



جامعة مولاي إسماعيل  
 UNIVERSITÉ MOULAY ISMAÏL



كلية العلوم  
FACULTÉ DES SCIENCES

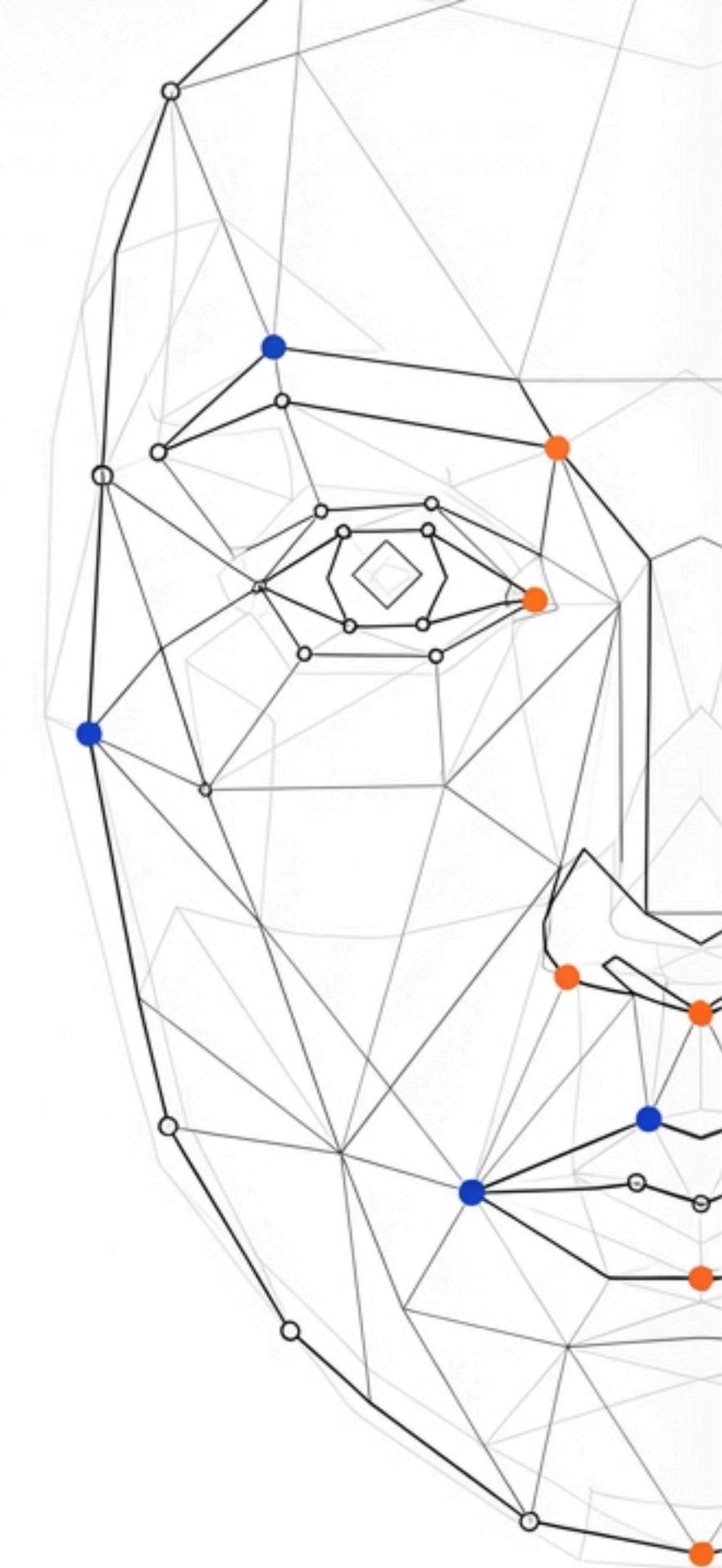
# Entraînement de reconnaissance faciale multi-modèle sur LFW

---

Sujet : Évaluation et optimisation d'architectures pour l'inférence vidéo

Présenté par : Zouitni Salah Eddine, Chergui Yassir

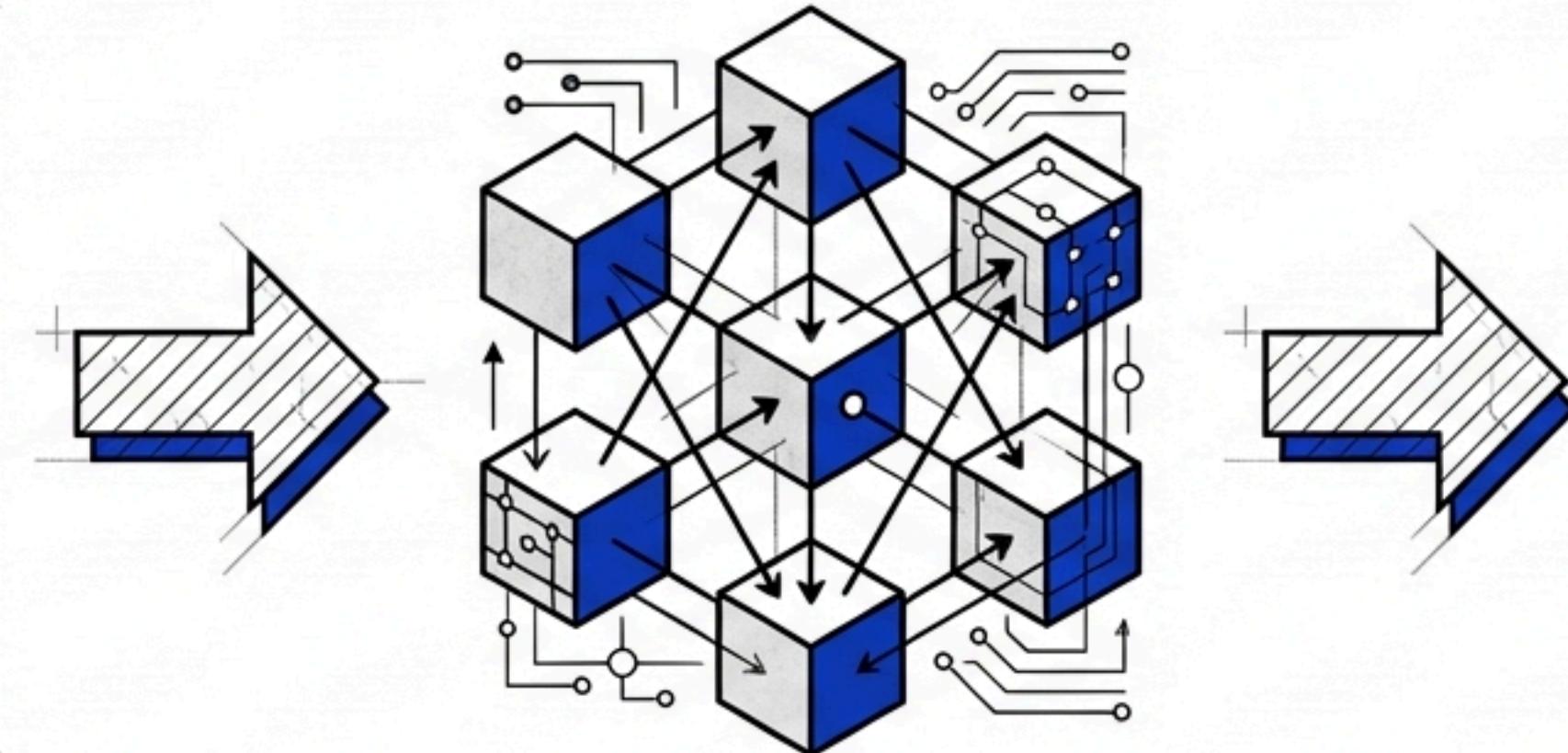
Supervisé par : Pr Ismaili Alaoui El Mehdi



# Objectifs et Tâche : Vers une Inférence Robuste



Labeled Faces in the Wild (LFW)



Neural Network Layer  
(Backbone + Head)



Modèle Optimal

## Tâche Principale

Entraîner des modèles multi-classes pour l'identification faciale.

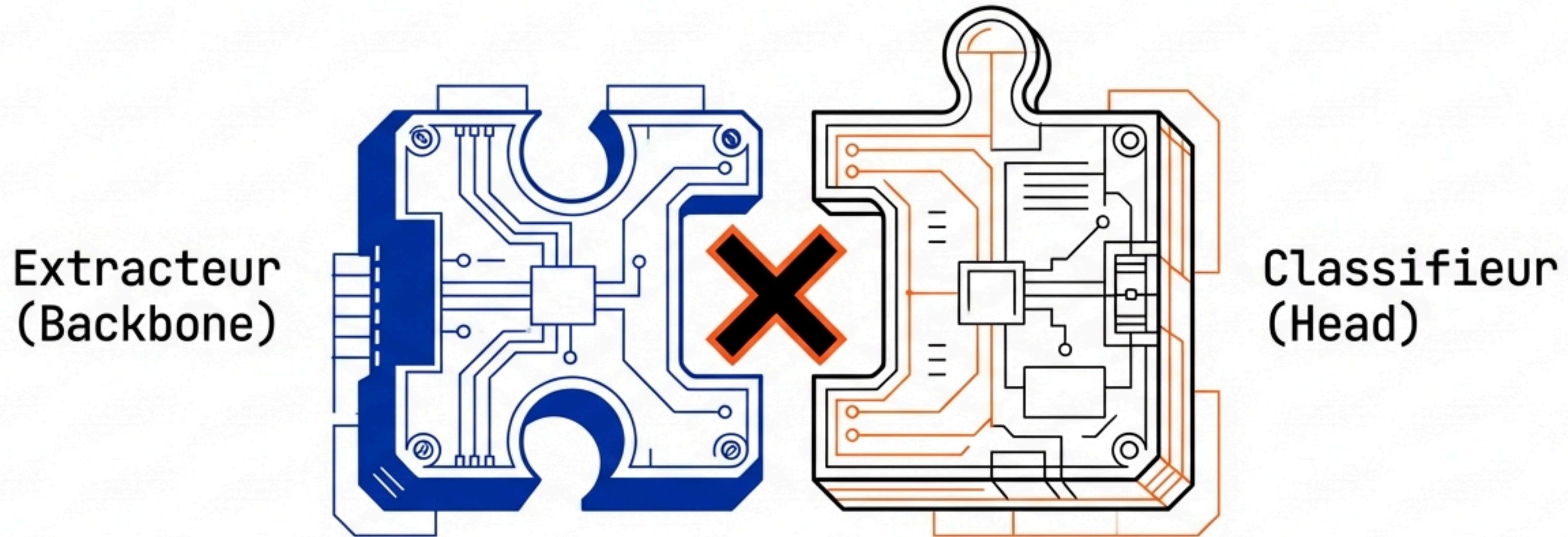
## Le Défi

Identifier la meilleure architecture (Backbone + Head) pour la vidéo.

## Contrainte

Équilibre entre fluidité temps réel et précision.

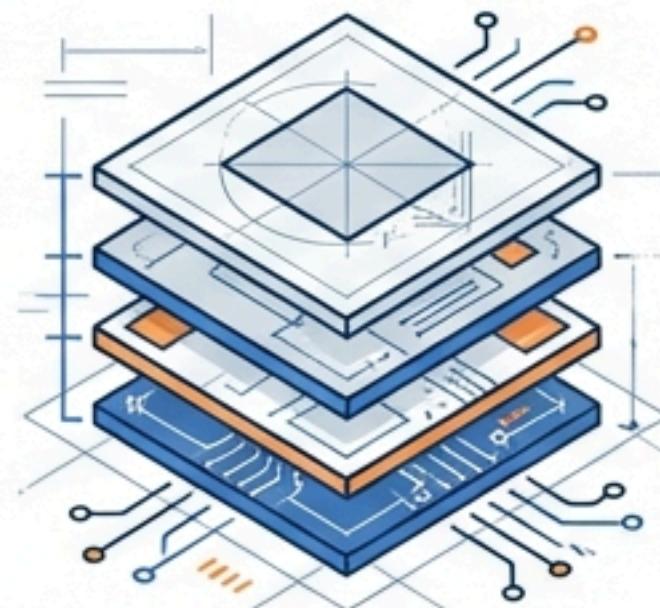
# Stratégie d'Approche : La Modularité



Exploration du produit croisé (Cross-Product)  
pour isoler la performance des composants.

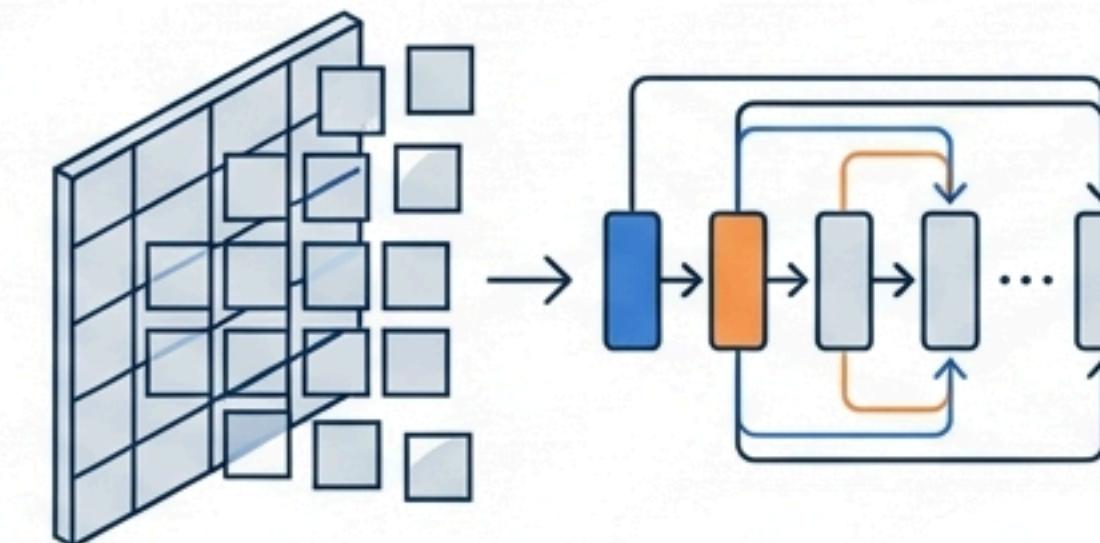
# Backbones : CNNs vs. Vision Transformers

## Les Réseaux Convolutifs (CNNs)



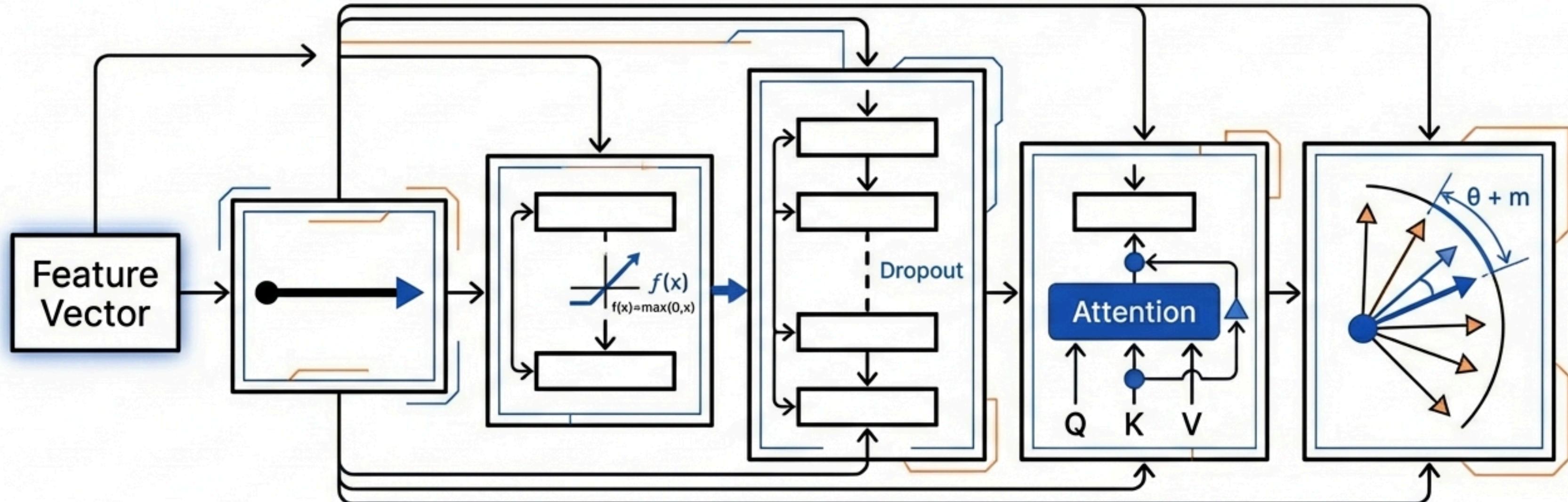
- ResNet (18, 34, 50)
- EfficientNet B0
- MobileNetV3
- ConvNeXt Tiny

## Les Transformers (L'Attention)



- ViT (Vision Transformer)
- Swin Transformer

# Têtes de Classification (Heads)



**Simple**

Couche linéaire  
unique

**MLP**

Activation  
ReLU

**Deep**

Dropout +  
Profondeur

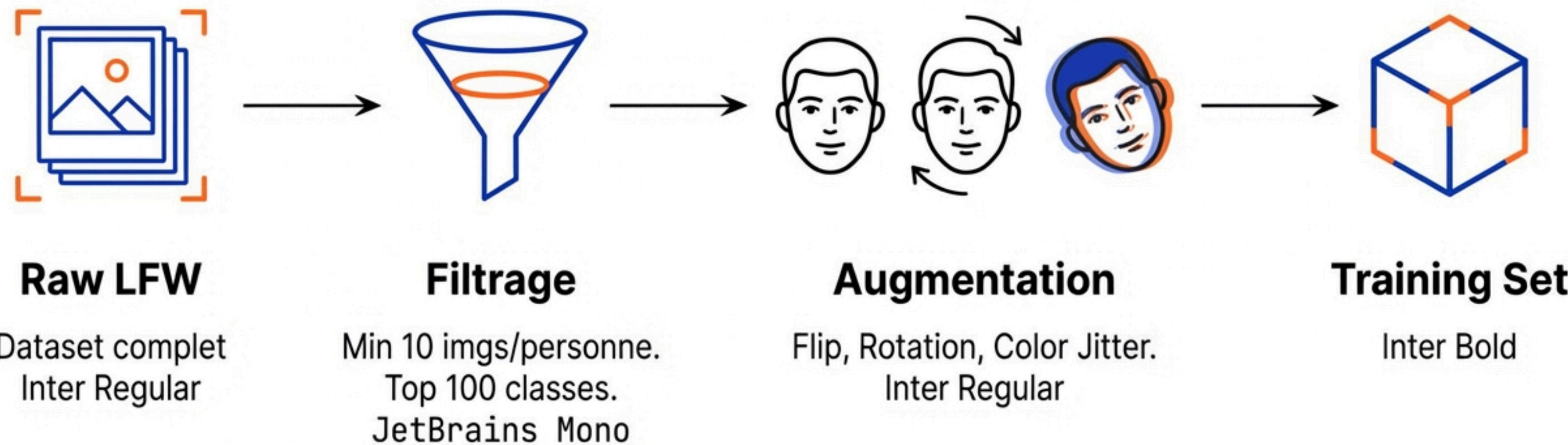
**Attention**

Self-Attention  
Pooling

**CosFace**

Marge et  
Séparabilité

# Méthodologie : Préparation des Données



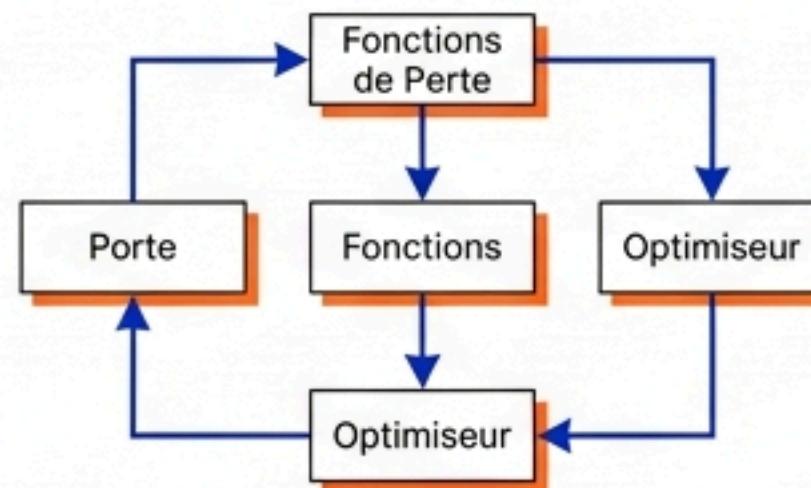
# Configuration de l'Entraînement

## Optimisation & Perte

Fonctions de Perte :

CrossEntropyLoss,  
CosineEmbeddingLoss

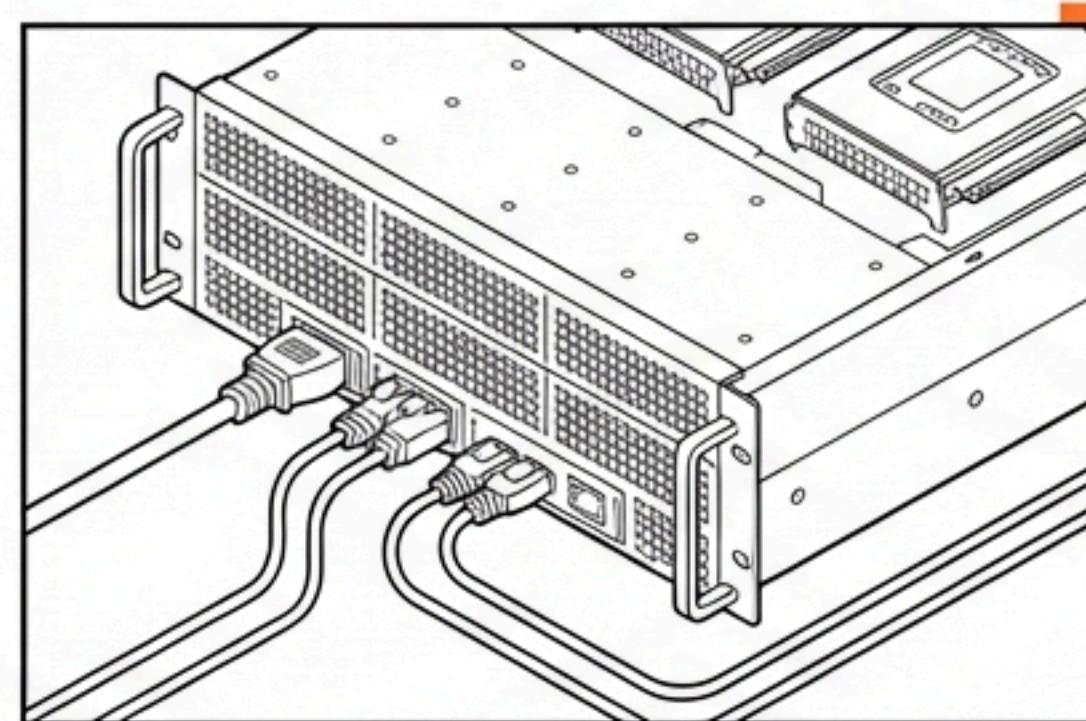
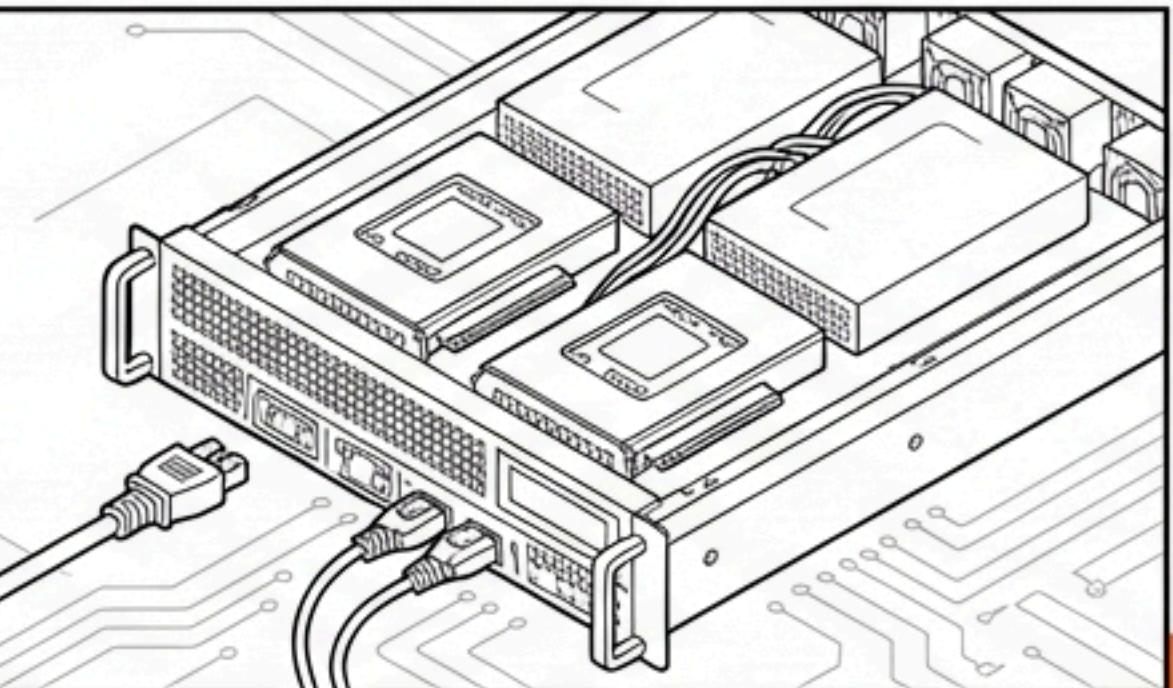
Optimiseur : AdamW avec  
Weight Decay



## Matériel (Hardware)

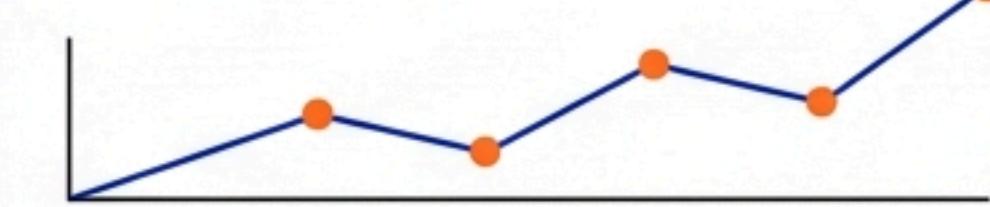
GPU : 2x NVIDIA T4

Parallélisme : DataParallel

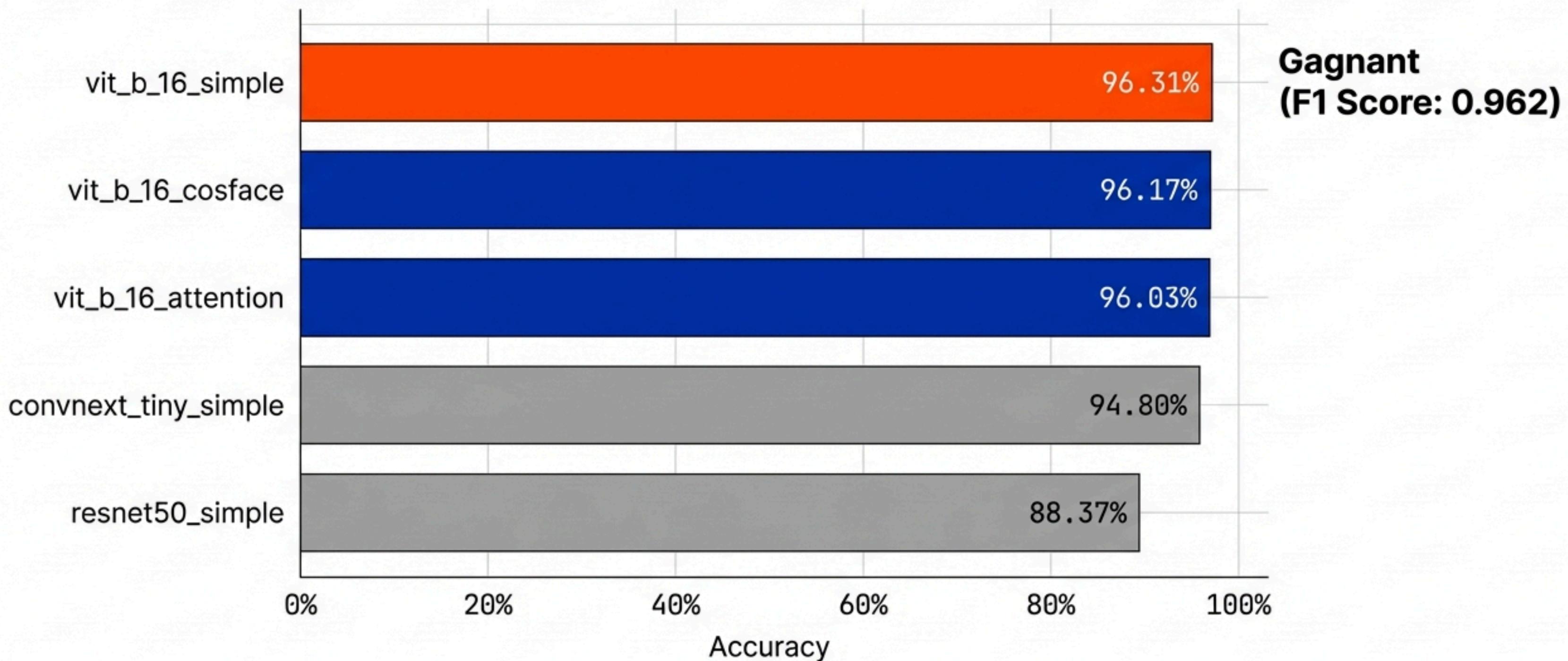


## Métriques

Accuracy, F1-Score, Temps d'inférence



# Résultats Quantitatifs : Le Classement



# Analyse Comparative : ViT vs. CNN

**96.31%**

ViT Best Accuracy



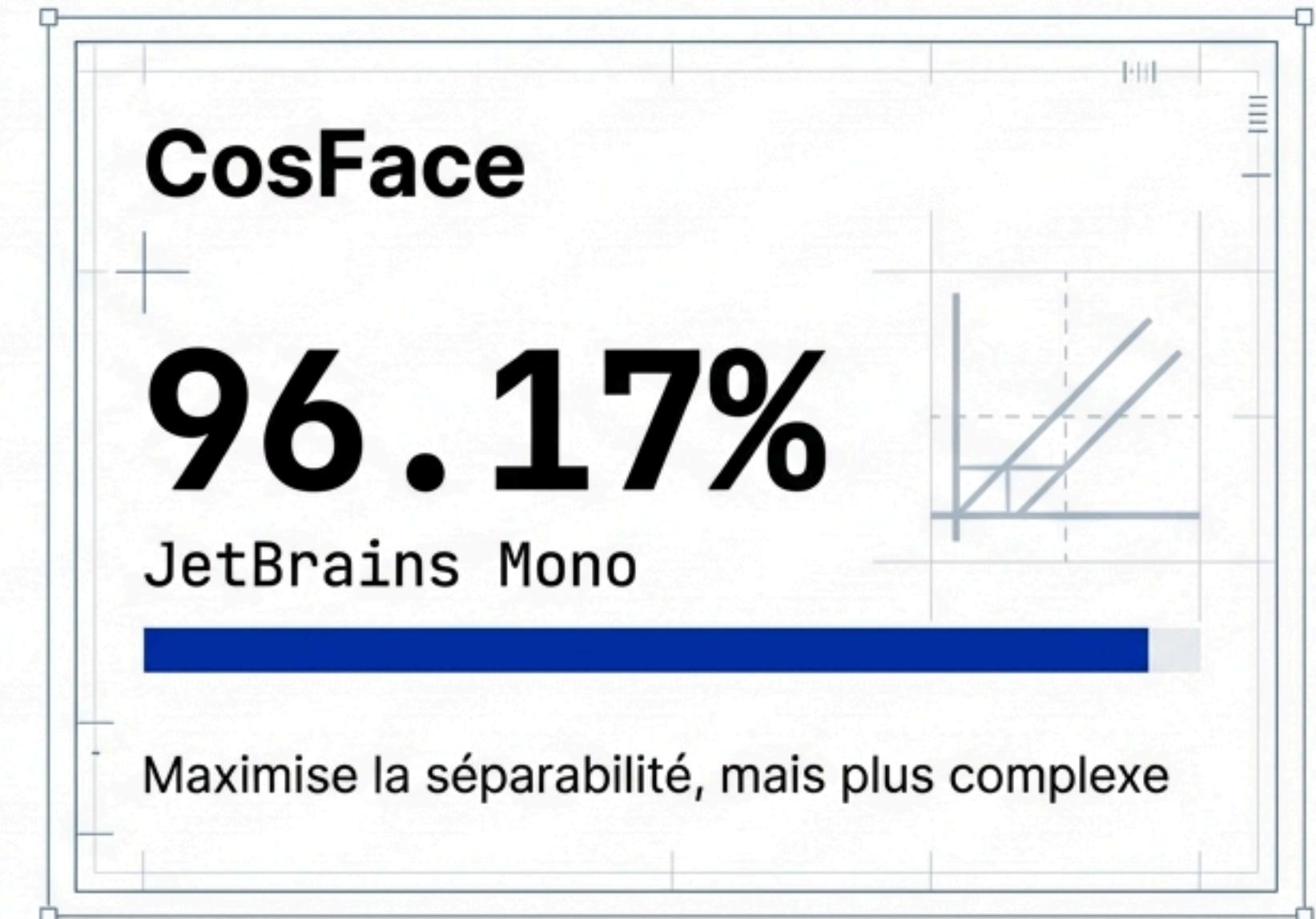
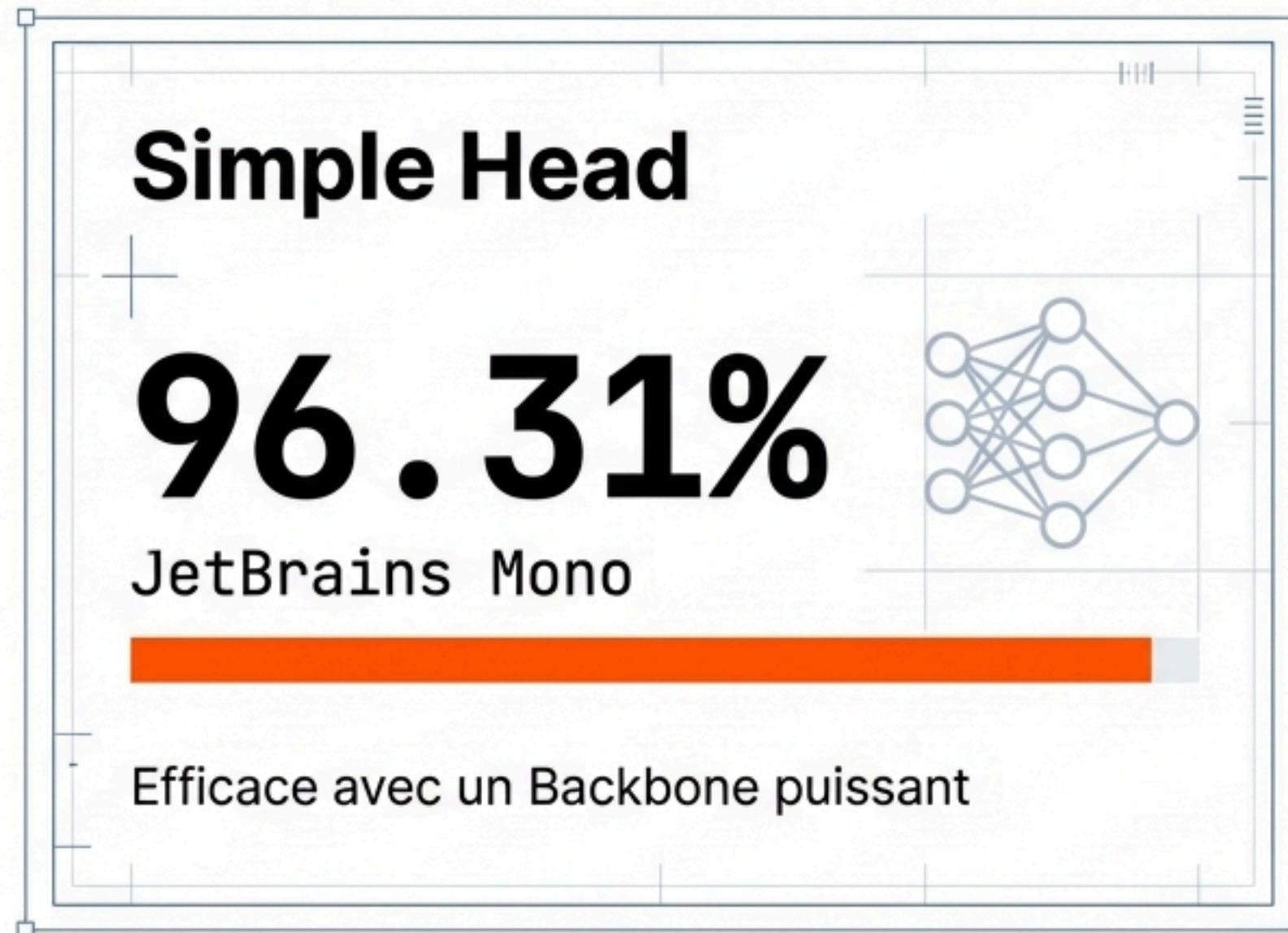
**88.37%**

ResNet50 Baseline

**Le Constat :** Les modèles Vision Transformer surpassent systématiquement les CNNs.

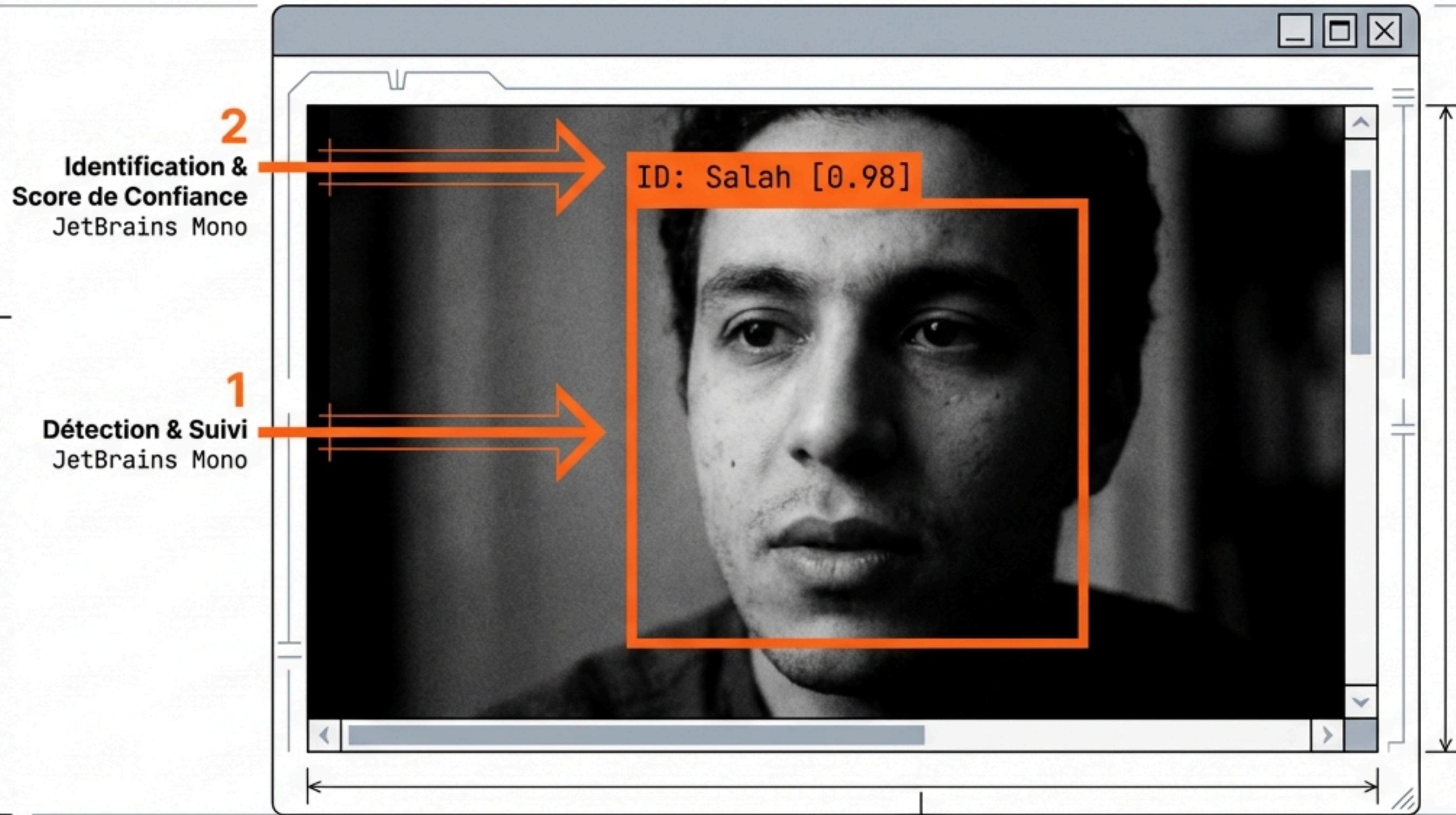
**Interprétation :** Les Transformers capturent des relations globales plus robustes pour l'identification faciale que les convolutions locales.

# Impact des Têtes de Classification



Observation : La puissance du Backbone (ViT) rend les têtes complexes moins critiques.

# Démonstration en Temps Réel : Configuration



Scénario :  
Inférence vidéo  
sur le modèle  
vit\_b\_16\_simple.

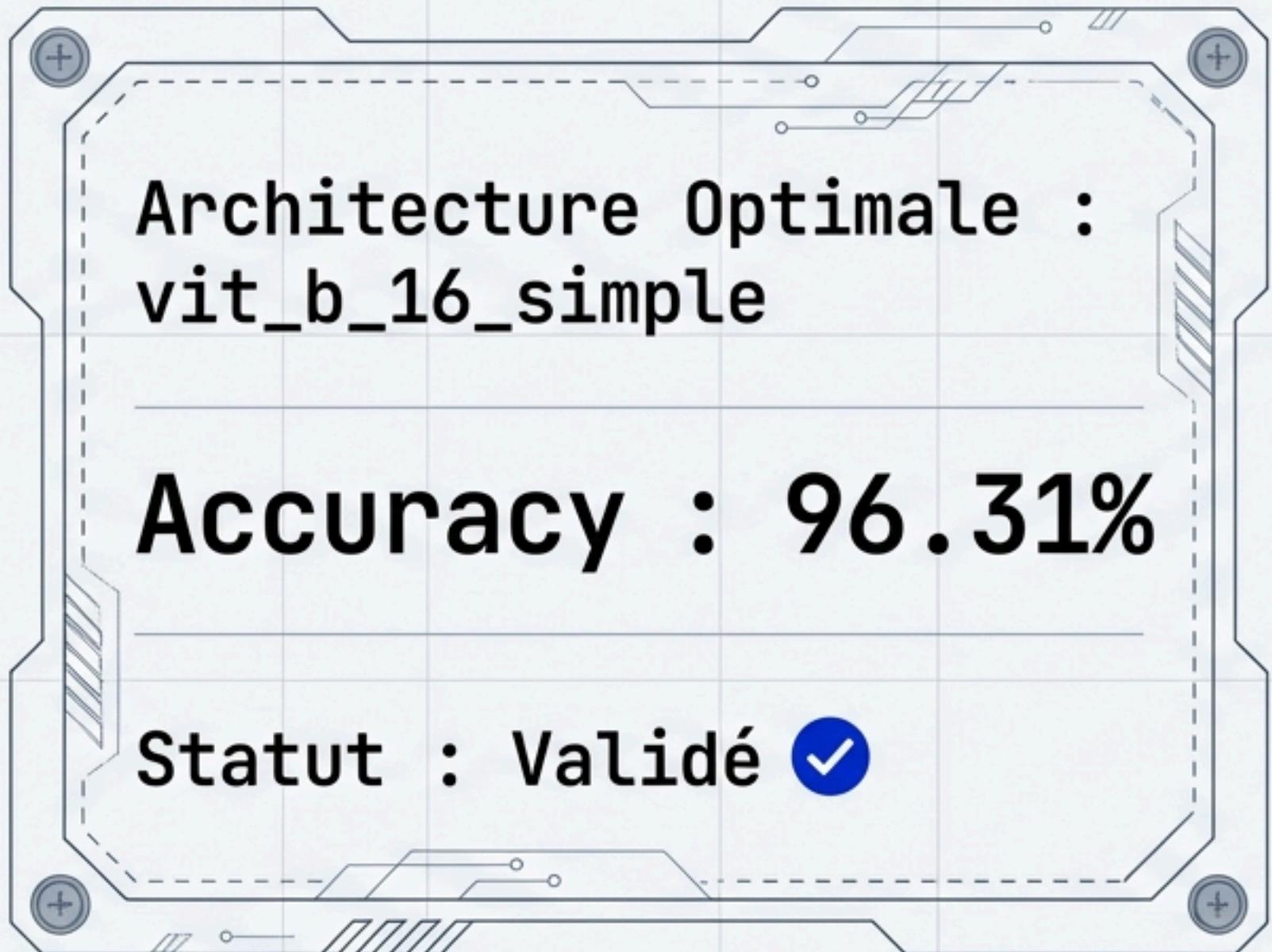


# Démonstration Live

Inférence du modèle vit\_b\_16\_simple

Inférence du modèle vit\_b\_16\_simple

# Conclusion et Synthèse



- L'approche multi-modèle a démontré la supériorité des Transformers (ViT) sur LFW.
- Robustesse et précision accrues par rapport aux CNNs traditionnels.
- Objectif atteint : Solution déployable pour l'inférence vidéo.