中图法分类号:TP 391.41　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-15

论文引用格式:Yang Rongyi, Qian Wenhua, Liu Peng. Color and Brush Stroke Features Integrated for Emotional Classification of Dongba Paintings[J/OL]. Journal of Image and Graphics, XXXX:1-15. DOI: 10.11834/jig.260014.(杨荣懿, 钱文华, 刘朋, 曹进德. 融合颜色笔触特征的东巴画情感分类[J/OL]. 中国图象图形学报,XXXX:1-15. DOI: 10.11834/jig.260014.)[DOI:10.11834/jig.260014]

融合颜色笔触特征的东巴画情感分类

杨荣懿[1],钱文华[1*],刘朋[1],曹进德[2]

1. 云南大学信息学院,昆明　650504; 2. 东南大学数学学院,南京　210009

摘　要:**目的**　东巴画作为纳西族珍贵的文化遗产,其绘画主体通过独特的符号体系、绘画色调与笔触技法传递复杂情感。现有视觉-语言模型的预训练范式偏向通用对象分类任务,面对东巴画这类具有特定文化属性的艺术形式时,难以有效建立多维度情感语义与绘画色彩结构、笔触纹理美学特征之间的映射关系,存在领域偏移引发的情感语义映射偏差问题。为此,本文提出一种融合颜色笔触特征的东巴画情感分类方法。**方法**　针对东巴画的色彩结构与笔触纹理特征,提出双域颜色笔触特征提取机制,通过空间域和频率域捕捉其线条走向与边缘细节信息;基于多模态模型的跨模态对齐空间,构建渐进式自适应融合架构,动态平衡通用视觉信息与特定艺术特征的贡献权重,以适配东巴画多元的绘画风格差异;针对标签不平衡场景下的多情感共存问题,引入绘画主体的对象类别先验知识,设计类别引导的级联问答推理机制,强化分类任务的逻辑性与准确性,有效提升模型在东巴画场景下的领域泛化能力。**结果**　结合纳西族文化典籍与历史文献考证,本文构建包含14维情感标签的东巴画多标签情感分类数据集,为东巴画情感分析研究提供数据支撑。在该数据集上的实验结果表明,本文所提方法的mAP指标达到91.77%,精确匹配准确率达到84.81%,相较于排名第二的TAI++模型分别提高了1.25%和8.86%。**结论**　本文提出的方法为民族艺术情感理解提供了新的技术路径,对文化遗产数字化保护与传承具有重要意义。

关键词:东巴画;情感分类;颜色笔触特征;自适应特征融合;级联推理

Color and Brush Stroke Features Integrated for Emotional Classification of Dongba Paintings

Yang Rongyi[1], Qian Wenhua[1*], Liu Peng[1]

1. School of Information Science and Engineering, Yunnan University, Kunming 650504; 2. School of Mathematics, Southeast University, Nanjing 210009

Abstract: **Objective** Dongba paintings, as a valuable cultural heritage of the Naxi people, embody rich and diverse emotional expressions through their distinctive symbolic systems, color compositions, and brushstroke techniques. These artworks integrate religious beliefs, mythological narratives, ritual functions, and collective cultural values, forming highly abstract, symbolic, and stylized visual representations that differ significantly from natural images and everyday visual scenes. In Dongba paintings, emotional meanings are often conveyed implicitly through color arrangements, line direc-

收稿日期:2026-01-07;**修回日期**:2026-03-11

*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"视觉与文化计算创新团队"(202505AS350009)云南省科技厅应用基础研究计划项目(202201AT070167)

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

tions, spatial structures, and brushstroke textures rather than explicit semantic objects. Such characteristics pose substantial challenges for automated emotion understanding and multi-label classification. However, most existing vision-language models are pre-trained under paradigms primarily oriented toward generic object recognition and natural image understanding, emphasizing object-level semantics while overlooking artistic aesthetics. When directly applied to culturally specific artworks such as Dongba paintings, these models struggle to establish reliable and fine-grained mappings between multi-dimensional emotional semantics and the aesthetic characteristics of color structures and brushstroke textures. This discrepancy leads to emotional semantic misalignment caused by domain shift, which severely restricts both classification performance and model interpretability in ethnic art scenarios. To overcome these limitations, this paper proposes an emotion classification method for Dongba paintings that explicitly integrates color and brushstroke features to enhance domain adaptability, representation robustness, and semantic interpretability. **Methods** Focusing on the unique visual characteristics of Dongba paintings, particularly their color structures and brushstroke textures, a dual-domain color-brushstroke feature extraction mechanism is designed. This mechanism jointly leverages spatial-domain representations and frequency-domain information to capture complementary aesthetic cues, including line orientations, edge distributions, texture patterns, and structural details that are closely related to emotional expression. Spatial-domain features emphasize global color composition and structural layout, while frequency-domain features highlight local texture variations and brushstroke details. By integrating information from both domains, the proposed feature extraction strategy enables a more comprehensive and discriminative representation of artistic aesthetics that are often ignored by generic visual encoders. On this basis, a progressive adaptive fusion architecture is constructed to dynamically balance the contributions of general visual features learned from large-scale pre-training and domain-specific artistic features extracted from Dongba paintings. Through progressive fusion, the model can gradually adjust feature importance across different layers, ensuring that culturally relevant aesthetic features are effectively emphasized while preserving the robustness of general visual representations. This adaptive strategy allows the model to accommodate stylistic diversity, heterogeneous painting patterns, and variations in artistic expression across different samples. Furthermore, to address the coexistence of multiple emotions under label-imbalanced conditions, object category prior knowledge of painting subjects is incorporated into the learning process. Since emotional expressions in Dongba paintings are often closely associated with depicted symbolic objects, a category-guided cascaded question-answering reasoning mechanism is introduced. This mechanism strengthens logical consistency, reinforces emotion–object associations, and improves classification accuracy, robustness, and reliability in complex multi-emotion scenarios. **Results** Extensive experiments conducted on a self-constructed multi-label Dongba painting emotion dataset demonstrate the effectiveness and stability of the proposed method. The model achieves a mean Average Precision (mAP) of 91.77% and an exact match accuracy of 84.81%, outperforming the second-ranked TAI++ model by 1.25% and 8.86%, respectively. Comparative experimental results show that explicitly modeling color-brushstroke features and incorporating category-guided reasoning effectively alleviate domain shift, enhance discriminative emotional representations, and produce more reliable and consistent emotion predictions in culturally specific art contexts. In the visualization experiments, the proposed dual-domain color-brushstroke feature extractor demonstrates a strong capability to efficiently capture the distinctive visual characteristics of Dongba paintings. Compared with ResNet-50 and VGG16, whose feature response patterns mainly focus on local texture extraction, the proposed model can effectively identify color schemes and brushstroke styles that are closely associated with specific emotional expressions, thereby validating the task-oriented design and effectiveness of the feature extraction strategy. Moreover, the generated attention heatmaps not only accurately localize emotionally salient regions within the paintings but also provide a quantitative representation of stylistic differences and color semantic connotations across different subject categories, further enhancing the interpretability of the model and its ability to distinguish subtle variations in artistic expression. **Conclusion** The proposed method provides an effective and interpretable technical pathway for emotion understanding in ethnic artworks. By integrating artistic aesthetic features with adaptive fusion and semantic reasoning mechanisms, the approach advances culturally aware visual emotion analysis and multi-label classification. Moreover, it offers meaningful technical support for the digital preservation, intelligent analysis, and sustainable inheritance of intangible cultural heritage, highlighting the importance of incorporating cultural and artistic knowledge into intelligent visual understanding systems.

Key words: Dongba paintings; emotion classification; color and brush stroke features; adaptive feature fusion; cascaded reasoning

0 引 言

多模态情感分类是融合视觉、文本、音频等多种模态数据，挖掘数据中蕴含的情感语义并完成类别判定的关键技术(Liu 等，2020)，其核心目标是突破单一模态数据的信息局限，构建跨模态情感语义的统一表征与精准映射，实现更全面、鲁棒的情感理解(Shou 等，2025)。目前，该技术已广泛应用于智能交互、舆情分析、文化遗产情感解读等领域(陶建华等，2024)，成为情感计算与人工智能领域的研究热点。

东巴画于2006年被正式列入国家级非物质文化遗产名录，是纳西族文化表达的核心载体。这类艺术品以木板画、卷轴画的艺术表达形式(钱文华等，2020)，描绘纳西族的文化信仰、仪式传统与自然崇拜，通过极具象征性的视觉元素和丰富多样的风格形态，传递多维度情感内涵(田梦溪等，2025)。如图1所示，其通过象征性构图与细腻色彩层次，具象化描绘了东巴守护神虎、马图腾样本赞许认同、沉静休憩的多元情感。

图1 东巴画实例

Fig. 1 Examples of Dongba Paintings

东巴画以多情感并存与符号叙事化情感表达为显著特征(潘森垒等，2023)，蕴含着深刻的文化寓意和民族情感。目前，针对东巴画的研究主要集中在艺术价值和绘画风格等方面，缺乏对其开展智能图像情感分类的工作。因此，开展东巴画智能图像情感分类研究，既能解读画作背后的文化内涵，又能大幅降低传统人工标注的高昂成本、减少对专家定量分析的强依赖(罗霜等，2025)。这不仅为纳西族文化遗产的保护提供有效技术支撑，更助力民族文化内涵的深度挖掘与活态传承(陶建华等，2024；Li 等，2021)。

多模态情感分类需融合视觉、文本等多源模态特征，通过跨模态语义对齐与情感特征融合，实现情感类别的精准判定(Shou 等，2025)。早期的多模态情感分类方法基于手工设计的单模态特征与简单融合策略实现情感判定(Krizhevsky 等，2012)。例如 Xu 等人(2017)使用卷积神经网络(convolutional neural network，CNN)与长短期记忆网络(long short-term memory，LSTM)分别编码文本和图像，提出多情感分析网络与层级自注意力网络两种方法；Yu 等人(2020)利用双线性交互层融合文本与视觉表示，提出实体敏感注意力和融合网络；Yang 等人(2016)则分别提出多视角注意力网络、多通道图神经网络，挖掘模态深层语义特征与复杂关系。这些方法的视觉特征提取多依赖预训练视觉模型，受限于局部感受野，难以捕捉东巴画细腻的色彩笔触特征与文化符号寓意；同时，特征融合方式缺乏深度语义建模，无法建立视觉特征与东巴画情感语义的有效映射，导致适配性与分类精度受限。

随着提示工程的发展，多模态提示耦合方法融合了视觉-语言预训练的跨模态对齐能力，并借助提示工程激活预训练知识、适配特定任务，为复杂多模态任务提供了高效解决方案(Zhou 等，2020)。该类方法以 CLIP(contrastive language-image pre-training)等大规模视觉-语言模型为基础(Kolesnikov 等，2019；Radford 等，2021)，结合文本提示(Zhou 等，2022)、视觉提示(Jia 等，2022)的单模适配优势，实现跨模态对齐与任务适配的协同优化。以 MaPLe(multi-modal prompt learning)(Khattak 等，2023)、TAI++(text as image for multi-label image classification by co-learning transferable prompt)(Wu 等，2024)、M^2PT(multimodal prompt tuning for zero-shot instruction learning)(Wang 等，2024)、ITHP(information-theoretic hierarchical perception)(Xiao 等，2024)、DASCO(dependency structure augmented scoping framework)(Liu 等，2025)为代表的相关方法，在跨模态协同、分类目标平衡、零样本适配及情感噪声消除等方面进一步拓展了技术边界。

随着多模态大型语言模型的兴起，标志着多模态情感识别进入端到端跨模态推理新阶段。此类模型以大型语言模型（large language models，LLMs）为核心（Li 等，2025；Yang，2025），引入视觉变换器（vision transformer，ViT）（Dosovitskiy 等，2021）等视觉编码器解析图像信息，通过跨模态投影层将视觉特征转化为语言模型可理解的语义表征，实现端到端的深度语义推理。其中，ChemVLM（chemistry vision language model）（Li 等，2025）凭借 LLM 强大的上下文建模能力，可有效关联视觉细节与文本语义，在复杂场景情感推理中展现出优异的逻辑连贯性；ConVis（contrastive decoding with hallucination visualization）（Park 等，2024）通过优化视觉-语言跨模态注意力机制，强化了图像特征与情感语义的映射精度。AffectVLM（affect vision language model）（Behzad 等，2025）通过情感感知的预训练与对比对齐构建情感嵌入空间，利用情感感知提示与跨模态对比学习增强对稀有情感类及多情感共存场景的判别能力。多模态大型语言模型依托强大的上下文建模与跨模态关联挖掘能力（Lin 等，2025），在视觉问答、多模态情感推理等任务中展现出巨大潜力（Gao 等，2025），为复杂情感分类任务提供了更高效的解决方案。

然而，上述方法应用于东巴画情感分类时存在以下问题：（1）颜色笔触特征提取不足。在色彩语义映射方面，当前预训练视觉模型基于大规模通用图像数据训练，其学习的色彩表征侧重于自然场景中的描述，虽能自动学习层级化视觉表征，但难以建立东巴画中色彩与情感之间的语义关联。在笔触纹理感知方面，受限于卷积网络的局部感受野特性，模型难以有效捕捉东巴画中绘画笔触的风格特征，而这些美学特征恰是承载情感意蕴的重要视觉线索。由于通用模型因难以捕捉东巴画中符号、色彩、笔触与文化语义的内在关联性，导致视觉特征与情感语义的映射缺乏可靠支撑，最终分类精准度与泛化能力受限。（2）领域偏移问题。在文化符号关联方面，现有多模态预训练大模型的预训练数据以自然图像与通用文本为主，缺乏东巴画相关的文化样本和针对纳西民族文化的语义对齐机制（潘森垒等，2023）。将其直接迁移至东巴画情感分类任务时，既难以建立东巴画抽象文化符号的情感寓意与绘画色彩结构、笔触纹理美学特征之间的语义关联，也无法对东巴画中多情感共存的复杂场景进行有效建模，最终引发领域偏移问题（Ge 等，2022）。

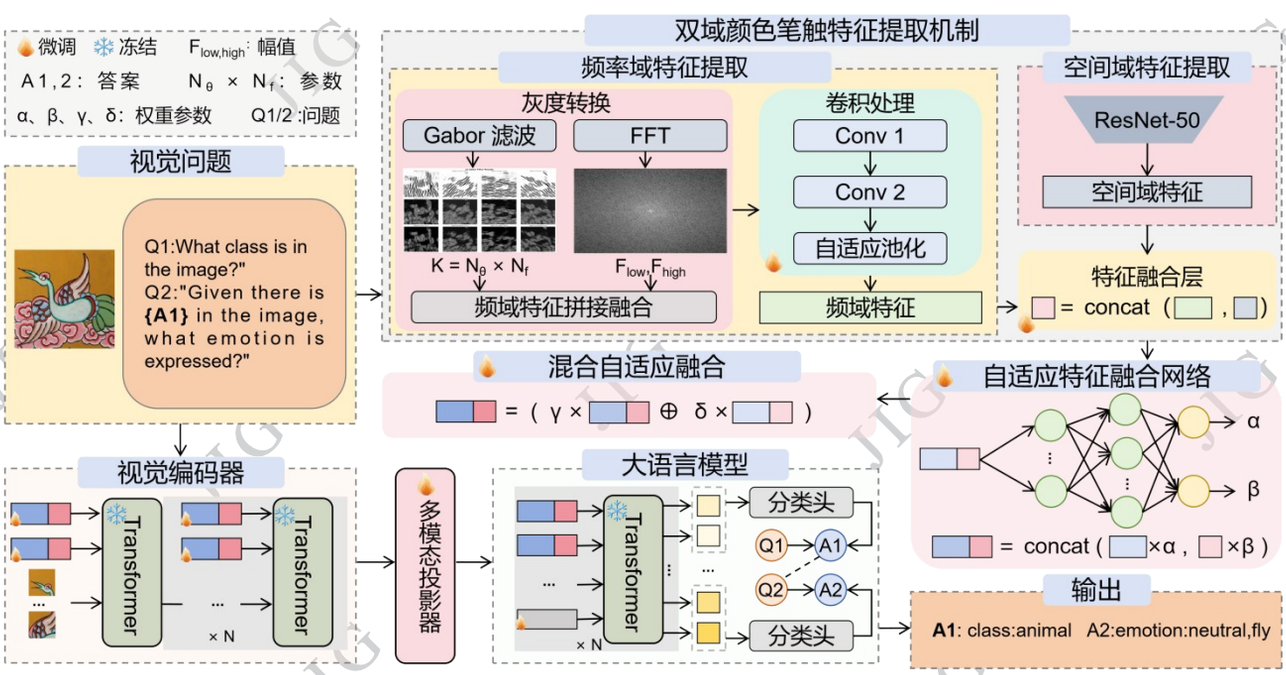

图2　模型总体结构图

Fig. 2　Overall Structure Diagram of the Model

为此，本文提出融合颜色笔触特征的东巴画情感分类方法，主要贡献在于：(1)建立了东巴画多标签情感分类数据集。参考丽江市博物馆馆藏实物与《纳西族东巴画》文献(张春凤，2018；刘芮冰，2024)，系统收集覆盖祈颂、自然崇拜、神话叙述等文化场景的东巴画样本；基于前期文化语义研究与典籍考证，先按题材将图像分为神祇、植物、动物、人物、鬼怪五类，再依视觉符号情感语义构建14维情感分类体系。(2)提出双域颜色笔触特征提取机制。从空间域与频率域提取东巴画的色彩结构与笔触纹理特征，捕获东巴画的线条走向与边缘细节信息，缓解因特征表达不足导致领域偏移的问题。(3)提出渐进式自适应融合模块。针对性平衡通用视觉信息与特定艺术特征的贡献权重，适配东巴画多元绘画风格与差异化特征分布，提升特征融合的适配性。(4)引入类别引导的级联问答推理机制。基于类别先验与情感表达的内在关联，通过大语言模型推理得到东巴画主体类别，将其作为上下文约束融入情感推理过程，有效降低领域偏移引发的情感语义映射偏差。

1 本文方法

本文提出一种融合颜色笔触特征的东巴画情感分类方法，有效提升模型的跨领域适配能力，整体框架如图2所示。具体而言，本方法设计双域颜色笔触特征提取机制，从空间域与频率域提取东巴画的色彩结构信息与笔触纹理细节，为后续处理提供东巴画艺术特质的特征支撑；针对东巴画多元绘画风格的差异，本文构建渐进式自适应融合架构，有效缓解单一融合模式导致的特征适配偏差，提升特征融合的精准度与场景适配性；考虑到标签不平衡场景下绘画主体类别与情感的内在依赖关系，本文提出类别引导的级联问答推理机制，引入类别先验知识作为上下文约束参与情感推理，既强化分类结果的逻辑性与准确性，又有效降低领域偏移引发的情感语义映射偏差。

1.1 双域颜色笔触特征提取机制

为有效捕获东巴画特有的笔触纹理质感、线条走向规律等细粒度艺术特征。本文设计双域颜色笔触特征提取机制，分别从空间域和频率域提取互补性视觉信息，实现对东巴画颜色与笔触特征的全面表征，缓解多模态模型在东巴画领域泛化性不足的问题。

1.1.1 空间域特征提取

东巴画常采用红、黄、黑等对比强烈的色彩，且画面布局蕴含独特的文化符号与叙事结构。为此，空间域分支选用预训练残差网络(residual network-50，ResNet-50)(He等，2016)，提取其最终残差阶段的深层卷积特征。相较于视觉变换器(vision transformer，ViT)等Transformer架构(Dosovitskiy等，2021)，ResNet-50的卷积归纳偏置更适合捕获东巴画中局部色块边界与区域性色彩分布模式(He等，2016；潘森垒等，2023)。现有研究表明，其层次化特征金字塔结构能够有效保留从细粒度纹理到全局布局的多尺度信息(Li等，2025)，且在中小规模数据集上具有更优的收敛稳定性与计算效率(Targ等，2016；Zhou等，2024)。为此，空间域分支选用预训练ResNet-50网络(He等，2016)提取输入图像的空间特征。

1.1.2 频率域特征提取

为捕获东巴画的笔触细节、纹理方向等细粒度视觉特征，频率域分支设计Gabor滤波器组(Luan等，2018)与快速傅里叶变换(fast fourier transform，FFT)(He等，1996)的协同提取机制。首先将输入图像转换为灰度图，通过Gabor滤波器组的多方向、多频率卷积操作，得到局部纹理特征。

$$\boldsymbol{F}_{\mathrm{g}}=Conv(\boldsymbol{I}_{\mathrm{g}};\boldsymbol{\Theta}_{\mathrm{g}})\#(1)$$

式中，$\boldsymbol{F}_{\mathrm{g}}$为Gabor滤波器组提取的局部纹理特征，Conv(·)为卷积操作，$\boldsymbol{I}_{\mathrm{g}}$为输入图像转换后的灰度图，$\boldsymbol{\Theta}_{\mathrm{g}}$为构建Gabor滤波器组的超参数集合。

同时，对$\boldsymbol{I}_{\mathrm{g}}$执行FFT以提取图像频域上的全局频率分布特征，计算为

$$\boldsymbol{F}_{\mathrm{gf}}=\log(|\mathit{fftshift}(\mathcal{F}(\boldsymbol{I}_{\mathrm{g}}))|+\epsilon)\#(2)$$

式中，$\boldsymbol{F}_{\mathrm{gf}}$为全局频率分布特征，log(·)为压缩频谱幅值的动态范围，*fftshift*(·)用于将零频分量移至频谱中心，$\mathcal{F}(\cdot)$为二维FFT算子，ϵ为数值稳定常数。

为进一步挖掘特征间的深层关联，提升频域特征的判别能力，将$\boldsymbol{F}_{\mathrm{g}}$和$\boldsymbol{F}_{\mathrm{gf}}$沿通道维度拼接后，通过两层卷积网络提取高阶频域特征，计算为

$$\boldsymbol{F}_{\mathrm{f}}=Pool\left(Conv\left(Concat(\boldsymbol{F}_{\mathrm{g}},\boldsymbol{F}_{\mathrm{gf}});\boldsymbol{\Theta}_{\mathrm{f}}\right)\right)\#(3)$$

式中，$\boldsymbol{F}_{\mathrm{f}}$为提取得到的高阶频域特征，*Pool*(·)为自适应平均池化算子，$\boldsymbol{\Theta}_{\mathrm{f}}$为频域特征提取模块的可学习参数集合。

卷积层聚合Gabor响应图，将局部纹理响应转为通道级语义特征；池化层压缩空间尺寸、降低特征维度并保留粗粒度空间分布，完成像素级响应至区域级纹理语义的抽象。

频域分支的设计目标是提取图像级纹理描述符。通过展平操作结合两层全连接网络对$\boldsymbol{F}_{\mathrm{f}}$进行维度压缩，计算为

$$\boldsymbol{F}'_{\mathrm{f}} = MLP(Flatten(\boldsymbol{F}_{\mathrm{f}}); \boldsymbol{\Theta}_m) \#(4)$$

式中，$\boldsymbol{F}'_{\mathrm{f}}$为压缩后的频域特征，$MLP(\cdot)$为多层感知机，$Flatten(\cdot)$为特征展平操作，$\boldsymbol{\Theta}_m$为多层感知机模块中所有线性层的可学习参数集合。

1.1.3 双域特征融合与降维

为充分利用空间域全局表征与频域纹理细节的互补优势，将空间特征与$\boldsymbol{F}'_{\mathrm{f}}$沿特征维度拼接，沿特征维度拼接后，通过三层全连接网络进行降维和特征整合，计算为

$$\boldsymbol{F}_{\mathrm{CS}} = MLP\left(Concat(\boldsymbol{F}_{\mathrm{s}},\boldsymbol{F}'_{\mathrm{f}}); \boldsymbol{\Theta}_{\mathrm{fr}}\right) \#(5)$$

式中，$\boldsymbol{F}_{\mathrm{CS}}$为压缩后的颜色笔触特征，$\boldsymbol{F}_{\mathrm{s}}$为空间特征，$\boldsymbol{\Theta}_{\mathrm{fr}}$为双域特征融合与降维的参数集合。最终，得到的$\boldsymbol{F}_{\mathrm{CS}}$通过颜色分类与纹理分类任务的联合训练，同时编码空间语义信息与频域纹理细节，为后续自适应融合提供丰富的特征表示。

1.2 渐进式自适应融合框架

为适配东巴画多元的绘画风格差异，本文提出渐进式自适应融合架构，如图3所示，通过三层渐进式融合策略，动态平衡通用视觉信息与特定艺术特征的贡献权重。

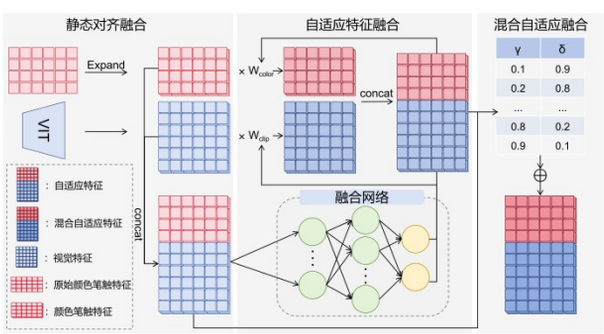

图3 渐进式自适应融合架构

Fig. 3 Progressive Multi-level Adaptive Fusion Architecture

1.2.1 静态对齐融合

静态融合层通过特征拼接生成基础融合特征，为后续自适应调整提供稳定基线。视觉编码器输出的token序列保留逐patch空间对应关系。由于$\boldsymbol{F}_{\mathrm{CS}}$为图像级表征，而CLIP视觉特征为序列级token表征，为使颜色笔触特征维度与视觉序列保持一致，需通过Expand操作沿序列维度扩展$\boldsymbol{F}_{\mathrm{CS}}$，计算为

$$\boldsymbol{F}_{\mathrm{CS}}^{\mathrm{ex}} = Expand(\boldsymbol{F}_{\mathrm{CS}},S) \#(6)$$

式中，$\boldsymbol{F}_{\mathrm{CS}}^{\mathrm{exp}}$为扩展后颜色笔触特征，$S$为CLIP视觉编码器输出的序列长度。该策略将图像级颜色笔触信息均匀分配至所有视觉tokens，为后续自适应调整提供统一先验信息。

为减小不同来源特征的尺度差异，对$\boldsymbol{F}_{\mathrm{VE}}$与$\boldsymbol{F}_{\mathrm{CS}}^{\mathrm{ex}}$分别进行归一化处理后，沿特征维度拼接生成静态融合特征。

1.2.2 Token-wise自适应融合

为自适应响应东巴画风格差异，引入token-wise自适应权重学习网络。该网络采用轻量级设计，为各视觉token独立计算特征权重，动态调整视觉特征与颜色笔触特征的贡献度。以静态融合前的原始拼接特征为输入，经两层全连接网络学习自适应权重，计算为

$$\begin{aligned} \boldsymbol{F}_{\mathrm{c}} &= Concat(\boldsymbol{F}_{\mathrm{VE}},\boldsymbol{F}_{\mathrm{CS}}^{\mathrm{ex}}) \#(7) \\ \boldsymbol{u} &= ReLU\left(Linear_1(\boldsymbol{F}_{\mathrm{c}})\right), \\ \boldsymbol{\omega} &= Softmax(Linear_2(\boldsymbol{u})) \end{aligned} \#(8)$$

式中，$\boldsymbol{F}_{\mathrm{c}}$为原始拼接特征，$\boldsymbol{F}_{\mathrm{VE}}$为CLIP提取的token级视觉特征，$\boldsymbol{u}$为全连接层中间特征，$\boldsymbol{\omega}$为自适应权重向量，$ReLU$、$Softmax$、$Linear$为激活函数与全连接层。

从$\boldsymbol{\omega}$中分离归一化的视觉特征权重α与颜色笔触特征权重β，二者均实现token级独立适配且满足$\alpha_{\mathrm{b,s}} + \beta_{\mathrm{b,s}} = 1(\forall \boldsymbol{b} \in [1,B], \boldsymbol{s} \in [1,S])$。

随后，逐元素相乘加权调制$\boldsymbol{F}_{\mathrm{VE}}$与$\boldsymbol{F}_{\mathrm{CS}}^{\mathrm{ex}}$，经通道拼接完成自适应特征融合，得到自适应融合特征。

1.2.3 混合策略融合

为平衡静态融合的稳定性与自适应融合的灵活性，采用线性组合策略对两层融合特征进行混合优化：

$$\boldsymbol{F}_{\mathrm{fo}} = \gamma \boldsymbol{F}_{\mathrm{st}} + \delta \boldsymbol{F}_{\mathrm{ap}} \#(9)$$

式中，$\boldsymbol{F}_{\mathrm{fo}}$为融合优化后的特征，$\boldsymbol{F}_{\mathrm{st}}$为静态融合特征，$\boldsymbol{F}_{\mathrm{ap}}$为自适应融合特征，$\gamma$为静态融合权重，$\delta$为自适应融合权重，且满足$\gamma + \delta = 1$。为保证融合特征尺度一致，本文动态平衡通用视觉与艺术特征权重，自适应适配东巴画风格差异，缓解过拟合并提升泛化性能。1.2.4 多模态投影与语言模型编码

首先，通过多模态投影器将 $\boldsymbol{F}_{\text{fo}}$ 映射至 LLM 的 embedding 空间，得到投影后的视觉向量，实现视觉与语言的跨模态语义对齐。最后，将该向量输入冻结的 LLM，完成语义生成任务。

LLM 冻结策略：训练时冻结语言模型参数，仅更新投影器、自适应融合层及颜色笔触特征提取器参数。该策略显存高效、训练稳定、计算成本低，有效降低训练复杂度。

对每个 batch 的序列，将多模态投影器输出的视觉向量作为 LLM 的 prefix embeddings，与文本 tokens 的 embeddings 拼接后送入冻结的大语言模型，并分配连续的位置编码。

Attention Mask 规则：对于序列索引 $i,j\in[1,L]$，其中前 S 个位置对应视觉 prefix，后续位置对应语言 tokens，query i，key j 的注意力掩码定义如下：

$$\boldsymbol{mask}(i,j)=\begin{cases}1, i\leqslant S, j\leqslant S(\text{prefix仅能关注自身})\\ 1, i>S, j\leqslant i(\text{语言}token\text{可关注}\\ \qquad\qquad \text{prefix前文tokens})\\ 0,\text{otherwise}\end{cases}\tag{10}$$

该规则防止视觉 prefix 访问后续语言 tokens，避免信息泄漏，同时保证视觉 prefix 可见性与语言 tokens 的自回归约束。最终输入序列送入冻结的 LLM 执行自回归解码，输出语义特征：

$$\boldsymbol{h}_{\text{LM}}=LLaMA(\boldsymbol{F}_{\text{pj}},\boldsymbol{E}_{\text{tx}};\boldsymbol{\Theta}_{\text{LM}})\tag{11}$$

式中，$\boldsymbol{F}_{\text{pj}}$ 为多模态投影器输出的视觉向量，$\boldsymbol{E}_{\text{tx}}$ 为文本 token embeddings，$\boldsymbol{\Theta}_{\text{LM}}$ 为语言模型参数。本文采用 LLM 自带的 RoPE（rotary position embedding）编码，视觉 prefix 与文本 tokens 共享连续的位置编码空间。将视觉特征映射到 *LLaMA*（large language model meta AI）的 embedding 空间，实现东巴画视觉特征与语言语义的跨模态对齐，为情感分类提供特征支撑。

1.3 类别引导的级联问答推理机制

本文设计类别引导的级联问答推理机制，如图 4 所示，首先利用问答式大语言模型推理得到东巴画的主体类别，再将该类别作为上下文约束融入情感推理过程，有效降低领域偏移带来的情感语义映射偏差。该机制针对训练样本构造级联对话序列 $[Q_1,A_1,Q_2,A_2]$，具体形式为：

$$\begin{cases}Q_1=\ \text{“What class is in the image?”}\\ Q_2=\ \text{“Given there is } c^{*}\text{ in the image,}\\ \qquad\ \text{what emotion is expressed?”}\\ A_1=c^{*}\\ A_2=\boldsymbol{e}_{\text{tx}}\end{cases}\tag{12}$$

式中，Q_1 为固定查询指令，Q_2 为融入类别先验的动态生成的查询指令，A_1 和 A_1 为对应指令的回答，c^{*} 为预测的主体类别，$\boldsymbol{e}_{\text{tx}}$ 为情感标签文本。

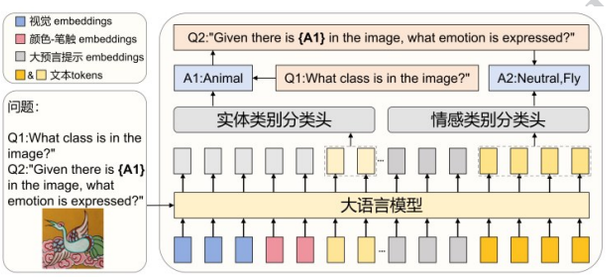

图 4　类别引导的级联问答推理机制

Fig. 4　Category-Guided Cascaded Question-Answering Reasoning Mechanism

1.3.1 主体类别先验获取

模型输入为投影后的视觉特征与 Q_1 的联合向量，通过 LLM 自回归解码生成类别回答。由于 LLM 生成的回答为自然语言描述，需通过后处理模块提取标准类别 token。通过正则匹配与关键词映射相结合的方式，从生成文本中解析出唯一标准类别，计算为

$$c^{*}=\underset{c\in\mathcal{C}}{\text{argmax}}\,P\left(c\,\middle|\,\boldsymbol{I}_{\text{pj}},Q_1;\boldsymbol{\Theta}_{\text{LM}}\right)\tag{13}$$

式中，c 为主体类别标签文本，$\boldsymbol{I}_{\text{pj}}$ 为输入的视觉特征，$\mathcal{C}=\{\text{animal},\text{god},\text{flower},\text{tree},\text{spirit},\text{human}\}$ 为东巴画主体类别集合。类别预测通过 LLM 自回归解码实现，其概率分布可分解为序列生成的条件概率乘积：

$$P(c|\boldsymbol{I},Q_1)=\prod_{t=1}^{|c|}P\left(c_t|c_{<t},\boldsymbol{I}_{\text{pj}},Q_1;\boldsymbol{\Theta}_{\text{LM}}\right)\tag{14}$$

式中，c_t 表示类别序列中的第 t 个 token，$c_{<t}$ 表示第 t 个 token 之前的生成序列。该分解形式确保了类别预测的概率完整性与逻辑合理性，为后续情感推理提供可靠的先验支撑。

1.3.2 类别约束的情感推理

输入 $\boldsymbol{I}_{\text{pj}}$、$c^{*}$ 与情感查询 Q_2，以识别的主体类别 c^{*} 为引导，结合视觉特征与类别先验的联合约束完成多标签情感预测，其数学表达为：

$$e^{*}=\underset{e\in\{0,1\}^{K}}{\operatorname{argmax}}\,P\left(e\left|\boldsymbol{I}_{\mathrm{pj}},c^{*},Q_{2};\boldsymbol{\Theta}_{\mathrm{LM}}\right.\right)\#(15)$$

式中，$e^{*}\in\{0,1\}^{K}$为最终预测的多标签情感向量，e为表示情感标签的多标签向量，K为情感类别数。

情感预测通过自回归文本生成实现，模型生成多情感标签的文本字符串，再解析为多标签独热编码向量。该策略充分利用LLM的文本生成能力，可自然建模多情感标签的共存关系，相较于传统并行建模策略更具语义表达的合理性与准确性。

1.4 损失函数与优化策略

为保障模型各模块的协同优化与性能均衡，本文先对颜色笔触特征提取器进行预训练，强化其视觉特征捕获能力。然后再开展端到端多模态联合训练，实现视觉与语言跨模态对齐与级联推理优化。

1.4.1 颜色笔触特征提取器预训练

颜色笔触特征提取器通过颜色分类与纹理分类两个辅助任务进行预训练，损失函数采用自适应权重调度策略动态平衡双任务优化优先级，具体损失函数$\mathcal{L}_{\mathrm{CT}}$计算如下：

$\mathcal{L}_{\mathrm{CT}}=\omega_{\mathrm{cl}}\mathcal{L}_{\mathrm{cl}}+\omega_{\mathrm{tx}}\mathcal{L}_{\mathrm{tx}}\#(16)$式中，$\mathcal{L}_{\mathrm{cl}}$和$\mathcal{L}_{\mathrm{tx}}$均采用交叉熵损失，$\omega_{\mathrm{cl}}$和$\omega_{\mathrm{tx}}$为双任务自适应权重。初始化权重为$\omega_{0}$，训练过程中基于验证集准确率差异动态调整。当双任务准确率差值$|\Delta_{\mathrm{acc}}|>\tau_{\mathrm{acc}}$时，对准确率较低的任务提升权重，单次调整幅度不超过$\Delta\omega_{\max}$，且权重取值被限制在$[\omega_{\min},\omega_{\max}]$区间内。该策略可有效避免单任务主导训练过程，确保颜色特征与纹理特征的均衡学习与协同表征。

1.4.2 颜色笔触特征提取器预训练

在端到端训练阶段，损失函数通过自回归生成机制统一优化类别识别与情感推理任务，使模型先学习主体类别识别，再进行东巴画情感推理：

$$\mathcal{L}_{\mathrm{LM}}=-\frac{1}{T}\sum_{t=1}^{T}\log P\left(y_{t}\left|y_{<t},\boldsymbol{I}_{\mathrm{pj}},x_{<t};\boldsymbol{\Theta}_{\mathrm{LM}}\right.\right)\#(17)$$

式中，$\mathcal{L}_{LM}$为损失函数，T为输入序列总长度，y_{t}为第t个位置的目标token，$y_{<t}$表示前$t-1$个位置的生成序列，x为$\boldsymbol{I}_{\mathrm{pj}}$与级联对话序列的联合向量。

2 实验结果与分析

2.1 实验数据

本文通过丽江市博物馆馆藏实物与《纳西族东巴画》相关文献考证（张春凤，2018；刘芮冰，2024），系统收集木板画、纸牌画和卷轴画典型纳西族东巴画真实样本。构建的样本库覆盖传统祈颂、自然崇拜、神话叙述等核心文化场景，呈现了纳西族对自然万物、仪式传统与社会生活的认知体系。基于研究团队前期文化语义解析成果（潘森垒等，2023；罗霜等，2025），结合纳西族文化典籍与历史文献考证，首先按创作题材将东巴画图像划分为神祇、植物、动物、人物、鬼怪五类；进一步依据绘画视觉符号所承载的情感语义，构建14维情感分类体系，具体包括赞许认同、绽放欣喜、平静安宁、乖戾逆反、昂扬振奋、警觉护守、消极否定、中立素净、正向积极、沉静休憩、萎靡消沉、勤勉务实、担忧焦虑及怀有戒心。其对应实验数据中的英文标签为：approval、bloom、calm、devil、fly、guard、negative、neutral、positive、rest、wilt、work、worried和defensive。上述情感类别为东巴画数据集的固有属性，每幅图像的情感标签均通过符号元素、绘画色调、笔触风格、构图场景及主体行为等多维度视觉信息综合提取，如图5和图6所示，该标准在五大题材类别中呈现差异化适配特征。

图6 东巴画风格数据集示例

Fig. 6 Examples of Dongba Painting Style Dataset

在人物题材画作中，情感标签通过色调氛围与与主体动作综合确定。例如，红色背景下手持武器且神情肃穆的人物形象标注为怀有戒心与警觉护守；黄色背景下躬身行礼的人物形象则归类为正向积极与勤勉务实。在鬼怪题材画作中，以冷色调为主且采用纯线条勾勒的鬼怪图腾传递消极否定与乖戾逆反情感；而暖色调背景下手持武器、呈凶猛姿态的牛魔形象则表达正向积极与警觉护守情感，主体行为差异可导致情感表达截然相反。

在神祇题材多承载祈福纳祥与仁慈庇佑的文化内涵，勤勉务实为其共性情感标签，结合神灵姿态与神情可进一步细分为正向积极或中立素净；当画作以红色为背景且主体为金翅大鹏鸟等守护神时，标注为正向积极与警觉护守动物与植物题材遵循独特的情感映射逻辑。在动物题材中，红色背景下姿态颤栗的白牦牛守护神传递警觉护守与担忧焦虑的复合情感，色调柔和的飞鸟鸣叫图则定义为中立素净与昂扬振奋。植物题材中，暖色调背景下细腻描绘的繁茂花卉传达正向积极与绽放欣喜情感，简约勾勒或粗犷写意的植物画作传递中立素净与平静安宁情感，暗红背景下的枯萎花瓣则呈现消极否定与萎靡消沉特征。

图5　东巴画多标签情感数据集示例

Fig. 5　Examples of Dongba Painting Multi-label Emotion Dataset

考虑到绘画色调与笔触风格是东巴画情感表达的核心视觉载体，本文针对二者展开系统性分类研究，将绘画色调分为暖色为主、冷色为主、色彩丰富三类，笔触风格分为粗犷写意、线条勾勒、细腻描绘三类，具体分类统计结果如表1所示。

神祇与人物题材作为东巴文化核心精神的具象化表达，多以细腻笔触刻画主体，色调涵盖冷色、暖色和多彩配色，通过不同色调传达情感内涵。鬼怪题材以冷色调为主导，通过色彩张力强化情感氛围渲染，部分作品采用细腻笔触结合暖色调形成差异化情感传递路径。动物题材普遍以细腻笔触刻画主体，色彩随姿态差异而变化，如神情严肃的白牦牛守护神采用冷色基调契合其情感属性。植物题材以线条勾勒为主要表现形式，色彩与笔触随情感表达需求动态调整。当传递正向情感时采用暖色调结合细腻笔触，表达中立情感时以简约勾勒为主，与前文情感分类体系形成呼应。

2.2　数据集构建及扩充

基于14维情感分类体系与标注标准，本文构建了东巴画多模态多标签情感数据集。文本模态参考纳西族文化典籍（张春凤，2018；刘芮冰，2024），经人工标注形成情感描述集；视觉模态为采集的东巴画真实样本，各样本对应一对情感标签，实现情感语义的多维度表达。为提升模型多标签情感分类性能，采用抠除图像背景、保留绘画主体的方式扩充数据。最终构建的数据集共包含858个文本-图像对，其中动物、花朵、灌木、人物、神祇和鬼怪：200、208、109、126、139、76对文本-图像对。本文实验数据集按照7∶2∶1划分训练集、验证集和测试集，各包含569、183和79个文本-图像对。

表1　东巴绘画数据集色调与笔触分类统计

Table 1　Classification Statistics of Hue and Brushstroke for the Dongba Painting Dataset

类别划分	具体类型	数目
绘画色调	色彩丰富	163
	冷色为主	248
	暖色为主	447
绘画笔触	粗犷写意	78
	线条勾勒	287
	细腻描绘	493

2.3　实验设置

本实验采用LLaVA（large language and vision assistant）架构构建多模态情感分类模型，基础权重加载自llava-pretrain-vicuna-7b-v1.3预训练版本。模型主干选用32层、隐藏单元维度为4096的LLaMA大模型，视觉分支采用CLIP-ViT-Large-Patch14，负责提取东巴画的高阶视觉特征。采用本文提出的渐进式自适应融合架构，混合策略中设置静态融合权重$\gamma = 0.8$、自适应融合权重$\delta = 0.2$；自适应分支采用128维隐藏层设计，并对融合权重施加正则化约束，以提升融合特征的稳定性与判别力。

为捕获东巴画的色彩与笔触纹理特征，本实验设计双域颜色笔触特征提取机制。该提取器中滤波器长宽比设为0.5，高斯包络标准差为2，相位偏移量为0。针对笔触纹理的多尺度、多方向表征需求，设计包含12个滤波器核的Gabor滤波器组，覆盖4个方向 $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ 和3个频率 $f \in \{0.1, 0.3, 0.5\}$。在颜色笔触特征提取器预训练的权重初始值 $\omega_0 = 2.2$，准确率差异调整阈值 $\tau_{ac} = 0.08$，权重单次最大调整幅度 $\Delta\omega_{mx} = 0.05$，权重取值区间为 $\omega_{mn} = 1.5$、$\omega_{mx} = 3.5$。基于NVIDIA RTX 3090显卡开展训练过程，采用AdamW优化器，初始学习率设置为5e-4，采用余弦退火调度策略调节学习率，权重衰减系数为0.01，warmup阶段占比为0.03。训练批次大小设为2，梯度累积步数为4，总训练轮数为3。最大序列长度限制为2048。评估与模型保存策略设置为每200步执行一次，最多保留3个最优checkpoint。为平衡训练效率与模型性能，训练过程中启用梯度裁剪、梯度检查点技术，并采用TF32与BF16混合精度加速方案。

2.4 实验评价指标

针对东巴画多标签情感分类的任务，本文将严格匹配准确率（exact match accuracy，EMA）作为核心评估指标，其功能是量化模型预测标签集合与真实标签集合在样本级的一致性，适配多标签场景下标签全集匹配的评估需求。

对于第 i 个样本 $i \in \{1, 2, \cdots, N\}$，定义真实情感标签集合为 Y_i，模型输出的预测标签集合为 $\hat{Y}_i$，二者均隶属于前文构建的14维情感标签空间。考虑到数据集中可能存在无真实标签样本，即 $|Y_i| = 0$。此类样本若纳入统计将导致评估结果出现偏差，为保障指标计算的科学性与合理性，本文明确界定有效样本集合为 $I = \{i \mid |Y_i| > 0\}$，并基于该有效集合完成EMA指标的计算。

$$\mathrm{EMA} = \frac{1}{|I|}\sum_{i \in I} 1\left[Y_i = \hat{Y}_i\right] \#(18)$$

式中，$1[\cdot]$ 为指示函数，当且仅当预测集合 $\hat{Y}_i$ 与真实集合 Y_i 在元素构成和集合基数上完全一致时取1，否则取0。该指标能准确反映模型在多标签场景中的样本级严格匹配能力。

为全面评估模型性能，本文同时采用多标签分类任务常用的补充评估指标：平均精度均值量化各类别情感预测的精准度与召回率综合表现，反映模型在类别级的整体排序能力；Precision-Micro、Recall-Micro及F1-Micro（Zhang等，2013）过全局汇总所有类别预测结果，适配东巴画情感标签分布可能存在的不平衡特性，为模型性能提供更全面、多维度的评价依据。

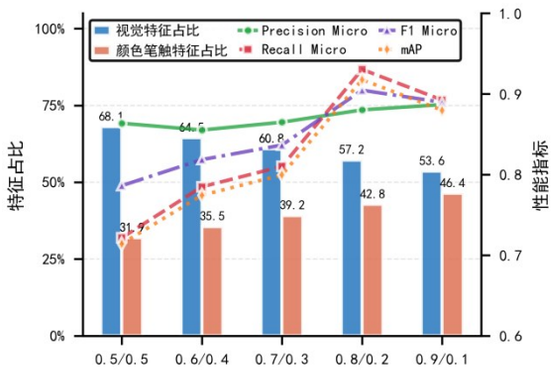

图7 静态与自适应融合比例对比实验参数可视化

Fig. 7 Visualization of Experimental Parameters for Static and Adaptive Fusion Ratio Comparison

2.5 特征融合参数实验

本文对混合策略中特征融合参数进行讨论，对不同静态融合权重与自适应融合权重比例组合下的情感分类指标进行了对比，实验结果如表2所示。本实验设置自适应融合权重 $f_{adaptive}$ 与静态融合权重 f_{static} 的取值组合从（0.1，0.9）到（0.5，0.5）逐步调整，对应的视觉特征与颜色笔触特征的占比变化区间为31.9%至68.1%。二者的具体配比及对应特征占比，如图7所示。

通过对比不同参数组合下的EMA评估指标，最终确定最优融合方案为静态融合权 $\gamma = 0.8$、自适应融合权重 $\delta = 0.2$，此时视觉特征占比57.2%、颜色笔触特征占比48.2%。实验结果表明，以视觉特征为主导、颜色笔触特征为辅助的融合模式可使EMA指标达到最优。颜色笔触特征的引入能够强化模型对东巴画情感语义的理解，而过度增加其权重则会引发特征冗余，导致模型性能下降。同时，图7中的指标折线图直观呈现了不同特征融合比例对模型情感分类性能的影响，进一步验证了上述最优参数组合的合理性。

2.6 与其他算法比较

为验证本文模型在东巴画多标签情感分类任务中的有效性，选取ITHP、DASCO、VPT、CoOP、

表2　不同特征比例融合下的指标对比

Table 2　Comparison of indicators under fusion of different feature proportions

静态融合比例	自适应融合比例	mAP(%)	EMA(%)	Precision-Micro(%)	Recall-Micro(%)	F1-Micro(%)
0.5	0.5	71.41	64.56	86.36	72.15	78.62
0.6	0.4	77.43	70.89	85.52	78.48	81.85
0.7	0.3	79.96	74.68	86.49	81.01	83.66
0.8	**0.2**	**91.77**	**84.81**	88.02	**93.04**	**90.46**
0.9	0.1	87.97	82.28	**88.68**	89.24	88.96

注：加粗字体为每列最优值。

表3　本文方法与其他方法指标对比

Table 3　Comparison of Metrics Between the Method in This Paper and Other Methods

Methods	mAP(%)	EMA(%)	Precision-Micro(%)	Recall-Micro(%)	F1-Micro(%)
VPT(Jia等，2022)	52.96	15.19	47.62	12.74	20.10
CoOp(Zhou等，2022)	59.18	53.16	71.50	59.69	63.24
CoCoOp(Zhou等，2022)	67.59	63.29	77.78	61.15	68.47
MaPle(Khattak等，2023)	65.30	56.96	80.91	56.33	66.42
ITHP(Xiao等，2024)	52.57	41.77	86.25	44.52	58.72
TAI++(Wu等，2024)	90.52	75.95	85.00	86.62	85.80
DASCO(Liu等，2025)	71.18	82.28	87.18	86.62	86.90
Ours	**91.77**	**84.81**	**88.02**	**93.04**	**90.46**

注：加粗字体为每列最优值。

MaPLe、CoCoOP、TAI++模型进行比较。基于mAP、EMA、Precision-Micro、Recall-Micro及F1-Micro五项指标开展对比实验，从表3中数据可知，本文模型在各指标分别达到了91.77%、84.81%、88.02%、93.04%和90.46%的性能。相较于排名第二的TAI++模型分别提高了1.25%、8.86%、3.02%、6.42%和4.66%。其中，核心指标EMA的提升幅度最为突出，由此可见，本文所提的融合颜色笔触特征的分类方法，能够有效解决跨领域多模态任务中的领域偏移问题。

同时，为验证本文提出的双域颜色笔触特征提取机制在东巴画绘画色调和笔触分类任务中的有效性，选取Resnet-50、视觉几何组网络（visual geometry group network 16，VGG16）、基础型视觉变换器（vision transformer-base/16，ViT-B/16）在本数据集上微调后，在测试集中进行定量比较。基于绘画色调、笔触分类准确率和精准匹配率指标开展对比实验，从表4中数据可知，本文模型在各指标分别达到了82.28%、92.41%、77.22%的性能。其中核心指标精准匹配率的提升幅度最为突出，由此可见，本文提出的双域颜色笔触特征提取机制，能有效捕获东巴画的绘画色调、笔触纹理质感、线条走向规律等细粒度艺术特征。

表4　不同方法颜色笔触分类性能对比

Table 4　Comparison of Color Stroke Classification Performance with Different Methods

Methods	色调分类准确率(%)	笔触分类准确率(%)	精准匹配率(%)
Resnet-50	73.42	89.87	64.56
VGG16	53.16	60.76	30.38
ViT-B/16	75.95	89.87	68.35
Ours	**82.28**	**92.41**	**77.22**

注：加粗字体为每列最优值。

表5　消融实验

Table 5　Ablation Experiment

渐进式自适应融合	双域颜色笔触特征	类别引导的级联问答推理	mAP(%)	EMA(%)	Precision-Micro(%)	Recall-Micro(%)	F1-Micro(%)
×	×	×	77.22	69.62	83.67	77.85	80.66
×	×	√	85.76	83.54	87.18	86.08	86.62
√	√	×	85.02	75.95	80.95	86.08	83.44
×	√	√	86.71	81.01	88.39	86.71	87.54
√	√	√	**91.77**	**84.81**	**88.02**	**93.04**	**90.46**

注:加粗字体为每列最优。

2.7　消融实验

为验证本文提出的渐进式自适应融合模块、双域颜色笔触特征提取机制及类别引导的级联问答推理机制的有效性,本文设计了系列消融模型并开展对比实验,结果如表5所示。双域颜色笔触特征提取机制与渐进式自适应融合模块的组合应用,能够有效捕获东巴画特有的色彩结构与笔触纹理信息,通过动态调控两类特征的融合占比,使模型灵活适配东巴画的多元风格差异。在此基础上引入类别引导的级联问答推理机制后,模型相较于未添加任何模块的基准模型的mAP、EMA、Precision-Micro、Recall-Micro及F1-Micro指标上分别提升了14.55%、15.19%、4.35%、15.19%和9.8%。由此可见,类别先验与视觉特征提取机制互补,添加各模块的本模型可更好地将东巴画情感语义与视觉特征的对齐,共同促进了模型性能的提升

2.8　可视化实验分析

为直观验证本文双域颜色笔触特征提取器对东巴画视觉特征的捕获能力,通过模型内生空间热力图量化呈现其关键关注区域,清晰反映模型对图像不同区域的关注强度分布规律。实验选取东巴画中动物、花朵、神祇、灌木、人物、鬼怪六大典型主体对象。对比经典基准模型与本文方法的热力图,分析二者对原图特征的响应差异,实验结果如图8所示。

从可视化结果可见,本方法生成的热力图能够聚焦各类主体对象的关键区域,且对不同主体的特征响应具有针对性。在动物中,热力图集中覆盖动物轮廓及细腻描绘的毛发、鳞片区域,体现模型对主体形态特征的有效捕获;在花朵中,热力图的感兴趣区在花瓣色彩及使用线条勾勒的花瓣纹理区域,而在灌木中,热力图突出灌木的色彩分布区域与粗犷笔触刻画部分,表明模型能够有效区分差异化绘画风格特征;但在神祇、人物和鬼怪的主体中,热力图重点关注其面部细节、装饰纹样和整体色彩氛围,这些区域恰是东巴画中承载情感语义的载体。如图8所示,在本数据集上微调得到的ResNet-50(He等,2016)、VGG16(Tammina等,2019)和ViT-B/16模型主要呈现局部纹理驱动的响应。其中,VGG16模型在画面复杂的动物题材中容易被背景云纹干扰,而ViT-B/16模型提取特征过度依赖全局图像块关联,易引发特征结构破碎化问题。相比之下,本文模型可有效识别与特定情感关联的色彩与笔触风格,验证了特征提取的针对性与有效性。

综上,可视化实验展现了本文双域颜色笔触特征提取器可高效捕获东巴画的视觉特征。模型生成的热力图不仅实现对关键区域的定位,更能量化表征不同主体对象的绘画风格差异与色彩语义内涵,为后续多阶段情感分类任务提供了兼具判别力与适配性的图像特征支撑,印证了该特征提取器在东巴画情感分类任务中对色彩与纹理特征的捕获适配性。

3　结　论

本文提出一种融合颜色笔触特征的东巴画情感分类方法,通过双域颜色笔触特征提取、渐进式自适应融合架构与类别引导的级联问答推理机制,有效捕捉东巴画的符号化语义、绘画色调与笔触技法。实验结果表明,相较于现有方法,本文所提方法在细粒度情感识别与跨模态语义对齐上更适配东巴画艺术表征特性,有效缓解了领域偏移引发的情感语义映射偏差。不仅为东巴画等民族艺术品的多标签情

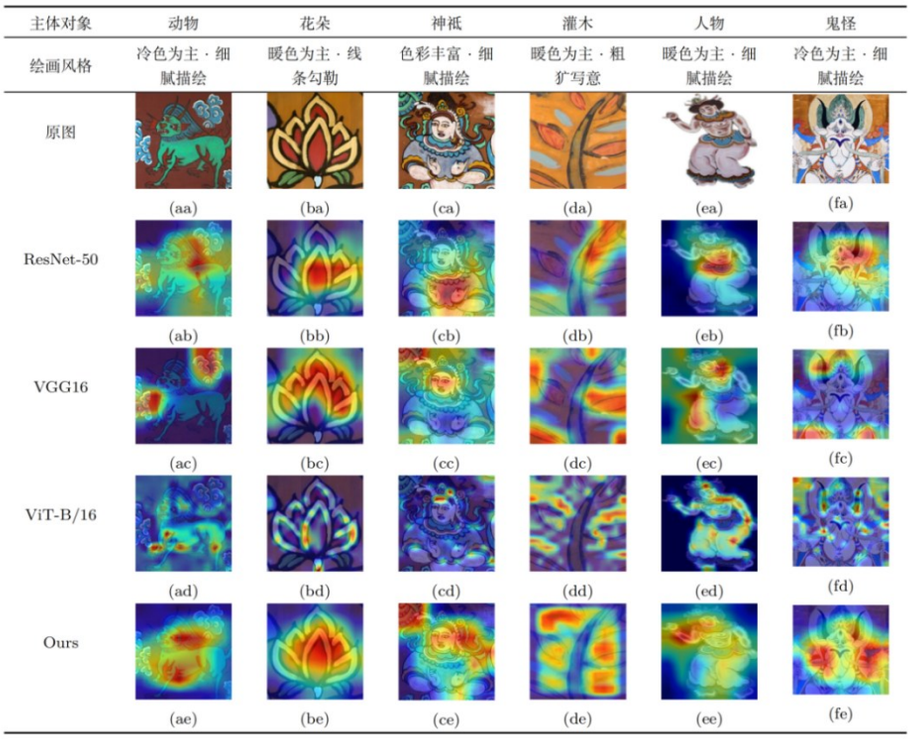

图8　本文方法与经典模型的东巴画颜色笔触特征提取可视化对比

Fig. 8　Visualization of Feature Extraction Comparison for Dongba Paintings′ Color and Brush Stroke Between the Proposed Method and Classical Models

感分类与跨模态理解提供了新的技术路径，更以该技术为支撑推动文化遗产数字化保护实践，同时助力民族文化内涵的深度挖掘。然而，本研究仍存在一定局限。受东巴画背景复杂度的制约，数据集扩充仍依赖人工标注，数据制备效率较低，且尚未引入语义分割等自动化技术。在未来工作中，将优化东巴画数据集的扩充策略，结合语义分割技术提升样本构建效率与多样性，进一步完善模型框架，以验证所提方法的广泛通用性及跨民族艺术场景迁移能力。

参考文献(References)

Behzad M, Zhao G. 2025. Contrastive language-image learning with augmented textual prompts for 3D/4D FER using vision-language model//Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition. Kuala Lumpur: IEEE: [DOI: 10.1109/FG61629.2025.11099228]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. 2021. An image is worth 16×16 words: transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. Addis Ababa: ICLR: [DOI: 10.48550/arXiv.2010.11929]

Gao J, Qiao Q, Wu T, Wang Z, Cao Z and Li W. 2025. Aim: let any multimodal large language models embrace efficient in-context learning//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 3077-3085 [DOI: 10.1609/aaai.v39i3.32316]

Ge C, Huang R, Xie M, Lai Z, Song S and Li S. 2022. Domain adaptation prompt learning for cross-domain image emotion classification//Proceedings of the IEEE Signal Processing Letters. Piscataway, USA: IEEE: 2106-2110 [DOI: 10.1109/TNNLS.2023.3327962]

He S and Torkelson M. 1996. A new approach to pipeline fft processor//Proceedings of the International Conference on Parallel Processing. Bloomington, USA: IEEE: 766-770 [DOI: 10.1109/IPPS.1996.508145]

He K, Zhang X, Ren S and Sun J. 2016. Deep residual learning for image recognition//Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference. Las Vegas, USA: IEEE: [DOI: 10.1109/CVPR.2016.90]

Li Z, Wang Y and Chen X. 2025. Design of Chinese painting style classification model based on multi-layer aggregation CNN. Computational Intelligence and Neuroscience, 2025: 1-12 [DOI: 10.1155/2025/7896452]

Jia M, Tang L, Chen B C, Cardie C, Belongie S, Hariharan B, et al. 2022. Visual prompt tuning//Proceedings of the European Conference on Computer Vision. Cham, Switzerland: Springer: 709-727

[DOI: 10.48550/arXiv.2203.12119]

Khattak M U, Rasheed H, Maaz M, Khan S and Khan F S. 2023. Maple: multimodal prompt learning for affective computing//Proceedings of the IEEE Transactions on Affective Computing. Piscataway, USA: IEEE: 1567-1580 [DOI: 10.1109/CVPR52729.2023.01832]

Kolesnikov A, Beyer L, Zhai X, Puigcerver J, Yung J, Gelly S and others. 2019. Large scale learning of general visual representations for transfer [DOI: 10.48550/arXiv.1912.11370]

Krizhevsky A, Sutskever I and Hinton G E. 2012. Imagenet classification with deep convolutional neural networks//Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe, USA: NIPS: [DOI: 10.1145/3065386]

Li J, Chen D L, Yu N, Zhao Z P and Lv Z H. 2021. Emotion recognition of chinese paintings at the thirteenth national exhibition of fine arts in china based on advanced affective computing//Proceedings of the Frontiers in Psychology. Lausanne, Switzerland: Frontiers Media: 741665 [DOI: 10.3389/fpsyg.2021.741665]

Li J X, Zhang D, Wang X Z, Hao Z Y, Lei J D, Tan Q, et al. 2025. Chemvlm: exploring the power of multimodal large language models in chemistry area//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 415-423 [DOI: 10.1609/aaai.v39i1.32020]

Lin Z, Lin M, Lin L and Ji R R. 2025. Boosting multimodal large language models with visual tokens withdrawal for rapid inference//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 5334-5342 [DOI: 10.1609/aaai.v39i5.32567]

Li Z, Wang Y and Chen X. 2025. Design of Chinese painting style classification model based on multi-layer aggregation CNN. Computational Intelligence and Neuroscience, 2025: 1-12 [DOI: 10.1155/2025/7896452]

Liu H, He L J, Liang J X, Ren Z H, Bi H X and Li F. 2025. Dependency structure augmented contextual scoping framework for multimodal aspect-based sentiment analysis. arXiv Preprint [DOI: 10.48550/arXiv.2504.11331]

Liu M and Zhang H. 2020. Image emotion classification: a survey. Pattern Recognition, 105: 107232 [DOI: 10.1016/j. patcog. 2020.107232]

Liu R B. 2024. Interpretation and Reconstruction of Dongba Script Symbols in Modern Dongba Paintings. Art Literature, (12): 1-3 (刘芮冰. 2024. 现代东巴画中东巴文字符号的解读与重组。美术文献, 2024 (12): 1-3) [DOI: 10.16585/j.cnki.mswx.2024.12.011]

Luan S Z, Chen C, Zhang B C, Han J G and Liu J Z. 2018. Gabor convolutional networks//Proceedings of the IEEE Transactions on Image Processing. Piscataway, USA: IEEE: 4357-4366 [DOI: 10.1109/TIP.2018.2835143]

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: 3361-3376 (罗霜, 钱文华, 刘朋. 2025. 结合提示学习和视觉语义生成的东巴画图像描述。中国图象图形学报, 30: 3361-3376) [DOI: 10.11834/jig.240601]

Pan S L, Qian W H, Cao J D and Xu D. 2023. Learning Attention for Dongba Paintings Emotion Classification. Journal of Graphics, 44 (1): 59-66 (潘森垒, 钱文华, 曹进德, 徐丹. 2023. 基于注意力机制的东巴画情感分类。图学学报, 44 (1): 59-66) [DOI: 10.11996/JG.j.2095-302X.2023010059]

Park Y, Lee D, Choe J and Chang B. 2025. Convis: contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 6434-6442 [DOI: 10.1609/aaai.v39i6.32689]

Qian W H, Xu D, Xu J, He L and Zhang B. 2020. Artistic Style Rendering of Dongba Paintings. Journal of System Simulation, 32 (7): 1349-1358 (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): 1349-1358) [DOI: 10.16182/j.issn1004731x.joss.19-VR0530]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the International Conference on Machine Learning. Baltimore, USA: PMLR: 8748-8763 [DOI: 10.48550/arXiv.2103.00020]

Shou Y, Meng T, Ai W, Zhang X, Chen Y, Liu Z and Li S. 2025. Multimodal large language models meet multimodal emotion recognition and reasoning: A survey [EB/OL]. [2025-09-29]. https://arxiv.org/pdf/2509.24322.pdf

Tao J H, Fan C H, Lian Z, Lyu Z, Shen Y and Liang S. 2024. Development of Multimodal Sentiment Recognition and Understanding. Journal of Image and Graphics, 29 (6): 1607-1627 (陶建华, 范存航, 连政, 吕钊, 沈莹, 梁山. 2024. 多模态情感识别与理解发展现状及趋势。中国图象图形学报, 29 (6): 1607-1627)

Tammina S, Nagalla S and Devarakonda N. 2019. Transfer learning using vgg-16 with deep convolutional neural network for classifying images. International Journal of Scientific and Research Publications (IJSRP), 9: 143-150 [DOI: 10.29322/IJSRP. 9.10.2019.p9420]

Targ S, Almeida D and Lyman K. 2016. Resnet in resnet: generalizing residual architectures [EB/OL]. [2016-03-25]. https://arxiv.org/pdf/1603.08029.pdf

Tian M and Ding S. 2025. The Road to Divine Land: Iconography, Deity, and Aesthetic Style. Arts, 14 (2): 22 (田梦溪, 丁少华. 2025. 神圣之地之路: 图像学、神祇与审美风格。艺术, 14 (2): 22 [DOI: 10.3390/arts14020022]

Wang T O, Liu Y Y, Liang J C, Zhao J H, Cui Y M, Mao Y N, et al. 2024. M2pt: multimodal prompt tuning for zero-shot instruction learning//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). Miami, Florida, USA: Association for Computational Linguistics: 3723-3740

［DOI：10.18653/v1/2024.emnlp-main.218］

Wu X Y，Jiang Q Y，Yang Y，Wu Y F，Chen Q G and Lu J F. 2024. Tai++ text as image for multi-label image classification by co-learning transferable prompt//Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. Macao，China：IJCAI：5226-5234［DOI：10.24963/ijcai.2024.578］

Xu N and Mao W. 2017. Multisentinet：a deep semantic network for multimodal sentiment analysis//Proceedings of the 2017 ACM Conference on Information and Knowledge Management. New York，USA：ACM：2399-2402［DOI：10.1145/3132847.3133142］

Xiao X，Liu G，Gupta G，Cao D，Li S，Li Y，Fang T，et al. 2024. Neuro-inspired information-theoretic hierarchical perception for multimodal learning//Proceedings of the 12th International Conference on Learning Representations. Addis Ababa，Ethiopia：ICLR：［DOI：10.48550/arXiv.2404.09403］

Yang Z，Yang D，Dyer C，He X，Smola A J and Hovy E H. 2016. Hierarchical attention networks for document classification//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. San Diego，USA：ACL：1480-1489［DOI：10.18653/v1/N16-1174］

Yu J F，Jiang J and Xia R. 2020. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. IEEE/ACM Transactions on Audio，Speech，and Language Processing，28：429-439［DOI：10.1109/TASLP.2019.2957872］

Zang C F. 2018. The Discovery and Value of the Naxi Dongba Scripture The Classic of the Five Directions and Five Emperors. Journal of Northwest Minzu University（Philosophy and Social Sciences Edition），6：181-184（张春凤，2018. 汉语东巴经《五方五帝经》的发现及其价值。西北民族大学学报（哲学社会科学版），6：181-184）［DOI：10.14084/j.cnki.cn62-1185/c.2018.06.026］

Zhang M L and Zhou Z H. 2013. A review on multi-label learning algorithms. IEEE Transactions on Knowledge and Data Engineering，26：1819-1837［DOI：10.1109/TKDE.2013.39］

Zhou K Y，Yang J K，Loy C C and Liu Z W. 2022. Conditional prompt learning for vision-language models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans，USA：IEEE：16816-16825［DOI：10.48550/arXiv.2203.05557］

Zhou Y，Li J and Zhang H. 2024. Efficient Training of ResNet on Small Datasets with Transfer Learning. IEEE Access，12：56789-56802［DOI：10.1109/ACCESS.2024.3389765］

Zhou K，Yang J，Loy C C and Liu Z. 2022. Learning to prompt for vision-language models. International Journal of Computer Vision，130：2337-2348［DOI：10.1007/s11263-022-01653-1］

Zhou L，Palangi H，Zhang L，Lu J，Hu H and Gao J. 2020. Unified vision-language pre-training for image captioning and VQA//Proceedings of the AAAI Conference on Artificial Intelligence. New York，USA：AAAI：13041-13049［DOI：10.1609/aaai.v34i07.7005］

作者简介

杨荣懿，2002年生，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：12024115003@stu.ynu.edu.cn

钱文华，1980年生，通讯作者，男，教授，博士生导师，CCF高级会员（38886D）主要研究方向为计算机视觉、文化计算等。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

曹进德，男，教授，主要研究方向为人工智能、数学。E-mail：jdcao@seu.edu.cn