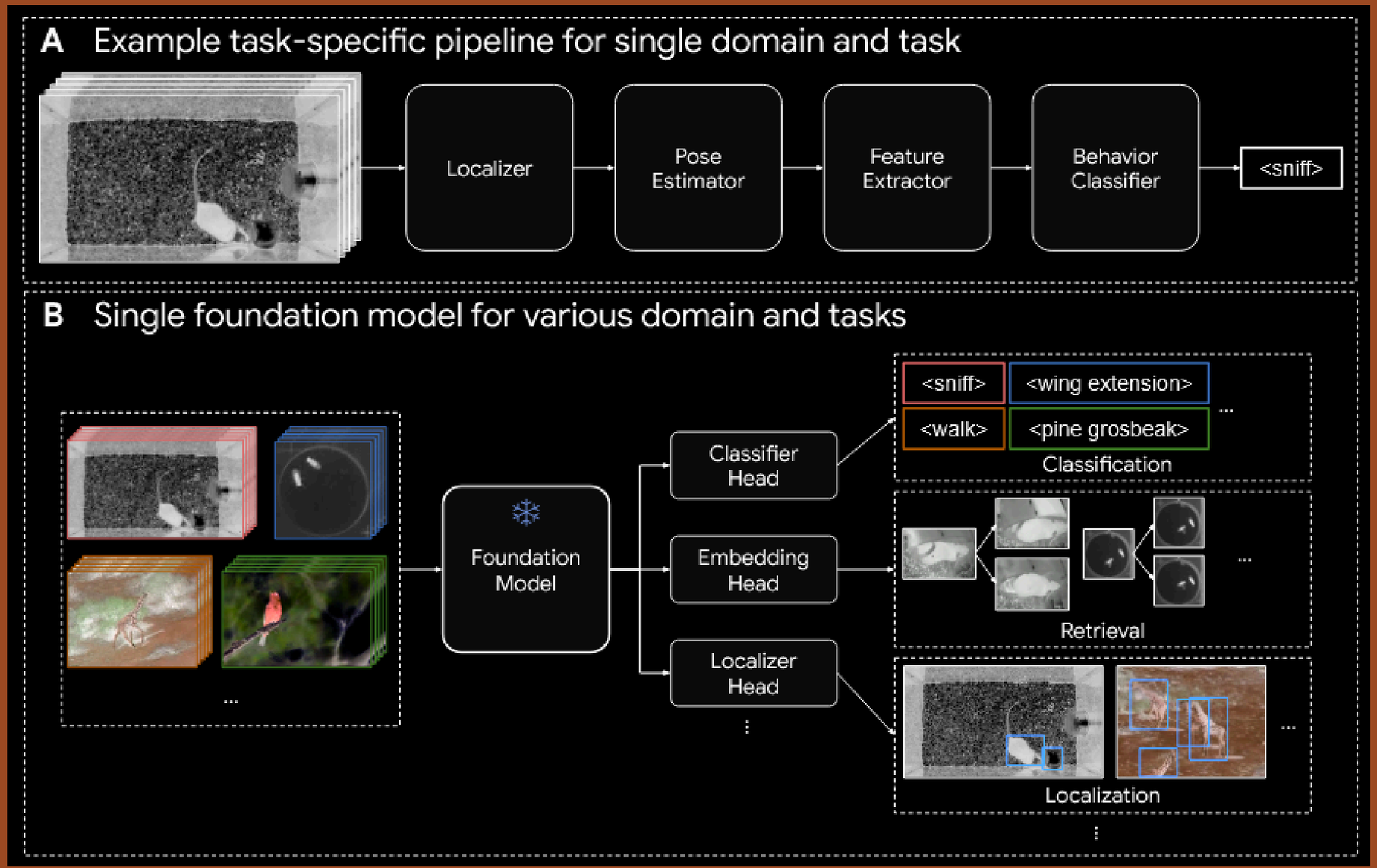
Video Foundation Models for Animal Behavior Analysis

Jennifer J. Sun, Hao Zhou, Long Zhao, Liangzhe Yuan, Bryan Seybold, David Hendon, Florian Schroff, David A. Ross, Hartwig Adam, Bo Hu, Ting Liu

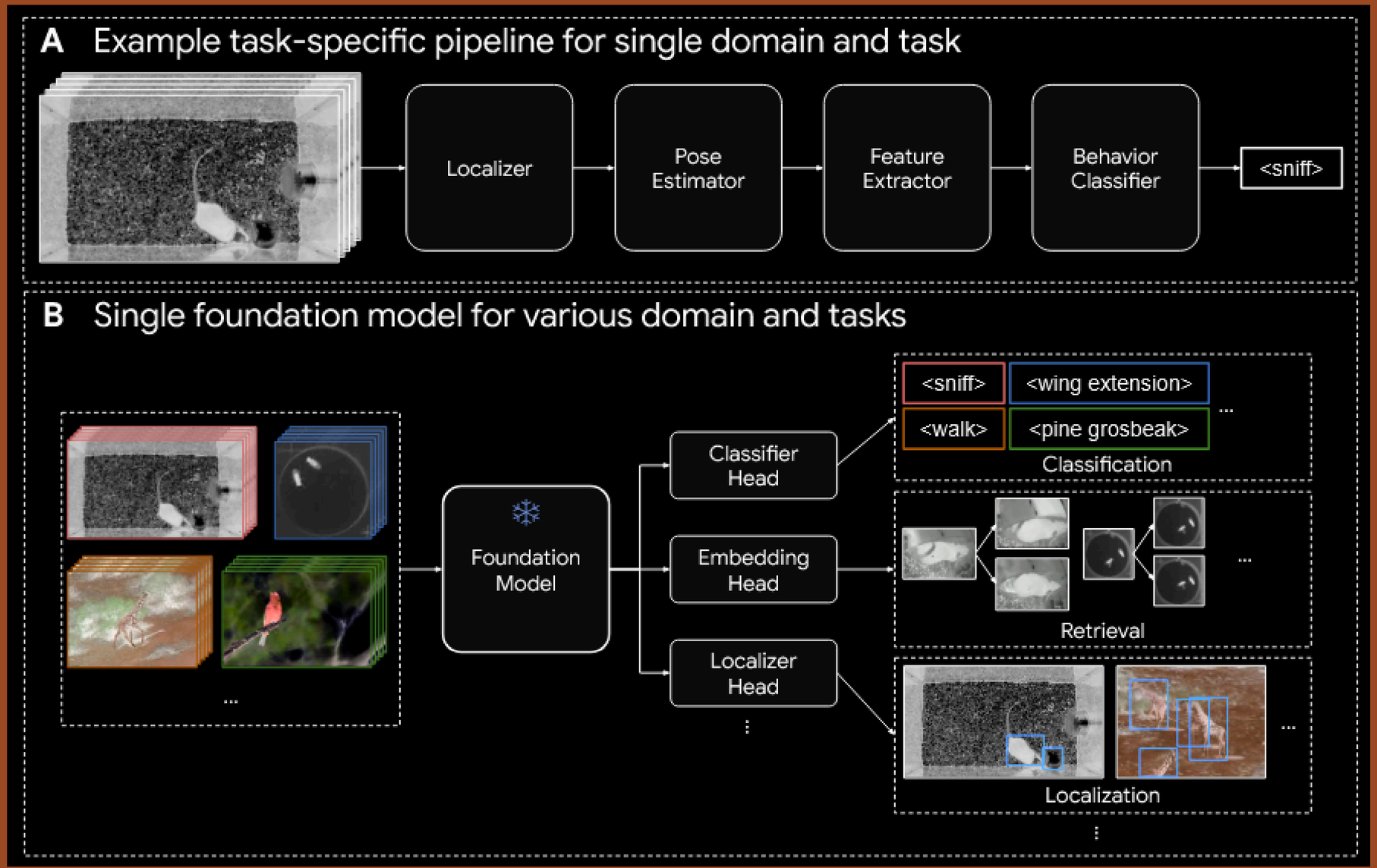
Foundation models



- Pré-entrainés sur des datasets énormes
- Apprentissage de features généraux
- Pas besoin de les adapter

En apprentissage, le modèle a déjà été '*pré-entrainé*' et est *frozen* : les poids ne changent pas, on entraine uniquement une tête (de classification, de localisation, etc...) et on utilise le modèle en tant que '*backbone*'. Il crée l'espace latent, mais d'une certaine manière 'ne sait pas quoi en faire'.

Foundation models



Avantages:

Généralisation forte
Moins d'annotations expertes nécessaires
Moins de fine-tuning nécessaire

DONC, plus efficace, moins de temps de calcul nécessaire qu'un modèle spécialisé entrainé uniquement sur une tâche précise, pour un type de dataset précis

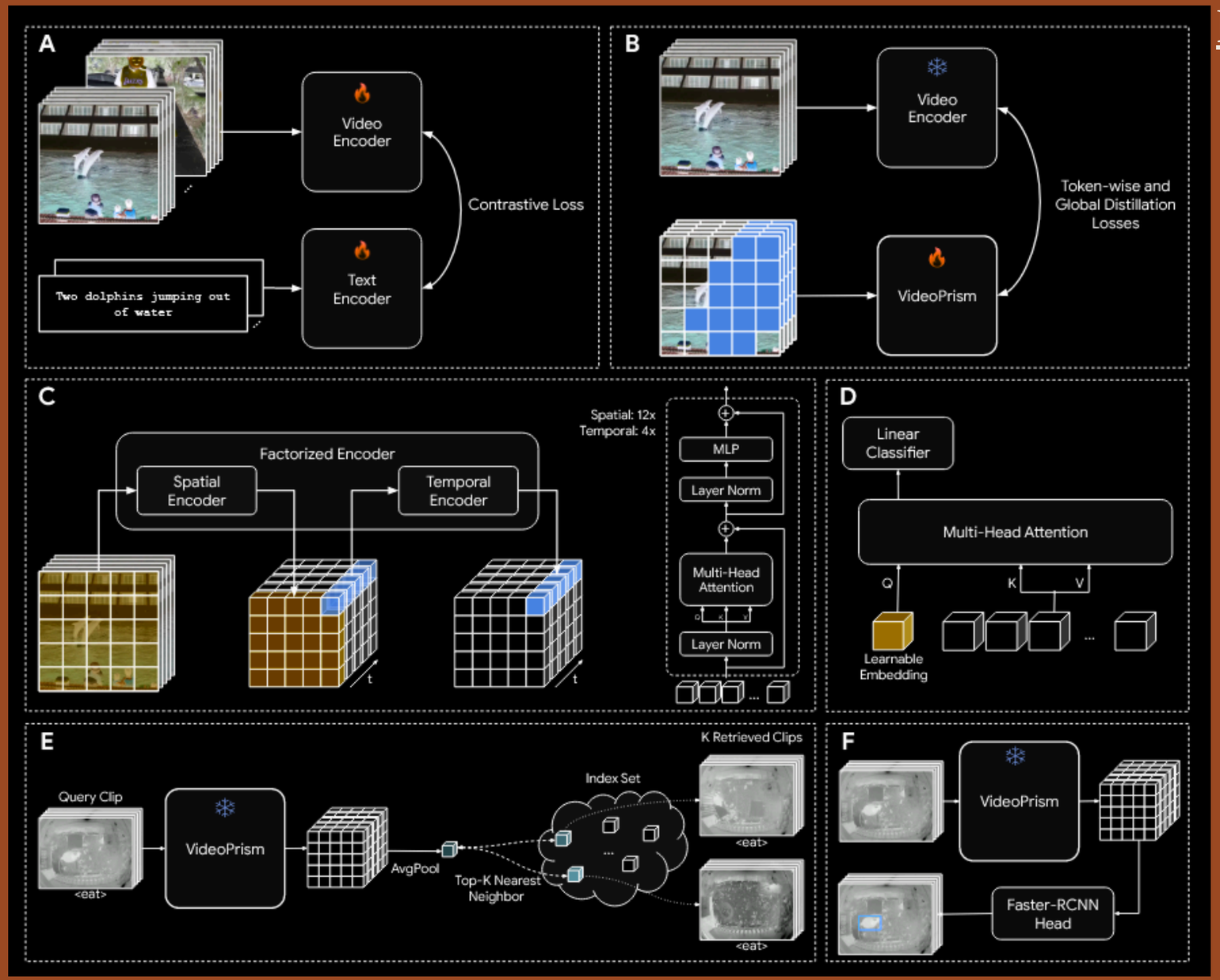
Modèle testé: VideoPrism

Le but de l'article est de démontrer l'intérêt des foundations models vidéos pour l'utilisation en écologie, éthologie et tout domaine de la biologie relatif aux animaux. Pour cela, un foundation model a été choisi : VideoPRISM.

C'est un modèle vidéo généraliste entraîné sur ~618 M clips web avec une petite partie annotée. Toutes les vidéos choisies sont relatives aux animaux. Il s'agit d'une architecture Vision Transformers, avec un encodeur spatio-temporel 'factorisé'.

Au final, on obtient des features spatio-temporels

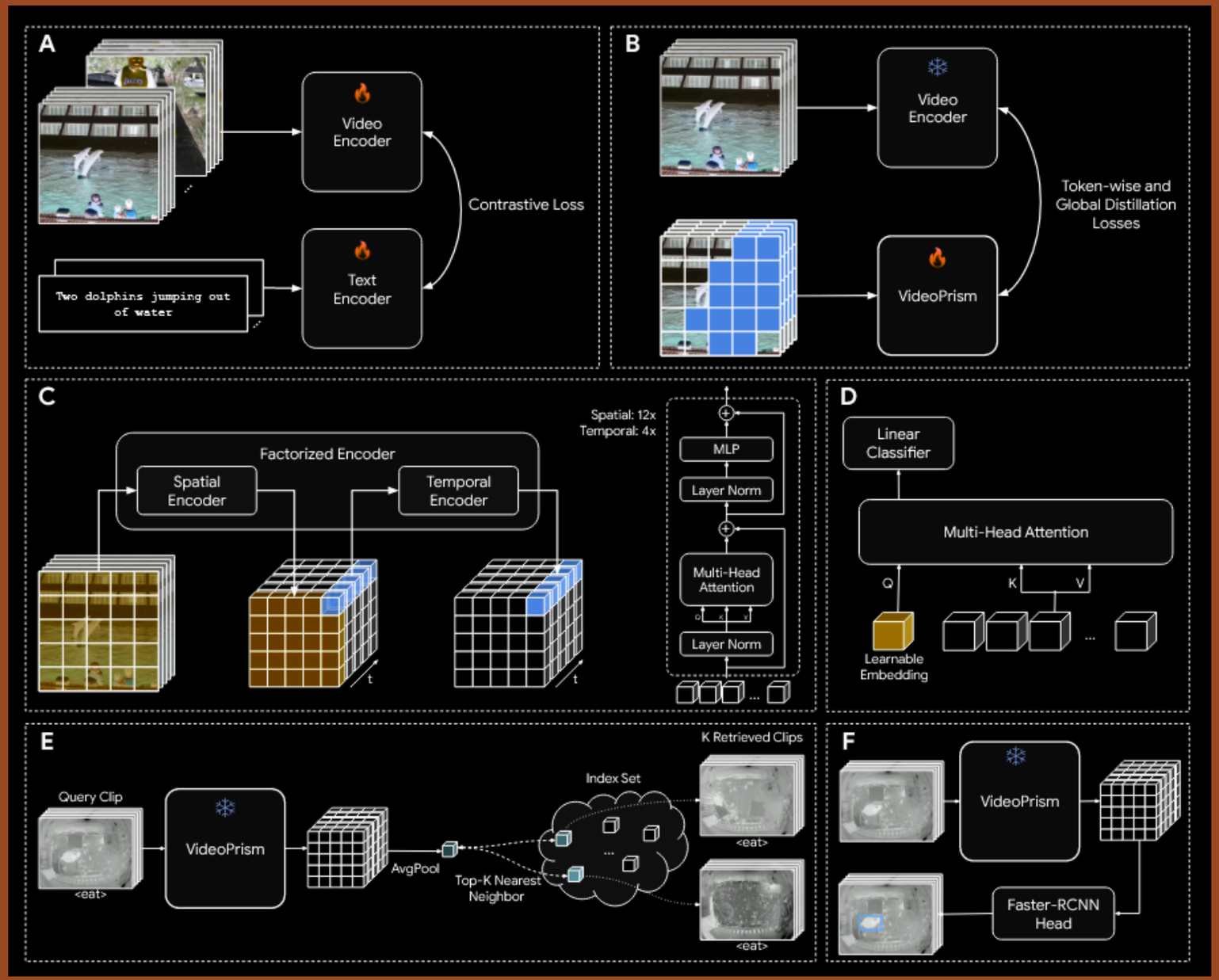
Modèle testé: VideoPrism



Pré-entrainement en 3 parties:

1. contraste vidéo-texte (A): on prend uniquement les vidéos annotées et on cherche à créer un encodeur qui crée +- une même représentation pour le texte et pour l'image pris séparément
2. Factorisation temporelle (B): On entraine le modèle sur tous les clips, avec une partie de l'encodage sur lla partie spatial (on prend sur une frame tous les patchs) et une autre sur le temporel (on suit 1 patch au fil du temps). Une partie de l'image est masquée
3. Distillation (C): le 1er modèle est le teacher et le deuxième le student. On force le modèle spatio-temporel à avoir un espace latent qui prend en compte la sémantique, c'est à dire qui ressemble à celui du 1er modèle

Modèle testé: VideoPrism



Pour l'utilisation de ce modèle, on peut soit :

- Ajouter une tête de classification, figer les poids du backbone (foundation model) et entrainé rapidement le modèle
- Prendre directement les features pour faire des k-neighbours
- Ajouter une tête permettant de trouver des bounding boxes à partir de l'attention du vision transformer

Tâches évaluées

Ici, VideoPrism est évaluéquasi-uniquement sur des datasets de souris. Il est comparé :
1- à CLIP, le meilleur modèle généraliste trouvé (entrainé sur des vidéos de tout types, pas uniquement des animaux contrairement à VideoPrism)
2- aux modèles spécifiquement entrainés sur la tâche évaluée

4 évaluations ont été réalisées:

- Classification de comportement animal
- Classification “few-shot”: apprentissage de la tête avec uniquement 10 à 100 images/classe
- Retrieval: recherche de comportements rares par nearest-neighbour pour rendre la recherche de ces comportements sur la vidéo plus rapide (par ex: recherche d’un comportement qui n’arrive qu’une fois par jour pendant 5s sur une vidéo de 24h)
- Localisation (bounding boxes)
- Généralisations à d’autres types d’animaux (mouches, girafes, oiseaux...)

Tâches évaluées

Voici les datasets et les métriques utilisées pour évaluer les modèles:

Results Section	Dataset	Metric	# Annotations (Train/Test)
Behavior classification	Calico [55]	mAP	1,253 / 400 clips
	CalMS21 [18]	mAP	507,660 / 262,107 frames
	CRIM13-Top [9]	mAP	1,630,114 / 1,994,885 frames
	CRIM13-Side [9]	mAP	1,612,327 / 1,892,817 frames
Few-shot behavior classification	Calico [55]	mAP	(10/50/100)-shot / 400 clips
	CalMS21 [18]	mAP	(10/50/100)-shot / 262,107 frames
Behavior retrieval	Calico [55]	mean Hit@K	1,253 / 400 clips
	CalMS21 [18]	mean Hit@K	507,660 / 262,107 frames
Localization	Calico-Detection [55]	AP@IoU	5,460 / 2,124 frames
	MARS-Top [5]	AP@IoU	10,000 / 5,000 frames
	MARS-Side [5]	AP@IoU	10,000 / 5,000 frames
Broader Applications	Fly vs. Fly [19]	mAP	1,067,329 / 322,393 frames
	KABR [21]	Macro-Acc.	1,545,513 / 290,021 frames
	SSW60 [20]	Accuracy	3,462 / 1,938 videos

Quelques métriques:

- Hit@K est utilisé pour évaluer l'accuracy du retrival sur K queries
- AP@IoU : average precision sur Intersection Over

Tâches évaluées

Sur toutes les tâches, VideoPrism était toujours a minima aussi bon que CLIP et les modèles spécifiques, et très souvent largement meilleur.

Classification

Method	Calico	CalMS21	CRIM13 Side	CRIM13 Top
Specialized models	52.6	88.9	-	-
CLIP	51.8	62.7	34.0	14.1
VideoPrism	71.2	91.1	64.5	64.9

Classification en few-shot

Dataset	Method	10-shot	50-shot	100-shot
Calico	CLIP	39.8 ± 2.1	42.6 ± 1.0	47.5 ± 0.1
	VideoPrism	50.5 ± 3.9	66.3 ± 0.9	60.2 ± 0.5
CalMS21	CLIP	27.5 ± 4.7	42.0 ± 1.8	49.1 ± 4.0
	VideoPrism	55.5 ± 3.1	69.8 ± 2.5	75.0 ± 2.2

Retrieval

Dataset	Method	mean Hit@1	mean Hit@5	mean Hit@10
Calico	Random	16.0	47.1	65.8
	CLIP	28.4	58.4	75.6
	VideoPrism	36.4	73.6	85.8
CalMS21	Random	19.7	47.4	62.2
	CLIP	28.5	56.1	69.9
	VideoPrism	42.0	66.2	75.1

Localisation

Dataset	Method	mAP	AP@50IoU	AP@75IoU
Calico-Detection	CLIP	94.0	98.9	97.5
	VideoPrism	95.2	98.9	96.0
MARS-Top	CLIP	58.2	97.9	64.6
	VideoPrism	61.1	98.3	71.2
MARS-Side	CLIP	43.9	91.1	35.3
	VideoPrism	46.3	92.5	40.5

Autres animaux

Method	Fly vs. Fly	SSW60	KABR
Specialized models	88.6	71.9	61.9
CLIP	61.7	48.3	32.2
VideoPrism	89.1	70.1	61.6

Limites et axes d'amélioration

– Classification : erreurs sur comportements rapides (peu de frames) → besoin de fenêtres temporelles plus larges.

– Few-shot : potentiel via active learning (intéraction avec un expert pour la vérification des prédictions).

– Retrieval : sensibilité aux variations lumineuses → nécessité d'augmentations adaptées.

– Coûts computationnels : distillation et architectures efficaces pour réduire mémoire/latence.

– Ouverture future : couplage foundation model + LLM multimodal pour apprentissage in-context et adaptation rapide à de nouveaux comportements. On peut expliquer au llm ce que l'on veut sous forme de texte + potentiellement une vidéo d'exemple et le LLM permet d'optimiser le fine-tuning du foundation model de façon plus simple