉丰富的上下文信息和深层的语义信息。本文采用预训练的BERT模型对解码器进行初始化。

首先，对解码器的输入进行处理，将内容提示模块的输出作为描述文本的后置提示。例如，对于描述文本为“神像端坐于莲花台上，绿色光环映照，慈眉善目，给人一种温暖而亲切的感觉”的东巴画，内容提示模块提取的内容提示信息为“这是关于神像的东巴画”，并将其作为引导文本传递给解码器，为生成的描述提供背景语境和图像特征引导。其次，对结合prompt的图像描述文本进行分词处理，得到的token序列Y再输入嵌入层。最后，将嵌入信息传递至编码层进行处理，编码层的结构如图4所示，包括两个子层：多头自注意力（multi-head attention）和前馈网络（feed forward network），中间通过残差连接（residual connection）和层归一化（layer normalization）相连。其中，多头自注意力的主要作用是捕捉输入序列中各个词之间的依赖关系。在描述神像图时，模型需同时理解“神像”和“莲花台”之间的空间关系，以及“绿色光环”和“温暖亲切”之间的情感联系。多头注意力机制计算每个词与其他词之间的注意力分数，有效地捕捉全局上下文信息，其具体计算为

$$A(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=f_{\text{softmax}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\times\boldsymbol{V} \tag{1}$$

式中，$\boldsymbol{Q}$为查询矩阵，表示当前词的表示向量；$\boldsymbol{K}$为键矩阵，表示所有词的表示向量；$\boldsymbol{V}$为值矩阵，用于计算加权求和后的输出。d_k为查询矩阵的维度。

前馈神经网络对每个位置的向量进行两次线性变换，并在两个变换之间应用ReLU（rectified linear unit）激活函数。通过两次线性变换，增强模型的非线性表示能力，使其能捕捉复杂的文本特征。例如，在生成描述“神像端坐”的文本时，前馈神经网络层能够识别出“端坐”的动作特征，并在生成中保留这一细节。两个层归一化和残差连接用于增强模型的稳定性和训练效果。为了有效地融合图像和文本信息，将图像特征整合到层归一化。图像中的“莲花台背景”和“绿色光环”等视觉特征与文本中的“端坐”

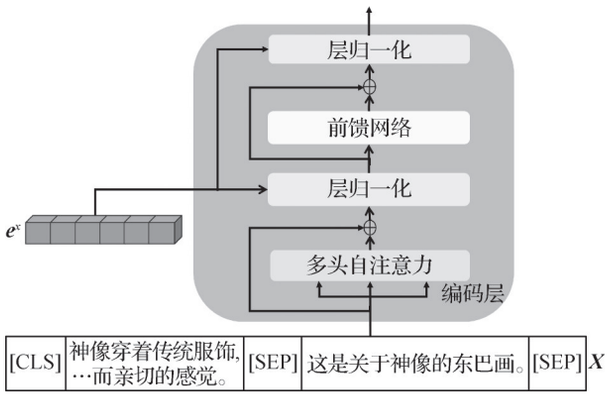


图4　编码层结构

Fig. 4　Structure of coding layer

和“温暖的氛围”相结合，共同作用于模型的隐藏状态。通过信息融合，生成文本能够更准确地反映图像的具体细节，使文本描述与图像内容更加一致。例如，图像中的神像特征会通过层归一化引导文本生成，确保描述中能够体现出“神像慈眉善目”的细腻情感表达。第1处层归一化和残差连接计算为

$$\boldsymbol{x}'=\frac{\boldsymbol{x}+A(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})+\boldsymbol{I}-mean\left(\boldsymbol{x}+A(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})+\boldsymbol{I}\right)}{\gamma\sqrt{var\left(\boldsymbol{x}+A(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})+\boldsymbol{I}\right)+\varepsilon}}+\beta \tag{2}$$

式中，$\boldsymbol{I}$为图像特征向量，$\boldsymbol{x}$为文本特征向量。$A(Q,K,V)$为多头自注意力机制的输出。γ和β为可学习的参数，ε为常数。

将第1处层归一化后的输出$\boldsymbol{x}'$送入前馈神经网络，进一步捕捉复杂特征，得到$FFN(\boldsymbol{x}')$。再进行第2个层归一化和残差连接，具体为

$$\boldsymbol{x}''=\frac{\boldsymbol{x}'+FFN(\boldsymbol{x}')+\boldsymbol{I}-mean\left(\boldsymbol{x}'+FFN(\boldsymbol{x}')+\boldsymbol{I}\right)}{\gamma\sqrt{var\left(\boldsymbol{x}'+FFN(\boldsymbol{x}')+\boldsymbol{I}\right)+\varepsilon}}+\beta \tag{3}$$

经过10层的编码层的处理，得到了编码后的隐藏状态表示H。随后，利用线性层将隐藏状态H转换为预测词汇表中每个位置单词的分数，并通过softmax函数将这些分数转换为概率分布。模型使用交叉熵损失（cross entropy loss）计算模型预测的概率分布与实际文本之间的差异。交叉熵损失函数为

$$L_{\text{text}}=-\frac{1}{bs}\sum_{t=1}^{T}\log p\left(y_t\left|y_{1:t-1},\boldsymbol{x}\right.\right) \tag{4}$$

式中，T为生成文本的长度，y_t为时间步t的目标词，即实际的文本描述中的词。$y_{1:t-1}$为在时间步t之前生成的所有词，$\boldsymbol{x}$为输入的文本特征向量，

$P\left(y_t \middle| y_{1:t-1}, \boldsymbol{x}\right)$为模型在时间步$t$生成目标词$y_t$的条件概率。$bs$为批量大小。

1.4 内容提示模块

为了充分利用东巴画主题明确的特点，本文在编码器之后引入了内容提示模块，其结构如图5所示。通过该模块，得到东巴画中宗教图案、神像精灵和地狱鬼怪等文化背景和特定的场景信息。解码器结合提示信息和文本信息，能更有效地理解和利用图像的内容信息，从而提升生成描述文本与图像主题的相关性和准确性。

首先将东巴画图像特征$\boldsymbol{e}^x$映射到全连接层(fully connected layers)，并采用归一化指数函数计算各个内容提示的概率分布。归一化后概率分布中的最大概率值即为模型所预测的提示信息，归一化指数函数为

$$S_i = \frac{\boldsymbol{e}^i}{\sum_{k=1}^{n} \boldsymbol{e}^k} \tag{5}$$

式中，S_i为某一类别的归一化数值，n为提示的类别数，$\boldsymbol{e}^i$和$\boldsymbol{e}^k$分别表示第i类和第k类提示的原始得分。计算$\max(S_i)$得到模型预测的东巴画内容标签。

以神像的东巴画为例，模块通过归一化图像特征得出与其内容相关的提示信息“神像”。其他可能的提示信息包括“署神”、“鬼怪”、“射箭”、“垂钓”或“吹笛”等帮助模型理解图像中潜在的主题和动作。

该模块将内容提示信息构建成prompt提示文本P，并将提示信息进行分词得到提示序列$\boldsymbol{e}^p$。以“神像”提示信息为例，该提示内容构建的提示文本为“这是关于神像的东巴画”，具体形式为

$$P = \text{这是关于}[\boldsymbol{X}, \boldsymbol{X}, \cdots, \boldsymbol{X}]\{b\}\text{的东巴画} \tag{6}$$

式中，$\boldsymbol{X}$为可学习的提示向量，b为图像的主体和动作信息。

为了提升内容提示模块的准确性，模型通过计算真实提示信息与预测提示信息之间的交叉熵损失，逐步优化预测输出。当模型预测提示信息为“署神”，而真实标签是“神像”时，交叉熵损失将用于量化两者的差异，并最小化损失函数来逐步调整其参数，生成符合目标提示的结果。内容提示模块的交叉熵损失函数为

$$L_{\text{prompt}} = -\frac{1}{bs}\sum_{i=1}^{bs}\log p(y_i|x_i) \tag{7}$$

式中，y_i为第i个样本真实的提示内容，x_i为第i个样本输入的提示内容，$p\left(y_i \middle| x_i\right)$为第$i$个样本属于真实提示内容$y_i$的概率。

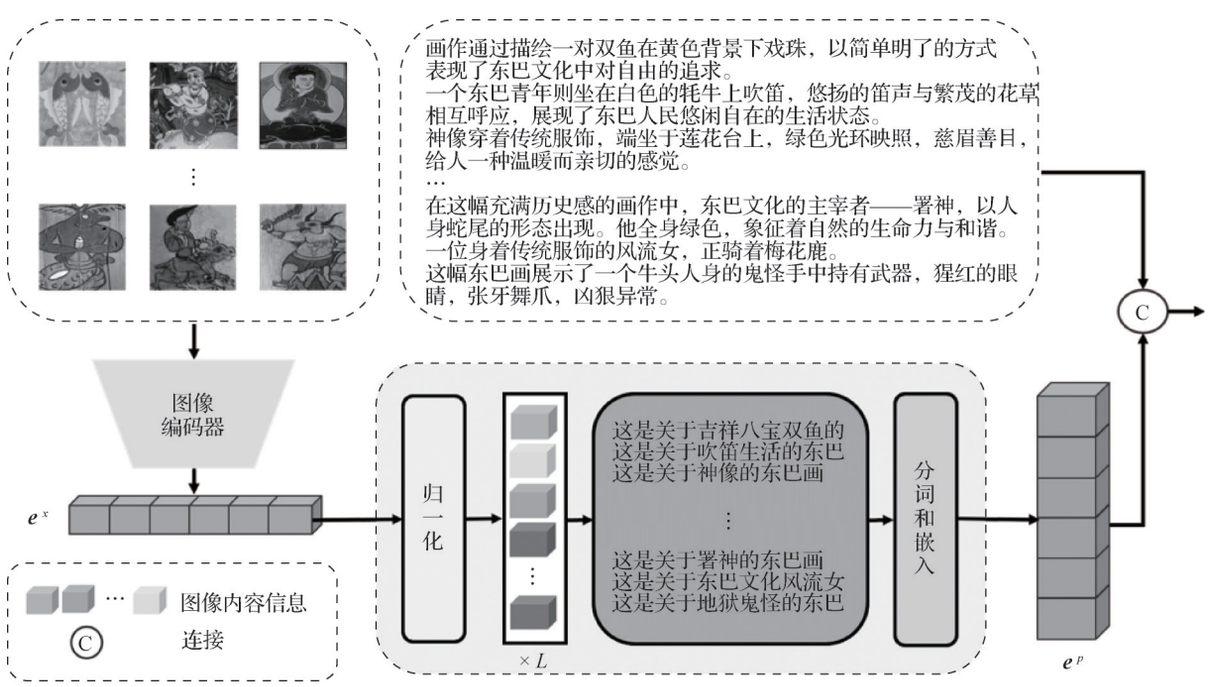


图5 内容提示模块结构

Fig. 5 Structure of content prompt module

1.5 视觉语义—生成融合损失

为了生成以内容为中心的图像描述，本文设计了内容提示模块，将内容提示作为后置prompt加入描述文本中。同时，为了确保生成的描述与图像高度相关，编码器提取的图像特征被输入至解码器的归一化层，以控制描述的生成过程。通过这种机制，模型能够增强对东巴画中图案和场景信息的识别能力，进而生成准确的内容中心图像描述。为了进一步提升模型描述的准确性，本文引入了视觉语义—生成融合损失。通过优化损失，模型能更有效地提取东巴画中的关键信息，从而生成与图像保持高度一致的描述文本。具体而言，视觉语义—生成融合损失包括提示模块损失L_{prompt}和模型生成损失L_{text}。具体为

$$L_{\text{fusion}} = \alpha L_{\text{text}} + \lambda L_{\text{prompt}} \tag{8}$$

式中，α和λ为加权参数。

2 实验结果与分析

2.1 实验数据

东巴画是纳西族的艺术瑰宝，主要有经卷画、木牌画、纸牌画和卷轴画等形式，内容丰富、包罗万象，再现了纳西族自然崇拜的观念和丰富多彩的社会生活。从东巴画相关文献（钱文华 等，2020）收集真实纳西族东巴画，将图像数据按内容分为神像精灵、地狱鬼怪、鸟兽虫鱼、花草树木、骑马垂钓、吹奏歌舞以及宗教图案等，并描述了每一类题材画面的内容和其在东巴文化中的含义。例如，在东巴画中，动物常被描述为东巴文化中神祇的坐骑或是有独特意义的神禽异兽。如：牦牛是纳西族的“门神”和东巴教祭坛的守护神，象征着骁勇威猛；白鹤是吉祥鸟，常被视为传递爱情的使者；白蝙蝠机智灵敏，在东巴文化的神话传说中，它充当信使，以神雕为坐骑飞抵天界，为人类求取卜书。同时，东巴画中常出现纳西八宝等传统文化元素的宗教图案，包括净水瓶、如意结、法螺、莲花和双鱼等，体现了人们对美好生活的向往与追求。如：净水瓶在东巴文化中象征着吉祥、财运和清净；吉祥结在东巴文化中寓意着团结、和睦、祥和、连续不断和福寿无疆；双鱼能在水中自由生长，因此在东巴文化中象征着无限生机。另一方面，东巴画也常以鬼怪为主题。在东巴文化中，灵魂回归祖先之地的途中，有各种鬼怪阻止它们去往神地和人间。如：在东巴布卷画中，描绘了鬼怪在地狱中对偷盗耕牛、滥杀野兽和污染河流的人进行惩罚的场景。东巴画图像描述主要描述画中对象的动作、背景、色彩以及对象与对象之间的关系等特征。本文东巴画图像描述数据集部分示例如图6所示。

2.2 数据集扩充

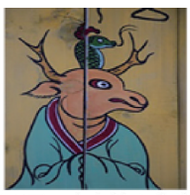
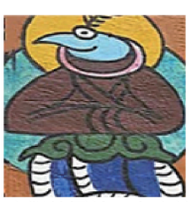
深度学习模型训练通常需要大量的东巴画图像描述数据集。为了缓解过拟合问题，本文对东巴画数据集采取了几何变换、噪声添加和像素操作3种数据增强方法对数据集进行扩充。通过几何变换的方式增强数据，帮助模型学习图像在不同角度下的特征，增强模型的泛化能力。通过加入随机噪点的方式增强数据，使模型在有干扰的情况下仍能准确描述图像内容，减少对无关或噪声特征的依赖，增强对主要描述信息（如对象动作和背景色彩）的关注。通过像素操作和锐化的方式增强数据，模拟东巴画在自然条件下（如氧化等过程）可能出现的腐蚀效果。利用数据增强技术有效扩展了数据集的规模和多样性，提升了模型在未见数据上的表现。经过数据增强处理，原始数据集的规模扩展至9 408幅图像。

2.3 实验设置

实验环境基于TensorFlow和Keras深度学习框架，显卡型号为NVIDIA RTX 2080Ti，模型的编码器为MobileNetV2，解码器为10层Transformer结构，并引入内容提示模块辅助图像描述生成。为加速东巴画图像描述生成任务的收敛，各模型均采用了其相对应的预训练模型进行微调，训练过程中权重衰减为0.8/8 epoch，batch_size设置为8，初始学习率为0.000 01，优化器为AdamW（Loshchilov和Hutter，2017），并使用交叉熵损失函数对模型进行微调。

2.4 评价指标

为验证模型的有效性，本文采用领域内常用的评估策略，包括BLEU（bilingual evaluation understudy）（Papineni 等，2002）、METEOR（metric for evaluation with explicit ordering）（Banerjee和Lavie，2005）、ROUGE（recall-oriented understudy for gisting evaluation）（Lin，2004）和CIDEr（consensus-based image description evaluation）（Vedantam等，2015）4种图像描述指标。通过计算4种描述指标的得分情况，评估模型生成的描述文本与参考文本之间的相

鬼怪图像			
标注语句	画面中展现了**一个鹿首人身的鬼怪**。它身着绿色衣服，置于黄色背景中。作品生动地表现了**东巴文化中神秘元素以及对自然神灵的敬畏**。	图中的**鼠头人身的鬼怪**身着绿色衣服和黄色裤子，置于蓝色背景之中。画面色彩对比强烈，突出了东巴画的艺术特色。	这幅东巴画展示了一个**牛头人身的鬼怪**，身着蓝色裤子。背景为棕色且有云纹装饰，这种组合展示了东巴文化的独特艺术风格。
署神图像			
标注语句	这幅画作展现了**东巴文化中主宰自然的署神形象**，他**头戴珠宝冠，人身蛇尾**，气宇轩昂。署神背部的花纹装饰，凸显其**威严与神圣**。	身穿棕色衣服的**署神，人身蛇尾**，形象生动传神。背部的黄色和蓝色背光，既丰富了画面的色彩层次，又凸显了**署神的威严与神圣**，令人心生敬畏。	在这幅充满历史感的画作中，**东巴文化的主宰者——署神，以人身蛇尾的形态**出现。他全身绿色，**象征着自然的生命力与和谐**。
吉祥八宝图像			
标注语句	**海螺，在东巴文化中代表好运**。画面中的海螺被一只手托着，其黄色背景增添了画面的活力。	画作通过描绘一对双鱼在黄色背景下戏珠，以简单明了的方式表现了**东巴文化中对自由的追求**。	画中的吉祥结以粉红和红色为主，搭配云纹与绿色饰带。**绳结盘曲，首尾相连，寓意着团结、和睦、福寿无疆**。

注: 加粗字体为该东巴画在东巴文化中的传统形象或独特含义。

图 6　东巴画数据集部分示例

Fig. 6　Examples of Dongba painting datasets

似性。

BLEU 通过 n 元组匹配度评估生成文本与参考文本间的准确性，反映生成图像描述的准确性和流畅性。n 可取 1、2、3、4 等自然数。指标的具体计算为

$$P_n = \frac{\sum_{z=1}^{Z}\sum_{k=1}^{n}\min\left(h_k(c_z), \max\limits_{m\in \boldsymbol{M}} h_k(s_{zm})\right)}{\sum_{z=1}^{Z}\sum_{k=1}^{n}h_k(c_z)} \tag{9}$$

$$BP = \begin{cases} 1 & g > r \\ \exp\left(1-\dfrac{r}{g}\right) & g \leqslant r \end{cases} \tag{10}$$

$$f_{\mathrm{BLEU}} = BP \times \exp\left(\sum_{n=1}^{N} w_n \log\left(P_n\right)\right) \tag{11}$$

式中，P_n 为生成文本与参考文本的 n 元组精确匹配比例。BP 为惩罚因子，用于避免生成过短的文本，N 为 n 元组的阶数，Z 为候选文本数。g 为生成文本长度，r 为参考文本长度，w_n 为权重。$h_k(c_z)$ 为第 z 个生成文本 c_z 中 k 元组的计数，$h_k(s_{zm})$ 为第 z 个生成文本对应的参考文本 s_{zm} 中 k 元组的计数，$\sum_{z=1}^{Z}\sum_{k=1}^{n}h_k(c_z)$ 以及 $\max\limits_{m\in \boldsymbol{M}} h_k(s_{zm})$ 分别是参考文本中 n 元组的总数量和参考文本集 $\boldsymbol{M}$ 中 k 元组的最大计数。

METEOR 通过词汇匹配、词形变化匹配和语义相似性评估生成文本的质量，即评估图像描述的词语和语义。其计算为

$$f_{\mathrm{METEOR}} = (1-P_{en}) \times \frac{P_m R_m}{\eta P_m + (1-\eta) R_m} \tag{12}$$

式中，η 为评价时的默认参数，P_{en} 为词序不匹配时的惩罚因子，P_m 为匹配的词汇数量占生成文本总词汇的比例，R_m 为匹配的词汇数量占参考文本总词汇的比例。

ROUGE_L基于最长公共子序列评估生成文本与参考文本之间的相似度，侧重于评估图像描述的召回率和覆盖度。其计算为

$$f_{\text{ROUGE\_L}}=\frac{\left(1+\mu^2\right)R_{\text{lcs}}P_{\text{lcs}}}{R_{\text{lcs}}+\mu^2P_{\text{lcs}}} \tag{13}$$

式中，μ为常数。R_{lcs}为准确率，表示生成文本与参考文本的公共子序列在生成文本中的比例。P_{lcs}为召回率(recall)，表示生成文本与参考文本之间的公共子序列在参考文本中的覆盖比例。

CIDEr通过计算生成文本与多个参考文本之间的n元组相似度，并加权词频—逆文档频率(term frequency-inverse document frequency，TF-IDF)，来衡量生成文本与参考文本的相关性以及多样性。

TF-IDF通过降低常见词权重，提升稀有词权重，提高描述文本中独特词汇的权重。CIDEr计算为

$$f_{\text{CIDEr}}(C_i,S_i)=\frac{1}{J}\sum_{j=1}^{J}\sum_{n=1}^{N}\frac{\boldsymbol{g}_n(c_i)\times\boldsymbol{g}_n(s_{ij})}{\left\|\boldsymbol{g}_n(c_i)\right\|\times\left\|\boldsymbol{g}_n(s_{ij})\right\|} \tag{14}$$

式中，J为每个生成文本对应的参考文本数量。$\boldsymbol{g}_n(c_i)$和$\boldsymbol{g}_n(s_{ij})$分别为生成文本c_i和参考文本s_{ij}的TF-IDF向量。

2.5 不同编码器对模型性能的影响

在本文模型中，编码器用于提取图像特征，并将其输入解码器的编码层中，以控制东巴画描述的生成。为了验证编码器类型对图像描述质量的影响，本文分别选取ResNet-50(residual network)、VGG16(Visual Geometry Group network)、EfficientNetB0(efficient network)、DenseNet121(dense convolutional network)和MobileNetV2作为编码器对东巴画进行特征提取，并保持模型其他结构不变，比较不同编码器下模型的评价指标，对比结果如表1所示。当模型编码器为MobileNetV2时，参数量为8.73 M，图像描述指标BLEU_1、BLEU_2、BLEU_3、BLEU_4、METEOR、ROUGE和CIDEr分别为0.603、0.426、0.317、0.246、0.256、0.403和0.599。将编码器替换为EfficientNetB0和VGG16后，参数量分别增加了6.82 M和47.44 M，但各项指标均有不同程度的下降。采用DenseNet121进行替换后，参数量增加了18.2 M，仅BLEU_4指标有所提升(增幅为0.006)，而其他描述指标均有所下降。当模型参数量显著增加时，例如表1中ResNet-50模型参数量增加了81.43 M，仅BLEU_1略有提升，其余图像描述指标(BLEU_2、BLEU_3、BLEU_4、METEOR、ROUGE和CIDEr)均出现不同程度的下降，分别下降了0.004、0.003、0.003、0.004、0.004和0.054。

综上所述，综合考虑各个编码器模型的参数量及其在各项图像描述指标上的表现，MobileNetV2能够在保持较低参数量的同时，提供较为平衡的图像描述性能。因此，本文选择MobileNetV2作为模型的编码器。

表1 不同编码器模型的图像描述评价指标对比

Table 1 Comparison of model image caption evaluation indexes of different encoders

模型	参数量/M	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE	CIDEr
MobileNetV2	8.73	0.603	**0.426**	**0.317**	0.246	**0.256**	**0.403**	**0.599**
EfficientNetB0	15.55	0.588	0.410	0.311	0.247	0.251	0.399	0.571
DenseNet121	26.93	0.581	0.408	0.313	**0.252**	0.249	0.396	0.549
VGG16	56.17	0.554	0.352	0.251	0.187	0.223	0.365	0.386
ResNet-50	90.16	**0.606**	0.422	0.314	0.243	0.252	0.399	0.545

注：加粗字体表示各列最优结果。

2.6 解码器编码层数量对模型性能的影响

为了验证解码器中编码层数量对东巴画描述质量的影响，本文在保持模型其他结构不变的前提下，分别采用不同编码层数量的解码器对图像特征向量进行解码，并分析其对图像描述评价指标的影响。图7展示了编码层层数分别为6、8、10、12时，解码器在图像描述任务中的表现。

从图7(a)(b)可以观察到，当模型解码器编码层数量增加时，图像描述准确性和流畅性评估指标以及相关性和多样性评估指标均有所提升，说明增

加解码器编码层的数量一定程度上提高了生成文本在描述图像细节方面的准确性，增强了描述语句的多样性，且句子结构自然连贯、清晰易懂。从图7(c)观察到，随着解码器编码层数量的增加，召回率和覆盖率评估指标增幅减缓，说明增加解码器编码层的数量在改进生成的描述在覆盖参考描述的关键信息方面有一定的改进。

综上所述，解码器编码层数量的增加能够有效提升东巴画图像描述文本的质量，但随着层数的增加，性能提升逐渐减缓，甚至出现下降。因此，本文模型选择10层作为解码器编码层数量，避免堆叠解码器—编码层带来过拟合问题。

2.7 与其他算法比较

为了验证本文模型的有效性，选取BLIP、ViECap和MacCap等图像描述模型进行比较。

2.7.1 客观质量评价

本文对不同模型生成的东巴画图像描述进行了评估，对比结果如表2所示。从表2可以观察到，本文PVGF-DPC模型的图像描述客观评价指标BLEU_1、BLEU_2、BLEU_3、BLEU_4、METEOR、ROUGE和CIDEr分别达到0.603、0.426、0.317、0.246、0.256、0.403和0.599。

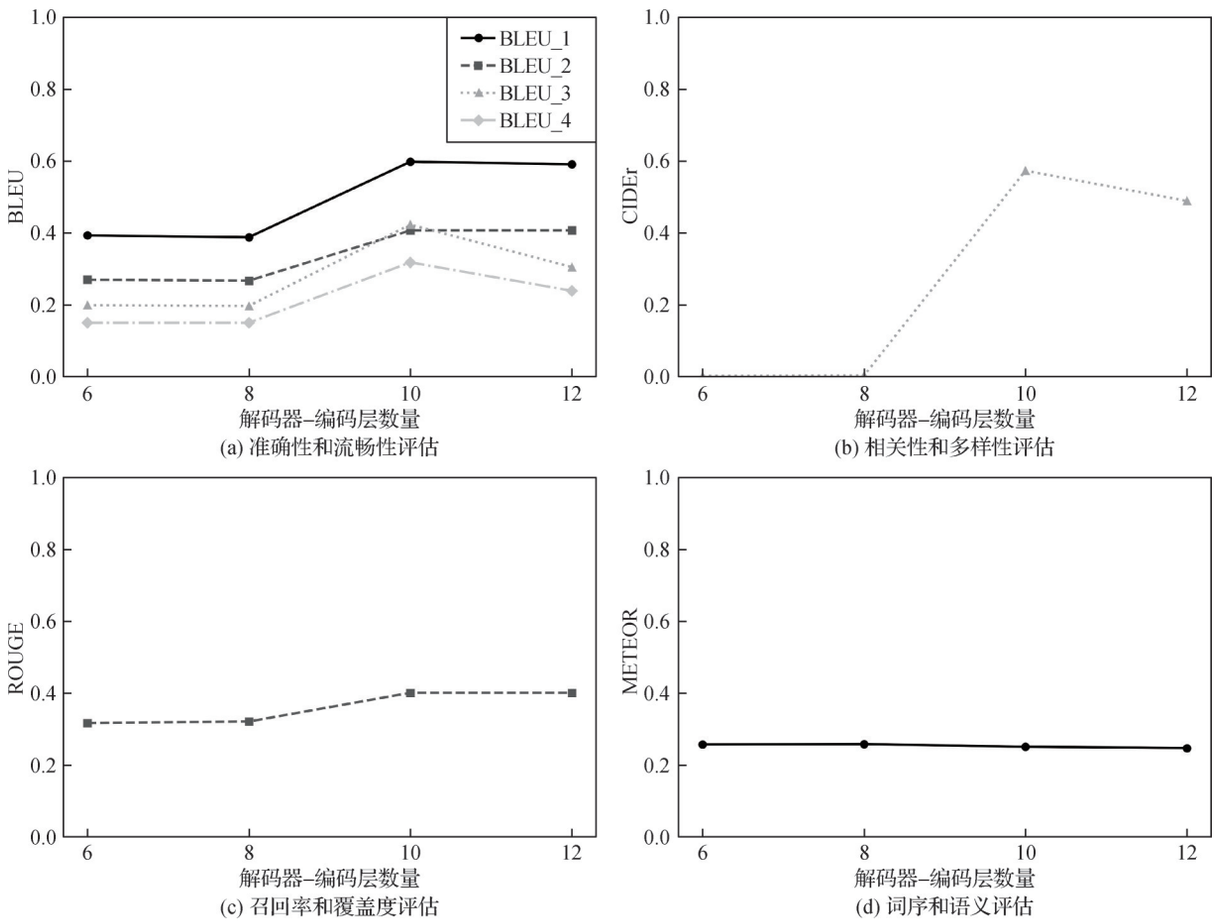


图7 不同数量的解码器编码层图像描述指标对比

Fig. 7 Comparison of different numbers of decoder coding layer image caption index ((a) evaluation of accuracy and fluency; (b) evaluation of relevance and diversity; (c) evaluation of recall rate and coverage; (d) evaluation of word order and semantics)

本文方法与其他方法在各评价指标上的对比分析如图8所示。通过雷达图分析，本文提出的PVGF-DPC模型在东巴画测试集上的图像描述客观指标相较于其他图像描述模型都有不同程度的提高。在准确性和流畅性方面，评估指标BLEU_1、BLEU_2、BLEU_3和BLEU_4相较于排名第2的ClipCap(clip prefix for image captioning)模型分别提高0.106、0.217、0.178和0.142。在相关性和多样性方面以及CIDEr指标相较于排名第2的ViECap提高0.416。在词序和语义方面，METEOR指标相较于排名第2的OFA(one for all)模型提高0.005。

表2　东巴画测试集上本文方法与其他方法图像描述指标对比

Table 2　Comparison of image caption index between the proposed method and other methods on the Dongba painting testset

模型	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE	CIDEr
ATT + LSTM	0.381	0.104	0.063	0.045	0.178	0.400	0.046
ClipCap(Mokady等,2021)	0.497	0.209	0.139	0.104	0.239	**0.472**	0.142
BLIP(Li等,2022)	0.365	0.102	0.054	0.030	0.180	0.361	0.026
OFA(Wang等,2022)	0.066	–	–	–	0.251	0.191	–
DeCap(Li等,2023)	0.208	–	–	–	0.104	0.280	–
CgT-GAN(Yu等,2023)	0.372	0.288	0.091	0.033	0.148	0.299	0.038
ViECap(Fei等,2023)	0.457	0.285	0.185	0.127	0.148	0.308	0.183
MeaCap(Zeng等,2024)	0.119	0.056	0.026	0.011	0.072	0.185	0.014
MacCap(Qiu等,2024)	0.186	0.123	0.079	0.056	0.083	0.245	0.012
PVGF-DPC(本文)	**0.603**	**0.426**	**0.317**	**0.246**	**0.256**	0.403	**0.599**

注:加粗字体表示各列最优结果,“–”表示实验数据失真或值低于检测限。

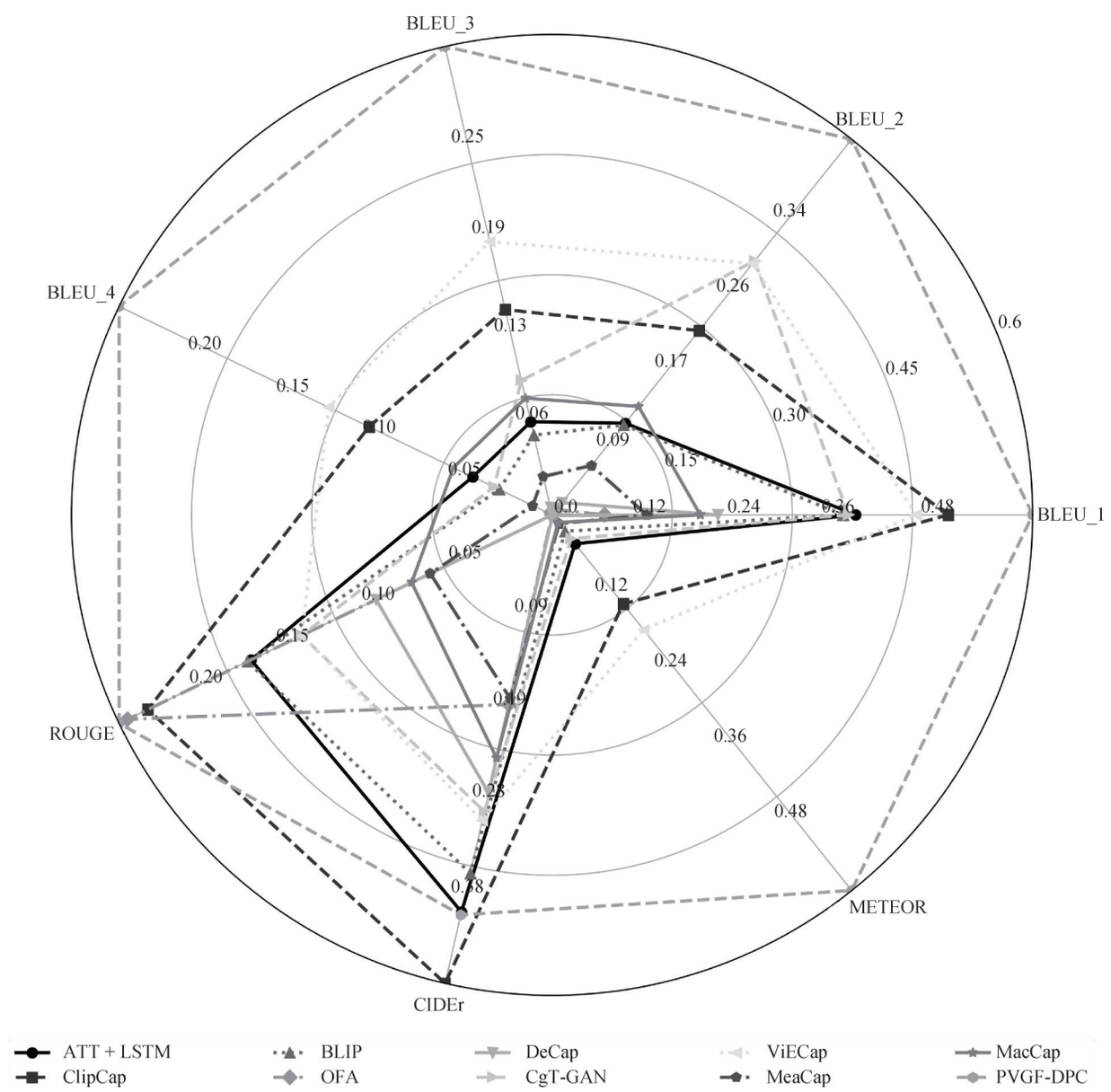


图8　本文方法与其他方法雷达图

Fig. 8　Index radar map of proposed method and other methods

2.7.2　主观质量评价

本文模型在各项图像描述评价指标上均取得了较好的结果，但为了更直观地说明其生成的图像描述结果与东巴画内容的匹配程度，并展现其背后的东巴文化内涵，本文选取了东巴画测试集中的图像描述语言、ClipCap、BLIP、ViECap 和 MacCap 模型生成的描述文本与本文实验结果进行比较。

根据实验结果显示，本文 PVGF-DPC 模型能够更好地识别图像中的物体，更全面地描述东巴画中的细节以及文化内涵。在图 9 示例 1 中，本文提出的 PVGF-DPC 模型对图像的描述更准确和流畅。其不仅能识别图像中关键的主体信息及其动作，例如“白蝙蝠挥动翅膀”，还能描述其在东巴文化中的神话传说，如“在东巴文化的神话传说中，亨英津白盘（白蝙蝠）充当信使，以阿罗休贡（神雕）为坐骑飞抵天界，求取卜书”，且描述的语言更加流畅自然。而 BLIP、ViECap 和 MacCap 模型的描述结果则出现了主体描述错误的情况，分别将蝙蝠识别成与其类似的白鹤、白龙以及鸟，ClipCap 模型则因为模型的限制语言描述不够流畅。在图 9 的示例 2 中，本文的 PVGF-DPC 模型对东巴文化中署神的细节和传统形象描述更加准确。具体而言，“黄色衣服的署神”描述出了图中主体的服饰信息，“头戴璀璨的珠宝冠，人身蛇尾，形象庄重而威严”则对署神在东巴文化中的传统形象进行了描述。在图 9 示例 3 中，本文的 PVGF-DPC 模型对图像主体在东巴文化中的内涵方面的描述更加准确。净水瓶是纳西族的吉祥八宝之一，在东巴文化中象征着吉祥、财运和清净。本文 PVGF-DPC 模型详细地描述了图像中净水瓶的外观细节，如“蓝色净水瓶”以及“绿色饰带点缀”，并全面阐述了其在东巴文化中的具体含义，即“吉祥和财运，以及精神的清净”。

示例 1		真实描述	在东巴文化的神话传说中，亨英津白盘 (白蝙蝠) 充当信使，以阿罗休贡 (神雕) 为坐骑飞抵天界，求取卜书。画面中白蝙蝠在空中飞翔，背景是流动的黄色云纹。
		ClipCap	在东巴文化的神话传说中，亨英津白盘 (白蝙蝠) 充当信使，以阿罗休贡 (神雕) 为坐骑飞抵天界，求取卜书。画面中白蝙蝠在空中展翅翱翔，它在其。
		BLIP	画中的白鹤，作为纳西族的吉祥鸟，优雅地站在岩石上展翅飞翔，背景色为温馨的
		ViECap	在这幅东巴画中，一条绿色鬃毛和尾巴的白龙活了起来，它圆圆的大眼睛和锋利的爪子增加了凶猛，棕色的背景增强了画面的深度感。
		MacCap	在这幅画的中央，一只有着绿色身体和尾巴的鸟
		PVGF-DPC	**在东巴文化的神话传说中，亨英津白盘 (白蝙蝠) 充当信使，以阿罗休贡 (神雕) 为坐骑飞抵天界，求取卜书。画面中白蝙蝠挥动翅膀，在空中翩翩起舞。**
示例 2		真实描述	黄色衣服的署神，在画面中格外醒目。他头戴珠宝冠，人身蛇尾，形象生动传神。背景中云纹装饰，不仅丰富了画面的视觉层次，也为署神的形象增添了一抹神圣与庄严。
		ClipCap	画面中的署神，头戴珠宝冠，人身蛇尾，形象生动而传神。画面整体色彩饱满，线条流畅，将署神的威严与神秘展现得淋漓尽致。
		BLIP	画中的男子，头戴花饰，身着蓝色长裤，坐在草地上吹奏笛子，背景是宁静的云纹
		ViECap	在图片的中央，一个男人戴着一顶白色的帽子
		MacCap	这位神身穿棕色衣服，头上戴着一顶明亮的珠宝王冠，他的蛇尾在画面中尤为突出。
		PVGF-DPC	身穿**黄色衣服的署神，头戴璀璨的珠宝冠，人身蛇尾，形象庄重而威严**。整幅画作色彩鲜明，线条流畅，将东巴文化中署神的形象刻画得栩栩如生。
示例 3		真实描述	这幅画中，一个蓝色的净水瓶位于棕色背景之中，瓶身由绿色饰带装饰点缀，象征着东巴文化中的吉祥和财运。
		ClipCap	在这幅东巴画作中，一只蓝色净水瓶被绿色的饰带点缀，置于棕色背景中，不仅表现了视觉的美感，也象征着纳西文化中的清净和吉祥。
		BLIP	画中，一个蓝色净水瓶在棕色背景下展示，瓶身的绿色饰带装饰象征着吉祥和清净。
		ViECap	这幅画以棕色为背景，展示了一个蓝色的净水瓶，上面装饰着绿色的腰带，在东巴文化中代表着好运、财富和清洁。
		MacCap	在画面中央，人物戴着一顶蓝色的帽子。
		PVGF-DPC	在棕色背景中，一只**蓝色净水瓶**以其**绿色饰带点缀，象征纳西文化中的吉祥和财运，以及精神的清净**。

注：加粗字体为本文方法描述更符合真实描述处。

图 9　东巴画测试集上本文方法与其他方法图像描述文本主观质量对比

Fig. 9　Comparison of subjective quality of image description text between the proposed method and other methods on the Dongba painting testset

根据图9呈现的结果显示，本文提出的PVGF-DPC模型在东巴画描述语言上表现更为流畅自然。其在画面主体和细节的识别方面表现也更加准确，并且能更全面地表达画面主体所蕴含的文化内涵。

2.8 消融实验

为了证明内容提示模块和视觉语义—生成融合损失对东巴画描述生成任务的有效性，本文在保持编码器和解码器结构的基础上，设置了相应的消融模型。实验结果如表3所示。表3中DBC（Dongba captioning）表示不使用视觉语义—生成融合损失训练模型，VGF-DPC（visual semantic-generation fusion-based Dongba painting captioning）表示仅采用视觉语义—生成融合损失训练模型，不采用内容提示模块增强解码器对文本的学习，PVGF-DPC则是本文提出的模型。

根据实验结果显示，在图像描述的准确性和流畅性方面，本文模型相较于DBC模型和VGF-DPC模型都有一定程度的提升，说明内容提示模块和视觉语义—生成融合损失能提高东巴画描述语言的准确性和流畅性，使得生成的东巴画图像描述文本更加自然流畅，易于理解。在图像描述的召回率和覆盖度方面，本文模型相较于DBC模型和VGF-DPC模型ROUGE指标分别提高0.011和0.002，说明提示模块在覆盖参考描述的关键信息方面有所改进。在图像描述的相关性和多样性方面，本文模型相较于DBC模型和VGF-DPC模型CIDEr指标分别提高0.073和0.11，说明本文模型能够更精确地捕捉图像中的细节，生成的东巴画图像描述更丰富多样。

表3 消融实验

Table 3 Ablation experiment

模型	BLEU_1	BLEU_2	BLEU_3	BLEU_4	METEOR	ROUGE	CIDEr
DBC	0.575	0.392	0.291	0.227	0.247	0.392	0.526
VGF-DPC	0.591	0.407	0.305	0.239	0.247	0.401	0.489
PVGF-DPC	**0.603**	**0.426**	**0.317**	**0.246**	**0.256**	**0.403**	**0.599**

注：加粗字体表示各列最优结果。

综上所述，通过引入视觉语义—生成融合损失和内容提示模块，可以有效增强编码器在训练过程中对图像主题相关特征的捕捉能力，有助于提升解码器对图像和文本信息的理解和利用，进而显著提高生成描述文本与图像主题的相关性和准确性。

3 结 论

针对东巴画视觉元素丰富、样本较少的特点，本文提出一种结合提示学习和视觉语义—生成融合的东巴画图像描述模型。引入了内容提示模块，通过图像特征向量得到图像的主体、动作等信息，并将其构建成描述文本的后置提示。同时，本文引入了视觉语义—生成融合损失，有效引导模型提取东巴画中的关键视觉元素，进而生成了与东巴画语义高度相关且富含文化内涵的图像描述文本。实验结果表明，在东巴画图像描述生成任务中，结合内容提示模块和视觉语义—生成融合损失训练模型，相较于单独采用编码器—解码器结构的模型，取得了更优的效果。与其他方法对比，本文方法在客观评价指标和文本主观质量上均获得了更佳的结果，进一步表明了本文模型在生成东巴画描述文本方面的有效性。然而，当目标风格相近时，编码器对图像中主体的识别可能存在混淆，导致生成的描述存在不确定性。此外，由于东巴画数据集的规模和多样性有限，模型生成的描述可能会受到约束。未来，将通过优化模型的提示任务来捕捉图像更精细和多样性特征，并扩大东巴画数据集。

参考文献（References）

Anderson P, He X D, Buehler C, Teney D, Johnson M, Gould S and Zhang L. 2018. Bottom-up and top-down attention for image captioning and visual question answering//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6077-6086 [DOI: 10.1109/CVPR.2018.00636]

Banerjee S and Lavie A. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evalua-

tion Measures for Machine Translation and/or Summarization. Ann Arbor, USA: ACL: 65-72

Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA: Association for Computational Linguistics: 4171-4186 [DOI: 10.18653/v1/N19-1423]

Devlin J, Gupta S, Girshick R, Mitchell M and Zitnick C L. 2015. Exploring nearest neighbor approaches for image captioning [EB/OL]. [2024-08-01]. https://arxiv.org/pdf/1505.04467.pdf

Fei J J, Wang T, Zhang J R, He Z Y, Wang C J and Zheng F. 2023. Transferable decoding with visual entities for zero-shot image captioning//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3113-3123 [DOI: 10.1109/iccv51070.2023.00291]

Johnson J, Karpathy A and Li F F. 2016. DenseCap: fully convolutional localization networks for dense captioning//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4565-4574 [DOI: 10.1109/CVPR.2016.494]

Li J N, Li D X, Xiong C M and Hoi S. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR: 12888-12900

Li S M, Kulkarni G, Berg T L, Berg A C and Choi Y. 2011. Composing simple image descriptions using web-scale N-grams//Proceedings of the 15th Conference on Computational Natural Language Learning. Portland, USA: Association for Computational Linguistics: 220-228 [DOI: 10.5555/2018936.2018962]

Li W, Zhu L C, Wen L Y and Yang Y. 2023. DeCap: decoding CLIP latents for zero-shot captioning via text-only training [EB/OL]. [2024-08-01]. https://arxiv.org/pdf /2303.03032.pdf

Lin C Y. 2004. ROUGE: a package for automatic evaluation of summaries//Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics: 74-81

Loshchilov I and Hutter F. 2017. Decoupled weight decay regularization [EB/OL]. [2024-08-01]. https://arxiv.org/pdf/1711.05101.pdf

Mokady R, Hertz A and Bermano A H. 2021. ClipCap: CLIP prefix for image captioning [EB/OL]. [2024-08-01]. https://arxiv.org/pdf/2111.09734.pdf

Papineni K, Roukos S, Ward T and Zhu W J. 2002. BLEU: a method for automatic evaluation of machine translation//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA: Association for Computational Linguistics: 311-318 [DOI: 10.3115/1073083.1073135]

Qian W H, Xu D, Xu J, He L and Zhang B. 2020. Simulation of Dongba art style painting. Journal of System Simulation, 32(7): 1349-1359 (钱文华, 徐丹, 徐瑾, 何磊, 张波. 2020. 东巴画艺术风格绘制. 系统仿真学报, 32(7): 1349-1359) [DOI: 10.16182/j.issn1004731x.joss.19-VR0530]

Qiu L T, Ning S and He X M. 2024. Mining fine-grained image-text alignment for zero-shot captioning via text-only training//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 4605-4613 [DOI: 10.1609/aaai.v38i5.28260]

Sandler M, Howard A, Zhu M L, Zhmoginov A and Chen L C. 2018. MobileNetV2: inverted residuals and linear bottlenecks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4510-4520 [DOI: 10.1109/cvpr.2018.00474]

Su Q. 2022. The blending and coexistence of aesthetic imagery: on the evolution of Dongba painting style. Gansu Social Sciences, (5): 98-104 (苏泉. 2022. 审美意象的交融共生: 论东巴画造型风格的衍进. 甘肃社会科学, (5): 98-104) [DOI: 10.3969/j.issn.1003-3637.2022.05.011]

Tang P J, Tan Y L and Li J Z. 2017. Image description based on the fusion of scene and object category prior knowledge. Journal of Image and Graphics, 22(9): 1251-1260 (汤鹏杰, 谭云兰, 李金忠. 2017. 融合图像场景及物体先验知识的图像描述生成模型. 中国图象图形学报, 22(9): 1251-1260) [DOI: 10.11834/jig.170052]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates, Inc.: 6000-6010

Vedantam R, Zitnick C L and Parikh D. 2015. CIDEr: consensus-based image description evaluation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 4566-4575 [DOI: 10.1109/CVPR.2015.7299087]

Vinyals O, Toshev A, Bengio S and Erhan D. 2015. Show and tell: a neural image caption generator//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3156-3164 [DOI: 10.1109/CVPR.2015.7298935]

Waghmare P M and Shinde S V. 2022. Image caption generation using neural network models and LSTM hierarchical structure//Das A K, Nayak J, Naik B, Dutta S and Pelusi D, eds. Computational Intelligence in Pattern Recognition. Singapore, Singapore: Springer: 109-117 [DOI: 10.1007/978-981-16-2543-5_10]

Wang N, Xie J H, Wu J H, Jia M B and Li L L. 2023. Controllable image captioning via prompting//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI: 2617-2625 [DOI: 10.1609/aaai.v37i2.25360]

Wang P, Yang A, Men R, Lin J Y, Bai S, Li Z K, Ma J X, Zhou C, Zhou J R and Yang H X, 2022. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework//Proceedings of the 39th International Conference on

Machine Learning. Baltimore, USA: PMLR: 23318-23340

Yang H R. 2021. Study on the artistic features of Naxi Dongba pictographic scripts. Chinese Minzu Art, (3): 60-63 (杨鸿荣. 2021. 纳西族东巴画谱艺术特征研究. 中国民族美术, (3): 60-63)

Yu J R, Li H R, Hao Y B, Zhu B, Xu T and He X N. 2023. CgT-GAN: CLIP-guided text GAN for image captioning//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: Association for Computing Machinery: 2252-2263 [DOI: 10.1145/3581783.3611891]

Zeng Z Q, Xie Y, Zhang H, Chen C Y, Chen B and Wang Z J. 2024. MeaCap: memory-augmented zero-shot image captioning//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 14100-14110 [DOI: 10.1109/cvpr52733.2024.01337]

Zhao Y Q, Jin Z, Zhang F, Zhao H Y, Tao Z W, Dou C F, Xu X H and Liu D H. 2023. Deep-learning-based image captioning: analysis and prospects. Journal of Image and Graphics, 28(9): 2788-2816 (赵永强, 金芝, 张峰, 赵海燕, 陶政为, 豆乘风, 徐新海, 刘东红. 2023. 深度学习图像描述方法分析与展望. 中国图象图形学报, 28(9): 2788-2816) [DOI: 10.11834/jig.220660]

Zheng Y, Li Y L and Wang S J. 2019. Intention oriented image captions with guiding objects//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 8395-8404 [DOI: 10.1109/cvpr.2019.00859]

作者简介

罗霜,女,硕士研究生,主要研究方向为深度学习和计算机视觉。E-mail:luoshuang@stu.ynu.edu.cn

钱文华,通信作者,男,教授,博士生导师,主要研究方向为计算机视觉和文化计算。E-mail:whqian@ynu.edu.cn

刘朋,男,博士研究生,主要研究方向为计算机视觉和图像处理。E-mail: pengliu0606@gmail.com