MEGC2026：面向视觉问答的微表情竞赛

Xinqi Fan[1]，Jingting Li[2]，John See[3],Moi Hoon Yap[1],Su-Jing Wang[2],Adrian K. Davison[1]1 曼彻斯特城市大学计算与数学系[2] 中国科学院大学心理学系、中国科学院心理研究所认知科学与心理健康国家重点实验室[3] 赫瑞-瓦特大学马来西亚校区数学与计算机科学学院

arXiv:2603.08927v1 [cs.CV] 9 Mar 2026

摘要—面部微表情（ME）是面部不自主运动，通常出现在高风险环境中：当一个人体验到某种情绪，却试图抑制或压抑面部表情时，面部会自发地产生这种运动。近年来，微表情识别、微表情检测和微表情生成领域取得了长足进展。多模态大语言模型（MLLM）和大型视觉语言模型（LVLM）的出现，凭借强大的多模态推理能力，为提升微表情分析带来了新的发展途径。2026年微表情大挑战（MEGC）设置了两项任务，体现了这些不断发展的研究方向：（1）微表情视频问答（ME-VQA），通过对较短视频序列进行视觉问答来探索微表情理解，并利用多模态大语言模型或大型视觉语言模型处理各类与微表情相关的问题；（2）微表情长视频问答（ME-LVQA），将视觉问答扩展至真实场景中的长时段视频序列，要求模型在较长时间跨度内进行时序推理并检测细微的微表情。所有参赛算法均须在公开排行榜上提交结果。更多详情请访问 https://megc2026.github.io。

I. 引言

当一个人试图抑制面部表情时，通常是在高风险情境下，脸部可能会发生不由自主的动作，即面部微表情（微表情） [3]。因此，微表情的持续时间非常短，通常不超过500毫秒（毫秒），这是将其与正常面部表情区分开来的明显特征 [9]。对微表情进行计算分析和任务自动化是多媒体研究中的新兴领域。然而，直到最近，少量自发诱发的面部微表情数据集的出现，才推动研究从计算层面进一步发展。

多年来，我们组织了一系列微表情大挑战（MEGC），以探索新兴方向并推动微表情研究达到最先进水平。迄今为止，已举办七届微表情竞赛：FG'18 [20],聚焦微表情识别；FG'19 [16]，涵盖微表情检测与识别；FG'20 [8]，以微表情检测为核心；MM'21 [10],首次引入微表情生成，并继续关注微表情检测；MM'22 [11]，继续关注微表情检测与生成；MM'23 [6]，聚焦微表情检测；MM'24 [15],首次引入先检测后识别范式和跨文化微表情检测；MM'24 [7]，首次引入视觉问答。所有微表情竞赛的概览见图1。

今年迎来了第9届MEGC，该赛事与Face and Gesture (FG) 2026联合举办。本届赛事设置了两项任务：（1）微表情视频问答（微表情视觉问答），以及（2）微表情长视频问答（微表情长视频问答）。

任务1：微表情视觉问答。多模态大语言模型（多模态大语言模型）和大型视觉语言模型（大型视觉语言模型）的兴起，凭借其先进的多模态推理能力，为通过这些模型开展微表情分析带来了新的机遇。在该任务中，传统的微表情标注（如情绪类别和动作单元）被重新表述为问答（QA）对。给定一段时长相对较短的视频序列以及自然语言提示，模型需要生成答案，描述所观察到的微表情、其属性以及相关信息。该任务为微表情理解引入了新颖的多模态方向，通过自然语言促进可解释性、灵活性以及与人类意图一致的交互。微表情视觉问答已在MEGC 2025中成功引入，本届MEGC 2026将继续开展该任务，并进行部分调整。

任务 2：微表情长视频问答。微表情长视频问答是MEGC 2026 引入的一项新任务。该任务在微表情视觉问答的基础上进行扩展，旨在应对现实、自然场景中长时长视频序列内的微表情分析挑战。模型需要在较长时间跨度内进行时序推理，在自然的面部运动和表情中检测并理解细微的微表情。参赛者必须回答有关包含多个表情事件的长视频中微表情的问题，因此，这一更具挑战性的任务变体将检验微表情分析方法在真实场景中的鲁棒性和实际适用性。

II. 任务 1：视频问答

A. 任务定义

在微表情视频问答设置中，模型的输入包括：(i) 一段捕捉与微表情相关的细微面部运动的视频帧序列，以及 (ii) 一个与该序列的情绪或属性相关的对应自然语言问题。模型需要依据其对视觉线索和微表情问题的理解，生成自然语言答案。这些问题可以涵盖广泛的属性，从二分类（例如"面部是否出现降口角肌动作单元？"）到多分类（例如"什么

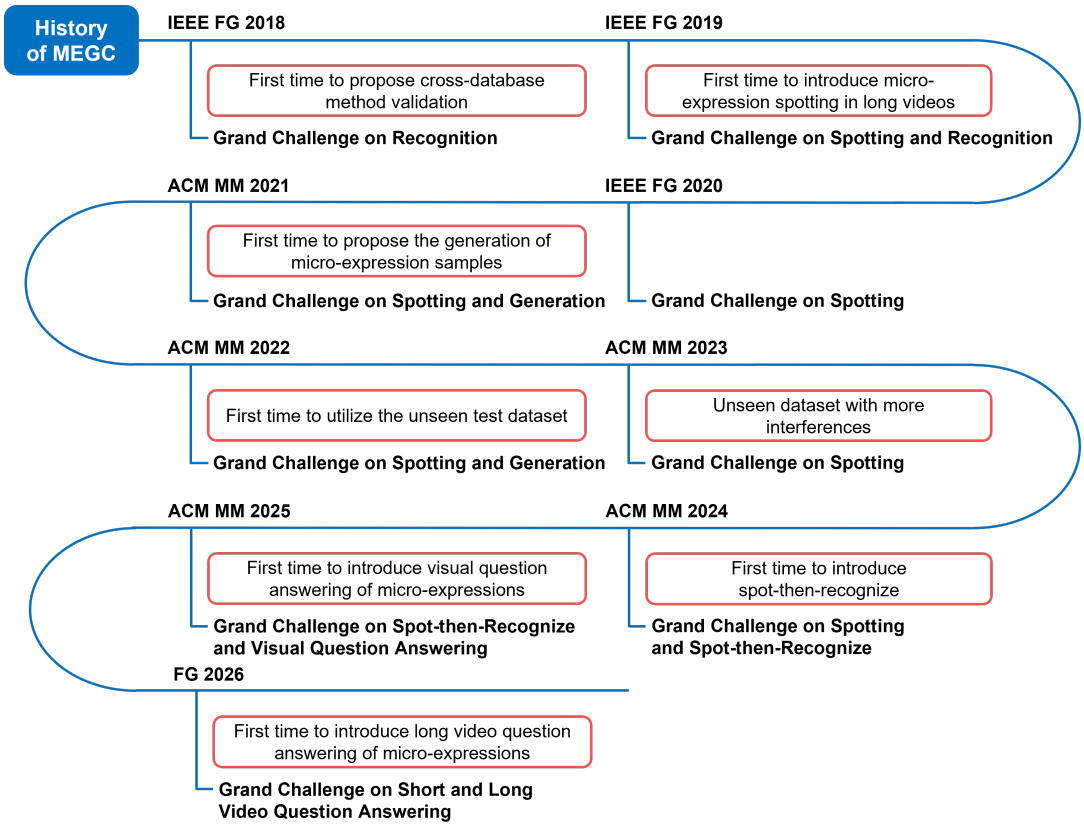


图1。微表情竞赛（MEGC）概览。

是表情类别？”），再到更复杂的询问，例如“请对这段视频片段中观察到的微表情进行详细分析。”。

B. 数据集

我们不限制训练集，但建议参与者使用 SAMM [5], CASME II [17], SMIC [13], CAS(ME)[3] [9], 或4DME [12] 数据集进行训练。

为便于任务训练，我们根据 SAMM、CASME II 和 SMIC 的标注整理了一个微表情视觉问答（ME-VQA-v1）数据集[1] [7]。此外，在 MEGC 2026 中，我们对部分问题和答案进行了优化，因此提供了 ME-VQA-v2 数据集[2]，适用于训练 ME-VQA 模型。

ME-VQA-V2 测试集包含 24 个微表情片段，其中 7 个片段来自 SAMM 挑战赛数据集 [19], [5]，17 个片段来自中国科学院（CAS(ME)）中不同视频的裁剪片段[3]（此前未公开发布）。SAMM 的帧率为 200 帧/秒，中国科学院（CAS(ME)）[3] 的帧率为 30 帧/秒。参与者应在该未见过的数据集上进行测试。

[1]ME-VQA 版本数据集可在以下网址获取：https://megc2025.github.io/challenge.html[2]ME-VQA-v2 数据集可在以下网址获取：https://megc2026.github.io/challenge.html

C. 评估指标

为评估我们的微表情视觉问答基线的性能，我们分别报告情绪分类指标和整体语言回答质量指标。

对于粗粒度和细粒度情绪分类，我们采用非加权 F1 分数（UF1）和非加权平均召回率（UAR），以确保对各类别进行均衡评估。粗粒度情绪类别包括积极、消极和惊讶。细粒度情绪类别包括快乐、惊讶、恐惧、厌恶、愤怒和悲伤。这些指标的详细信息见 [16]。

对于所有视觉问答答案，我们报告双语评估替补（BLEU） [14]以及面向召回率的摘要评估替补（ROUGE）-1 [4]，以评估生成文本的质量。BLEU 按如下方式评估预测答案与参考答案之间的 n 元语法精确率：

$$\mathrm{BLEU} = \exp\left(\min\left(1-\frac{r}{c},0\right)+\sum_{n=1}^{N} w_n \log p_n\right), \quad (1)$$

其中，r 是参考答案的长度，c是候选答案的长度，p_n 是 n-元语法精确率，w_n 是权重。ROUGE-1 分数定义为候选答案 C 与参考答案 R 之间一元语法重叠的召回率：

$$\text{ROUGE-1} = \frac{\sum_{w\in V} \min\left(\mathrm{N}_p(w),\ \mathrm{N}_r(w)\right)}{\sum_{w\in V} \mathrm{N}_r(w)}, \quad (2)$$

表 I

基于 ME-VQA-V2 训练集和测试集的 ME-VQA 任务基线结果。使用在精选多模态情绪视觉问答数据集上训练的 QWEN2.5VL-3B 和 QWEN3VL-4B（基线方法）进行 ME-VQA 任务测试集评估。粗粒度情绪类别指积极、消极和惊讶。细粒度情绪类别指快乐、惊讶、恐惧、厌恶、愤怒和悲伤。ZS 指零样本性能，FT 指微调性能。

方法	SAMM					中国科学院(ME)[3]					总体				
	粗粒度		细粒度			粗粒度		细粒度			粗粒度		细粒度		
	UF1	非加权平均召回率	UF1	非加权平均召回率	BLEU ROUGE	UF1	非加权平均召回率	UF1	非加权平均召回率	BLEU ROUGE	UF1	非加权平均召回率	UF1	非加权平均召回率	BLEU ROUGE
Qwen2.5VL-3B（零样本）	0.242	0.333	0	0	0.090	0.295	0.262	0.333	0.057	0.048	0.089	0.315	0.256	0.333	0.057	0.048	0.065	0.312
Qwen2.5VL-3B（微调）	0.206	0.25	0.057	0.1	0.234	0.331	0.376	0.515	0.044	0.111	0.109	0.375	0.328	0.422	0.059	0.117	0.122	0.362
Qwen3VL-4B（零样本）	0.242	0.333	0	0	0.068	0.229	0.262	0.382	0.114	0.222	0.065	0.382	0.256	0.333	0.079	0.200	0.066	0.339
Qwen3VL-4B（微调）	0.242	0.333	0.178	0.250	0.128	0.416	0.262	0.333	0	0	0.236	0.413	0.256	0.333	0.052	0.064	0.204	0.415

其中， V 是参考答案中的唯一词语集合，$N_C(w)$ 和 $N_R(w)$ 分别是词语 w 在候选答案和参考答案中的出现次数。

D. 基础方法

对于ME-VQA任务，我们采用Qwen2.5VL-3B模型 [2] 作为基线方法。Qwen2.5VL-3B是一种较新的大型视觉语言模型，基于Qwen 2.5 架构，包含视觉编码器、语言模型骨干网络和跨模态融合模块。该模型在大规模图文数据集上进行预训练，因此能够执行各种视觉问答任务，并具备较强的零样本和少样本泛化能力。

为建立通用基线，我们考察了Qwen2.5VL-3B模型的零样本（ZS）和微调（FT）设置。微调时，我们对视觉编码器、视觉与语言之间的投影层，以及语言模型中的查询和键应用了QLoRA。模型的所有输入均为视频。

E. 基线结果

表I 中的结果表明，Qwen2.5VL-3B和Qwen3VL-4B在ME-VQA上的表现都有限，尤其是在细粒度情绪识别方面。在零样本设置下，两种模型的粗粒度UF1/非加权平均召回率表现中等（约为0.24–0.33），但细粒度性能极低，大多数情况下UF1分数接近于零。这表明，尽管大型视觉语言模型具备一定的粗粒度情绪区分能力，但仍难以捕捉细粒度微表情分类所需的细微差异。

微调带来了稳定但幅度有限的改进，尤其是在CAS(ME)^3，上，Qwen2.5VL-3B 的粗粒度非加权F1分数/非加权平均召回率有明显提升。然而，在各数据集上，细粒度识别仍然较弱，表明细微情绪区分仍是一个主要挑战。相比之下，微调后语言质量指标（BLEU 和 ROUGE）的提升更为明显，尤其是对于 Qwen3VL-4B，这表明模型在生成语言上合理的答案方面适应得更为有效，而在实现准确的微表情分类方面则不然。

III. 任务2：长视频问答

A. 任务定义

微表情长视频问答将微表情视觉问答从短片段和预分割片段扩展到更能反映真实场景面部行为分析的长视频序列。给定一段包含自发面部活动的长视频和一个自然语言问题，微表情长视频问答的目标是生成关于序列中存在的微表情的答案，例如其发生情况、情绪类别、时间特征以及相关动作单元。与微表情视频问答设置不同，微表情长视频问答的任务形式要求模型在单一推理框架内统一时间定位和问答。微表情长视频问答任务包含多种问题类型，旨在评估对长视频中微表情事件的全面理解。这些问题涵盖事件计数，例如表情、微表情或宏表情（MaE）事件的总数；事件分类，即判断某个特定索引事件是微表情还是宏表情；以及属性级分析，即列出贯穿整个视频的不同动作单元。此类问题要求模型对整个时间序列进行推理，区分不同的表情类型，整合多个事件中的信息，并基于长时程视频证据提供结构化且自然语言化的答案。参与者必须使用多模态大语言模型或视觉语言模型来完成此任务。

B. 数据集

我们对训练集没有限制。对于微表情长视频问答任务，参与者可以使用任何标准微表情长视频数据集，例如 SAMM-LV [18], CAS(ME)^3 [9],CAS(ME)^2, 4DME [12],或其他长时程自然情境视频数据集。

为便于任务训练，我们整理了一个 ME-LVQA训练集[3]，其中包含 SAMM-LV 和 CAS(ME)^3, ，并添加了相关问答对。

微表情长视频问答的测试集包含 30 个长视频，其中包括来自 SAMM 挑战赛数据集的 10 个长视频 [19], [5]，以及从中国科学院 (CAS(ME))^3 的不同视频中截取的 20 个片段（此前尚未发布）。SAMM 挑战赛数据集的帧率为 200 帧/秒，而

[3]ME-LVQA数据集可在以下网址获取:
https://megc2026.github.io/challenge.html

表 II 微表情长视频问答任务的基线结果。使用在经整理的微表情长视频问答数据集上训练的 Qwen2.5VL-3B 和 Qwen3VL-4B（基线方法）得到的微表情长视频问答任务测试集结果。微调使用了来自 SAMM-LV 的 5 名受试者以及来自中国科学院(ME)^3 的 5 名受试者。ZS 表示零样本性能，FT 表示微调性能。

(a) SAMM

	表情		ME		MaE		AU		表情类型	
方法	MAE	RMSE	MAE	RMSE	MAE	RMSE	F1	杰卡德系数	UF1	UAR
Qwen2.5VL-3B（零样本）	5.80	6.48	3.10	3.89	2.10	2.70	0.161	0.098	0.373	0.490
Qwen2.5VL-3B（微调）	5.20	6.05	3.20	4.02	2.10	2.70	0.169	0.110	0.560	0.636
Qwen3VL-4B（零样本）	6.20	6.84	3.60	4.27	3.10	3.54	0.075	0.048	0.479	0.525
Qwen3VL-4B（微调）	5.50	6.35	3.30	4.11	3.00	3.44	0.167	0.103	0.487	0.591

(b) 中国科学院（微表情）^{3}

	表情		ME		MaE		AU		表情类型	
方法	MAE	RMSE	MAE	RMSE	MAE	RMSE	F1	杰卡德系数	UF1	UAR
Qwen2.5VL-3B（零样本）	16.35	19.67	8.35	11.05	5.95	7.10	0.189	0.110	0.355	0.500
Qwen2.5VL-3B（微调）	12.15	15.58	9.35	12.56	5.95	7.10	0.152	0.099	0.400	0.505
Qwen3VL-4B（零样本）	16.65	19.82	10.20	13.17	6.80	7.84	0.205	0.132	0.481	0.538
Qwen3VL-4B（微调）	15.85	19.25	9.80	12.73	6.80	7.80	0.246	0.154	0.355	0.500

(c) 总体（中国科学院（微表情）^{3} + SAMM）

	表情		ME		MaE		AU		表情类型	
方法	MAE	RMSE	MAE	RMSE	MAE	RMSE	F1	杰卡德系数	UF1	UAR
Qwen2.5VL-3B（零样本）	12.83	16.49	6.60	9.30	4.67	6.00	0.180	0.106	0.369	0.501
Qwen2.5VL-3B（微调）	9.83	13.19	7.30	10.51	4.67	6.00	0.158	0.102	0.467	0.553
Qwen3VL-4B（零样本）	13.17	16.66	8.00	11.03	5.57	6.72	0.161	0.104	0.487	0.539
Qwen3VL-4B（微调）	12.40	16.14	7.63	10.66	5.53	6.67	0.219	0.137	0.413	0.535

中国科学院 (CAS(ME))^3 的帧率为 30 帧/秒。参与者应在此未公开数据集上进行测试。

C. 评估指标

为评估微表情长视频问答任务的性能，我们根据各类问题的性质采用任务特定指标。

对于需要预测表情事件数量的问题，包括表情事件总数以及微表情和宏表情的数量，我们将其建模为回归任务。我们报告平均绝对误差（MAE）和均方根误差（RMSE）。平均绝对误差衡量预测误差的平均幅度，而均方根误差对较大的偏差给予更重的惩罚。其定义如下:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|, \tag{3}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{4}$$

其中， y_i 和$\hat{}y_i$ 分别表示第 i个样本中事件数量的真实值和预测值， N表示参与评估的样本数。

对于询问视频中出现了哪些不同动作单元的问题，输出是一组动作单元描述。由于答案中动作单元的顺序不会影响正确性，我们将其视为集合预测问题，并使用基于集合的 F1分数和 Jaccard指数进行评估。这些指标衡量预测动作单元集合与参考动作单元集合之间的重合程度，并且不受动作单元描述顺序的影响。给定预测动作单元集合 $\hat{A}$ 和真实动作单元集合 A，基于集合的 F1分数定义如下:

$$F1_{\mathrm{AU}} = \frac{2|\hat{A} \cap A|}{|\hat{A}| + |A|}, \tag{5}$$

杰卡德系数定义如下:

$$\mathrm{Jaccard}_{\mathrm{AU}} = \frac{|\hat{A} \cap A|}{|\hat{A} \cup A|}. \tag{6}$$

对于询问特定表情事件类型（微表情或宏表情）的问题，我们将其视为二元分类任务。遵循微表情分析中常用的评估协议，我们报告非加权 F1 分数（UF1）和非加权平均召回率（UAR） [16]。

D. 基线方法

对于微表情长视频问答任务，我们采用 Qwen2.5VL-3B [2]和 Qwen3VL-4B [1] 作为基线模型。模型的所有输入均为视频。Qwen2.5VL-3B 是一款紧凑型视觉语言模型，包含视觉编码器、语言模型骨干网络和跨模态融合模块；Qwen3VL-4B 则是其增强版本，具备更强的多模态推理能力。

我们在零样本（ZS）和微调（FT）设置下评估这两个模型，以考察它们在长视频微表情推理任务中的效果。在微调实验中，我们采用量化低秩适配（QLoRA）来高效更新模型参数。具体而言，我们将 LoRA适配器应用于视觉编码器和语言模型组件。

微调使用 ME-LVQA 训练数据中经过筛选和精简的子集进行 5 轮次，其中包括从 SAMM 和 CASME3 中随机选取的各 5 名受试者。这样选择是出于计算资源有限的考虑。

E. 基线结果

表 II 展示了 ME-LVQA 在 SAMM 和 CAS(ME)^3 挑战赛测试集及其组合上的基线评估结果。结果显示，与短片段 ME-VQA 任务相比，性能显著下降，凸显了长时视频理解带来的巨大挑战。

结果证实，长视频微表情理解远比短片段分析更具挑战性。尽管微调能够提升性能，但在微表情计数和动作单元识别方面仍未能弥合差距，错误率依然很高。因此，时间定位和细粒度面部动作建模仍是主要瓶颈。

解读这些发现时，应考虑到微调设置有限：仅使用了 10 名受试者（5 名来自 SAMM，5 名来自 CAS(ME)^3）。如此有限的受试者多样性可能会限制身份不变学习，以及对未见个体的泛化能力。鉴于微表情具有个体特异性且细微，有限的提升表明模型可能只学到了部分依赖受试者的模式，而非稳健的时间动态；这也凸显了为参赛者探索提供规模更大、更多样化训练数据的必要性。

参考文献

[1] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge 等。Qwen3-vl 技术报告。*arXiv* 预印本 *arXiv:2511.21631*，2025年。[2] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang 等。Qwen2.5-VL 技术报告。*arXiv* 预印本 *arXiv:2502.13923*，2025年。[3] X. Ben, Y. Ren, J. Zhang, S.-J. Wang, K. Kpalma, W. Meng 和 Y.-J. Liu。基于视频的面部微表情分析：数据集、特征与算法综述。*IEEE*模式分析与机器智能汇刊，2021年。[4] L. Chin-Yew。ROUGE：一种用于自动评估摘要的软件包。载于文本摘要分支专题研讨会论文集，*2004*年，2004年。[5] A. K. Davison, C. Lansley, N. Costen, K. Tan 和 M. H. Yap。 [samm]：一个自发微面部运动数据集。*IEEE*情感计算汇刊，9(1)：116–129，2018年。[6] A. K. Davison, J. Li, M. H. Yap, J. See, W.-H. Cheng, X. Li, X. Hong 和 S.-J. Wang。MEGC2023：ACM多媒体2023微表情大挑战。载于第*31*届*ACM*多媒体国际会议论文集，第9625–9629页，2023年。[7] X. Fan, J. Li, J. See, M. H. Yap, W.-H. Cheng, X. Li, X. Hong, S.-J. Wang 和 A. K. Davison。MEGC2025：先检测再识别与视觉问答微表情大挑战。载于第*33*届*ACM*多媒体国际会议论文集，第13951–13956页，2025年。[8] L. Jingting, S.-J. Wang, M. H. Yap, J. See, X. Hong 和 X. Li。MEGC2020——第三届面部微表情竞赛。载于*2020*年第*15*届*IEEE*自动人脸与手势识别国际会议（*FG 2020*），第777–780页。IEEE，2020年。[9] J. Li, Z. Dong, S. Lu, S.-J. Wang, W.-J. Yan, Y. Ma, Y. Liu, C. Huang 和 X. Fu。CAS(ME)^3：第三代自发面部微表情数据库，具有深度信息和较高的生态效度。*IEEE*模式分析与机器智能汇刊，2022年。[10] J. Li, M. H. Yap, W.-H. Cheng, J. See, X. Hong, X. Li 和 S.-J. Wang。FME'21：首届面部微表情研讨会——面部表情生成与微表情检测的先进技术。载于第*29*届*ACM*多媒体国际会议论文集，第5700–5701页，2021年。[11] J. Li, M. H. Yap, W.-H. Cheng, J. See, X. Hong, X. Li, S.-J. Wang, A. K. Davison, Y. Li 和 Z. Dong。MEGC2022：ACM多媒体2022微表情竞赛。载于第*30*届*ACM*多媒体国际会议论文集，第7170–7174页，2022年。[12] X. Li, S. Cheng, Y. Li, M. Behzad, J. Shen, S. Zafeiriou, M. Pantic 和 G. Zhao。4DME：一个包含多模态的自发4D微表情数据集。*IEEE*情感计算汇刊，第1–18页，2022年。[13] X. Li, T. Pfister, X. Huang, G. Zhao 和 M. Pietik¨ainen。自发微表情数据库：诱发、采集与基线。载于*2013*年第*10*届*IEEE*自动人脸与手势识别国际会议暨研讨会（*FG*），第1–6页。IEEE，2013年。[14] K. Papineni, S. Roukos, T. Ward 和 W.-J. Zhu。BLEU：一种自动评估机器翻译的方法。载于第*40*届计算语言学协会年会论文集，第311–318页，2002年。[15] J. See, J. Li, A. K. Davison, G. B. Liong, M. H. Yap, W.-H. Cheng, X. Li, X. Hong 和 S.-J. Wang。MEGC2024：ACM多媒体2024面部微表情竞赛。载于第*32*届*ACM*多媒体国际会议论文集，第11482–11483页，2024年。[16] J. See, M. H. Yap, J. Li, X. Hong 和 S.-J. Wang。MEGC 2019——第二届面部微表情竞赛。载于*IEEE*自动人脸与手势识别国际会议，第1–5页。IEEE，2019年。[17] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen 和 X. Fu。CASME II：改进的自发微表情数据库及基线评估。*PLOS ONE*，9(1)：e86041，2014年。[18] C. H. Yap, C. Kendrick 和 M. H. Yap。SAMM长视频：一个自发面部微表情与宏表情数据集。载于*2020*年第*15*届*IEEE*自动人脸与手势识别国际会议（*FG 2020*），第771–776页。IEEE，2020年。[19] C. H. Yap, M. H. Yap, A. Davison, C. Kendrick, J. Li, S.-J. Wang 和 R. Cunningham。利用时间定向参考帧，通过3D-CNN检测长视频序列中的面部微表情与宏表情。载于第*30*届*ACM*多媒体国际会议论文集，第7016–7020页，2022年。[20] M. H. Yap, J. See, X. Hong 和 S.-J. Wang。2018年面部微表情大挑战总结。载于*IEEE*自动人脸与手势识别国际会议，第675–678页。IEEE，2018年。