Projet 4 : Vision-LLM pour la Reconnaissance Faciale des Émotions et des Expressions Composées (FER-CE)

1. Contexte et Motivation

La reconnaissance d'expressions faciales (FER – *Facial Emotion Recognition*) constitue l'un des domaines phares de la vision artificielle. Elle permet d'analyser des signaux non-verbaux essentiels en interaction humain-machine, psychologie comportementale, sécurité, robotique sociale et santé mentale.

Traditionnellement, les modèles FER reposaient sur des architectures CNN et se limitaient aux émotions basiques (colère, joie, tristesse…). Cependant, dans les interactions naturelles, les humains expriment fréquemment des émotions composées (*compound expressions*) telles que :

- happily surprised,
- sadly angry,
- fearfully disgusted,
- ou encore des mélanges subtils difficilement classifiables de manière discrète.

Ces états mixtes sont difficiles à classifier, car ils impliquent des micro-variations de plusieurs Action Units (AUs), souvent peu visibles ou ambiguës. Les approches récentes en IA multimodale, notamment les **Vision-LLM (Large Vision-Language Models)**, offrent une rupture : elles unifient **analyse d'image** et **raisonnement linguistique**, permettant non seulement de prédire une émotion mais aussi de **l'expliquer**, de produire une **description textuelle contextualisée**, et d'interpréter des nuances complexes.

Parallèlement, les modèles Vision-LLM (BLIP-2, LLaVA, Qwen-VL, InternVL…) permettent d'unifier perception visuelle et **raisonnement linguistique**, ce qui ouvre une nouvelle voie :

- **expliquer** les émotions composées,
- **justifier** la prédiction par les indices faciaux,
- **décrire** la nuance émotionnelle plutôt que la réduire à un label discret.

Ce projet vise à exploiter un Vision-LLM pour analyser, classifier et interpréter les émotions composées.

2. Jeux de Données Utilisés

Le dataset RAF-CE, disponible sur le site officiel (http://whdeng.cn/RAF/model4.html), consiste en :

- Images faciales en conditions réelles, annotées selon 14 catégories d'expressions composées.
- Annotations AU pour chaque visage, permettant d'explorer les corrélations entre mouvement musculaire et expression composée.

3. Méthodologie

3.1. Pipeline Vision-LLM

On peut conceptualiser le pipeline en trois couches successives :

1. **Couche 1 : Préparation et alignement des données**

   - Détection et recadrage du visage
   - Normalisation
   - Data augmentation (éclairage, rotations, occlusions légères)
   - Vérification de la distribution des classes

2. **Couche 2 : Entraînement Vision-LLM pour émotions composées**

   Les Vision-LLM combinent :

   - un **encodeur visuel** (CLIP, ViT-L, EVA)
   - un **LLM** (Vicuna, LLaMA, Qwen)
   - un module d'alignement vision-langage (Q-Former, projection, cross-attention)

   Modèles proposés :

   - **BLIP-2** (Zero-shot + fine-tuning LoRA)
   - **LLaVA 1.6 / 1.7**
   - **Qwen-VL 2.0** (fort en micro-expressions)
   - **InternVL 1.5** (vision puissante)

   Objectifs d'apprentissage :

   - Classification des émotions composées (RAF-CE) : Sortie = 1 label parmi les 14 classes composées.
   - Génération d'explications textuelles, comme :
     - "The person seems happily surprised: smiling mouth and raised eyebrows."
     - "The facial cues indicate a mixture of sadness and anger."

   N.B. Vous pouvez utiliser le **Prompt-Engineering Visuel.**

   Pour améliorer la qualité des prédictions et des explications générées, il est possible d'utiliser des prompts visuels guidant explicitement le modèle. Par exemple :
   "Describe the emotional state and explain which facial cues contribute to it (e.g., eyebrows, eyes, mouth, muscle tension)."
   Ce type de prompt renforce la cohérence entre la classification, la justification textuelle et les indices faciaux réellement observables sur l'image.

3. **Couche 3 : Interprétation Multimodale (Vision + Langage)**

   Interprétation visuelle

   - Grad-CAM sur l'encodeur visuel
   - Heatmaps des AUs implicites (bouche, yeux, sourcils)

Interprétation linguistique

- analyse des phrases générées
- cohérence entre la justification et les régions détectées
- extraction automatique des indices faciaux mentionnés

Alignement image ↔ concept émotionnel

Validation que les zones sollicitées (sourcils, commissures, orbicularis oculi…) correspondent réellement à la combinaison émotionnelle.

3.2. Modèles Implémentés

Baselines Vision Only

- ResNet
- ViT
- Swin Transformer

Vision-LLM : (au choix : un ou plusieurs)

- BLIP-2
- LLaVA
- Qwen-VL
- InternVL
- CLIP

Métriques

- Accuracy
- F1-Score macro (indispensable car classes déséquilibrées)
- Confusion matrix
- Score textuel : BLEU / ROUGE (qualité des explications)
- Optionnel :
  - **Faithfulness Score** pour mesurer l'alignemet entre la justification et la heatmap
  - ou **CLIPScore** pour la validation de la coherence entre vision et texte

3.3. Contributions

1. Pipeline complet alignant imagerie & texte.
2. Benchmark des modèles Vision-Only vs Vision-LLM.
3. Interprétation XAI pour relier les zones faciales avec les explications verbales.

3.4. Livrables

- Code propre (Kaggle, Colab ou repo GitHub)
- Rapport scientifique
- Visualisations XAI et explications
- Comparaison des modèles
- Optionnel : Interface Streamlit "Upload → Emotion Composée + Explication"