arXiv:2601.00156v1 [cs.CV] 1 Jan 2026

Focal-RegionFace：为任意选定的面部焦点区域生成细粒度多属性描述

郑凯文[a]、傅俊辰[a]、徐松培[a]、何耀清[a]、Joemon M. Jose[a]、韩虎[c]、葛旭日[b,*]

[a]格拉斯哥大学，格拉斯哥，英国[b]山东大学，中国山东[c]中国科学院计算技术研究所，中国北京

摘要

本文介绍了面部分析中一个尚未得到充分探索的问题：针对任意选定的面部区域，生成并识别包含面部动作单元（AU）、情绪状态和年龄估计的多属性自然语言描述（称为***FaceFocalDesc***）。我们认为，系统聚焦于单个面部区域的能力有助于提升理解与控制能力。为实现这一能力，我们构建了一个针对任意选定面部区域的全新多属性描述数据集，其中提供了丰富的区域级标注和自然语言描述。此外，我们提出了一种基于 Qwen2.5-VL 微调的视觉语言模型，称为 **Focal-RegionFace**，用于面部状态分析。该模型通过多个逐步推进的微调阶段，不断优化对局部面部特征的关注，从而实现可解释的年龄估计、面部动作单元检测和情绪检测。实验结果表明，在新基准上，Focal-RegionFace 无论采用传统且广泛使用的指标，还是新提出的指标，均取得了最佳性能。这充分验证了其在细粒度多属性面部区域焦点分析场景中的有效性和通用性。

关键词：多属性面部区域描述生成、面部区域描述生成、面部属性识别

*通讯作者。电子邮箱：k.zheng.1@research.gla.ac.uk（郑凯文），j.fu.3@research.gla.ac.uk（傅俊辰），s.xu.1@research.gla.ac.uk（徐松培），heyaoqin1009@gmail.com（何耀清），Joemon.Jose@glasgow.ac.uk（Joemon M. Jose），hanhu@ict.ac.cn（韩虎），xuri.ge@sdu.edu.cn（XuriGe）

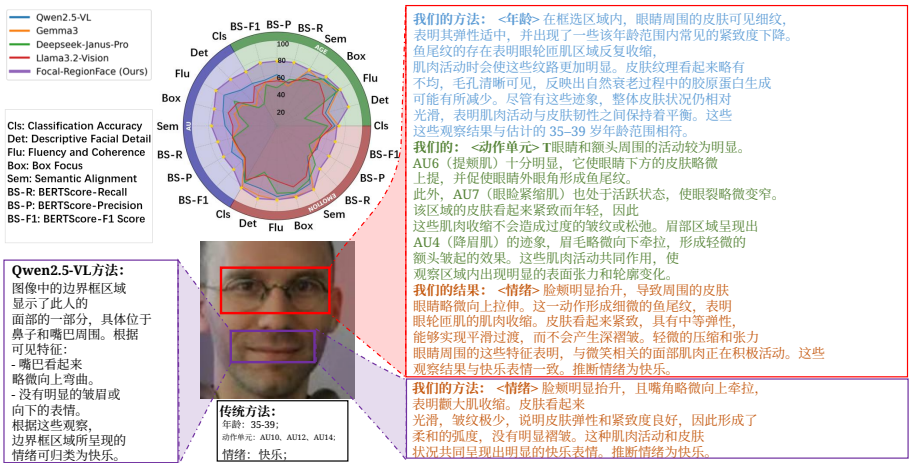

图1：主流多模态大语言模型与我们的模型在面部状态分析能力方面的比较，我们的模型在所有自然语言处理指标上均取得了更优性能。具体而言，我们展示了传统面部状态识别方法、多模态大语言模型 Qwen2.5-VL 以及我们的 Focal-RegionFace 模型的详细结果。我们的 Focal-RegionFace 模型能够针对任意选定的面部区域生成更加详细的多属性面部描述。

1. 引言

人脸分析是视觉语言研究的基础，为情感计算、医学诊断和人机交互等应用提供支撑。尽管传统方法 [1, 2] 能够预测结构化输出（例如动作单元或情绪类别），但其可解释性和灵活性通常较为有限。相比之下，自然语言描述能够提供更符合人类认知且更具可解释性的反馈，在医疗保健和监控等领域尤其具有价值 [3, 4]。现有大多数工作 [5, 6] 聚焦于全局人脸描述，而其他工作 [7, 8, 9] 则探索细粒度属性问答，忽视了对局部细粒度焦点理解的需求。在实际应用中，用户往往更加关注局部面部状态，例如眼睛或嘴部周围的皱纹状况，这对于美容或医疗建议等任务至关重要，也凸显了对细粒度、区域感知面部焦点分析的需求。在本研究中，我们针对一个尚未得到充分探索的面部分析任务提出了一种新颖的解决方案，即任意选定面部区域的状态描述生成（***FaceFocalDesc***）。

FaceFocalDesc 的能力*FaceFocalDesc*。如图1所示，我们提出的

FaceFocalDesc 通过支持对任意面部区域生成细粒度的多属性语言描述，改变了主流面部分析方法的范式。一方面，传统的基于视觉的模型 [10, 11] 专注于对面部状态进行结构化预测。例如， [12] 以黑箱方式直接预测给定面部图像的年龄，不提供任何可解释信息，因此缺乏可信度 [13]。另一方面，视觉语言模型被引入面部分析任务，旨在通过生成人类可读的面部状态描述来提高可解释性。例如，VL-FAU [14]通过整合语言生成分支，为面部动作单元（AU）状态生成粗略的基于规则的语言描述。多模态大语言模型的最新进展也推动了面部领域模型的发展，例如 Face-LLaVA [15], 和 Emotion-Llama [16], ，它们利用视觉语言预训练，将面部特征与全局层面的描述性语义相匹配。然而，这些方法仍存在两个根本局限。首先，它们仅依赖全局面部表征，无法处理用户任意指定的局部区域。其次，它们通常只针对单属性输出（例如情绪分类或图像描述生成），无法实现多属性、区域感知的面部状态建模。

.....面临的挑战*FaceFocalDesc*。尽管上述概念优势显著，构建一个可控且可解释的*FaceFocalDesc* 系统仍面临多项不容忽视的技术挑战。首先，与模型可依赖整体线索的全局面部描述生成不同，*FaceFocalDesc*必须在局部信息约束下运行，而这类约束往往缺乏完整的语义上下文。因此，模型必须学会基于不完整的视觉信号进行推理，同时保持语义完整性和语言流畅性。这要求模型具备高级空间焦点感知能力。其次，将多种面部理解任务（例如动作单元检测 [17, 18], 情绪识别 [16], 和年龄估计 [19],）整合到统一的语言生成框架中并非易事。这些任务在语义结构和视觉相关特征方面存在固有差异，简单地将它们结合起来可能会导致描述支离破碎或过于笼统。第三，现有的大规模面部描述数据集通常以全局描述为主，数据稀疏且针对特定任务，缺少针对特定区域的多任务语言输出标注。数据匮乏也因此成为训练和评估

FaceFocalDesc。

所提出的方法——*Focal-RegionFace*。为应对上述挑战，我们提出了一种基于广泛使用的 Qwen2.5-VL 模型 [20] 的新型 Focal-RegionFace 框架，用于全新的面部分析范式*FaceFocalDesc*，从而能够针对任意选定的面部区域生成细粒度、多属性的语言描述。Focal-RegionFace 旨在超越全局面部描述，迈向区域感知、可控且语义丰富的理解。

具体而言，我们首先构建了一个专为*FaceFocalDesc*设计的新基准数据集，其中包含区域级细粒度多属性标注及相应的多属性标签。该数据集为预训练基础多模态大语言模型 [21] 提供必要的监督微调数据，使其能够学习具有空间依据的多属性语言信息。

随后，我们为 Focal-RegionFace 提出了一种四阶段渐进式微调策略。首先，在全局面部属性识别任务上微调基础 Qwen2.5-VL 模型，使其具备基本的面部感知能力。接着，我们使用带有随机标注边界框的整脸图像进行区域引导描述生成，使模型能够学习初步的空间聚焦能力和区域感知语言生成能力。为了进一步提升区域聚焦的精确度，我们采用掩码区域微调，仅保留选定面部区域可见，迫使模型将语言与局部视觉内容进行对齐。最后，我们利用丰富的区域级描述进一步微调模型，进行显式的多属性分类，从而提升其预测动作单元、情绪和年龄的能力。这种渐进式设计有效地使模型具备强大的空间推理和多属性对齐能力，从而能够对任意位置进行细粒度、可解释的面部分析。

本文的主要贡献如下：

- 我们提出了一项全新且重要的面部分析任务，即针对任意选定区域生成面部区域-焦点多属性描述（称为***FaceFo-calDesc***）。

- 我们提出一种基于 Qwen2.5-VL 框架的新型多阶段微调方法，用于生成聚焦于面部区域的描述，称为 **Focal-**

RegionFace。面部区域可任意选取，Focal-RegionFace 能够生成属性描述，包括动作单元、情绪和年龄，并识别其相应类别。

- 我们构建了一个用于*FaceFocalDesc*训练和评估的新基准，其中包含多属性区域级面部状态描述及相应的属性标签。

- 除了传统的识别和自然语言处理评估指标外，我们还基于预训练多模态大语言模型，为*FaceFocalDesc* 提出了一种新颖且实用的评估方法，涵盖生成描述的分类准确率、细节描述能力、流畅度和自然度、局部关注以及语义相关性。

在新的*FaceFocalDesc*基准上进行的大量实验验证了我们提出的 **Focal-RegionFace** 模型的动机和有效性，有助于推动细粒度交互式面部状态分析的未来研究。与主流多模态大语言模型（如Qwen2.5-VL、Deepseek-Janus-Pro [22] 和Llama3.2-Vision [23], ）相比，我们提出的模型在开源和闭源评估模型上的生成与识别测试中均取得了最佳性能。

2. 多模态面部区域聚焦数据集

如图1所示，尽管传统人脸数据集（例如BP4D [24], AffectNet[25], UTKFace [26], 等）推动了人脸分析任务的发展，但仍存在三个主要局限：（1）聚焦于黑箱任务（例如动作单元和情绪识别），可解释性有限，例如基于皮肤纹理进行推理；（2）MERR [15] 和FaceInstruct-1M [16] 等注重可解释性的数据集提供了全局描述，但缺少对任意面部区域的标注；（3）很少有数据集能够同时为细粒度面部感兴趣区域提供多属性标注（动作单元、情绪、年龄）[27]。

为弥补这些不足，我们针对 *FaceFocalDesc* 任务引入了多模态面部区域聚焦数据集（MFRF）。该数据集支持以感兴趣区域为中心的细粒度分析

涵盖动作单元、情绪和年龄，并配有丰富的语言描述，以实现交互式和区域感知的面部理解。

数据收集。 为实现高质量的区域-焦点面部描述标注，我们构建了一个整合四个成熟多属性数据集的新基准：用于动作单元识别的 BP4D、用于情绪识别的 Aff-Wild2 [28] 和 RAF-DB [29] ，以及用于年龄估计的 UTKFace。对于年龄任务，我们将原始年龄标签重新映射为 12 个区间（[0–4], [5–9], . .. , [50–59], 60+），以反映渐进的面部变化 [30, 31], ，而动作单元和情绪标签保持不变。

在过滤冗余和低质量样本后，我们获得了 10,000 张图像（其中 3,000 张来自 BP4D、2,000 张来自 Aff-Wild2，5,000 张来自 UTKFace），每张图像均带有属性标签。对于每张图像，我们依据面部关键点 [32] 选取 12 个大小不一的面部区域，以确保它们与关键面部区域的重叠度至少达到 80%。随后，使用属性驱动提示，由 GPT-4o [33] 为每个区域添加标注，并进行人工修订。该流程生成了 120,000 张区域-焦点面部图像，并附有细粒度多属性标注。此外，我们还构建了 60,000 个图像-描述对，用于多属性微调。

为进行全面评估，测试集包含 1,000 张图像（其中 300 张来自 BP4D、200 张来自 RAF-DB、500 张来自 UTKFace），每张图像均包含 12 个随机区域，共计 12,000 个区域级样本。基于关键点的区域融合策略进一步支持多区域联合描述，并作为多属性识别微调的先验知识（见方法，第一阶段）。

标注策略。 不同于传统的全局层面面部分析，我们的方法引入区域焦点描述，明确将结构化标注与可解释的模型推理联系起来。我们在随机选取的 ROI 中标注面部动作单元、情绪和年龄，重点关注局部肌肉运动、与年龄相关的皮肤线索，以及局限于边界框区域内的表情。

MFRF 提示词设计遵循三项原则：上下文聚焦、区域约束和结构化生成（详见附录）。上下文聚焦要求 GPT-4o 扮演属性专家，关注 ROI 内的细粒度纹理和肌肉活动。区域约束要求排除边界框外的信息，并与真实标签保持一致，以确保空间语义准确性。结构化生

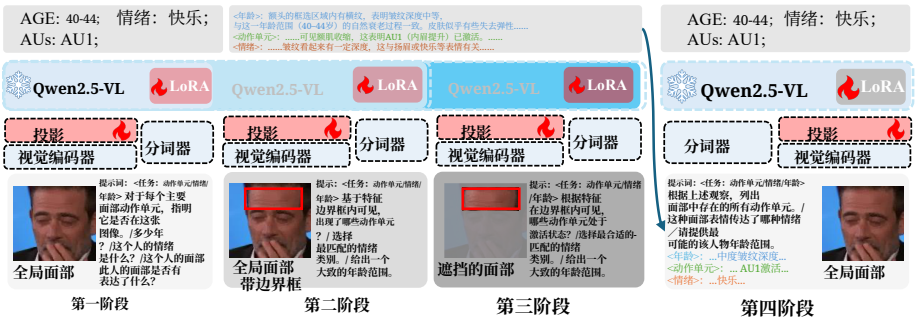

图2：采用多阶段微调的 Focal-RegionFace 概览。我们首先在第一阶段对 Qwen2.5-VL 进行全局面部多属性信息感知微调，包括年龄、情绪和动作单元识别。随后，在第二阶段和第三阶段，我们以渐进式微调的方式使模型聚焦于区域聚焦推理，从而获得具备细粒度多属性语言解释能力的 Focal-RegionFace 多模态大语言模型。接着，基于多区域视觉理解结果进行进一步的多模态推理微调（第四阶段），使模型具备细粒度多模态多属性识别能力。

成则确保输出连贯的段落式文本，将局部视觉细节与可解释性相结合。

这一设计产生了高质量、区域感知的基准，支持对动作单元、情绪和年龄估计的可解释模型进行微调和评估。

3. 提出的方法

3.1. 初步介绍

任务定义。 *FaceFocalDesc*被定义为一项条件式多属性描述生成与识别任务，涵盖动作单元、情绪和年龄，从而实现区域感知的可解释性。给定一张面部图像 I 和任意选定的区域（感兴趣区域，ROI），模型能够生成细粒度的多属性自然语言描述 $D^{<AU/EMO/AGE>}$。此后，还可以将历史区域描述 $D^{<AU/EMO/AGE>}$ 作为提示，进一步给出最终属性决策 $P^{<AU/EMO/AGE>}$ 。这种定义支持单轮和历史感知生成模式，有助于实现渐进式、可解释的面部分析。

Focal-RegionFace。 为解决上述任务，我们在图2中提出了Focal-RegionFace，这是一种四阶段渐进式微调框架，旨在增强对面部区域焦点的理解和多属性语言生成能力。具体而言，该框架包括：（第一阶段）全局感知人脸，使预训练基础模型能够获得全面的面部视觉表征感知；（第二阶段）区域感知视觉-语言对齐，建立感兴趣区域定位和语义推理的初步能力；（第三阶段）人脸区域聚焦对齐，增强模型关注空间定义的面部区域的能力；以及（第四阶段）区域聚焦引导的多属性识别，整合历史感兴趣区域可解释信息，执行最终的多属性决策。这种渐进式设计赋予模型空间感知、语义精确性以及在局部面部分析中的可解释决策能力。

网络架构。 Focal-RegionFace 构建于 Qwen2.5-VL 架构之上。我们采用多阶段 LoRA 微调 [34]，使基础模型具备面部区域-焦点视觉与语言推理能力。首先，每幅图像都由 Qwen 的视觉编码器进行处理，随后通过可学习的投影映射至大语言模型的词元嵌入空间。LoRA 模块应用于关键注意力层，从而增强区域特定表征和多属性推理能力。这种结构使模型能够有效捕捉局部面部动态并执行细粒度分析。

Name	描述	范围
Cls	面部细节描述与属性分类的匹配评估。	0–100
Det	描述性面部细节 — 面部细节描述的丰富度演变。	0–100
Flu	生成的语言描述的流畅性与连贯性。	0–100
Box	区域描述与目标区域（边界框）之间的相关性。	0–100
Sem	生成的描述与视觉内容的语义对齐程度。	0–100
Win%	模型获得最高分的样本比例。	0–100

表1：基于多模态大语言模型的评估指标描述及对应的分数范围。

3.2. 训练策略

在实验中，我们发现单阶段微调缺乏从感知到理解再到表达的语义学习顺序。这导致区域级属性学习与语言生成之间脱节，降低了细粒度

可解释性和一致性。因此，我们提出了一种新颖的多阶段微调策略来解决这一局限。在所有训练阶段，Qwen2.5-VL 骨干网络始终完全冻结，仅对 LoRA 和投影层进行微调。

第一阶段：全局感知人脸。在第一阶段，模型使用不含边界框的预处理图像，根据全局面部线索预测基本面部属性，例如动作单元（AUs）、情绪和年龄范围。输入查询旨在提取全局信息。输出采用简单标签的形式，例如 AU3、AU4、愤怒、30–34。为增强泛化能力和鲁棒性，我们针对每种面部属性在不同阶段构建五种不同的查询提示，并将其随机分配给每张图像。这种多样化提示策略提升了模型在各种面部情境下的适应能力（详见附录）。这一阶段建立了对面部特征的一般感知，使模型能够在聚焦特定区域之前，全面理解面部属性。

第二阶段：区域感知人脸视觉语言对齐。在第二阶段，我们引入区域级视觉语言对齐。输入包括添加了随机生成边界框的预处理图像。在此阶段，查询针对局部区域，引导模型仅关注每个边界框内的视觉内容。我们使用面部属性的详细自然语言描述进行监督微调。这一过程使模型初步理解局部区域及其语言关联，为后续更精确的定位任务奠定基础。

第三阶段：人脸区域聚焦对齐。为进一步增强区域聚焦，第三阶段引入了感兴趣区域（ROI）微调策略。第三阶段的图像经过掩码处理，仅保留模型关注的目标区域，而被掩盖的区域则转换为灰度图。这种有意的掩码处理迫使模型仅根据目标内容生成描述，而忽略全局上下文。训练过程中保留了与第二阶段相同的结构化查询和图像描述。此阶段提升了模型捕捉局部表情、细纹和细微肌肉变化的能力。

第四阶段：区域聚焦引导的多属性识别。在最后阶段，

区域聚焦引导的多属性识别强调多区域聚合和整体评估。输入由一张标注了多个方框标注区域的预处理人脸图像构成，这些区域对应于阶段II和III中定义的区域。对于每个区域，模型利用先前学到的细粒度描述进行多区域推理。结果采用第一阶段的简洁真实标注结构呈现（例如，AU3、愤怒、30–34）。这一阶段有两个主要目的：首先，验证模型整合多个区域的详细观察结果的能力；其次，模拟真实应用场景，即同时查询多个面部区域，以获得统一的解读。这一步最终完善了模型在局部和整体情境下进行多属性推理的能力。

4. 实验

4.1. 实验设置

实现细节。 在每个阶段，我们都使用批量大小为 16、学习率为 2e-5 的设置，对 4 位量化的 Qwen2.5-VL-32B 进行微调，并采用余弦学习率调度器，训练 10 轮。我们启用了梯度检查点以降低内存消耗，并应用 0.01 的权重衰减进行正则化。更多细节见附录。

评估指标。

我们采用三类指标对 Focal-RegionFace 进行全面评估。(1) 基于多模态大语言模型的评估指标（表1）专为新的*FaceFocalDesc*任务设计。借助开源和闭源多模态大语言模型的多模态推理能力，我们让它们担任评审，对生成的区域焦点描述从多个方面进行评分。这为模型推理和生成质量提供了客观且抗偏差的衡量方式。(2) 主流自然语言处理指标，包括 BERTScore [35]（精确率、召回率、F1分数）、语法问题 [36]（性别身份）和专家评分（ER）。30 名经验丰富的标注员分为六个小组，对描述质量和语义对齐进行评分，并汇总评分以达成共识[16]。(3) 传统识别指标，包括情绪和年龄预测的 AU F1 和准确率。

模型	Gemini-2.5-Pro							GPT-4o						
	Cls	Det	Flu	Box	Sem	胜率/%	排名	Cls	Det	Flu	Box	Sem	胜率/%	排名
Qwen2.5-VL	52.69	47.35	74.49	73.22	51.88	13.51	2	67.28	64.62	78.20	74.73	69.68	14.57	3
Gemma3	59.04	47.69	71.40	76.53	58.01	12.40	3	67.35	60.10	72.15	71.70	68.16	19.47	2
Deepseek-Janus-Pro	44.33	13.76	79.83	80.11	43.78	1.66	5	55.20	39.91	73.41	68.19	55.98	7.55	4
Llama3.2-Vision	51.01	33.18	74.33	68.70	45.96	4.87	4	63.61	51.16	60.78	67.85	61.53	3.70	5
Focal-RegionFace（我们的方法）	**70.46**	**82.91**	**93.83**	**91.81**	**74.70**	**67.56**	1	**74.38**	**83.86**	**84.72**	**81.33**	**79.51**	**57.71**	1

表 2：使用闭源模型评估的 Focal-RegionFace 与不同多模态大语言模型的比较。

基于多模态大语言模型的评估详情。为评估细粒度语言质量、区域特异性和语义对齐，我们针对闭源和开源模型采用不同策略。对于闭源模型评估，Gemini-2.5-Pro[1] 和 GPT-4o[2] 担任评审，在统一的图像条件评估提示下，对五个模型——Focal-RegionFace 和四个基线：Llama3.2-Vision、Qwen2.5-VL、Deepseek-Janus-Pro 和 Gemma3 [37]——生成的描述进行联合评分。两位评审均在同一提示下同时评估全部五条描述，以确保公平。对于开源模型评估，Llama3.2-Vision、Qwen2.5-VL 和 Deepseek-Janus-Pro 使用相同的评估提示，分别将 Focal-RegionFace 与每个基线进行一对一比较。这种双重策略确保了两种设置下的评估公平、标准化且可复现。提示详情见附录。

4.2. 实验结果

I. 基于多模态大语言模型的评估进行定量比较。为评估我们的多阶段训练策略（图2）的有效性，我们使用闭源和开源多模态大语言模型作为智能专家评估器，对 Focal-RegionFace 的性能进行比较。受预算限制，我们对闭源模型采用全局排名（表2），而对开源模型则进行一对一评估（表3）。

总体而言，我们的结果一致表明，渐进式结构化微调策略显著提升了多模态面部理解能力，这一点体现在

[1]https://deepmind.google/technologies/gemini/pro/[2] https://openai.com/index/hello-gpt-4o/

	Qwen2.5-VL					
比较	Cls	Det	Flu	Box	Sem	胜率/%
Qwen2.5-VL	65.50	73.12	56.45	75.45	80.36	**64.68**
Focal-RegionFace	**75.08**	**81.32**	**89.67**	**81.63**	**91.04**	35.32
Deepseek-Janus-Pro	47.80	56.99	35.74	70.34	81.35	14.99
Focal-RegionFace	**78.34**	**85.14**	**92.78**	**83.50**	**98.13**	**85.01**
Llama3.2-Vision	58.90	61.19	40.58	70.68	81.50	22.05
Focal-RegionFace	**77.06**	**83.98**	**89.40**	**84.26**	**94.82**	**77.95**
	Deepseek-Janus-Pro					
比较	Cls	Det	Flu	Box	Sem	胜率/%
Qwen2.5-VL	76.30	78.05	79.71	78.42	75.40	0.00
Focal-RegionFace	**89.55**	**89.54**	**89.51**	**88.33**	**87.37**	**100.00**
Deepseek-Janus-Pro	70.81	71.96	73.79	72.88	71.38	0.00
Focal-RegionFace	**89.87**	**89.96**	**89.99**	**89.38**	**88.48**	**100.00**
Llama3.2-Vision	70.09	71.23	77.63	77.61	77.67	0.00
Focal-RegionFace	**89.61**	**89.87**	**89.89**	**88.14**	**87.66**	**100.00**
	Llama3.2-Vision					
比较	Cls	Det	Flu	Box	Sem	胜率/%
Qwen2.5-VL	59.95	67.90	54.88	70.65	54.51	15.46
Focal-RegionFace	**80.29**	**82.72**	**73.29**	**82.38**	**76.12**	**84.54**
Deepseek-Janus-Pro	44.83	52.85	43.68	54.73	41.75	7.97
Focal-RegionFace	**83.48**	**80.25**	**75.38**	**83.91**	**80.42**	**92.03**
Llama3.2-Vision	62.97	68.25	59.13	71.74	64.84	13.87
Focal-RegionFace	**82.77**	**87.42**	**79.03**	**86.69**	**81.29**	**86.13**

表3：开源多模态大语言模型评估器对不同多模态大语言模型与 Focal-RegionFace 的比较。

所有评估指标上的性能始终更优。

在基于闭源多模态大语言模型的评估中，我们的模型始终优于有竞争力的基线模型。值得注意的是，在所有模型中，Qwen2.5-VL 和 Gemma3 的性能最强，而 Deepseek-Janus-Pro 和 LLaMA3.2-Vision 的表现相对较差，这表明它们可能不太适合面部理解任务。

在开源模型评估中，我们使用相应的开源多模态大语言模型，将我们的模型与各个基线逐一比较。我们的方法总体上始终表现更好，只有一个例外：与

模型	BS-P	BS-R	BS-F1	性别身份 (↓)	ER
Deepseek-Janus-Pro	57.45	46.65	51.16	0.7802	34.84
Llama3.2-Vision	53.63	52.57	52.09	2.9200	55.23
Gemma3	51.46	53.95	52.53	1.9133	78.50
Qwen2.5-VL	51.67	58.62	54.84	1.6333	76.38
Focal-RegionFace（我们的模型）	**75.55**	**75.76**	**75.98**	**0.4318**	**86.72**

表 4：基于自然语言处理指标（即 BERTScore（%）和语法问题（↓ 越好））对描述质量进行的定量评估。

模型	区域-焦点			完整面部		
	Emo	Age	AU	Emo	Age	AU
Deepseek-Janus-Pro	35.21	31.92	9.21	41.20	36.43	14.26
Llama3.2-Vision	18.42	25.18	11.56	38.48	37.46	18.43
Gemma3	37.77	38.88	21.31	45.86	50.14	32.61
Qwen2.5-VL	35.64	38.11	10.06	45.73	47.84	24.16
Focal-RegionFace（我们的方法）	**40.35**	**43.65**	**23.12**	**53.74**	**64.37**	**40.22**

表5：使用面部区域焦点图像与完整面部图像进行多属性识别的定量评估。

与 Qwen2.5-VL 相比，我们的模型胜率略低。我们推测，这可能是由于评估偏差所致：正如 [38] 中所讨论的，模型往往更倾向于自己的输出，而非其他模型生成的内容。生成单条描述的平均响应时间约为 0.6 秒，表明该模型有望用于实时交互应用。

II. 主流自然语言处理指标评估。为提高评估的全面性，我们还采用了主要的自然语言处理指标来评估描述生成。如表 4 所示，Focal-RegionFace 在所有指标上均表现出更强的性能。这进一步表明，与标准注释相比，Focal-RegionFace 生成的描述具有更好的一致性，且语法错误更少。

III. 传统多属性识别评估。表5展示了我们的模型与其他预训练多模态大语言模型通过传统分类评估进行的比较，

包括情绪和年龄预测的识别准确率，以及动作单元识别的 F1 分数。当我们仅将选定区域作为图像输入时（模拟面部遮挡情况），我们的 Focal-RegionFace 模型能够以更高的准确率和更强的鲁棒性进行识别，优于主流多模态大语言模型。当聚焦于完整面部信息时，我们的方法在所有属性识别任务中仍保持最佳性能。

4.3. 消融研究

I. 多阶段从第一阶段到第三阶段的影响： 为了解 Focal-RegionFace 中每个微调阶段的影响，我们针对结果进行消融研究，相关结果见表 6 和图 3。与基线 Qwen2.5-VL-32B 相比，表 6 显示，第一阶段的面部感知微调提升了多属性识别的性能。对于多属性描述生成，图 3 显示，在我们的多阶段渐进式面部区域聚焦微调对齐下，我们的模型生成的多属性描述在闭源评估器 GPT-4o 的评估中，在多个方面均取得了显著提升。具体而言，在区域聚焦程度的得分方面，我们的模型得分稳步且显著提高，从第一阶段的 59.9% 提升至第二阶段微调后的 79.8%，并在最终三阶段区域聚焦微调后达到 89.7%。此外，如表 6 所示，对描述质量指标的进一步分析表明，在三个阶段中，BS-P、BS-R 和 BS-F1（BERTScore）均取得了持续提升。从第一阶段到第三阶段，平均 F1 分数从 31.2% 提升至 76.0%，表明随着模型区域感知能力的增强，其语言复杂度和流畅度也得到了提升。第一阶段不考虑 GI（语法问题）分数，因为该阶段不会生成句子。第三阶段后的 GI 分数低于基线，表明多阶段微调还提升了句子质量。

这些结果表明，我们的渐进式微调使 Focal-RegionFace 能够更有效地捕捉细节丰富的区域特定面部属性。各项指标的详细分解见附录。

II. 第四阶段的效果。为进一步验证第四阶段的有效性，我们在传统的多属性识别任务下进行了消融研究。如

模型/阶段	Emo	Age	AU	BS-P	BS-R	BS-F1	GI
Qwen2.5-VL	35.64	38.11	10.06	62.91	64.73	63.77	0.58
第一阶段	36.27	38.92	12.25	46.82	23.67	31.19	N/A
第二阶段	37.62	38.98	12.76	72.63	71.93	72.24	**0.31**
第三阶段	38.33	39.35	13.17	**75.55**	**75.76**	**75.98**	0.43
第四阶段	**40.35**	**43.65**	**23.12**	75.02	73.33	74.17	0.36

表 6：消融实验：多属性和自然语言处理指标。

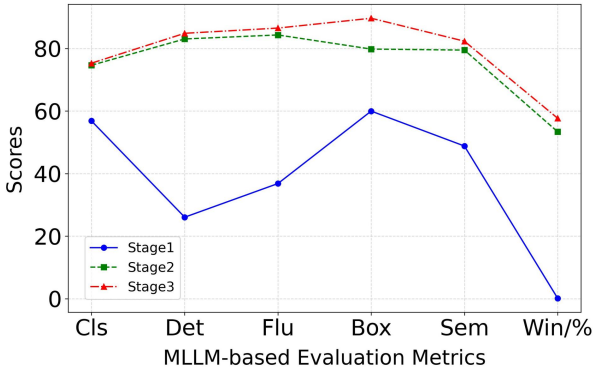

图 3：消融实验：多模态大语言模型评估

表 6 所示，第三阶段和第四阶段在所有属性上的表现均显著优于基线 Qwen2.5-VL。第三阶段引入区域焦点对齐，增强了局部特征提取，从而显著提升了动作单元和情绪识别的表现。在第四阶段，区域聚焦引导的多属性识别进一步提升了性能，动作单元识别率提高至 23.12%，年龄预测准确率达到 43.65%，较前几个阶段实现了大幅跃升。这种渐进式改进证实了多模态多区域聚合对细粒度属性识别的有效性。然而，自然语言处理指标略有下降，这可能是由于引入了额外的分类信息，略微影响了模型的文本生成能力。

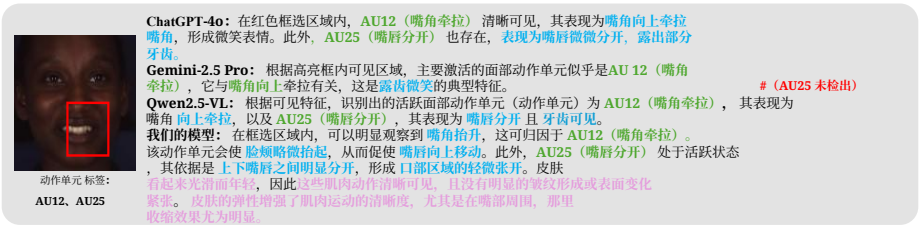

图4：不同面部状态描述生成器在多种面部属性上的可视化比较，包括面部动作单元、情绪和年龄。红色框为随机选取的区域。红色文字中的描述是错误或与区域无关的生成结果。（蓝色：动作单元描述；绿色：肌肉描述；紫色：皮肤细节综合分析）

5. 可视化

为展示 Focal-RegionFace 在 *FocalDec*任务上的有效性，图4对多个受试者的随机选定区域生成的多属性描述进行了可视化比较，并与 ChatGPT-4o、Gemini-2.5-Pro 和 Qwen2.5-VL 的结果进行了评估。

我们的模型擅长局部面部分析，能够通过详细评估皮肤纹理、弹性和肌张力，更准确地估计年龄；同时，它还能通过精准识别细微的肌肉运动，实现更出色的动作单元检测。这些以生理学为依据的细粒度解读，使预测既准确又可解释，展现了模型在区域感知、高精度面部理解方面的优势。

6. 结论

我们提出*FaceFocalDesc*，这是一项针对任意面部区域进行细粒度多属性识别与描述生成的新任务，并推出 MFRF 基准，其中包含 12 万条区域级标注和基于多模态大语言模型的语义评估指标。为解决这一任务，我们提出了 Focal-RegionFace，这是一种基于 Qwen2.5-VL 的模型，采用四阶段渐进式微调策略进行训练，逐步构建全局感知、区域感知对齐、区域焦点细化和多属性识别能力。实验结果表明，Focal-RegionFace 显著优于

当前最先进的多模态大语言模型（例如 Llama3.2-Vision），在生成和识别任务中均表现出色，能够提供更优异的以区域为中心的面部描述性能。

尽管取得了这些令人鼓舞的结果，仍存在若干局限性。受计算资源限制，我们的研究主要聚焦于开源和闭源多模态大语言模型；Gemini-2.5-Pro 和 GPT-4o 等更大规模的模型仅作为评估器，而非微调骨干模型。此外，MFRF 中的细粒度区域标注由 GPT-4o辅助半自动流程生成，尽管经过人工优化，仍可能引入风格不一致或标注偏差。此外，我们的评估依赖开源和闭源多模态大语言模型的判断，而这些判断可能受到模型特有的语言偏好的影响。未来的工作可以探索将 Focal-RegionFace 扩展至更大规模的模型，通过人机协同标注提高标注可靠性，并制定更稳健的跨模型评估协议，以实现更公平、可解释性更强的评估。

附录A. FRFM数据集设计方法详情

附录 *A.1.* 面部区域选择方法

参数	描述
L	面部关键点集合，表示为 $L = \{(x_i, y_i) \mid i \in [1, N]\}$
N	面部关键点总数
B_r	随机生成的边界框集合，表示为 $B_r = \{(x_1, y_1, x_2, y_2)\}$
N_b	所需的边界框数量
IOU_{thresh}	IoU 的最大重叠阈值
W_f，H_f	面部区域的宽度和高度
W_{min}，W_{max}	生成的框的最小和最大宽度
H_{min}，H_{max}	生成的框的最小和最大高度
M	生成不重叠框的最大尝试次数
S_f	最终成功生成的边界框集合

表 A.7：面部区域选择方法中所用参数的详细信息。

根据表 A.7 中的参数，我们按照以下步骤对框区域进行随机划分。

面部区域估计。给定面部标志点集合 L，面部区域的宽度 W_f 和高度 H_f 计算如下：

$$W_f = \max_{x_i \in L}(x_i) - \min_{x_i \in L}(x_i), \qquad H_f = \max_{y_i \in L}(y_i) - \min_{y_i \in L}(y_i) \tag{A.1}$$

面部区域的边界坐标由以下方式确定：

$$(fx_1, fy_1) = (\min_{x_i \in L}(x_i), \min_{y_i \in L}(y_i)), \qquad (fx_2, fy_2) = (\max_{x_i \in L}(x_i), \max_{y_i \in L}(y_i)) \tag{A.2}$$

随机框生成。随机生成的边界框的最小和最大尺寸定义如下：

$$W_{min} = 0.2 \times W_f, \quad W_{max} = 0.4 \times W_f \tag{A.3}$$

$$H_{min} = 0.2 \times H_f, \quad H_{max} = 0.4 \times H_f \tag{A.4}$$

对于每个生成的边界框，其坐标按如下方式计算：

$$x_1 = \text{rand}(fx_1,\ fx_2 - W_{rand}), \qquad y_1 = \text{rand}(fy_1,\ fy_2 - H_{rand}) \tag{A.5}$$

$$x_2 = x_1 + W_{rand}, \quad y_2 = y_1 + H_{rand} \tag{A.6}$$

其中，W_{rand} 和 H_{rand} 分别从 $[W_{min}、W_{max}]$ 以及 $[H_{min}、H_{max}]$ 中采样。

交并比（IoU）。对于任意两个边界框 $B_1 = (x_1, y_1, x_2, y_2)$ 和 $B_2 = (x_3, y_3, x_4, y_4)$：

$$\text{IoU}(B_1, B_2) = A_{ov} \,/\, (A_1 + A_2 - A_{ov}) \tag{A.7}$$

其中：

$$\begin{aligned} \text{重叠面积} &= \max(0, \min(x_2, x_4) - \max(x_1, x_3)) \\ &\quad \times \max(0, \min(y_2, y_4) - \max(y_1, y_3)) \end{aligned} \tag{A.8}$$

迭代逻辑。每次生成一个边界框时，都会检查它与 S_f 中所有边界框的交并比：

$$S_f = \{B_i \mid \text{IoU}(B_i, B_j) < IOU_{thresh}, \forall B_j \in S_f\} \tag{A.9}$$

如果所有交并比值都低于 IOU_{thresh}，则将新边界框添加到 S_f。

最终生成过程。

1. **初始化：**根据面部标志点 L估计面部区域 W_f、H_f。

2. **随机采样：** 最多生成 M 个随机边界框：

   - 在面部区域内随机采样坐标。
   - 计算与现有边界框的交并比。
   - 如果满足交并比约束，则将边界框添加到 S_f 中。

3. **终止：** 重复执行，直到 $|S_f| = N_b$。

附录 *A.2. GPT-4o* 生成提示词详细信息。

在本节中，我们将介绍所采用的提示设计，以确保 GPT-4o[3]能够可靠地遵循我们的指令。正如正文所述，FRFM 提示分为三个连续阶段：上下文聚焦、区域约束和结构化生成。

在上下文聚焦阶段，GPT-4o [33] 被赋予专家角色（例如，具备深厚 FACS 知识的法医年龄估算与面部皮肤病学专家），如图 A.5a 所示。角色设定使模型聚焦于恰当的领域，并抑制无关推理。随后，区域约束阶段严格限制模型仅关注框选的面部区域（例如，“仅检查框选区域”），确保整体面部线索不会影响分析。最后，在结构化生成阶段，指示 GPT-4o 生成一段逻辑清晰的文字，其中：(1) 描述框内的表面特征和肌肉相关线索；(2) 不涉及框外特征或明确提及动作单元引用；(3) 以与所提供的真实标签相符的年龄估计作为结尾。

对于情绪与年龄提示词，GPT-4o 会明确依据框选区域及相应的真实标签进行引导。相比之下，AU 提示缺少框选区域 AU 标注。为解决这一问题，我们通过两步构建区域级动作单元真实映射：首先，GPT-4o 从全局真实标注中选出在框选区域内呈现激活状态的动作单元；随后，如果某个动作单元的标准激活区域至少有 60% 落在框内，则将其纳入。所有选出的动作单元均视为区域级

[3]https://openai.com/index/hello-gpt-4o/

真实标注。由于没有框选区域 AU 标签，AU 提示的回复格式不受限制；不过，分析仍严格限定于指定区域和提示范围，与年龄和情绪设置保持一致。

<年龄 真实标注：XX> 你是一位在法医年龄估计和面部皮肤病学方面经验丰富的专业专家，深谙FACS 和年龄相关皮肤标志。

已知：
• 面部图像中的一个框选区域
• 完整的年龄真实标注（仅供内部参考）

仅检查框选区域。不要参考区域外的任何特征，也不要直接提及动作单元。请将以下内容整合起来：
1. **皮肤表面线索**（皱纹深度、细纹模式、毛孔可见度、整体弹性）
2. 2. **底层肌肉效应**（紧张或隆起区域会强化或柔化这些特征）

用一个流畅的英文段落详细描述你观察到的内容——肌肉收缩如何与皮肤质量相互作用。
结论要求：
- **将提供的真实标签作为确定的年龄。**
- - 不要推断或替换为其他年龄范围——你最终陈述的年龄范围**必须与真实标签完全一致**。

示例回答：
> “在框选区域内，细小的水平纹路横贯鼻唇沟区域，在颊肌轻微
收缩时会加深。皮肤有中度松弛，毛孔清晰可见，表明胶原蛋白密度有所降低。颧弓处的轻微隆起
表明口角提肌可能正在活动，但紧致的皮肤阻止了明显皱褶的形成。总体而言，这些
综合迹象表明，估计年龄范围为40–45岁。”

（a）使用 GPT-4o 生成年龄细粒度描述的提示词细节。

<情绪 groundtruth: XX> 你是一位接受过面部动作编码系统训练的面部表情分析专业专家
（FACS）和皮肤状况的皮肤科评估。
给定：
• 来自面部图像的框选区域
• 完整的情绪真实标签（仅供内部参考）

仅观察该边界框内的像素。不要推测边界框外的任何内容，也不要引用情绪标签。相反，
结合：
1. **可见的肌肉效应**（收缩、抬升、压缩、表面张力）
2. 2. **皮肤状况**（弹性、紧致度、皱纹深度、毛孔可见度、纹理）

用一个流畅的英文段落描述你所看到的内容——肌肉牵拉如何使皮肤变形，以及皮肤的
年龄/弹性如何调节这些变形。
结论要求：
- **使用所提供的真实标签作为确定的情绪。**
- - 不要推断或替换为其他类别——最终陈述的情绪必须与真实标注完全一致。
示例答案：
> “在框选区域内，下眼睑看起来抬起并略微起皱，表明眼轮匝肌收缩
所致。随着紧绷的皮肤被拉伸，外眼角的细小放射状纹路加深，这与颧部提升肌的活动相符。
与此同时，光滑、年轻的皮肤表面使明显的鱼尾纹褶皱不易形成，取而代之的是眼周轻微的
起伏。综合来看，这些线索表明此人正感到轻微惊讶。推断出的情绪
感到惊讶。”

（b）使用 GPT-4o 生成情绪细粒度描述的提示词细节。

<全局动作单元真实值映射：XX> 你是一位接受过面部动作编码
系统（FACS）培训的面部行为分析专家。给定一张面部图像中的框选区域和一幅完整的动作单元（AU）激活图，你的任务是仔细观察框选区域本身，并根据你掌握的动作单元知识，描述可见的面部肌肉活动和表情
编码于你的专业知识中。

你的分析必须严格局限于框选区域——在任何情况下都不得对此范围之外的内容进行推测。仅根据该区域内肉眼可见的情况，判断是否存在面部肌肉活动，例如肌肉收缩、抬升、压缩或表面张力变化。只有当某个动作单元（AU）所对应的肌肉效果明显处于活动状态，且其相关肌肉区域至少有 60% 位于框选区域内时，才应描述该动作单元。描述可见的肌肉活动时，必须明确指出哪个 AU 导致了框选区域内的哪种可观察变化（例如：“AU12 使脸颊抬起”）。你可以描述任何明显处于活动状态的 AU，即使它未列在提供的 AU 映射中；但不得仅依据映射或框外区域提及或引用任何 AU。AU 映射仅供内部参考；你的判断必须完全基于框选区域内的视觉线索。

除肌肉行为外，还必须考虑皮肤状况（例如年轻、紧致、老化、松弛），因为皮肤状况会影响肌肉激活在皮肤表面的表现。请思考皮肤的弹性或紧致度如何影响皱纹形成、表面张力或肌肉牵拉的可见程度。

不要提及情绪或整体面部表情。输出应为一段使用专业、自然英语撰写的流畅文字，内容严格限于框选区域内可见的部分。

（c）使用 GPT-4o 生成动作单元细粒度描述的提示词详情。

图 A.5：使用 GPT-4o 生成年龄、情绪和动作单元细粒度描述的提示词细节。

附录 B. 训练策略详情

附录 *B.1.* 多样化提示策略（第一阶段至第四阶段）。

如第 3.2 节第一阶段部分所述，对于 FRFM 数据集的每个阶段（第一阶段至第四阶段），我们针对三种属性分别设计了五种不同的查询，并将其随机分配给相应的属性图像，以增强模型对多样化查询环境的适应能力。值得注意的是，我们还在每个查询前添加了相应属性的识别标签： <任务：情绪>、 <任务：动作单元>或 <任务：年龄>。

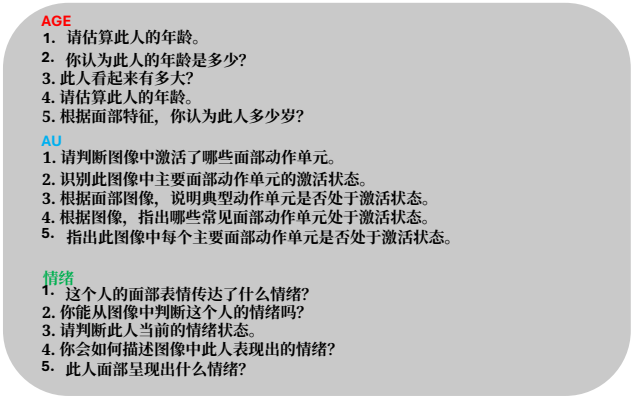

图 B.6：第一阶段使用的多样化提示详情。

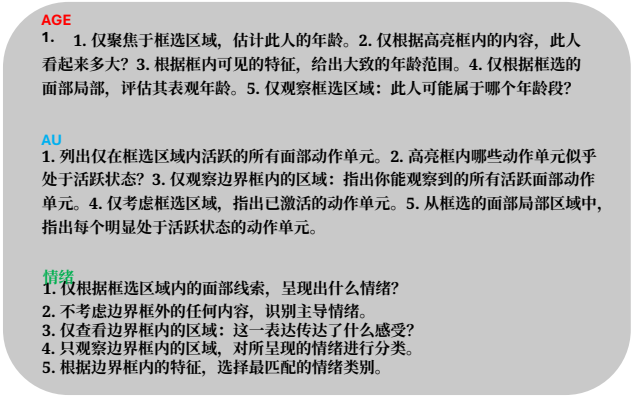

图 B.7：第二阶段和第三阶段使用的多样化提示详情。

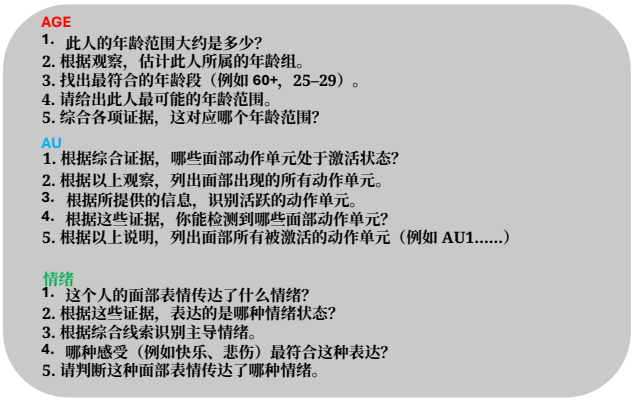

图 B.8：第四阶段使用的多样化提示详情。

附录 C. 实验细节

附录 *C.1.* 实现细节。

如论文正文第4.1节所述，我们简要介绍实验设置的实现细节。所有实验中使用的基线模型为 Qwen2.5-32B-VL [20], , 并始终采用 4 位量化。如表 C.8 所示，我们在第一阶段至第三阶段采用相同的参数设置。然而，在第四阶段，由于输入查询发生了显著变化，我们调整了截断阈长度，其余参数保持不变。

附录 *C.2.* 基于多模态大语言模型的评估提示设置细节。

第4.1节概述了我们基于多模态大语言模型的评估细节。我们精心设计了两种提示，分别用于评估开源模型和闭源模型。

如图 C.10 所示，开源模型的评估提示采用成对比较策略：对于相应的图像，每次评估都会包含两条图像描述，一条来自我们的模型，另一条来自其他模型。这一策略能够提高回答的稳定性和比较判断的准确性。在设计提示时，我们首先为大语言模型设定一个具体角色，然后介绍任务的五项关键评估标准：分类准确率、面部描述细节丰富度、语言流畅性与自然度、框内聚焦度和语义相关性。这种

表 C.8：实验参数。

多阶段 I–III			
参数	**值**	**参数**	**值**
训练轮数	10	权重衰减	0.01
预热比例	0.2	学习率	2×10^{-5}
批量大小	16	梯度累积步数	4
学习率调度器类型	余弦	截断长度	1024
LoRA秩 [34]	16	LoRA alpha [34]	128
LoRA dropout [34]	0.15		
Focal-RegionFace 第四阶段			
截断长度	2048		

结构使得我们能够从任务准确率和描述质量两个维度进行评分。最后，我们明确规定了回答格式，以便于后续解析和分析。

相比之下，闭源模型的提示词采用了多描述输入策略，如图 C.9 所示。这是因为 GPT-4o[33]和 Gemini-2.5-Pro [4]等闭源模型，即使面对长上下文和大量令牌输入，仍能保持出色的性能和稳定性。尽管提示词的结构组件与开源设置中的基本相同，但我们对部分词语和句子进行了修改，以符合闭源 API 所施加的隐私、伦理和敏感内容约束。

需要注意的是，两类提示在推理时都会包含对应的图像，以支持视觉基础分析。这一设计利用了 Gemini-2.5-Pro [39] 和 ChatGPT-4o [33]等高性能模型强大的视觉语言推理能力。通过联合输入图像和多条图像描述，我们可以获得更可靠、更细粒度的评估结果。不过，我们

[4]https://deepmind.google/technologies/gemini/pro/

强调，闭源评估的结果并非旨在取代使用开源模型开展的实验。相反，它们可作为补充参考，帮助我们更全面、均衡地理解不同模型范式下的描述质量。

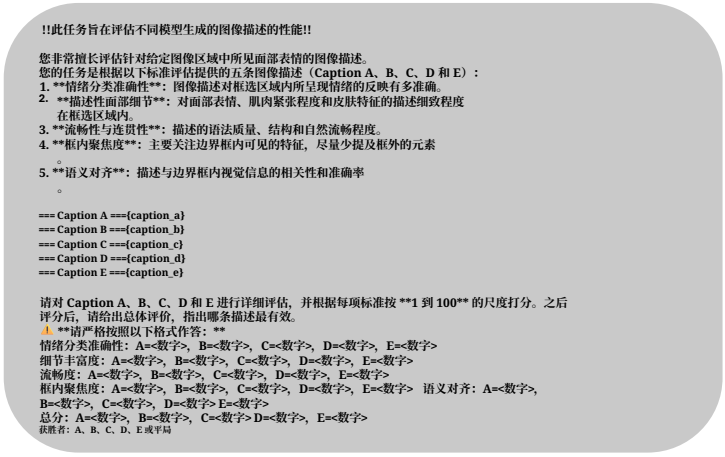

图 C.9：闭源评估提示的详细信息。

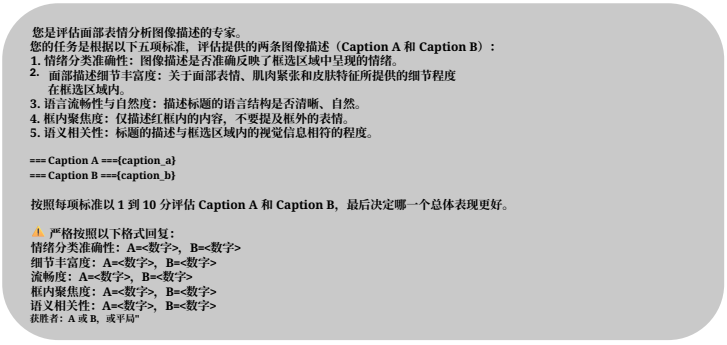

图 C.10：开源评估提示的详细信息。

附录 *C.3.* 基于多模态大语言模型的详细指标分解。

为补充主要评估结果，我们针对多个面部理解维度——年龄、动作单元和情绪——进行了更细粒度的分析，涵盖开源模型（表 C.10）和闭源多模态大语言模型（表 C.9）。我们报告了各项评估标准（例如分类、检测、流畅度等）的详细得分。我们的 Focal-RegionFace 在这三个维度的几乎所有指标和设置下都始终表现出色，进一步证明了我们方法的有效性。

值得注意的是，在开源和闭源模型评估中，都出现过这样的情况：某个模型在多项指标上的得分高于其他模型，但最终胜率却更低。Qwen2.5- VL [20]作为评估器、将自身与我们的模型进行比较时，这一现象尤为明显。主要原因在于，某些测试样本的得分存在离群值，且指标值有所波动。尽管最终平均分看起来可能很高，胜率仍可能较低。这是因为高性能大语言模型在充当评估器时，往往会偏向自身生成的图像描述。在我们的实验中，由于我们的模型生成的图像描述通常优于 Qwen2.5-VL [20], ，评分因此出现不一致——某些指标偏向 Qwen2.5-VL [20] ，而其他指标则偏向我们的模型——最终导致平均分较高，但单项获胜次数较少。

表 C.9：基于闭源多模态大语言模型评估，对多种属性预测（年龄、动作单元和情绪）各项指标的详细拆解。

	Gemini-2.5-Pro [39]							**GPT-4o** [33]						
AGE														
模型	Cls	Det	Flu	Box	Sem	胜率/%	Rank	Cls	Det	Flu	Box	Sem	胜率/%	Rank
Qwen2.5-VL [20]	70.73	62.80	88.33	79.51	66.57	14.89	3	78.20	82.86	86.93	81.18	82.68	7.32	4
Gemma3 [37]	**81.76**	66.76	88.99	80.63	78.04	15.96	2	85.83	84.24	86.67	**85.07**	**91.2**	39.61	2
Deepseek-Janus-Pro [22]	73.90	20.33	89.65	88.00	68.58	1.04	5	76.33	56.06	78.88	82.36	75.87	8.32	3
Llama3.2-Vision [23]	69.97	42.33	70.97	67.85	56.71	1.71	4	73.22	59.35	64.59	73.99	73.51	0.40	5
Focal-RegionFace	77.84	**87.21**	**94.50**	**93.56**	**78.56**	**66.40**	1	**86.55**	**88.06**	**92.43**	83.01	90.06	**52.30**	1
AU														
Qwen2.5-VL [20]	33.67	38.94	72.25	68.77	37.87	9.57	2	59.33	66.56	76.03	65.31	65.62	16.65	2
Gemma3 [37]	31.91	32.15	40.82	69.30	35.84	8.95	3	49.18	34.13	47.80	51.84	49.21	3.48	5
Deepseek-Janus-Pro [22]	8.08	9.11	59.55	61.67	12.18	0.75	5	29.97	30.24	58.05	47.49	35.28	8.72	3
Llama3.2-Vision [23]	32.02	23.45	65.77	67.48	33.89	7.73	4	53.13	38.96	53.59	59.08	56.79	7.61	4
Focal-RegionFace	**67.04**	**79.25**	**93.34**	**89.30**	**73.23**	**72.68**	1	**75.94**	**87.89**	**88.66**	**72.88**	**80.88**	**71.43**	1
情绪														
Qwen2.5-VL [20]	53.67	40.30	62.89	71.39	51.19	15.31	2	64.30	44.44	71.65	77.71	60.73	28.48	2
Gemma3 [37]	63.45	44.16	84.40	79.66	60.14	12.28	3	67.04	30.21	80.24	74.13	64.11	15.31	3
Deepseek-Janus-Pro [22]	51.01	11.84	90.29	90.67	50.59	3.19	5	56.23	20.42	82.41	74.71	56.80	11.61	4
Llama3.2-Vision [23]	51.03	33.75	86.24	70.77	47.29	5.17	4	49.48	31.13	83.87	74.47	54.28	3.1	5
Focal-RegionFace	**66.51**	**82.27**	**93.65**	**92.58**	**72.30**	**63.61**	1	**71.80**	**75.62**	**85.08**	**88.10**	**67.58**	**49.39**	1

表 C.10：基于开源多模态大语言模型评估，对多种属性预测（年龄、动作单元和情绪）的各项指标进行详细拆解。

	Qwen2.5-VL [20]						Deepseek-Janus-Pro [22]						Llama3.2-Vision [23]					
AGE																		
模型	Cls	Det	Flu	Box	Sem	胜率/%	Cls	Det	Flu	Box	Sem	胜率/%	Cls	Det	Flu	Box	Sem	胜率/%
Qwen2.5-VL [20]	68.41	80.38	55.12	76.35	80.46	**58.34**	79.49	79.90	81.44	78.92	75.24	0.07	58.09	67.77	53.25	74.49	62.27	18.91
Focal-RegionFace	73.64	73.47	91.45	81.37	92.23	41.66	88.82	91.05	91.38	89.81	89.67	**99.93**	77.80	84.02	70.98	80.60	72.24	**81.09**
Deepseek-Janus-Pro [22]	50.18	56.11	38.55	73.01	83.47	10.18	81.32	82.46	81.31	82.33	83.71	0.11	44.32	55.78	46.67	62.66	50.13	6.25
Focal-RegionFace	77.44	85.59	93.12	86.39	96.53	**89.82**	89.22	89.48	90.46	89.37	89.11	**99.89**	83.43	75.56	65.39	80.09	71.83	**93.75**
Llama3.2-Vision [23]	64.41	69.82	53.78	75.37	81.60	37.04	79.90	80.98	84.18	82.64	84.97	0.05	63.17	73.06	61.94	76.30	67.31	11.96
Focal-RegionFace	77.00	83.67	87.43	82.38	88.33	**62.96**	88.23	89.04	89.90	89.03	89.12	**99.95**	78.24	85.68	76.22	84.12	76.32	**88.04**
AU																		
Qwen2.5-VL [20]	64.20	73.99	57.13	75.20	80.72	**62.02**	76.90	78.29	81.23	78.55	75.01	0.00	60.81	69.17	54.17	70.40	62.05	20.00
Focal-RegionFace	76.82	81.25	79.10	80.77	92.99	37.98	87.16	89.28	89.23	86.92	86.62	**100.00**	79.70	83.43	72.12	82.32	76.34	**80.00**
Deepseek-Janus-Pro [22]	48.22	57.16	36.60	71.77	79.45	12.52	77.40	79.14	81.59	80.01	79.42	0.00	43.82	53.31	43.21	58.41	47.98	9.15
Focal-RegionFace	80.42	86.09	92.96	83.62	94.58	**87.48**	91.94	91.82	92.24	91.31	92.07	**100.00**	83.22	78.39	70.77	82.52	77.81	**90.85**
Llama3.2-Vision [23]	57.19	66.41	49.54	72.46	83.42	36.34	71.77	72.97	80.36	80.17	82.45	0.00	64.13	71.28	58.51	71.59	64.77	14.80
Focal-RegionFace	78.48	85.36	89.53	83.90	90.76	**63.66**	87.97	89.14	91.08	88.73	88.67	**100.00**	81.36	88.98	77.19	84.81	76.12	**85.20**
情绪																		
Qwen2.5-VL [20]	63.89	72.90	56.70	74.12	80.00	34.98	72.51	78.57	81.09	79.41	75.84	0.00	60.95	66.91	55.93	71.85	59.21	14.26
Focal-RegionFace	75.83	82.70	97.44	82.09	91.90	**65.02**	92.67	92.29	91.95	91.49	92.44	**100.00**	85.45	83.63	73.99	83.67	82.51	**85.74**
Deepseek-Janus-Pro [22]	45.01	56.31	26.86	73.32	82.32	8.10	70.18	80.05	81.79	79.61	80.12	0.00	46.33	53.76	36.56	56.96	44.23	3.78
Focal-RegionFace	81.25	86.29	93.01	84.37	94.21	**91.90**	92.50	92.43	92.64	91.78	92.38	**100.00**	87.25	80.74	71.22	81.90	75.35	**96.22**
Llama3.2-Vision [23]	54.10	65.32	41.56	73.10	79.86	17.62	67.57	75.04	83.77	82.33	82.90	0.00	60.29	69.63	58.74	76.80	62.41	8.98
Focal-RegionFace	78.85	81.13	88.47	76.90	88.77	**82.38**	89.56	89.67	90.77	89.86	89.69	**100.00**	76.60	85.70	74.53	81.63	74.11	**91.02**

附录 *C.4.* 传统多属性识别评估详情。

在第 4.3-III 节中，我们展示了我们的模型在动作单元、情绪和年龄识别任务中的平均性能，证明了其在传统多属性识别方面优于大多数现有开源大语言模型。本节将进一步提供支持这些结果的详细信息。如表 C.11 所示，在每个属性识别任务的大多数类别中，我们的模型均取得了最佳或次佳性能，表明其在不同类别中都保持了稳定的优势。

表 C.11：基于开源多模态大语言模型评估，对使用面部区域焦点图像进行多属性预测（年龄、动作单元和情绪）时各项指标的详细分解。评估指标为准确率和 F1 分数（%）。

模型	中性	愤怒	厌恶	Fear	快乐	悲伤	惊讶	平均值 (%)
Deepseek-Janus-Pro [22]	4.44	**85.28**	0.00	17.86	43.10	15.48	80.34	35.21
Llama3.2-Vision [23]	0.00	36.61	17.86	3.57	28.16	3.57	44.64	18.42
Gemma3 [37]	1.67	50.57	18.45	27.38	63.09	**78.57**	24.69	37.77
Qwen2.5-VL [20]	**95.89**	55.17	2.38	29.17	26.44	16.67	23.81	35.64
Focal-RegionFace	2.22	51.15	**22.62**	**32.14**	**68.97**	77.38	27.98	**40.35**

模型	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-59	60+	平均值（%）
Deepseek-Janus-Pro [22]	17.95	15.04	**67.73**	17.53	61.79	30.00	**46.03**	0.00	15.87	0.00	18.75	75.64	31.92
Llama3.2-Vision [23]	**96.83**	1.63	15.87	2.78	**92.28**	0.00	0.81	0.00	0.00	0.00	0.00	**93.50**	25.18
Gemma3 [37]	88.10	40.65	50.00	25.79	39.43	**33.33**	38.21	10.71	15.87	2.85	**59.13**	62.20	38.88
Qwen2.5-VL [20]	88.49	**63.82**	26.59	33.73	54.07	30.95	17.07	17.46	**18.65**	**14.63**	10.32	82.52	38.11
Focal-RegionFace	91.46	55.28	48.41	**49.21**	45.53	11.11	2.44	**59.52**	18.25	12.60	39.29	90.65	**43.65**

模型	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	F1 分数
Deepseek-Janus-Pro [22]	0.00	10.52	2.03	16.12	4.15	23.53	7.21	10.13	9.68	21.32	2.43	3.40	9.21
Llama3.2-Vision [23]	7.21	1.52	1.71	17.93	**22.80**	29.08	0.00	12.21	11.98	15.33	7.13	11.82	11.56
Gemma3 [37]	**14.85**	10.50	**27.63**	20.11	18.39	43.32	15.64	**20.05**	**16.53**	38.82	10.45	19.43	21.31
Qwen2.5-VL [20]	1.53	0.00	3.66	**38.67**	0.00	31.36	0.00	5.22	1.39	29.77	1.22	7.90	10.06
Focal-RegionFace	14.29	**29.80**	17.24	30.10	7.99	**44.08**	**27.15**	14.70	10.37	**40.17**	**13.81**	**27.74**	**23.12**

附录 *C.5*。消融研究：多阶段从 *I* 到 *III* 的影响——各项指标的详细分解。

在第 4.4 节中，我们针对第一阶段至第三阶段开展了消融研究，以评估模型在各阶段的性能。本节提供更多细节，以补充正文中呈现的图表。如表 C.12 所示，我们报告了三个阶段在五项评估指标上的详细得分以及胜率。结果表明，在第一阶段，我们的模型在所有指标上均不如其他模型。然而，进入第二阶段后，各项得分均显著提升，证明了我们的区域感知人脸视觉语言对齐策略的有效性和成功。在第三阶段，模型性能较第二阶段进一步提升，但增幅并不明显。这是

意料之中的，因为第三阶段主要侧重于强化框级注意力——掩码机制迫使模型严格基于目标区域生成描述——这进一步验证了我们设计的鲁棒性和相关性。

表 C.12：基于所提出的基于多模态大语言模型的评估进行的详细消融研究。

		GPT-4o [33]				
	指标	Focal-RegionFace	Qwen2.5-VL [20]	Gemma3 [37]	Deepseek-Janus-Pro [22]	Llama3.2-Vision [23]
第一阶段	**Cls**	56.85	77.11	**77.96**	67.37	70.15
	Det	26.06	77.68	**79.94**	49.94	58.36
	Flu	36.82	**84.98**	81.22	77.21	64.70
	Box	59.99	**82.43**	80.97	75.53	72.44
	Sem	48.82	79.54	**81.05**	67.06	67.07
	获胜/%	0.14	41.28	**49.96**	2.5	6.12
第二阶段	**Cls**	**74.65**	67.28	67.2	55.04	63.36
	Det	**83.06**	64.70	60.16	40.07	51.12
	Flu	**84.34**	78.17	72.18	73.41	60.81
	Box	**79.82**	74.66	72.83	67.90	67.57
	Sem	**79.50**	69.64	68.04	55.82	61.30
	获胜/%	**49.97**	18.09	27.96	0.76	3.22
第三阶段	**Cls**	**74.38**	67.28	67.35	55.20	63.61
	Det	**83.86**	64.62	60.10	39.91	51.16
	Flu	**84.72**	78.20	72.15	73.41	60.78
	Box	**81.33**	74.73	71.70	68.19	67.85
	Sem	**79.51**	69.68	68.16	55.98	61.53
	获胜/%	**54.71**	14.57	19.47	7.55	3.70

表 C.13：基于所提出的基于多模态大语言模型的评估进行的详细消融研究。

		GPT-4o [33]				
	指标	Focal-RegionFace	Qwen2.5-VL [20]	Gemma3 [37]	Deepseek-Janus-Pro [22]	Llama3.2-Vision [23]
单阶段	**Cls**	65.47	**75.47**	73.12	59.38	67.43
	Det	58.24	74.96	**71.55**	44.31	55.23
	Flu	38.57	**79.89**	77.51	73.96	62.77
	Box	73.22	**81.52**	78.94	68.42	70.82
	Sem	68.83	74.42	**80.72**	59.69	58.86
	获胜/%	23.80	32.72	**37.24**	2.50	3.74

表 C.14：基于所提出的基于多模态大语言模型的评估的详细消融实验。

		GPT-4o [33]				
	指标	Focal-RegionFace	Qwen2.5-VL [20]	Gemma3 [37]	Deepseek-Janus-Pro [22]	Llama3.2-Vision [23]
第三阶段	**Cls**	20.73	78.41	**79.13**	65.45	70.84
	Det	53.56	**79.33**	78.63	47.41	60.19
	Flu	73.42	81.48	**83.44**	74.23	63.42
	Box	53.20	78.90	**81.69**	71.78	70.77
	Sem	67.53	75.89	**80.65**	69.32	68.33
	获胜/%	11.67	38.64	**41.60**	3.20	4.89
第一阶段+III	**Cls**	66.48	**69.41**	68.23	54.26	64.41
	Det	**68.47**	67.84	67.85	41.98	55.26
	Flu	60.87	**77.23**	72.87	69.38	66.85
	Box	71.46	**76.91**	76.42	68.50	71.43
	Sem	**70.23**	70.01	69.10	60.41	65.43
	获胜/%	25.42	28.31	**29.94**	6.99	9.34
第二阶段+III	**Cls**	**71.56**	63.26	65.44	63.78	61.33
	Det	**80.31**	61.33	60.58	59.26	60.71
	Flu	**82.44**	78.43	67.65	66.42	61.28
	Box	**90.07**	72.76	74.70	52.17	70.14
	Sem	**81.08**	72.54	68.42	57.32	47.16
	获胜/%	**49.51**	19.23	20.21	8.23	2.82

附录 *C.6.* 消融研究：比较多阶段与单阶段微调

我们采用与前文相同的实验设置，对单阶段微调下的模型进行评估，并在表 C.13 中报告五项评估指标和胜率。如表所示，单阶段微调在所有指标上的表现均持续低于多阶段方法。这一差距表明，在单一阶段同时使用异构数据和多个目标进行训练，会妨碍有效的知识获取。相比之下，多阶段策略确保学习进程循序渐进，使模型能够先获取基础表征，再处理更复杂的任务，从而获得更稳健、泛化能力更强的性能。

附录 *C.7.* 消融研究：选择性微调组合。

为分析各个微调阶段的贡献，我们评估了三种部分配置：Base + Stage III、Base + Stage I +Stage III，以及 Base + Stage II + Stage III（表 C.14）。仅使用第一阶段训练会导致性能严重下降，

由于模型既缺乏全局感知，也缺乏区域感知对齐，这表明区域掩码微调依赖于先前的区域语义学习。引入第一阶段后结果有所改善，但若没有第二阶段，结果仍不稳定，这说明仅靠全局感知无法取代区域感知监督。相比之下，Base + 第二阶段 +第三阶段取得了优异的性能，证实第二阶段在建立局部语义理解方面发挥着核心作用，但其表现仍略逊于完整流程。总而言之，这些结果表明，渐进式微调至关重要：第二阶段对于第三阶段的实现不可或缺，而第一阶段则始终能够增强鲁棒性和泛化能力。

附录 D. 可视化：更多示例



图 D.11：可视化示例 1。

参考文献

[1] J. Mao, R. Xu, X. Yin, Y. Chang, B. Nie, A. Huang, Y. Wang, Poster++: 一种更简洁且更强大的面部表情识别网络，模式识别 157 (2025年) 110951。doi：https://doi.org/10.1016/j.patcog.2024.110951。URL https://www.sciencedirect.com/science/article/pii/ S0031320324007027

[2] R. Rothe、R. Timofte、L. Van Gool，从单幅图像中预测真实年龄和表观年龄的深度期望值，无需面部关键点，国际计算机视觉期刊 126 (2–4)（2018年）144–157。doi:10.1007/s11263-016-0940-3。URL https://doi.org/10.1007/s11263-016-0940-3

[3]H. Zhao、Y. Wang、X. Gao，将密码与脑电图模板相结合的脑电波验证系统，用于在线识别与认证，模式识别 169（2026年）112009。doi:https: // doi.org/10.1016/j.patcog.2025.112009。URL https://www.sciencedirect.com/science/article/pii/ S0031320325006697

[4] C. Lei、K. Dang、S. Song 等，医疗保健领域的人工智能辅助面部分析：从疾病检测到综合管理，Patterns 6 (2)（2025年）101175。doi:https://doi.org/10.1016/j.patter.2025.101175。URL https://www.sciencedirect.com/science/article/pii/ S2666389925000236

[5] G. Wang, F. Lin, T. Wu, Z. Liu, Z. Ba, K. Ren, 《Fsfm：通过自监督面部表征学习实现通用化的人脸安全基础模型》（2025年）。ar Xiv:2412.12032。URL https://arxiv.org/abs/2412.12032

[6]S. Wang、Y. Chang、Q. Li、C. Wang、G. Li、M. Mao，通过无监督多源域自适应实现姿态鲁棒个性化面部表情识别，《模式识别》150（2024年）110311。

doi:https://doi.org/10.1016/j.patcog.2024.110311. URL https://www.sciencedirect.com/science/article/pii/ S0031320324年000621

[7] X. Wang, X. Ma, X. Hou, M. Ding, Y. Li, J. Chen, W. Chen, X. Peng, L. Shen, 《Facebench：用于评测面部感知大语言模型的多视角、多层级面部属性视觉问答数据集》（2025年）。arXiv:2503.21457。网址：https://arxiv.org/abs/2503.21457

[8] K. Narayan, V. VS, V. M. Patel, 《Facexbench：评估多模态大语言模型的面部理解能力》（2025年）。arXiv:2501.10360。网址：https://arxiv.org/abs/2501.10360

[9] K. Zheng, X. Ge, J. Fu, J. Peng, J. M. Jose, 《全面面部状态分析的多模态表征学习技术》（2025年）。arXiv:2504.10351。网址：https://arxiv.org/abs/2504.10351

[10] Z. Wu、J. Cui，La-net：标签噪声下可靠面部表情识别的关键点感知学习（2023）。arXiv：2307.09023。URL https://arxiv.org/abs/2307.09023

[11] Y. Zhang、X. Zheng、C. Liang、J. Hu、W. Deng，通用面部表情识别（2024）。arXiv：2408.10614。URL https://arxiv.org/abs/2408.10614

[12] T. Ganel、C. Sofer、M. A. Goodale，人类对面部年龄感知的偏差是客观存在的，并且在当前人工智能技术中更加明显，《Scientific Reports》12（1）（2022）22519。doi：10.1038/s41598-022-27009-w。URL https://doi.org/10.1038/s41598-022-27009-w

[13] C. Rudin，停止为高风险决策解释黑箱机器学习模型，改用可解释模型，《Nature Machine Intelligence》1 (5) (2019年) 206–215。doi:10.1038/s42256-019-0048-x。URL https://doi.org/10.1038/s42256-019-0048-x

[14] X. Ge、J. Fu、F. Chen、S. An、N. Sebe、J. M. Jose，通过视觉语言联合学习实现端到端可解释面部动作单元识别，载于：MM，MM '24，ACM，2024年，第 8189–8198 页。doi:10.1145/3664647.3681443。URL http://dx.doi.org/10.1145/3664647.3681443

[15] A. Chaubey, X. Guan, M. Soleymani, 《Face-llava：通过指令微调实现面部表情和属性理解》（2025年）。arXiv:2504.07198。网址：https://arxiv.org/abs/2504.07198

[16] Z. Cheng, Z.-Q. Cheng, J.-Y.He, K. Wang, Y. Lin, Z.Lian, X. Peng, A. Hauptmann, 《Emotion-llama：通过指令微调实现多模态情绪识别与推理》，《神经信息处理系统进展》37（2024年），110805–110853。

[17] Y. Li, J. Zeng, S. Shan, X. Chen, 《基于视频的面部动作单元检测自监督表征学习》，载于：2019年 CVPR，2019年，第10916–10925页。doi:10.1109/CVPR.2019.01118。

[18] A. Romero, J. Leon, P. Arbelaez，多视角动态面部动作单元检测（2018年）。arXiv:1704.07863。网址：https://arxiv.org/abs/1704.07863

[19] Y. Shou, X. Cao, H. Liu, D. Meng，用于年龄估计的掩码对比图表征学习，《模式识别》158（2025年）110974。

[20] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z.Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, J. Lin，Qwen2.5-vl 技术报告，arXiv 预印本 arXiv:2502.13923（2025年）。

[21] J. Fu, X.Ge, X.Xin, A. Karatzoglou, I. Arapakis, K. Zheng, Y. Ni, J. M. J. Joemon, 《序列推荐中多模态基础模型的高效适配》，IEEE知识与数据工程汇刊（2025年）。

[22] X. Chen, Z. Wu, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, C. Ruan，《Janus-pro：通过数据和模型扩展实现统一的多模态理解与生成》，arXiv预印本 arXiv:2501.17811（2025年）。

[23] J. Lee、K.-U. Song、S. Yang、D. Lim、J. Kim、W. Shin、B.-K. Kim、Y. J. Lee、T.-H. Kim，通过裁剪交叉注意力视觉特征提升 llama-3.2-vision 的效率，arXiv 预印本 arXiv:2504.00557（2025年）。[24]X. Zhang、L. Yin、J. F. Cohn、S. Canavan、M. Reale、A. Horowitz、P. Liu、J. M. Girard，Bp4d-spontaneous：高分辨率自发三维动态面部表情数据库，IVC 32 (10)（2014）692–706，Automatic Face and Gesture Recognition 2013 最佳论文。doi:https://doi.org/10.1016/j.imavis.2014.06.002。URL https://www.sciencedirect.com/science/article/pii/ S0262885614001012

[25] A. Mollahosseini、B. Hasani、M. H. Mahoor，Affectnet：一个用于野外面部表情、效价和唤醒度计算的数据库，IEEE TAC 10 (1) (2019年) 18–31。doi:10.1109/TAFFC.2017.2740923。URL https://doi.org/10.1109/TAFFC.2017.2740923

[26] Z. Zhang、Y. Song、H. Qi，通过条件对抗自编码器实现年龄增长/回归，载于：IEEE计算机视觉与模式识别会议（CVPR），IEEE，2017年。

[27] P. Viola，M. Jones，使用简单特征增强级联进行快速目标检测，载于：CVPR 2001，第 1 卷，2001 年，第 I–I 页。doi:10.1109/CVPR.2001. 990517。
[28] D. Kollias，P. Tzirakis，A. Baird，A. Cowen，S. Zafeiriou，Abaw：效价-唤醒度估计、表情识别、动作单元检测与情绪反应强度估计挑战（2023）。arXiv:2303.01498。网址：https://arxiv.org/abs/2303.01498

[29] W.-J. Yan，S. Li，C. Que，J. Pei，W. Deng，Raf-au 数据库：包含主观情绪判断和客观 AU 标注的野外面部表情数据集，载于：亚洲计算机视觉会议，2020 年。

[30] I. Ahn、Y. Baek、B.-N. Seo、S. E. Lim、K. Jung、H. S. Kim、J. Kim、S. Lee、S. Lee，《韩国青年和中年成年人面部图像和人口统计数据的感知年龄估计》，Scientific Reports 14 (1) (2024年) 30084。doi:10.1038/s41598-024-78695-7。URL https://doi.org/10.1038/s41598-024-78695-7

[31] C. Trojahn、G. Dobos、A. Lichterfeld 等，《人类面部皮肤老化的特征：区分外在与内在生物学现象》，《生物医学研究国际》2015 (2015) 318586。doi:10.1155/2015/318586。URL https://doi.org/10.1155/2015/318586

[32] C. Sagonas、G. Tzimiropoulos、S. Zafeiriou、M. Pantic，《野外300张人脸挑战赛：首届面部关键点定位挑战赛》，载于：2013年国际计算机视觉大会研讨会，2013年，第397–403页。doi:10.1109/ICCVW.2013.59。

[33] P. P. Ray，《ChatGPT：关于背景、应用、主要挑战、偏差、伦理、局限性和未来前景的综合综述》，《物联网与信息物理系统》3（2023）121–154。doi:10.1016/j.iotcps.2023.04.003。URL https://www.sciencedirect.com/science/article/pii/ S266734522300024X

[34] Y. Zheng、R. Zhang、J.Zhang、Y.Ye、Z.Luo、Z. Feng、Y.Ma，《LlamaFactory：语言模型的统一高效微调》，计算语言学协会，曼谷，泰国，2024年。URL http://arxiv.org/abs/2403.13372

[35] T. Zhang、V. Kishore、F. Wu、K. Q. Weinberger、Y. Artzi，《BERTScore：使用 BERT 评估文本生成》，arXiv 预印本 arXiv:1904.09675（2019年）。

[36] N. Yushev，LanguageTool Python：LanguageTool 的 Python 封装，访问日期：

2025年5月15日（2019年）。网址：https://github.com/jxmorris12/language_tool_python[37] G. Team、A. Kamath、J. Ferret、S. Pathak、N. Vieillard 等，《Gemma 3 技术报告》（2025年）。arXiv:2503.19786。网址：https://arxiv.org/abs/2503.19786 [38] A. Panickssery、S. Bowman、S. Feng，《大语言模型评估器会识别并偏好自身的生成结果》，《神经信息处理系统进展》第37卷（2024年），68772–68802。[39] G. Team、R. Anil、S. Borgeaud、J.-B. Alayrac、J. Yu 等，《Gemini：一系列高性能多模态模型》（2025年）。arXiv:2312.11805。网址：https://arxiv.org/abs/2312.11805