东南大学-多模态与情感智能研究组

李勇-东南大学
计算机科学与工程学院
Email: yong.li@seu.edu.cn
研究组主页： https://mysee1989.github.io/

代表性应用

应用场景

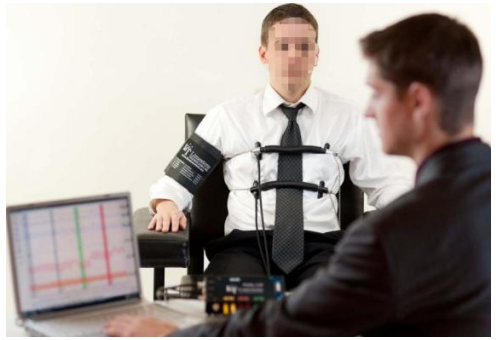
- 教育、医疗、刑侦、军事等领域

数字人

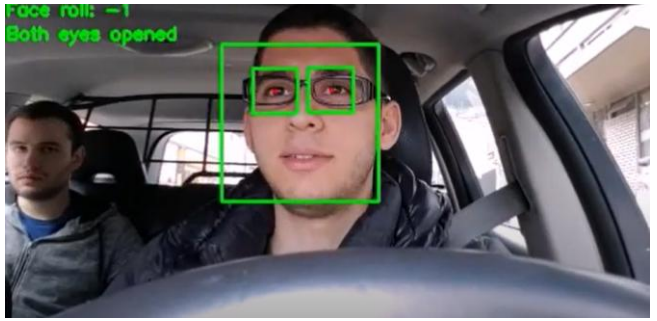
智慧养老

士兵作战意图同步

智慧医疗：患者恢复程度智能评估



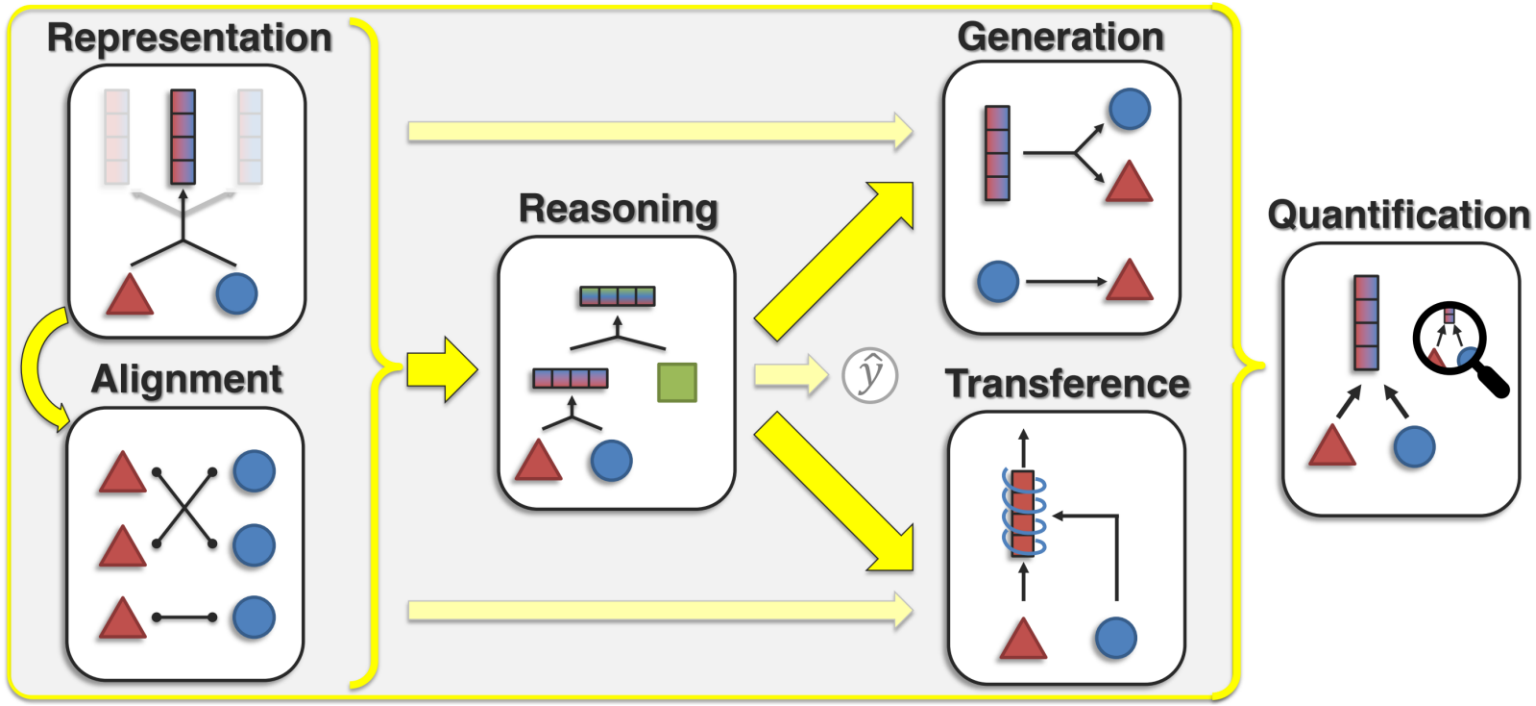
刑侦测谎



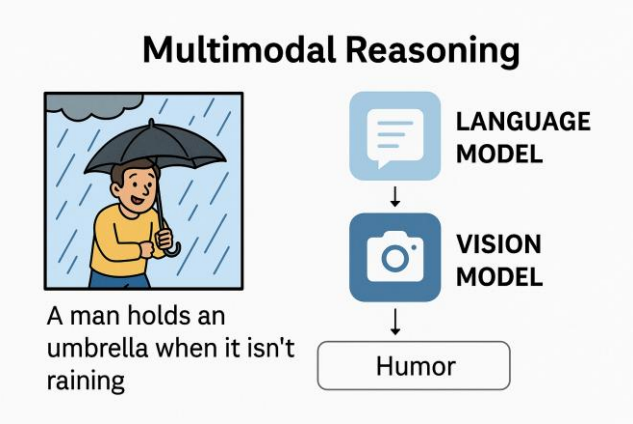
疲劳驾驶检测

挑战问题和研究内容

多模态学习：表征、对齐、推理、生成、迁移和量化



- **Louis-Philippe Morency**
- 卡内基梅隆大学计算机学院语言技术研究所，副教授

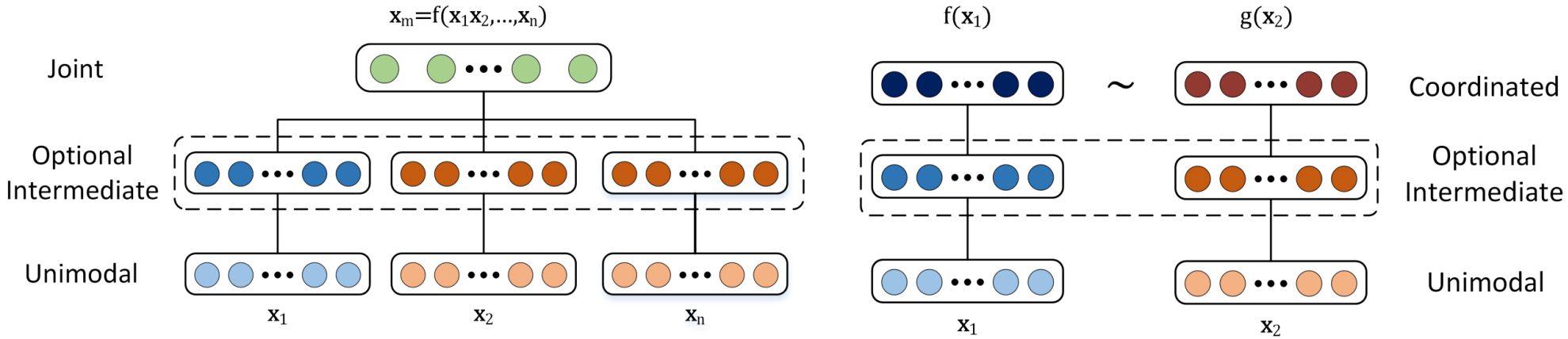


L.-P. Morency, A. Zadeh, P. Liang. Foundations & Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. *ACM CSUR*, 2024.

挑战问题和研究内容

多模态学习：表征（Representation）

- 研究多模态数据的表示方式，使其可以：（1）充分利用模态的互补性；（2）尽可能消除冗余。
  - 表示空间的相似关系应如实反映概念空间的相似性；
  - 即使部分模态的信息缺失仍然容易获取表示；
  - 根据已知模态的信息可填充或推算缺失模态的表示。



(a) Joint representation　　(b) Coordinated representations

联合表示深度融合模态信息，但对缺失模态敏感。协调表示处理缺失模态灵活，但难以捕获深层交互，且需额外对齐。

T. Baltrusaitis, C. Ahuja, L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI*, 2019. 引用4448次

挑战问题和研究内容

多模态学习：对齐（Alignment）

- 研究如何在多个模态中寻找并确定不同模态内子元素的直接对应关系。

显式对齐方法

以对齐为优化目标，核心问题是定义和计算相似性

- 无监督多模态对齐：以预设的序列关系作为约束条件

动态时间规整

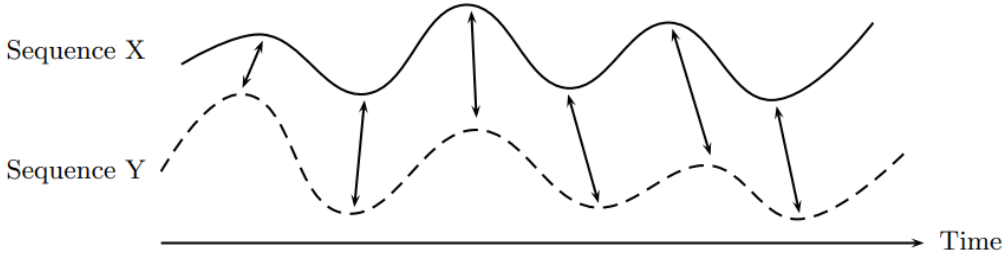
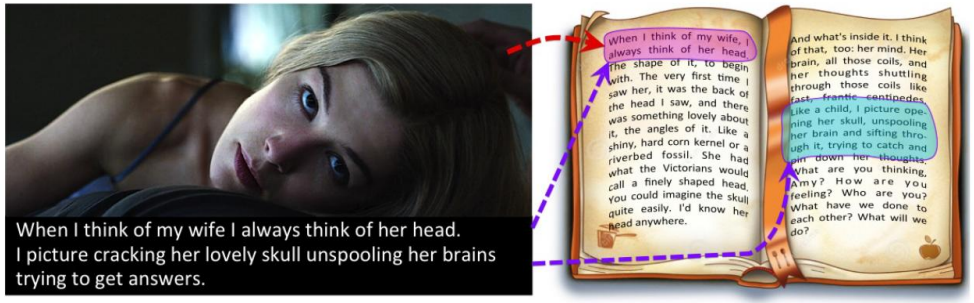


Fig. 4.1. Time alignment of two time-dependent sequences. Aligned points are indicated by the *arrows*

- 监督/弱监督多模态对齐：以全部/部分子元素对作为监督信号

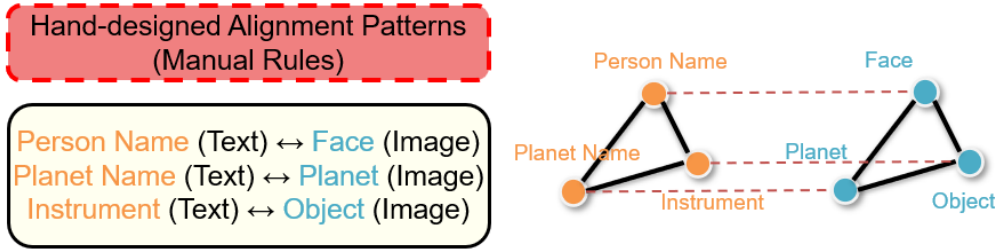
视觉和文本语义对齐



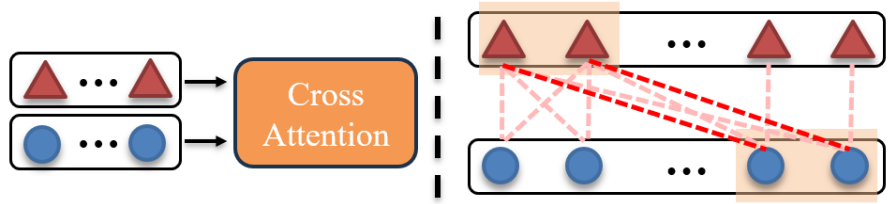
隐式对齐方法

对齐作为下游任务的中间步骤出现

- 基于图模型的多模态对齐：需手工设计子元素对齐模式



- 基于神经网络的多模态对齐：一般基于注意力机制实现对齐

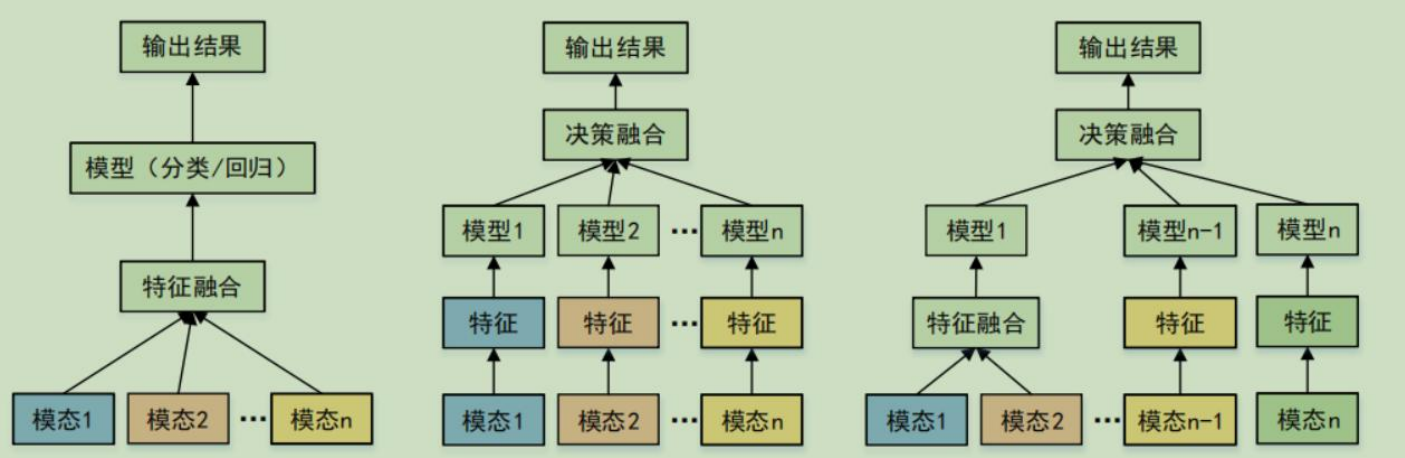


T. Baltrusaitis, C. Ahuja, L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI*, 2019.

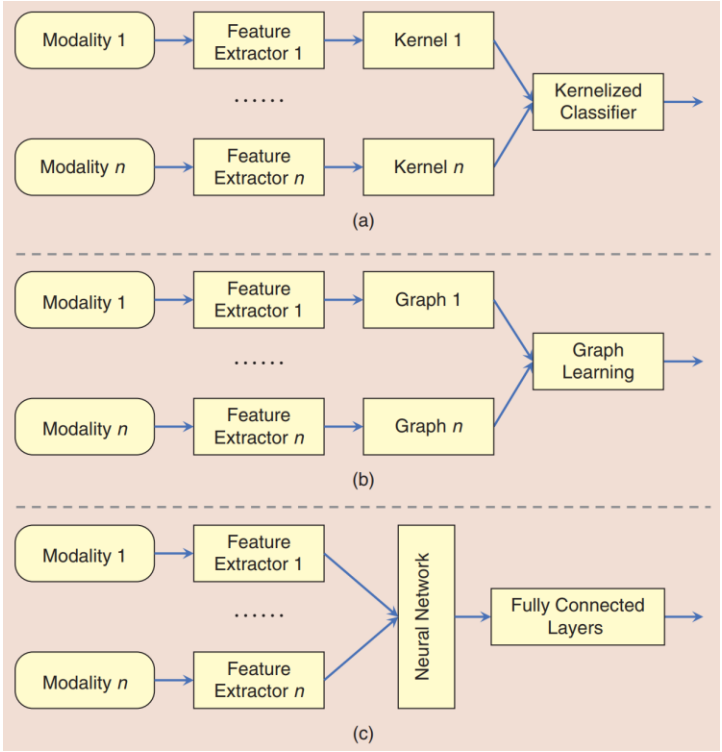
挑战问题和研究内容

多模态学习：融合（Fusion）

- 研究如何将不同模态的信息融合在一起以获得更准确的标签或连续值预测。
- 模型无关的融合
  - 前期融合（特征级融合）
  - 后期融合（决策级融合）
  - 混合式融合
- 模型相关的融合
  - 基于多核学习的融合
  - 基于图模型的融合
  - 基于神经网络的融合



模型无关融合：优点是通用性强，但可能丢失深层交互。



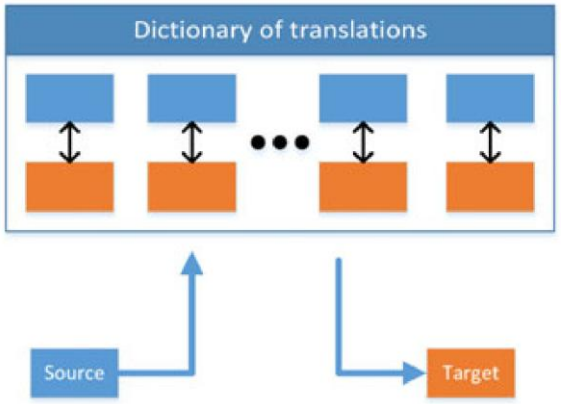
模型相关融合：优点是能深度捕获交互，但通用性差，依赖特定模型。

T. Baltrusaitis, C. Ahuja, L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI*, 2019.

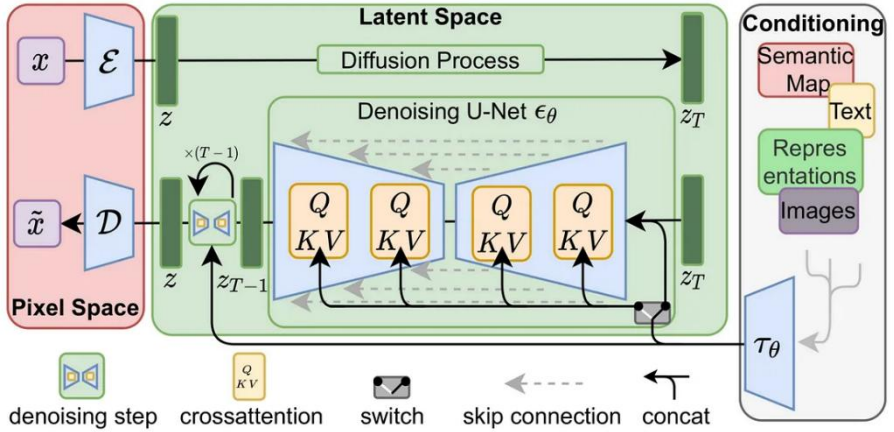
挑战问题和研究内容

多模态学习：翻译（Translation）

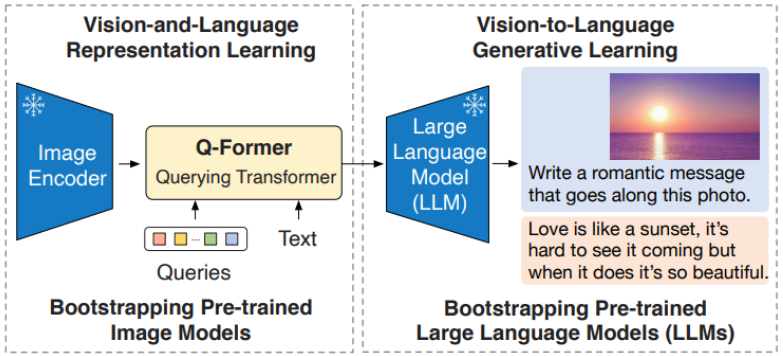
- 研究如何将数据从一种模态翻译（映射）到另一种模态。
- 基于实例的方法：（1）基于检索的方法；（2）基于检索结果混合的方法。
- 基于生成的方法：（1）基于语法的生成；（2）基于编解码生成；（3）连续生成。



(a) Example-based



(b) Generative-based, **text** to **image**



(c) Generative-based, **image** to **text**

T. Baltrusaitis, C. Ahuja, L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI*, 2019. 引用4448次

R. Rombach et al. High-resolution image synthesis with latent diffusion models. *CVPR* 2022

研究背景

国际前沿研究

- 美国工程院院士、IEEE Fellow、麻省理工学院Rosalind Picard教授

如果机器不具有感知和表达情绪的能力，那么它就无法通过图灵测试，也就意味着不具有真正意义上的智能。

- 美国工程院院士、ACM/AAAS Fellow、斯坦福大学李飞飞教授

下一步人工智能的发展，需要加强对情感、情绪的了解。情绪、情感，是人工智能未来的方向。

心智世界模型赋予AI代理理解人类内心状态的能力，是实现自然、情感、智能人机交互的核心机制。构建“心智世界模型（mental world model）”，包括对用户情绪和情感状态的建模，是未来人机协作的关键能力之一。

Meta

研究背景

行为模态情感分析相关数据集（部分）

数据库名称	模态	数据形式	数据规模	标注类型	采集机构	发布时间
IEMOCAP [1]	图像、文本、音频	视频片段	10039	情感类别	南加州大学	2008
ICT-MMMO [2]	图像、文本、音频	视频片段	370	情感类别	卡内基梅隆大学	2013
MSP-IMPROV [3]	图像、文本、音频	视频片段	8438	情感类别	德克萨斯大学达拉斯分校	2016
CMU-MOSI [4]	图像、文本、音频	视频片段	2199	情感强度	卡内基梅隆大学	2016
CHEAVD [5]	图像、音频	视频片段	7030	情感类别	中科院自动化所	2017
CMU-MOSEI [6]	图像、文本、音频	视频片段	23453	情感强度 情感类别	卡内基梅隆大学	2018
MELD [7]	图像、文本、音频	视频片段	13708	情感类别	新加坡国立大学	2018
CH-SIMS [8]	图像、文本、音频	视频片段	2281	情感类别	清华大学	2020
M^3ED [9]	图像、文本、音频	视频片段	24449	情感类别	中国人民大学	2022
MER2023 [10]	图像、文本、音频	视频片段	78178	情感强度 情感类别	中科院自动化所	2023
MERR [11]	图像、文本、音频	视频片段	33105	情感类别 情感描述	深圳技术大学	2024
EMER-Coarse [12]	图像、文本、音频	视频片段	115595	情感类别 情感描述	中科院自动化所	2024

发展趋势：

- （规模）从小到大
- （标注）由粗到精

研究背景

生理信号情感分析相关数据集（部分）

Database	Modality	Data	Emotion Annotation	Institute	Year	Publication Name
MAHNOB-HCI[1]	EEG, ECG, GSR, SKT, EOG, respiration, face body video, Audio	27 subjects;	9 classes, VA,	University of Geneva, Switzerland	2011	TAC
DEAP[2]	EEG, EOG, EMG, GSR, BVP, SKT, Respiration, Face Video	32 subjects;	V A liking, dominance, familiarity	Queen Mary University of London	2012	TAC
RECOLA[3]	ECG, GSR, Audio, Face Video	46 subjects;	V A	Université de Fribourg, Switzerland	2013	FG conference
DECAF[4]	MEG, EOG, ECG, EMG, Face Video	46 subjects	VA, dominance	University of Trento, Italy	2015	TAC
BP4D+[5]	ECG, GSR, BVP, Respiration, Face Video, Thermal	140 subjects	10 emotions, AU	Binghamton University, USA	2016	CVPR
DREAMER[6]	EEG, ECG	23 subjects	VA, dominance	University of the West of Scotland, UK	2018	IEEE JBHI
AMIGOS[6.1]	EEG, ECG, GSR, Audio, Video, Depth	40 subjects	V A, personality traits	Queen Mary University of London	2018	TAC
SEED-V[7]	EOG; ECG	20 subjects;	5 classes	上海交通大学	2019	Conference on Neural Engineering
CASE[8]	ECG, BVP, EMG, GSR, Respiration, Skin Temperature	30 subjects	V A	Institute of Robotics and Mechatronics, DLR, Germany	2019	Scientific data
BU-EEG[9]	EEG; Face Video	29 subjects	7 classes, AU, pain	Binghamton University, USA	2020	FG conference
MGEED[10]	Images, Depth; OMG, EEG, ECG	17 subjects; 150K facial images;	6 emotions, VA	University of Portsmouth, United Kingdom	2023	TAC
Mixed-ER [11]	EEG, Face Video, GSR, PPG	73 subjects	3 emotions	清华大学	2024	Scientific Data

发展趋势：

- EEG, ECG, GSR
- 行为与生理信号耦合

生理信号类别	英文名称	英文缩写
脑电图	Electroencephalogram	EEG
肌电图	Electromyogram	EMG
心电图	Electrocardiogram	ECG
眼电图	Electrooculogram	EOG
心率变异性	Heart rate variability	HRV
皮肤电反应	Galvanic skin response	GSR
皮肤电应答	Electrodermal response	EDR
皮肤电活动	Electrodermal activity	EDA
血压信号	Blood pressure	BP
皮肤温度	Skin temperature	ST
呼吸模式	Respiration pattern	RSP
光电容积脉搏波	Photoplethysmogram	PPG
眼动信号	Eye movement	EM
脉搏信号	Pulse rate	PR
血氧饱和度	Oxygen saturation	SpO2

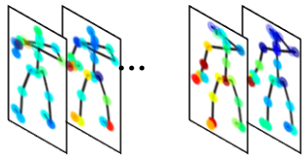
摘自《基于生理信号的情感计算研究综述》，自动化学报，2021

多模态情感计算-情感模态

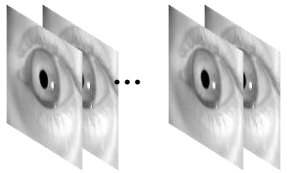
情感计算多模态数据-显性情感线索

- 人脸表情：一个或多个人脸区域/单元的孤立运动或运动组合
- 眼球运动：眼睛是心灵的窗户
- 语言语音：说话者通过使用不同文字、语调、声音大小和节奏来表达他们的意图
- 行为：将紧握的拳头推到空中，通常被视作表达胜利或欣喜的姿势
- 步态：与悲伤和满足等低激活度情感相比，愤怒和兴奋等高激活度情感与快速运动更相关
- 脑电
- 心电
- 体温
- 脉搏
- ……

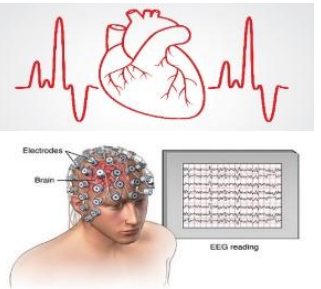
人脸表情



人体动作骨架点



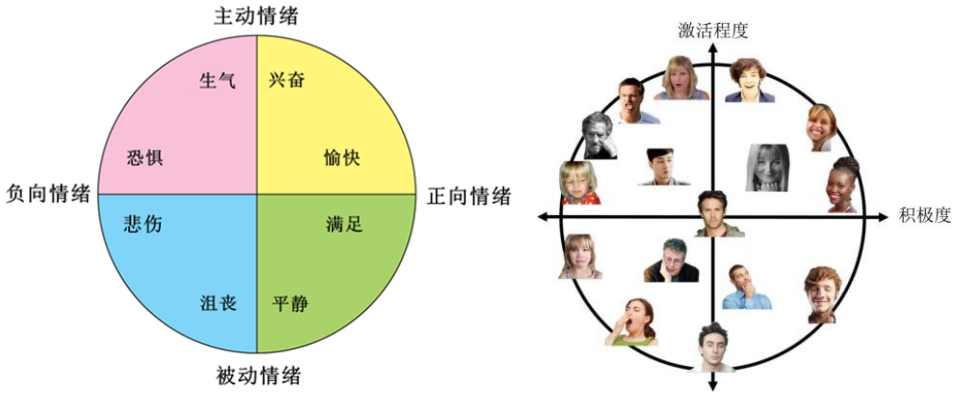
眼动信号



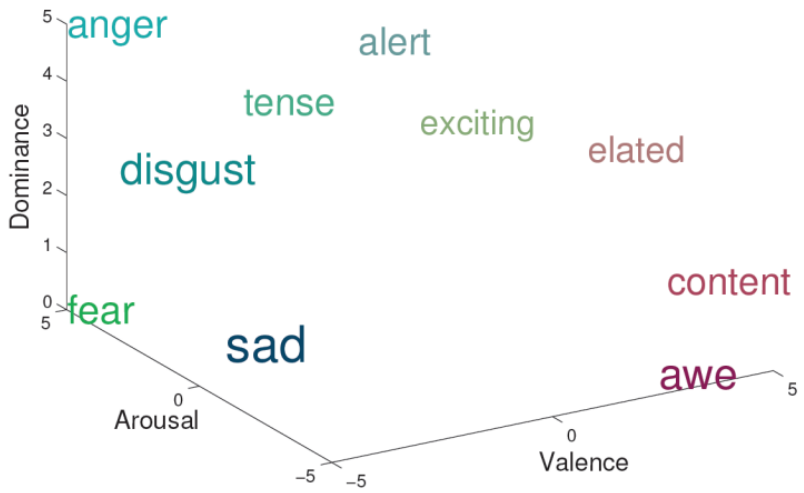
脑电信号

多模态情感识别-情感定义

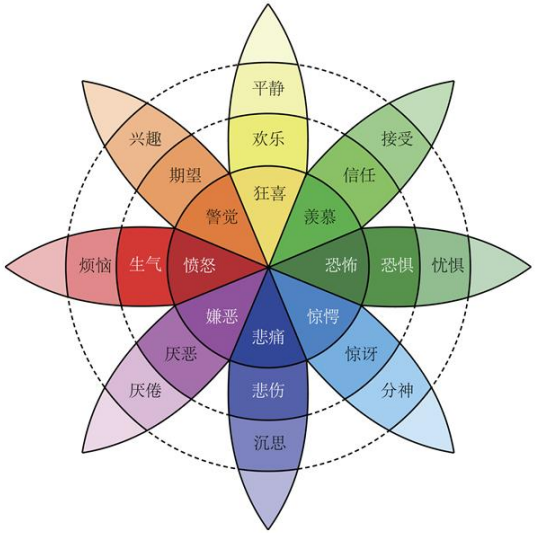
- 心理学对情感没有统一、严格的定义，多采用定性的分析方法。情感类别越来越多样化和细粒度。
- 心理学情感模型
  - 离散情感状态：Ekman六类（高兴、悲伤、恐惧、厌恶、愤怒、惊讶）
  - 连续情感空间：二维情感模型（愉悦度（Valence）和激活度（Arousal））



二维情感模型



三维情感模型



情绪轮模型

离散情感直观易懂但缺乏细微度，难表强度；**连续情感**捕捉细节和强度但标注复杂，理解不直观，维度选择有争议。

研究组相关进展（部分）

单模态情感识别

Yong Li, Jiabei Zeng, Shiguang Shan*. Learning Representations for Facial Actions from Unlabeled Videos, **IEEE TPAMI**, 2022. 视频情感分析

Yong Li, Menglin Liu et al. Decoupled Doubly Contrastive Learning for Cross Domain Facial Action Unit Detection, **IEEE TIP**, 2025.

Yi Ding **Yong Li**, et al. EmT: A novel transformer for generalized cross-subject EEG emotion recognition, **TNNLS,** 2025. 脑电情感分析

Yi Ding#, **Yong Li#**, et al. EEG-Deformer: A Dense Convolutional Transformer for Brain-computer Interfaces, **JBHI**, 2024. 脑电情感分析

Lifan Xia, **Yong Li***, et al. Collaborative Contrastive Learning for Cross-Domain Gaze Estimation, **PR,** 2024. 人脸视线估计

Yong Li, Menglin Liu et al. Counterfactual discriminative micro-expression recognition, **Visual Intelligence**,2024. 中文介绍链接：https://zhuanlan.zhihu.com/p/2014271498089678406

Yong Li, Shiguang Shan*. Contrastive Learning of Person-independent Representations for Facial Action Unit Detection, **IEEE TIP**, 2023.

Ximan Li, **Yong Li**, et al. Compound expression recognition in-the-wild with au-assisted meta multi-task learning, **CVPRW**, 2023

Yong Li, Antoni Chan, et al. Use of online therapy session data to develop behavioural markers for cognitive outcomes in non-pharmacological intervention, **Alzheimer's & Dementia**, 2023.

Yong Li, Shiguang Shan*. Meta Auxiliary Learning for Facial Action Unit Detection, **IEEE TAC**, 2023.

Yong Li, Yi Ren et al. Beyond Overfitting: Doubly Adaptive Dropout for Generalizable AU Detection, **IEEE TAC**, 2025. 图像情感分析

Yong Li, Jiabei Zeng, et al. Self-supervised representation learning from videos for Facial Action Unit Detection, **CVPR**, 2019.

Zili Wang, Lingjie Lao, Xiaoya Zhang, **Yong Li***. Context-dependent Emotion Recognition, **ChinaMM最佳海报奖**, 2022. 场景情感分析

李勇，曾加贝，山世光*. 面部动作单元检测方法进展与挑战，**中国图象图形学学报**，2020（入选 **2021** 年中国图象图形学报优秀论文）

Yong Li, Jiabei zeng, Shiguang Shan, Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism, **IEEE TIP**, 2019. **ESI 高被引**

Yong Li, Yufei Sun, ZhenCui, Pengcheng Shen, Shiguang Shan. Instance-Consistent Fair Face Recognition, **TPAMI**, 2025.**文章的图3展示了近10年代表性人脸识别技术演进路线**

多模态情感识别

Yong Li, Yuanzhi Wang, et al. Decoupled **Multimodal** Distilling for Emotion Recognition, **CVPR(Highlight)**, 2023. 多模态情感分析

Yuanzhi Wang, **Yong Li***, et al. Incomplete **multimodality-diffused** emotion recognition, **NeurIPS**, 2023.

Yuanzhi Wang, Zhen Cui*, **Yong Li***. Distribution-Consistent Modal Recovering for Incomplete **Multimodal** Learning, **ICCV**, 2023.多模态情感分析

Decoupled Hierarchical Distillation for **Multimodal** Emotion Recognition. **IEEE T-PAMI, 2026.** 中文介绍链接：https://zhuanlan.zhihu.com/p/2004516626947658969

Hierarchical **Vision-Language** Interaction for Facial Action Unit Detection. **IEEE TAC, 2026**

LAMA: Language as a Modality Anchor for Cross-Domain AU Detection, Under Review.

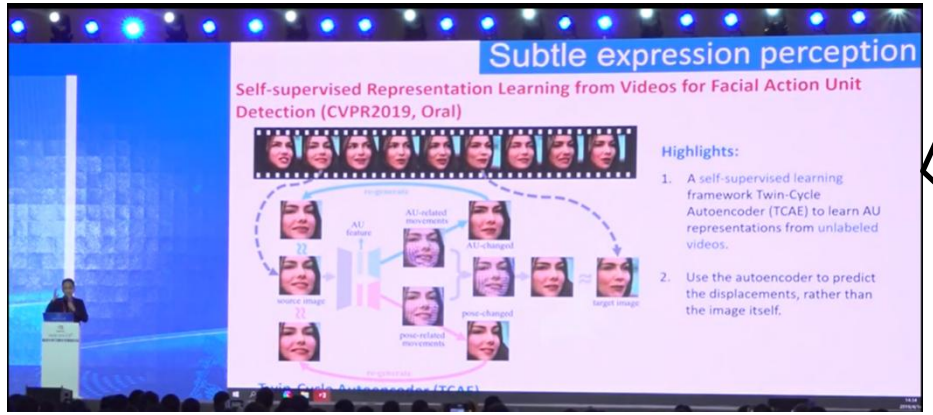
单模态情感识别

突出性成果

创新总结：构建基于大规模无标注视频的表情自监督学习框架，突破细微表情标注数据不足导致的泛化瓶颈

挑战 如何克服细微表情运动单元检测任务中标注数据匮乏难题

创新点 提出对近邻人脸视频帧之间的“刚体和柔性变换”进行解耦、重建及交叉验证，有效解决了监督表情数据匮乏问题(CVPR’19, T-PAMI’22)



在VALSE’19大会上，北京邮电大学教授在演讲中称:如何在这种非监督情况下，从海量视频里自动的挖掘出这种小的面部动作，这是个非常本质的问题。这个是CVPR 2019的一篇文章，它做出了特别突破性的工作。

a few dozen. Additionally, to achieve current state-of-the-art performance, most of the published methods have made adaptations to their CNN architectures to utilize additional features for the representation learning [9], [42], [31]. These

MIT博士，微软首席研究员Daniel McDuff称我们的工作是“目前的先进算法”。

Yong Li, Jiabei Zeng, Shiguang Shan*, Xilin Chen. Self-supervised representation learning from videos for facial action unit detection, ***CVPR 2019.***

Yong Li, Jiabei Zeng, Shiguang Shan*. Learning Representations for Facial Actions from Unlabeled Videos, ***IEEE TPAMI, 2022.***

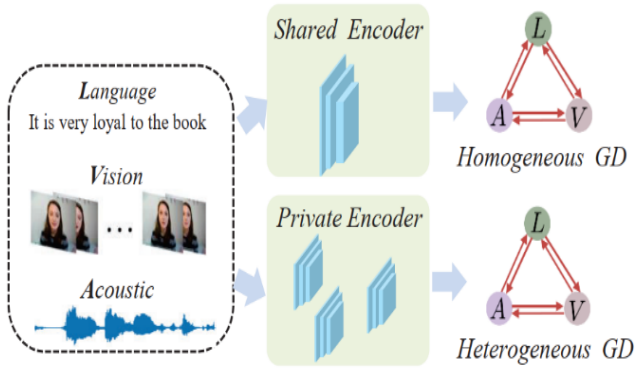
多模态情感识别

突出性成果

创新总结：构建自动化多模态互蒸馏框架，揭示异构多模态动态蒸馏机理，突破弱模态细微情感特征提取难题

挑战 如何自动、有效挖掘弱模态中蕴含的细微情感线索

创新点 提出 “自动化多模态互蒸馏”算法，设计图蒸馏策略实现任意模态间互蒸馏，揭示异构多模态动态蒸馏机理，有效提升识别精度并具备解释性（CVPR’23）



中科大联合云知声在 Multimedia 2023的论文称：受启发于我们的工作

3.2.4 Graph Distillation (GD). Inspired by the success of using multimodal approaches in emotion recognition task [22], transferring knowledge between different modalities is beneficial to the task. Therefore, we propose to apply it to the task of humor detection. We define the different modality as node in graph, and the distillation strength from modality i to modality j is denoted as edge $\omega_{i\to j}$ connecting the corresponding nodes. We consider the

丹麦哥本哈根大学在 COLING 2024的论文称：受启发于我们的工作

4.2.2. Subspace Constraint

Despite performing the aforementioned process, feature disentangling cannot be thoroughly guaranteed. There exists the potential for information to freely permeate between feature representations, whereby all modality information may be solely encoded in $\boldsymbol{H}_m^{hete}$, which renders homogeneous (modality-agnostic) multi-modal features meaningless. Inspired by Li et al. (2023), we introduce a consistency constraint in the modality-agnostic subspace to strengthen the commonality across modalities, which is formulated as follows,

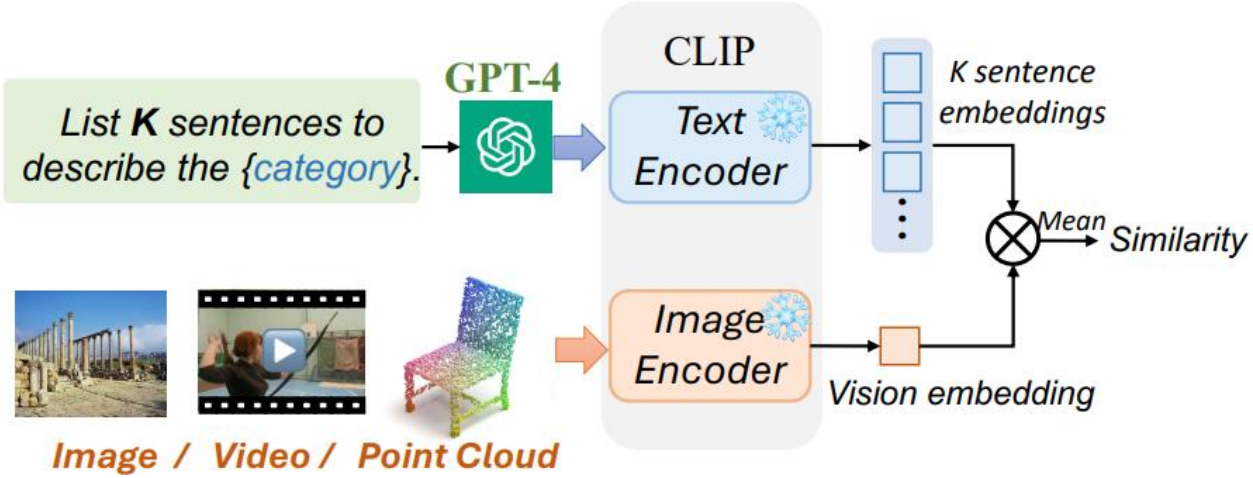
[1] **Yong Li** et al. Decoupled multimodal distilling for emotion recognition, CVPR 2023, Highlight，被引302次

[2] **Yong Li** et al. Hierarchical Distillation of Cross-Modal Knowledge for Robust Emotion Recognition, T-PAMI, 2026

未来研究方向

大模型驱动的情感分析

摘自文献[1]



Dataset Backbone (#Param)	Real-world Affective Faces (RAF-DB) Baseline	 GPT Prompts	 Top-1 Δ
CLIP ViT-B/32 (88M)	22.4 / 76.6	45.8 / 90.6	+23.4
CLIP ViT-B/16 (86M)	27.5 / 69.1	54.4 / 94.4	+26.9
CLIP ViT-L/14 (304M)	26.1 / 72.1	47.2 / 92.0	+21.1
EVA ViT-E/14 (4.4B)	31.0 / 90.9	54.9 / 93.7	+23.9
GPT-4V	68.7 / 93.8		

GPT-4V评测结论：

- AU识别精度高，评测来源单一
- 离散表情识别率低，采用Chain Of Thoughts有所提升
- 复合表情识别率一般

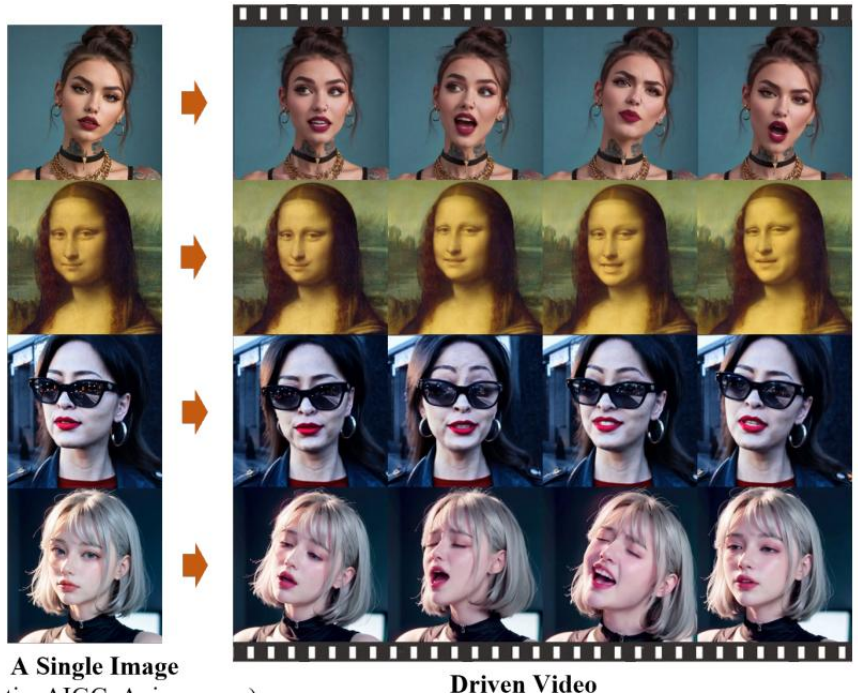
GPT-4V评测结论：

- 离散表情识别距离监督方法有很大差距

[1] GPT as Psychologist? Preliminary Evaluations for GPT-4V on Visual Affective Computing, *CVPRW 2024*
[2] GPT4Vis: What Can GPT-4 Do for Zero-shot Visual Recognition? *Arxiv, 2023*

未来研究方向

数字人表情克隆/生成



Emo (ECCV 2024)[1] 系统评估了人脸表情的真实性（新提出E-FID指标，无法反映表情的时序自然性

FantasyTalking[2] 分阶段学习“音频”与“视频”的对齐，通过手动区域抠图的方式控制唇部运动与音频对齐，基于序列人脸关键点方差控制表情

与**字节跳动**合作研发数字人

[1] EMO: Emote Portrait Alive Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions, *ECCV 2024*.
[2] FantasyTalking: Realistic Talking Portrait Generation via Coherent Motion Synthesis, *Arxiv 2025*.

未来研究方向

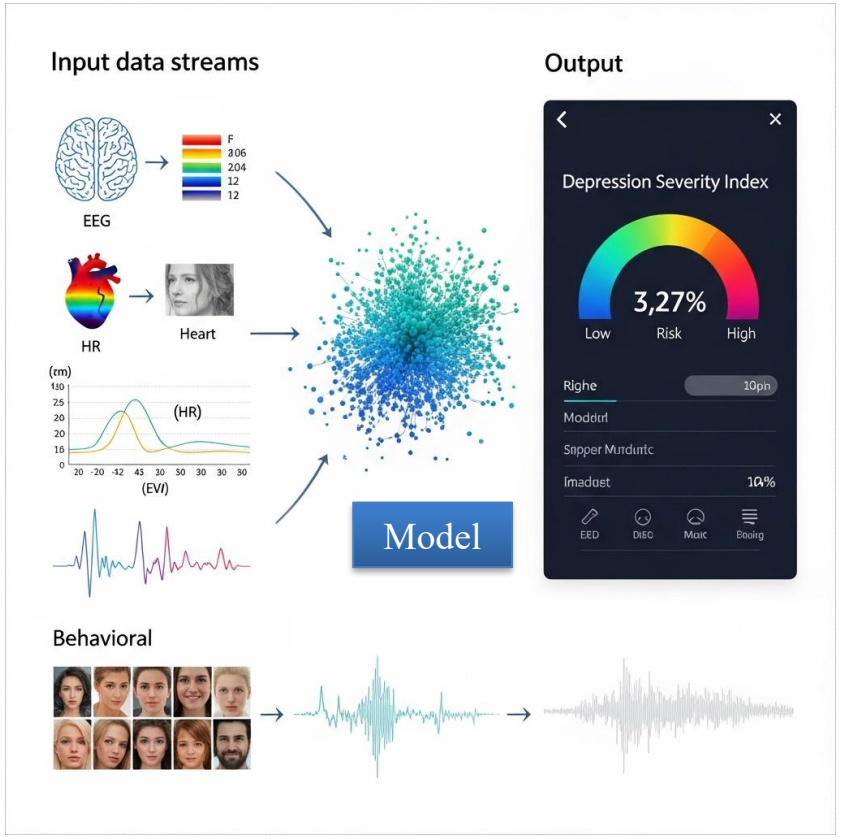
- **新的方法论**
  - 上下文和先验知识建模
    - 上下文信息，如会话和社会环境，会明显影响用户的情感体验。
    - 用户的先验知识，如个性和年龄，也与情感感知相关。
  - 从未标记的、不可靠的、不匹配的情感信号中学习
    - 探索先进的机器学习技术，如自监督表示学习、动态数据选择和平衡、领域自适应、嵌入情感的特殊属性
  - **行为和生理模态的紧密耦合**
    - 抑郁症、**焦虑症**的检测及后续个性化音乐治疗
  - 可信、可解释的情感克隆/生成
    - 表情/Gaze与其他信息（ID、姿态、年龄等）充分解耦
  - **心智世界模型：面向心理感知的多模态大模型**

未来研究方向

- **行为和生理模态的紧密耦合**
  - 抑郁症、焦虑症的检测及后续个性化音乐治疗
- **心智世界模型：面向心理感知的多模态大模型**

生理模态

行为模态



精神类疾病**无感诊断**及**免药物干预**

个性化音乐治疗

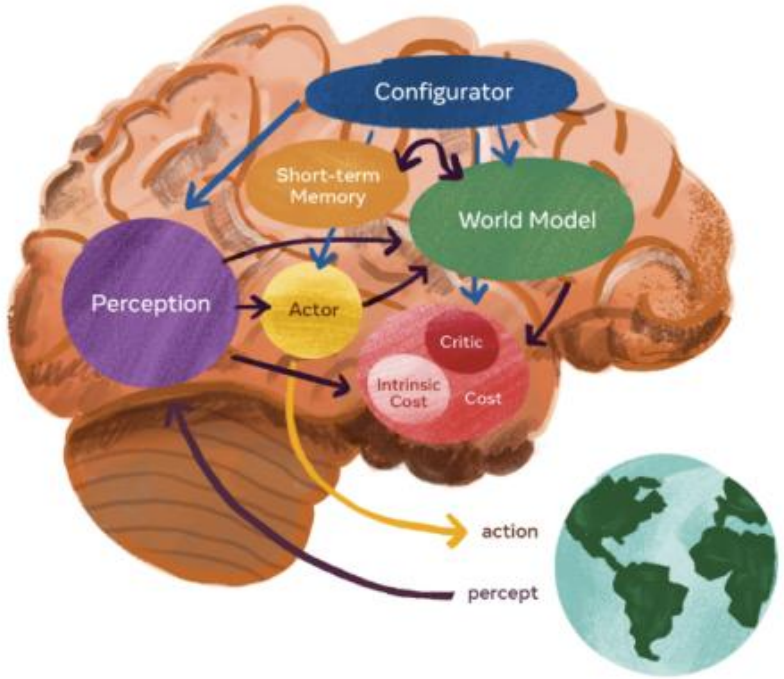


Figure 2 A modular system architecture for autonomous intelligence (LeCun, 2022)