



ILLUIN
TECHNOLOGY



CentraleSupélec

PRÉSENTATION INTERNE | Janvier 2025

NLP Course - Vision Language Models

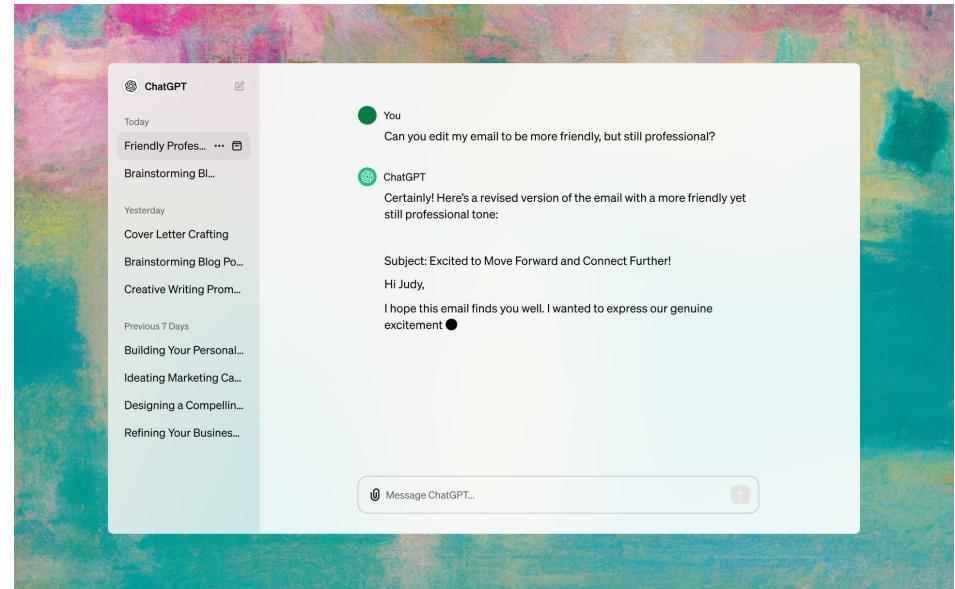
LLM : un des top business use cases aujourd’hui

Les Large Language Models (LLMs) sont devenus incontournables...

Les utilisateurs peuvent interagir de manière naturelle (texte) avec les LLMs qui peuvent les aider de par leurs connaissances et leurs capacités de raisonnement.

... mais sont limités à de la compréhension de texte

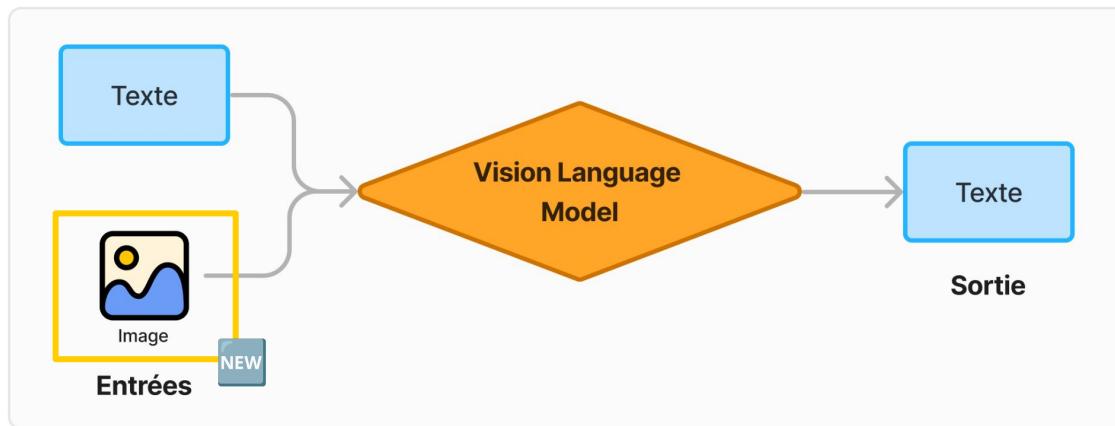
Comment faire lorsqu'on souhaite que notre modèle raisonne sur des images ?



VLM : vers du raisonnement au delà du simple texte

Vision Language Models (VLMs)

Un VLM fonctionne comme un LLM (modèle génératif) mais peut prendre une/des image(s) en entrée en plus du texte. On parle de modèle multimodal.



VLM : quels use cases ? (1/2)

Prompt: Caption this image.



“In this image we can see birds on the wall. In the background, we can see a building with windows and plants in the pots.”

Captioning

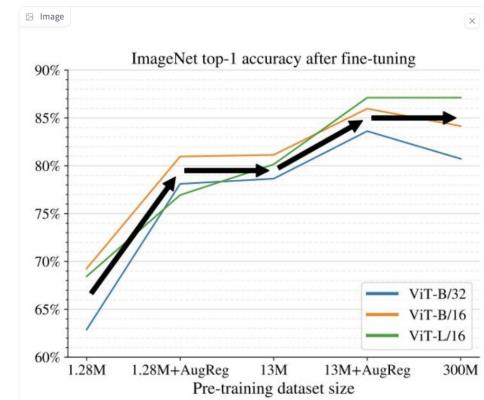
Prompt: Parse this receipt as JSON.



```
{
  "items": [
    {
      "name": "Bintang Bremer",
      "Quantity": 1,
      "Price": 59000
    },
    {
      "name": "Chicken H-H",
      "Quantity": 1,
      "Price": 190000
    },
    {
      "name": "Ades",
      "Quantity": 1,
      "Price": 10000
    }
  ],
  "Subtotal": 259000,
  "Service": 9600,
  "Tax": 52416,
  "Discount": 19000,
  "Total": 302016
}
```

**OCR + Layout
Understanding**

Prompt: What does this image show?



“vit-b/32, vit-1/16, and vit-b/16+augreg are trained on different sizes of training data.”

**Vision Question Answering
& Chart Understanding**

VLM : quels use cases ? (2/2)

Input Prompt



Describe this image in detail:

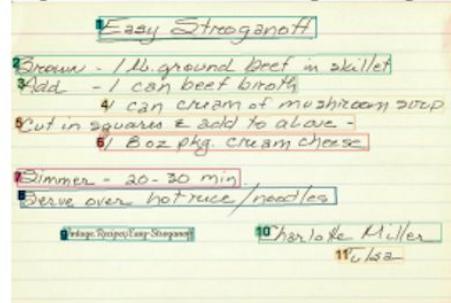
A snowman sits next to a campfire in the snow. He is wearing a hat, scarf, and mittens. There are several pots nearby, likely containing a hot meal for the snowman. The scene is serene and peaceful, with the snowman looking at the fire with a warm smile.



Grounded captioning

À gauche : Kosmos-2: Grounding Multimodal Large Language Models to the World, Peng et al. (2023)
 À droite : Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks, Xiao et al. (2024)

Prompt: What is the text in the image, with regions?



Easy Stroganoff¹
 Brown 1 lb. ground beef in skillet²
 Add 1 can beef broth³
 1 can cream of mushroom soup⁴
 Cut in squares & add to above⁵
 1/2 pkg. cream cheese⁶
 Simmer 20-3 min.⁷
 Serve over hot rice/noodles⁸
 Vintage Recipes/Easy-Stroganoff⁹ Charlotte Miller¹⁰
 Tulsa¹¹

Handwriting Recognition (HTR)





ILLUIN
TECHNOLOGY

Change the code

SOMMAIRE KEYNOTE VLM

01. Préliminaires

02. Concevoir son VLM

03. Exemples de VLMs récents

04. Limitations

05. Next steps



ILLUIN
TECHNOLOGY

Change the code

SOMMAIRE KEYNOTE VLM

01. Préliminaires

02. Concevoir son VLM

03. Exemples de VLMs récents

04. Limitations

05. Next steps

Historique des innovations liées aux VLMs

2020	2021	2022	2023	2024
Vision Transformer (ViT) 1ère utilisation des transformers dans les tâches de vision.	CLIP Classification et la recherche d'images en zéro-shot grâce des embeddings multimodaux texte/image.	Flamingo Un des premiers VLMs modernes. Modèle de DeepMind permettant à un utilisateur de chater sur un contexte texte/image/vidéo.	LLaVA Un des premiers VLMs open-source. Montre qu'il est possible de construire un VLM performant à partir de modèles pré-entraînés.	GPT 4 Vision Une version multimodale de GPT-4, avec des performances impressionnantes à sa sortie.
Innovations préliminaires		Principaux VLMs		



ILLUIN
TECHNOLOGY

Change the code

Vision Transformer (ViT)

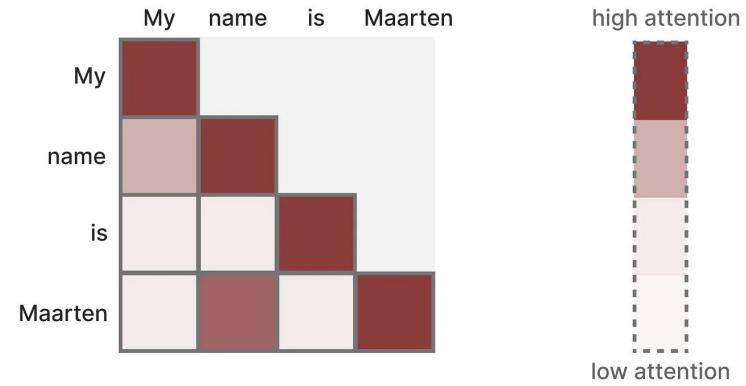


Vision Transformer (ViT)

Pourquoi utiliser les Transformers dans des tâches de Computer Vision ?

Le Transformer, de par sa **self-attention**, peut être entraîné avec de la parallélisation. Ceci fait que le Transformer scale avec le nombre d'exemples dans le train set.

Le Transformer ayant fait ses preuves dans le NLP, on peut imaginer que ces atouts sont généralisables à la Computer Vision.



Calcul de la self-attention en parallèle au training pour un Transformer

Vision Transformer (ViT)

1. Découpage en patches

L'image est divisée en petits patches de taille fixe. Chaque patch est aplati en un vecteur.

2. Embedding des patches

Chaque vecteur de patch est projeté dans un espace de dimension fixe à l'aide d'une projection linéaire.

3. Ajout des embeddings positionnels

Des informations sur la position relative des patches sont ajoutées aux embeddings.

4. Passe dans l'encodeur

Les embeddings enrichis envoyés à un Transformer encoder avec un token **[CLS]**. La représentation vectorielle de l'image correspond au vecteur d'output correspondant à **[CLS]**.

5. Classification finale

La représentation du token **[CLS]** est passée dans une couche entièrement connectée pour prédire les classes.

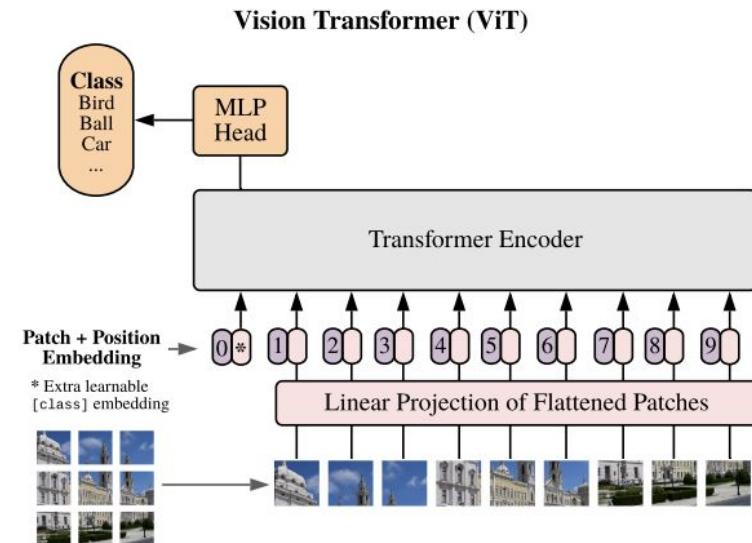


Schéma du fonctionnement du ViT



Vision Transformer (ViT)

Étapes du Vision Transformer (ViT)

1. Découpage en patches
2. Embedding des patches
3. Ajout des embeddings positionnels
4. Passe dans l'encodeur
5. Classification finale



An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, Dosovitskiy et al. (2020)



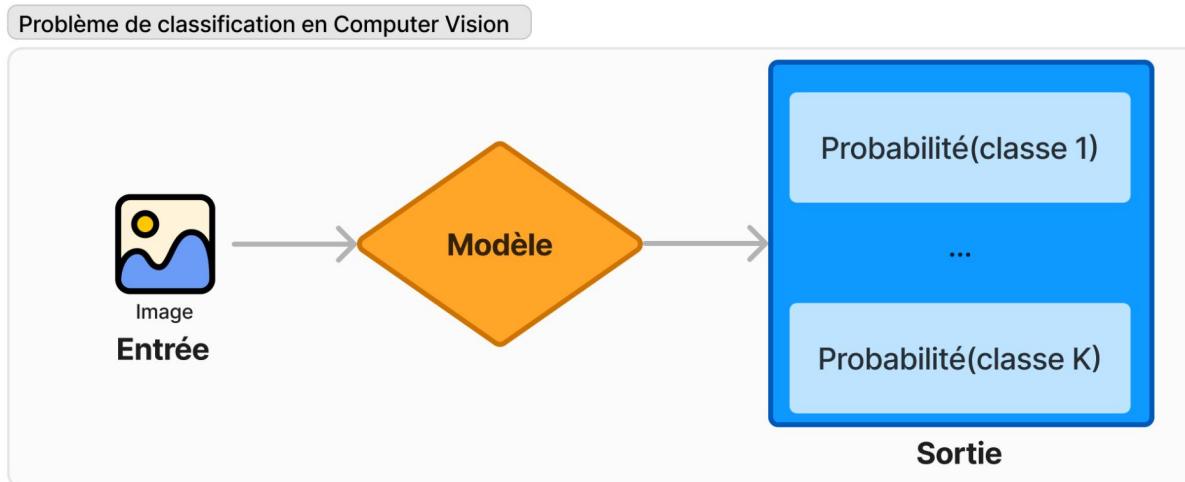
ILLUIN
TECHNOLOGY

Change the code

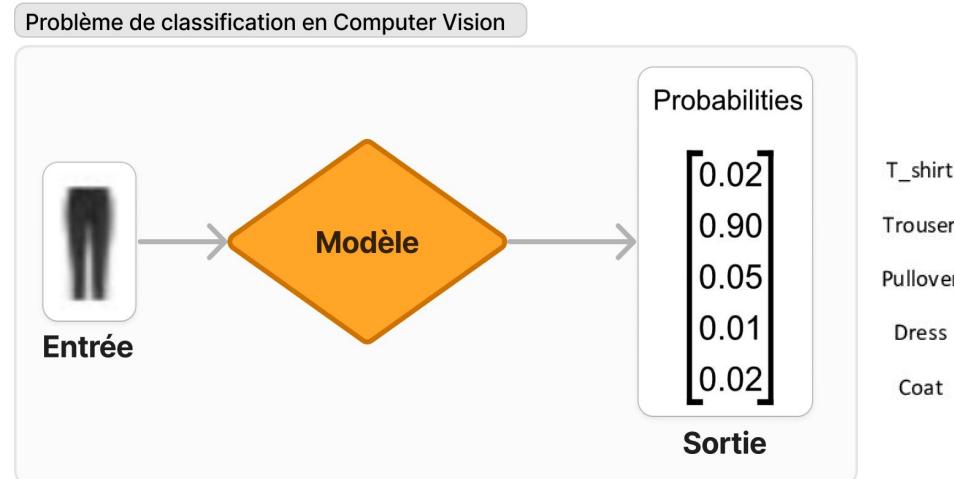
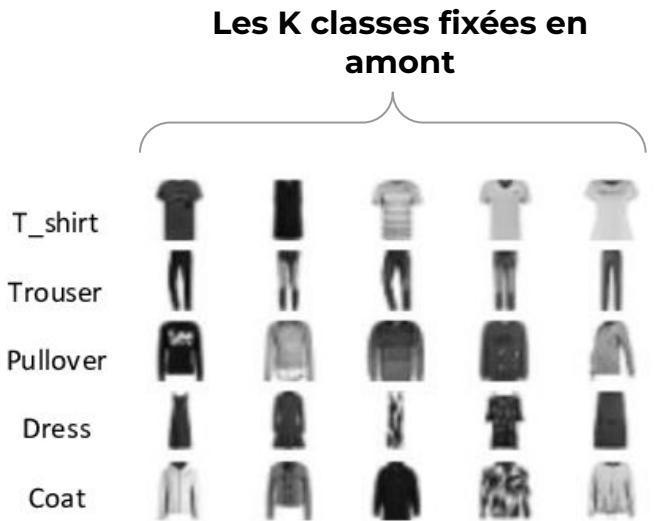
CLIP : Contrastive Language-Image Pretraining

Rappel : Classification classique

Soient K classes fixées. Pour une image d'entrée, le modèle renvoie un vecteur de taille K où la i-ème valeur correspond à la “probabilité” que l'input soit de la classe i.



Rappel : Classification classique

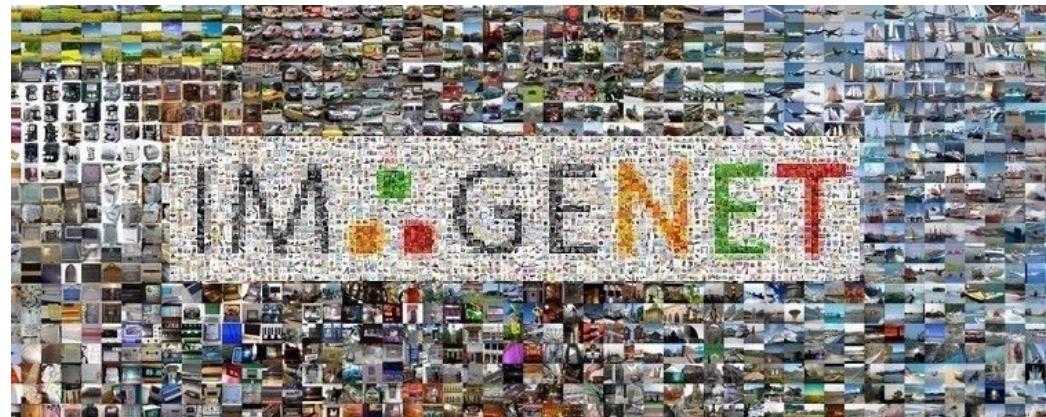


Exemple de classification avec un sous-ensemble de MNIST Fashion

Limites de la classification classique

Limites de la classification classique

- Avoir un modèle généraliste demande d'avoir de fixer à l'avance un large ensemble de classes.
- Seul l'ordre des classes dans le vecteur de sortie importe. Le nom des classes choisies n'a aucun impact sur le résultat final !



ImageNet contient 1000 classes et plus d'1 million d'exemples.

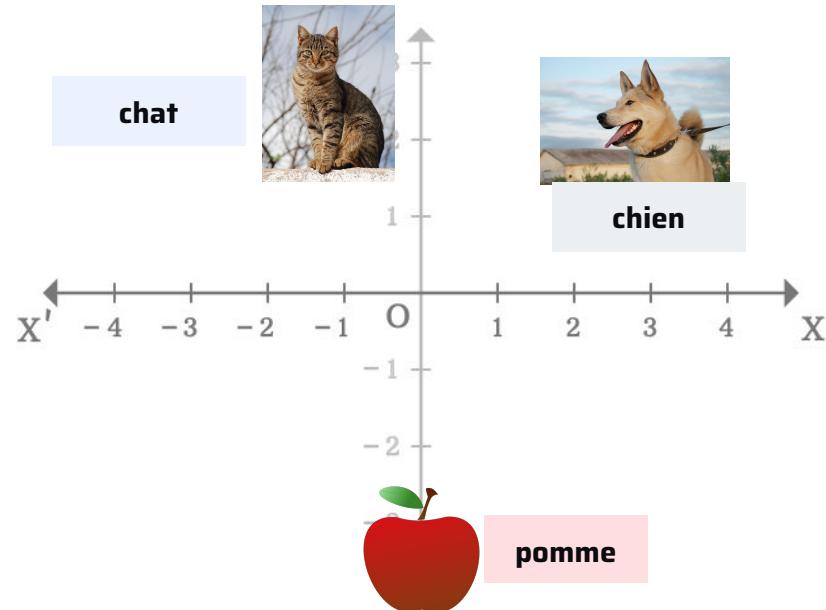
Classification zero-shot

Avec une classification zero-shot, l'utilisateur définit avant l'inférence les classes possibles.



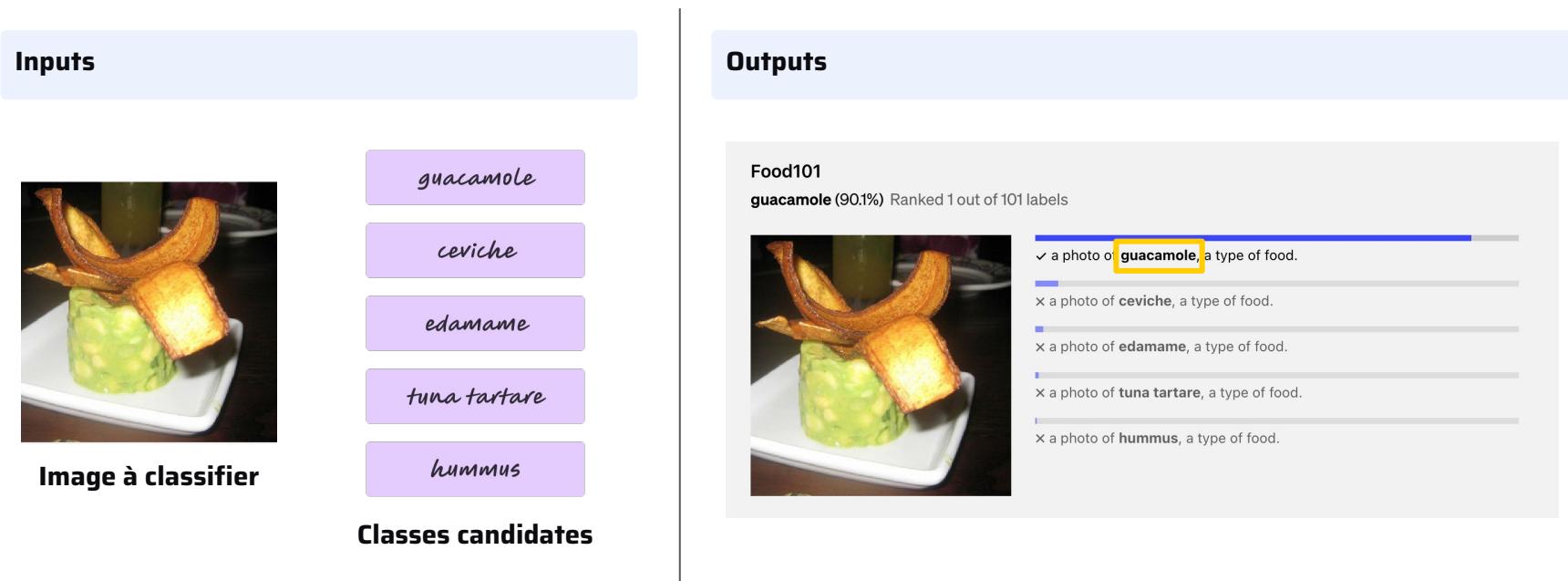
CLIP : embeddings multimodaux

CLIP propose un **nouveau paradigme pour la classification zero-shot**: générer des embeddings pour du texte et des images et faire en sorte que plus un texte et une image sont similaires et plus leurs embeddings seront semblables.



CLIP : inférence pour la classification zero-shot

Méthode : Prédire la classe avec la plus haute similarité avec l'image d'entrée.



CLIP : Architecture

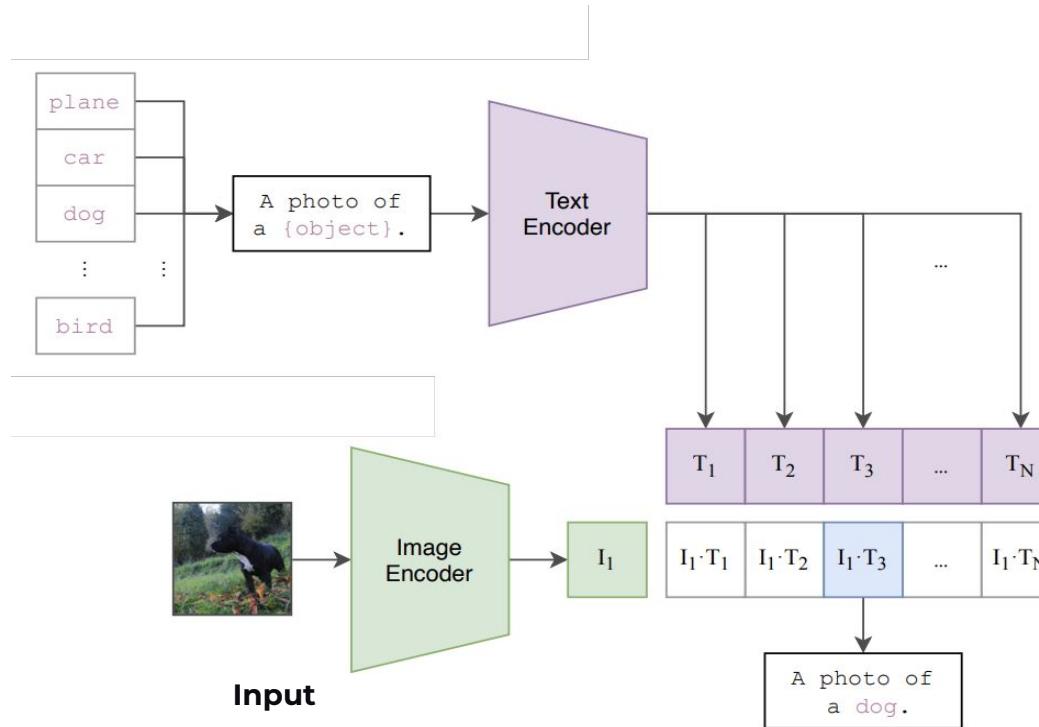


Schéma de l'architecture de CLIP

CLIP : comment ça s'entraîne ?

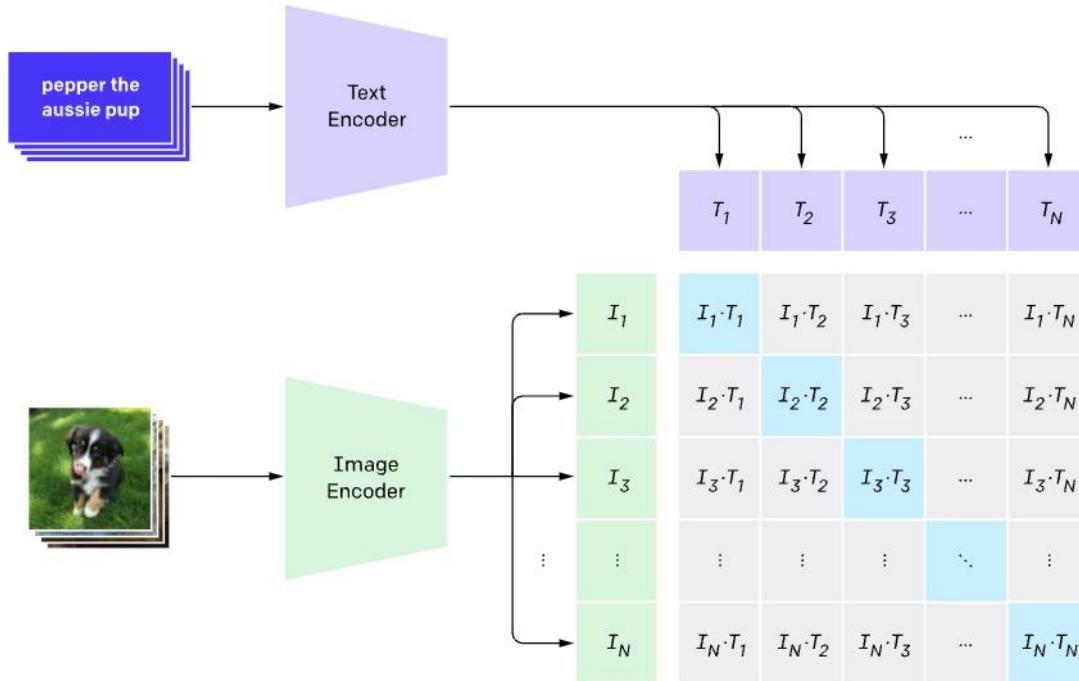


Schéma de l'entraînement de CLIP

Objectifs de l'entraînement contrastif

- ↑ Maximiser les scores de similarité pour les paires correctes
- ↓ Minimiser les scores de similarité pour les paires incorrectes

CLIP : comment ça s'entraîne ?

Objectifs de l'entraînement contrastif

- ↑ Maximiser les scores de similarité pour les paires correctes
- ↓ Minimiser les scores de similarité pour les paires incorrectes

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$...	$I_3 \cdot T_N$
:	:	:	:	:	:
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$...	$I_N \cdot T_N$

$$\mathcal{L}_{image} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^N \exp(S_{ij}/\tau)}$$

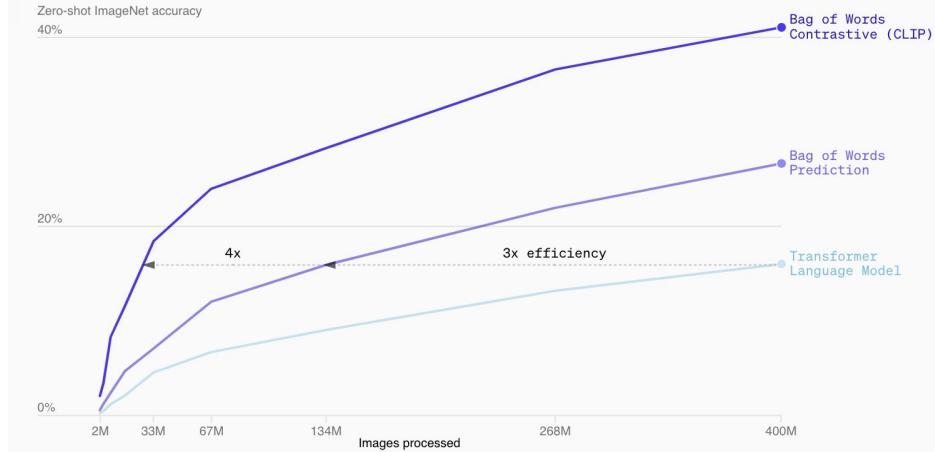
$$\mathcal{L}_{text} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^N \exp(S_{ji}/\tau)}$$

$$\mathcal{L} = \frac{1}{2}(\mathcal{L}_{image} + \mathcal{L}_{text})$$

CLIP : performances

Sur le graphe à droite :

- (1) Courbe du bas : CNN + Transformer decoder en mode génératif (VirText)
- (2) Courbe du milieu : Un CNN qui prédit un binary count vectorizer bag-of-word en output entraîné sur une tâche de classification
- (3) Courbe du haut : Pareil que la courbe du milieu mais la loss est cette fois contrastive



We originally explored training image-to-caption language models but found this approach struggled at zero-shot transfer. In this 16 GPU day experiment, a language model only achieves 16% accuracy on ImageNet after training for 400 million images. CLIP is much more efficient and achieves the same accuracy roughly 10x faster.

Accuracy en fonction du nombre d'exemples vus au training

Learning Visual Features from Large Weakly Supervised Data, Joulin et al. (2015)

VirText: Learning Visual Representations from Textual Annotations, Desai (2020)

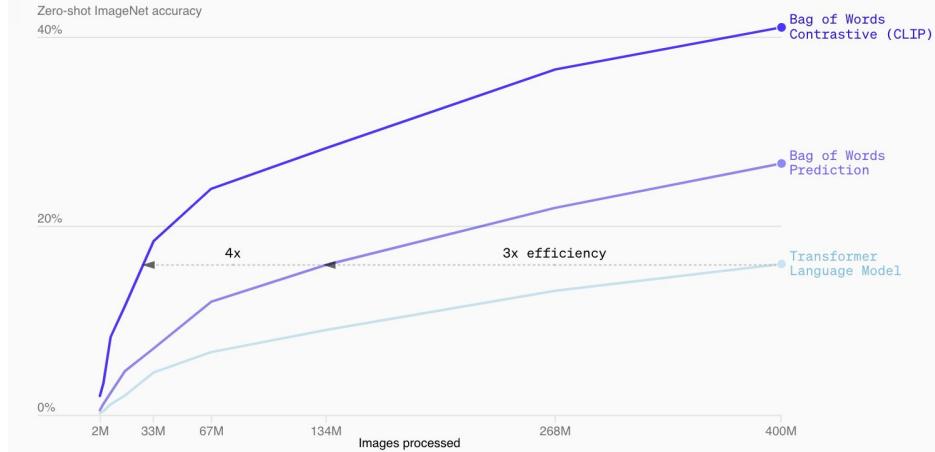
Learning Transferable Visual Models From Natural Language Supervision, Radford et al. (2021)

CLIP : performances

Pourquoi la contrastive loss aide ?

(1) et (2) poussent le modèle à apprendre de l'exact match pour chacun des mots de la classe d'image.

A contrario, (3) récompense le modèle pour chaque mot bien prédit : plus les embeddings se ressemblent et plus la loss est faible.



We originally explored training image-to-caption language models but found this approach struggled at zero-shot transfer. In this 16 GPU day experiment, a language model only achieves 16% accuracy on ImageNet after training for 400 million images. CLIP is much more efficient and achieves the same accuracy roughly 10x faster.

Accuracy en fonction du nombre d'exemples vus au training

Learning Visual Features from Large Weakly Supervised Data, Joulin et al. (2015)

VirTeX: Learning Visual Representations from Textual Annotations, Desai (2020)

Learning Transferable Visual Models From Natural Language Supervision, Radford et al. (2021)



CLIP donne-t-il de “meilleurs” embeddings ?

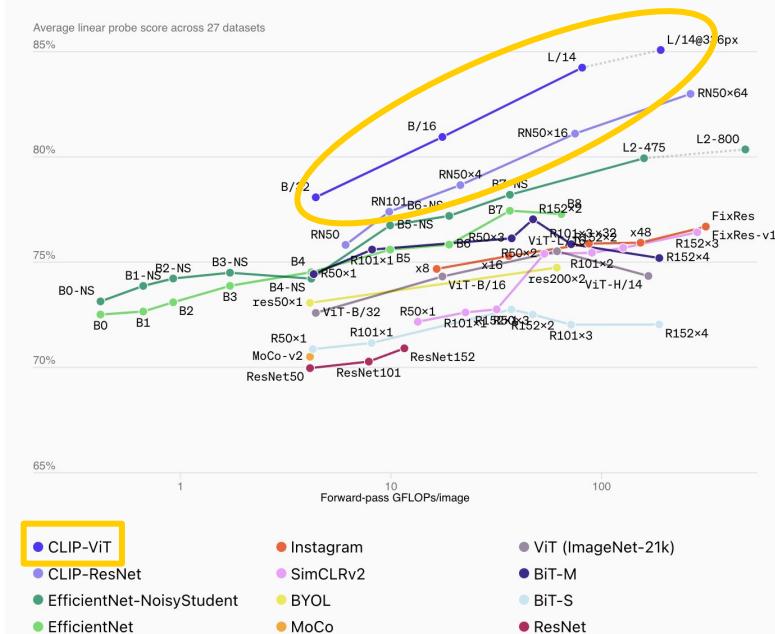
Objectif du linear probing

Entraîner un linear classifier qui prend en entrée les embeddings du modèle d'embedding à évaluer sur N tâches différentes.

Meilleur est le score moyen, meilleure est la capacité de représentation du modèle d'embedding.

Image encoder

L'image encoder utilisé dans l'architecture CLIP ayant donné les meilleures performances utilise un **ViT**.



Across a suite of 27 datasets measuring tasks such as fine-grained object classification, OCR, activity recognition in videos, and geo-localization, we find that CLIP models learn more widely useful image representations. CLIP models are also more compute efficient than the models from 10 prior approaches that we compare with.



ILLUIN
TECHNOLOGY

Change the code

SOMMAIRE KEYNOTE VLM

01. Préliminaires

02. Concevoir son VLM

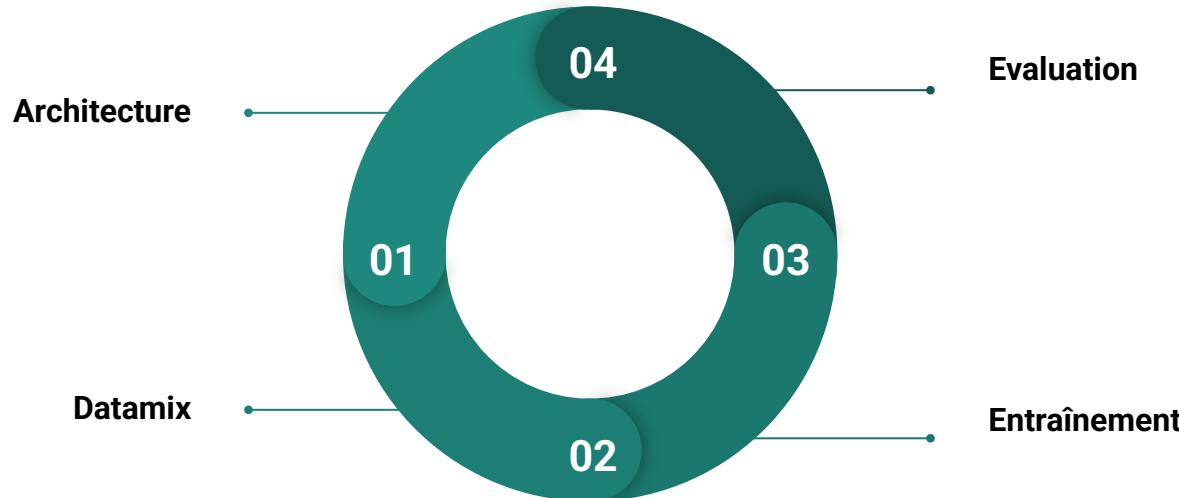
03. Exemples de VLMs récents

04. Limitations

05. Next steps

Concevoir son VLM

Entraîner un VLM comprend 4 étapes clés et n'est pas un processus linéaire car il est très courant d'entraîner plusieurs modèles avec différentes configurations pour avoir les meilleures performances.



1. Choisir l'architecture de son VLM

La plupart des VLM suivent l'architecture suivante :

- Vision encoder (CLIP ViT) → chaque patch d'image est transformé en vecteur
- Connecteur → aligne la représentation du vision encoder avec celle du Language Model
- Language Model → un LLM utilisé de manière auto-régressive

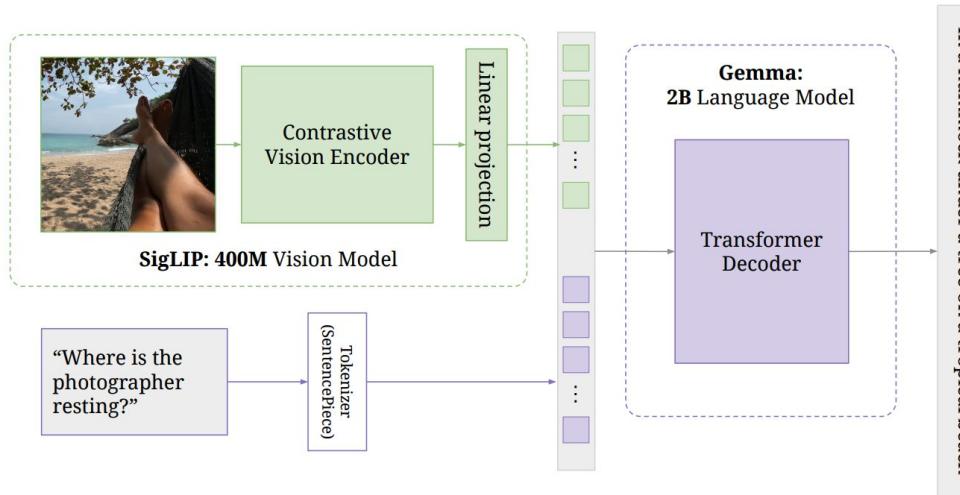


Figure 1 | PaliGemma's architecture: a SigLIP image encoder feeds into a Gemma decoder LM.

PaliGemma: A versatile 3B VLM for transfer, Beyer et al. (2024).

1. Choisir l'architecture de son VLM

La plupart des VLM suivent l'architecture suivante :

- Vision encoder (CLIP ViT) → chaque patch d'image est transformé en vecteur
- Connecteur → aligne la représentation du vision encoder avec celle du Language Model
- Language Model → un LLM utilisé de manière auto-régressive

Vision encoder



Le ViT choisi a été entraîné avec CLIP dans l'espoir que les embeddings visuels soient un minimum proche des embeddings textuels du Language Model.

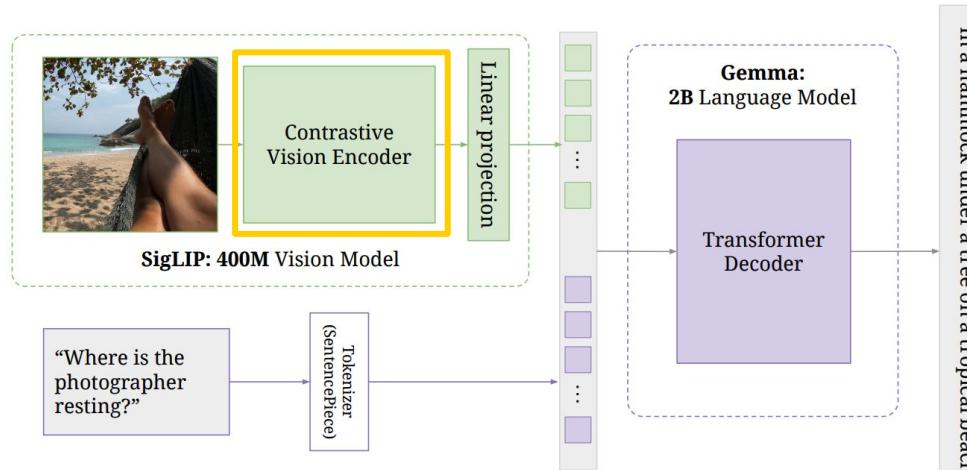


Figure 1 | PaliGemma's architecture: a SigLIP image encoder feeds into a Gemma decoder LM.

PaliGemma: A versatile 3B VLM for transfer, Beyer et al. (2024).



1. Choisir l'architecture de son VLM

La plupart des VLM suivent l'architecture suivante :

- Vision encoder (CLIP ViT) → chaque patch d'image est transformé en vecteur
- Connecteur → aligne la représentation du vision encoder avec celle du Language Model
- Language Model → un LLM utilisé de manière auto-régressive

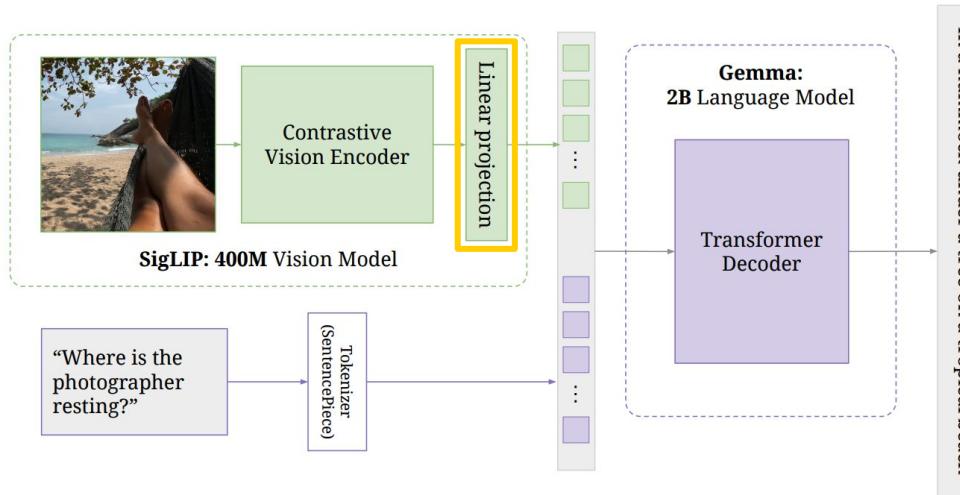


Figure 1 | PaliGemma's architecture: a SigLIP image encoder feeds into a Gemma decoder LM.

PaliGemma: A versatile 3B VLM for transfer, Beyer et al. (2024).

1. Choisir l'architecture de son VLM

La plupart des VLM suivent l'architecture suivante :

- Vision encoder (CLIP ViT) → chaque patch d'image est transformé en vecteur
- Connecteur → aligne la représentation du vision encoder avec celle du Language Model
- Language Model → un LLM utilisé de manière auto-régressive

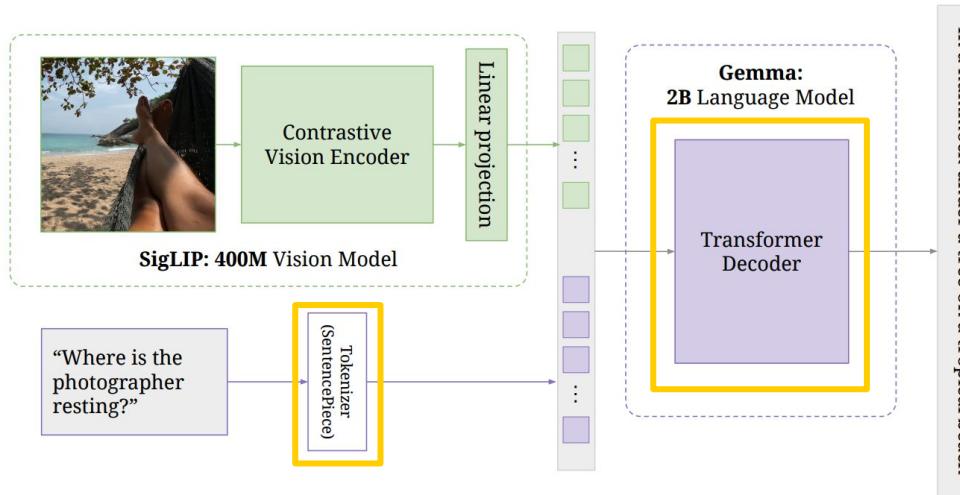


Figure 1 | PaliGemma's architecture: a SigLIP image encoder feeds into a Gemma decoder LM.

PaliGemma: A versatile 3B VLM for transfer, Beyer et al. (2024).



1. Choisir l'architecture de son VLM : quelques astuces

La résolution d'image est cruciale

La plupart des VLM travaillent à résolution fixée. Cette résolution est cruciale pour le traitement **des images de document**. Deux raisons à cela :

1. Une résolution trop faible rend la lecture du texte difficile.
2. À taille de patch fixée, une plus grande résolution permet d'avoir plus de token visuels et a fortiori plus plus de FLOPS pour l'étape de raisonnement dans le LLM.

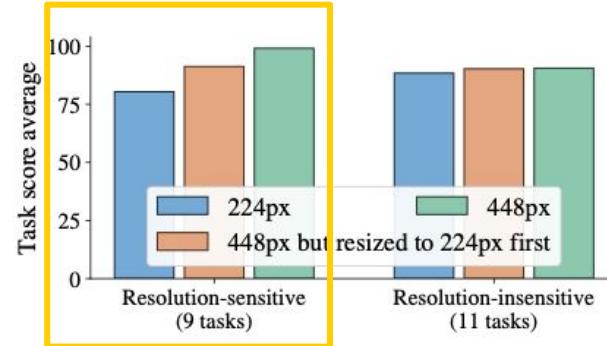


Figure 9 | Increasing resolution has two effects: increased information content of the input image, and increased model capacity via sequence length. For tasks that benefit from increased resolution, both of these effects contribute roughly equally to the overall gain.



1. Choisir l'architecture de son VLM : quelques astuces

La résolution d'image est cruciale

La plupart des VLM travaillent à résolution fixée. Cette résolution est cruciale pour le traitement **des images de document**. Deux raisons à cela :

1. Une résolution trop faible rend la lecture du texte difficile.
2. À taille de patch fixée, une plus grande résolution permet d'avoir plus de token visuels et a fortiori plus plus de FLOPS pour l'étape de raisonnement dans le LLM.

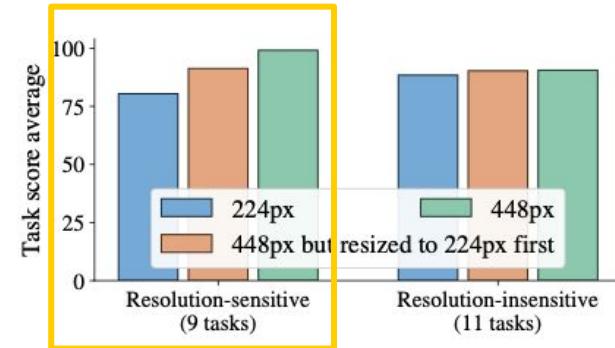
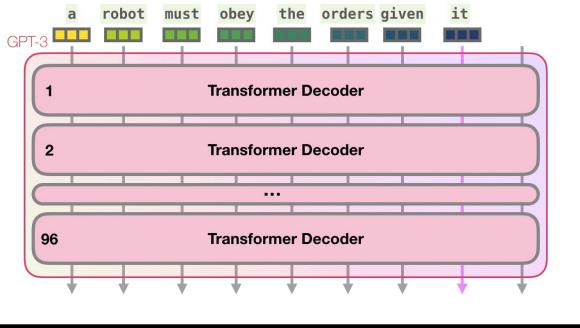


Figure 9 | Increasing resolution has two effects: increased information content of the input image, and increased model capacity via sequence length. For tasks that benefit from increased resolution, both of these effects contribute roughly equally to the overall gain.

MMI: Methods, Analysis & Insights from Multimodal LLM Pre-training. McKinzie et al. (2024)

PaliGemma: A versatile 3B VLM for transfer, Beyer et al. (2024)

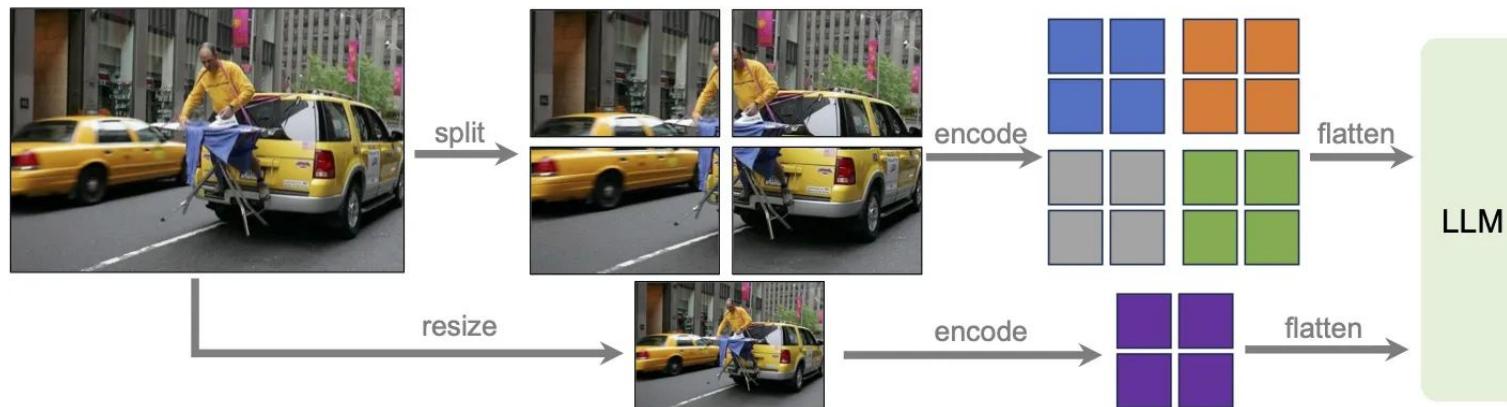
How GPT3 Works - Visualizations and Animations, <https://jalamar.github.io/how-gpt3-works-visualizations-animations/>



1. Choisir l'architecture de son VLM : quelques astuces

SPHINX : Augmenter la résolution du VLM sans changer de Vision Encoder

Le Vision Encoder traite séparément plusieurs sous-parties de l'image d'origine à la résolution native. Les représentations sont simplement concaténées avant d'être passés au LLM.



SPHINX: The Joint Mixing of Weights, Tasks, and Visual Embeddings for Multi-modal Large Language Models, Lin et al. (2023)
Improved Baselines with Visual Instruction Tuning, Liu et al. (2023)
What matters when building vision-language models?, Laurençon (2024)

1. Choisir l'architecture de son VLM : quelques astuces

Plus grand le backbone LLM, meilleure est la performance du VLM

À nouveau, l'impact de la taille du LLM impacte **surtout les tâches documentaires** (cf expérience ci-dessous sur la taille du backbone Gemma 2 utilisé pour PaliGemma 2).

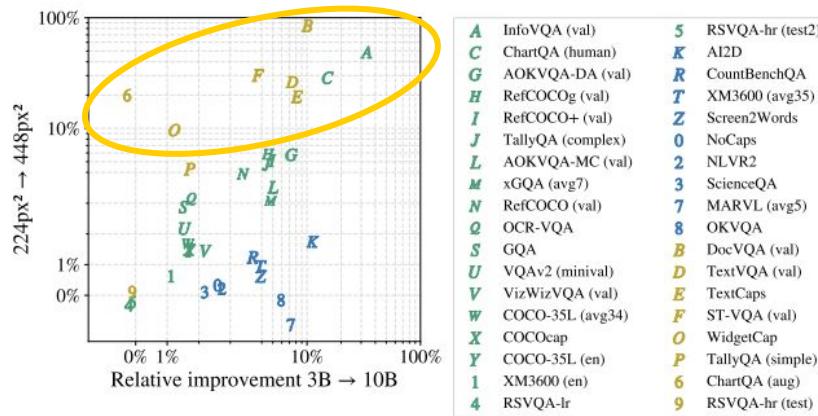
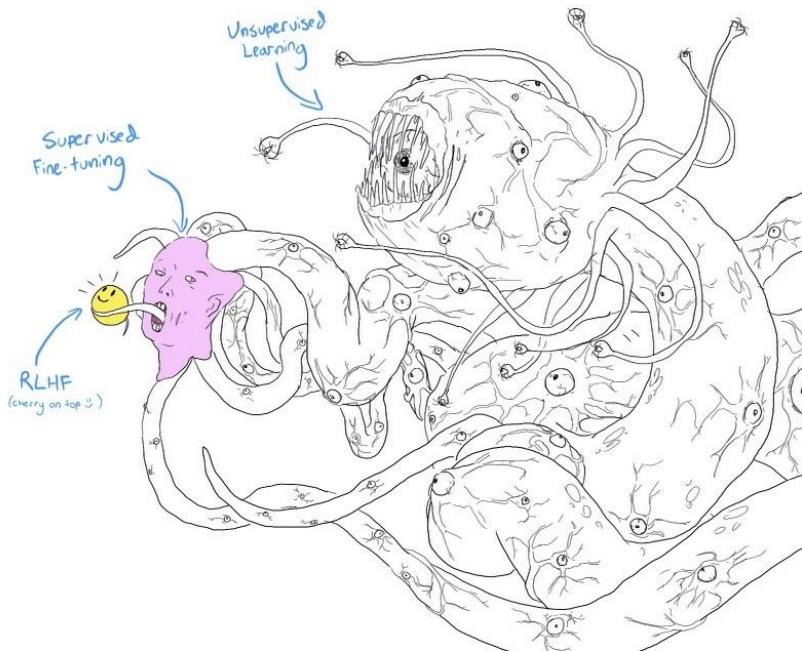


Figure 3 | Relative improvements of metrics after transfer, when choosing a pre-trained checkpoint with a larger LM, or with a higher resolution. The tasks are grouped into tasks sensitive to both model size and resolution (—), sensitive to model size (—), and sensitive to resolution (—). Note that some benchmarks are quite saturated (e.g. ScienceQA's relative improvement of 2.2% corresponds to an error reduction of 53.8% – see Figure 13). Data used to create this plot available in Table 13.

2. Datamix : rappel sur le training d'un LLM classique

Un LLM “chat” est entraîné en 3 étapes. À chaque étape est associée un datamix propre:

1. **Le Pre-Training** : Le LLM est entraîné à compléter une phrase basée sur le prochain token le plus probable.
2. **Le Supervised Fine-Tuning (SFT)** : Le LLM apprend à répondre à des instructions de l'utilisateur (e.g. résumé, traduction, etc...)
3. **Le Reinforcement Learning** : Le modèle apprend à répondre de la manière qui plaît le plus à l'Homme.



2. Choisir son datamix

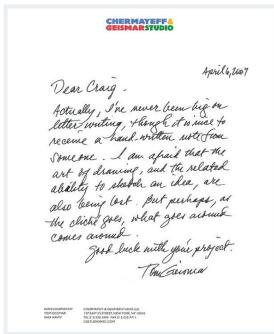
Pre-training

Tâches génériques “simples”, e.g. tâches d'OCR/HTR

Exemple utilisé par Idefics 2 :

Prompt

The content of the letter is:



Completion

Dear Craig,

Actually, I've never been big on letter writing, though it is nice to receive [...]

Supervised Fine-Tuning

Tâches spécifiques demandant plus de raisonnement.

Exemple pris dans le dataset *The Cauldron* :

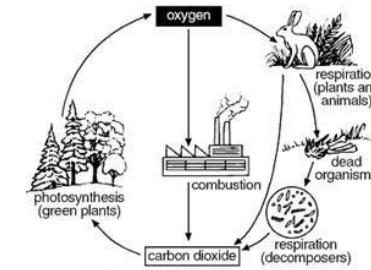
Prompt

What do respiration and combustion give out

Choices:

- A. Oxygen
- B. Carbon dioxide
- C. Nitrogen
- D. Heat

Answer with the letter.



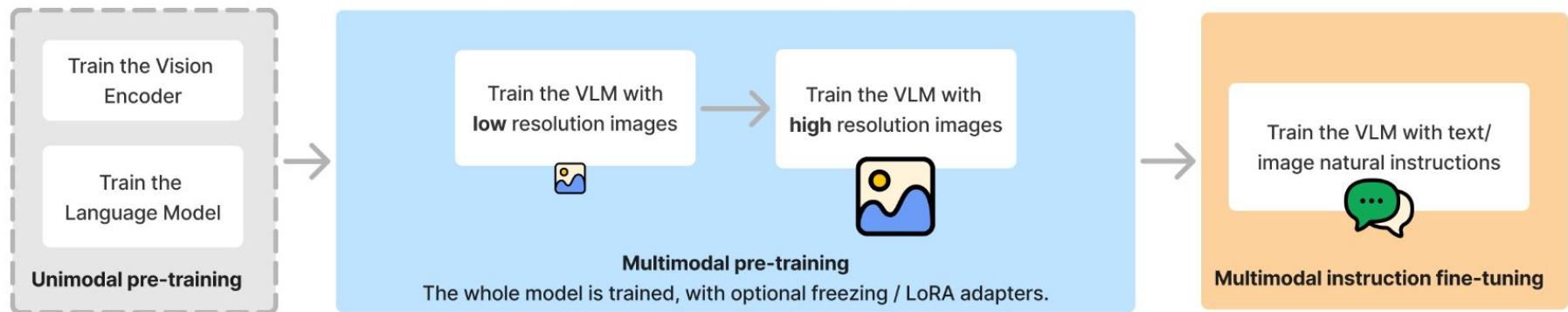
Completion

B



3. Entraînement multi-step

L'étape de **Unimodal pre-training** n'est pas requise si on part de modèles pré-trained. Comme les LLMs, on retrouve les phases de pre-training et d'instruction fine-tuning. Cependant, la phase de Reinforcement Learning est souvent omise du training.



3. Entraînement : quelques astuces au pretraining

Ajouter des exemples avec plusieurs images (interleaved)

Augmente les capacités de **few-shot**, avec un optimal atteint avec un ratio 50/50.

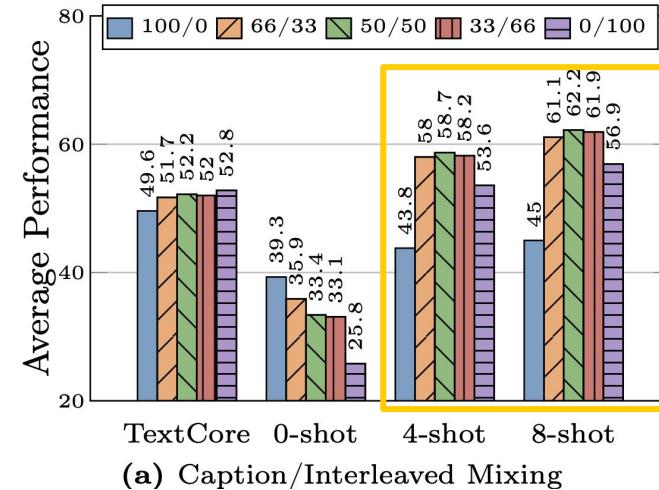
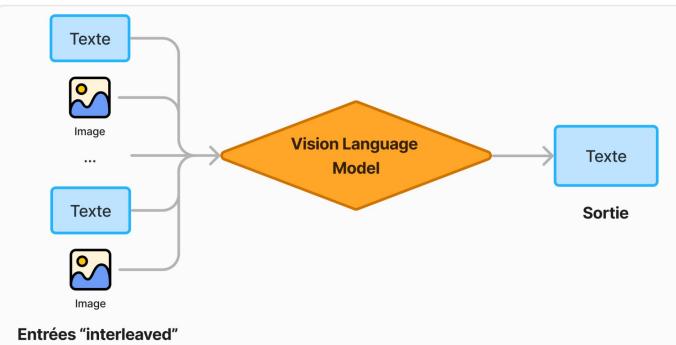


Fig. 5: Data Ablations. For each ablation, we present four different metrics: TextCore, 0-shot, 4-shot, and 8-shot. (a) Results with image data where we present five different mixing ratios between interleaved and captioned data. (b) Results with and without text-only data. We mix the text-only data separately with captioned and interleaved data. (c) Results with different mixing ratios between image data (caption and interleaved) and text-only data. (d) Results with and without including VeCap as part of caption data.

3. Entraînement : quelques astuces au pretraining

Entraîner son VLM sur des exemples text-only au pre-training

Permet de préserver les capacités de raisonnement du language model sur les tâches text-only. Utile si on veut un unique modèle génératif capable de jouer le rôle d'un LLM.

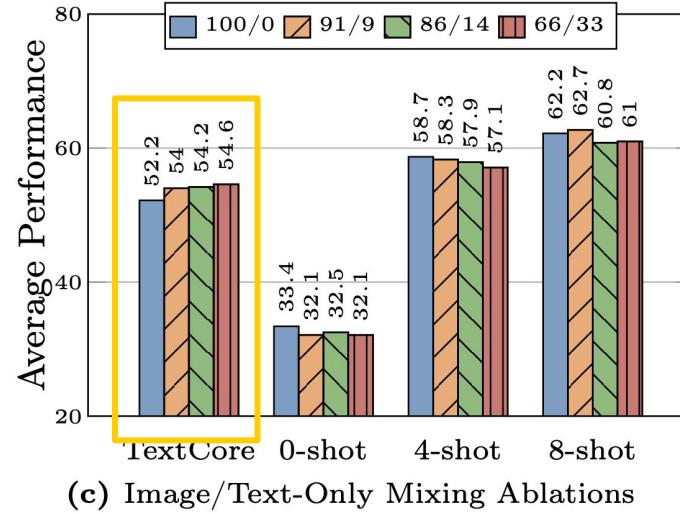


Fig. 5: Data Ablations. For each ablation, we present four different metrics: TextCore, 0-shot, 4-shot, and 8-shot. (a) Results with image data where we present five different mixing ratios between interleaved and captioned data. (b) Results with and without text-only data. We mix the text-only data separately with captioned and interleaved data. (c) Results with different mixing ratios between image data (caption and interleaved) and text-only data. (d) Results with and without including VeCap as part of caption data.

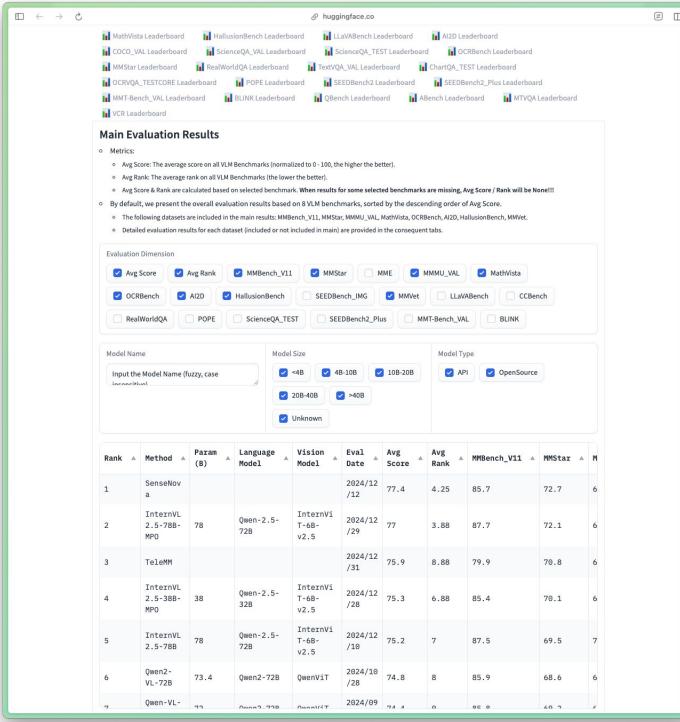
3. Evaluation : OpenVLM leaderboard

OpenVLM benchmark

- Collection de 30+ benchmarks orientés multimodal
- 200+ VLMs évalués

Attention

OpenVLM donne une idée de la performance **générale** d'un VLM mais ne remplace pas une évaluation sur la data distribution du use case !



3. Evaluation : OpenVLM benchmark

Benchmark Name	Description	Benchmark Name	Description
MMBench	Evaluates all-round capabilities of LMMs, including tasks such as image captioning, visual question answering, and text-based reasoning.	OCRBench	Assesses text understanding capabilities, particularly in Optical Character Recognition (OCR) contexts, evaluating the ability to extract and understand text from images.
MMT-Bench	Evaluates performance across multiple tasks, providing a comprehensive assessment of LMM capabilities across a range of challenges.	TextVQA	Evaluates text recognition capabilities within visual question answering tasks, assessing the ability to answer questions that require understanding text within images.
MMMU	Designed for multi-modal examination and assessment, testing the ability of LMMs to perform well in exam-like scenarios that involve both visual and textual information.	AI2D	Focuses on diagram understanding and interpretation, evaluating the ability to extract information and answer questions about diagrams.
MathVista	Tests capabilities in solving mathematical problems using multi-modal information, evaluating the ability to integrate visual and textual cues to arrive at solutions.	ChartQA	Assesses the ability to understand and answer questions about charts, evaluating the ability to interpret visual representations of data.
COCO Caption	Evaluates image captioning abilities, assessing the quality and accuracy of generated captions for images.	DocVQA	Evaluates document understanding capabilities, including information extraction and question answering, assessing the ability to understand and process complex documents.

3. Evaluation : Chatbot Arena Vision benchmark

Chatbot Arena Vision

- Les utilisateurs engagent des conversations avec différents chatbots et votent pour celui qui offre la meilleure expérience.
- Seuls les échanges avec au moins une image sont pris en compte dans ce benchmark vision.
- À chaque modèle est associé un score ELO. Ce score est moins explicable que les scores sur OpenVLM mais est en revanche holistique.

The screenshot shows the Chatbot Arena LLM Leaderboard page. At the top, there are tabs for 'Arena (battle)', 'Arena (side-by-side)', 'Direct Chat', 'Leaderboard' (which is selected), and 'About Us'. Below the tabs, the title 'Chatbot Arena LLM Leaderboard: Community-driven Evaluation for Best LLM and AI chatbots' is displayed, along with links to 'GitHub' and 'Kaggle Competition'. A brief description explains that Chatbot Arena is an open platform for crowdsourced AI benchmarking, developed by researchers at UC Berkeley SkyLab and LMArena. It mentions over 1,000,000 user votes and the Bradley-Terry model for generating leaderboards. Technical details are linked to a paper. A call to action encourages community engagement and voting. Below this, a message invites users to visit chatbot-arena.web.app for a new experimental UI and feedback. The main content area features a table titled 'Overall Questions' with the following data:

#models: 49 (100%) #votes: 158,343 (100%)								
Rank* (UB)	Rank (StyleCtr)	Model	Arena Score	95% CI	Votes	Organization	License	
1	1	Gemini-2.0-Flash-Thinking-Exp-1219	1271	+14/-17	1370	Google	Proprietary	
1	1	Gemini-2.0-Flash-Exp	1257	+11/-11	2440	Google	Proprietary	
2	1	Gemini-Exp-1206	1239	+11/-13	2774	Google	Proprietary	
3	3	ChatGPT-4o-latest (2024-11-20)	1227	+7/-7	4479	OpenAI	Proprietary	
3	3	Gemini-1.5-Pro-002	1221	+6/-6	7674	Google	Proprietary	



3. Evaluation : Leaderboard ILLUIN pour les tâches métiers

ILLUIN Leaderboard

- Les VLMs évoluent rapidement mais le choix d'un VLM pour un use case métier n'est pas clair.
- Le Leaderboard ILLUIN vise à évaluer les derniers VLMs sur des datasets représentatifs de l'industrie pour faciliter le processus de choix de modèle.
- Le leaderboard et le travail de mise à jour constituent un produit destiné à être vendu auprès de clients.

The screenshot shows the ILLUIN LLM Leaderboard on the Hugging Face platform. The interface includes sections for Models (listing various VLMs like deepseek-ai/DeepSeek-V3, gpt-3.5-turbo-0125, etc.) and Metrics (listing evaluation metrics like answer_relevancy, completeness, faithfulness, usefulness, eval_source_f1_score, eval_source_exact_match, eval_source_precision, eval_source_recall). A table at the bottom displays performance metrics for different models, such as Qwen/Qwen2.5-72B-Instruct, deepseek-ai/DeepSeek-V3, and gpt-3.5-turbo-0125, across the specified metrics.

model	answer_relevancy	completeness	faithfulness	usefulness
Qwen/Qwen2.5-72B-Instruct	4.75	4.26	0.66	0.98
deepseek-ai/DeepSeek-V3	4.83	4.52	0.8	1
gpt-3.5-turbo-0125	4.12	2.77	0.61	0.94
gpt-4o-2024-08-06	4.71	4.17	0.59	0.98
meta-llama/Llama-3.1-405B-Instruct	4.68	4.51	0.7	0.97
meta-llama/Llama-3.1-70B-Instruct	4.67	4.33	0.69	0.97
meta-llama/Llama-3.1-8B-Instruct	4.52	3.69	0.36	0.94
meta-llama/Llama-3.2-3B-Instruct	4.31	3.18	0.31	0.98
mistralai/Mistral-8x22B-Instruct-v0.1	4.28	4.72	0.59	1
vertex_ai/claudie-3-5-sonnet-v2@20241022	4.79	4.89	0.62	0.96

ILLUIN LLM Leaderboard (illuin/llm-leaderboard), ILLUIN Technology, <https://huggingface.co/spaces/illuin/llm-leaderboard>



ILLUIN
TECHNOLOGY

Change the code

SOMMAIRE KEYNOTE VLM

01. Préliminaires

02. Concevoir son VLM

03. Exemples de VLMs récents

04. Limitations

05. Next steps

PaliGemma

Prefix-LM

- PaliGemma est un Prefix-LM : l'attention n'est pas masquée sur l'input **{image + prompt utilisateur}** (full attention).
- Le raisonnement est que l'attention causale n'est pas nécessaire pour l'image d'input car celle-ci est toujours un input utilisateur. En outre, la full attention permet aux tokens visuels associés à l'image d'avoir connaissance de la tâche décrite dans le prefix

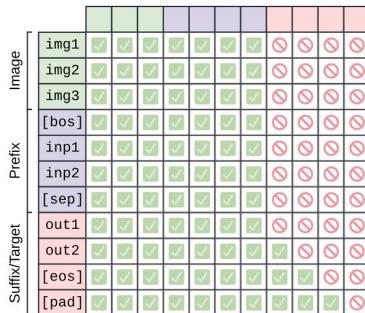


Figure 2 | PaliGemma's Prefix-LM masking: block attention throughout image and prefix, autoregressive attention on the suffix. Each square indicates whether the row can attend to the column.

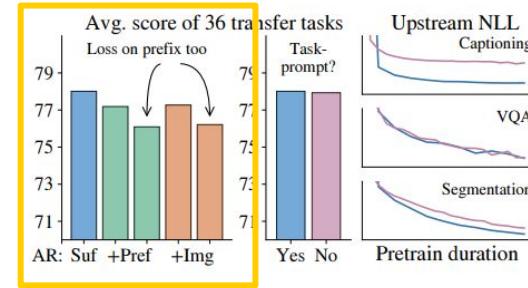


Figure 5 | Learning setup for Stage1. Left: Where to apply the auto-regressive mask and loss. Middle and right: whether to include a task indicator.

Qwen2-VL

Dynamic resolution/ratio + support vidéo

- Remplace les positional embeddings du ViT vanilla par RoPE-2D pour supporter les différentes résolutions/ratio d'image
- Remplace les positional embeddings du LLM vanilla par un RoPE-3D pour supporter les inputs vidéo.
- Compression des tokens visuels par groupe de 2x2 patch disjoints avec un MLP pour réduire le nombre de visual tokens pour permettre de traiter des inputs vidéos plus longs.

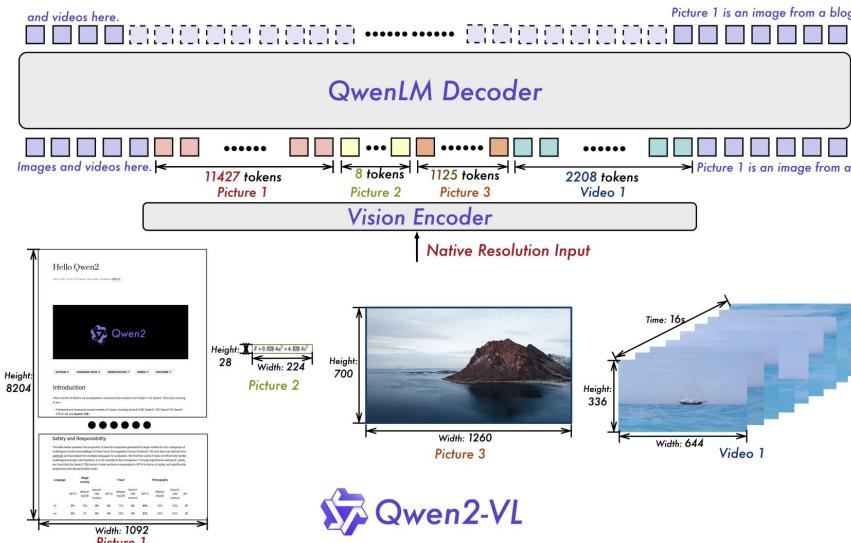


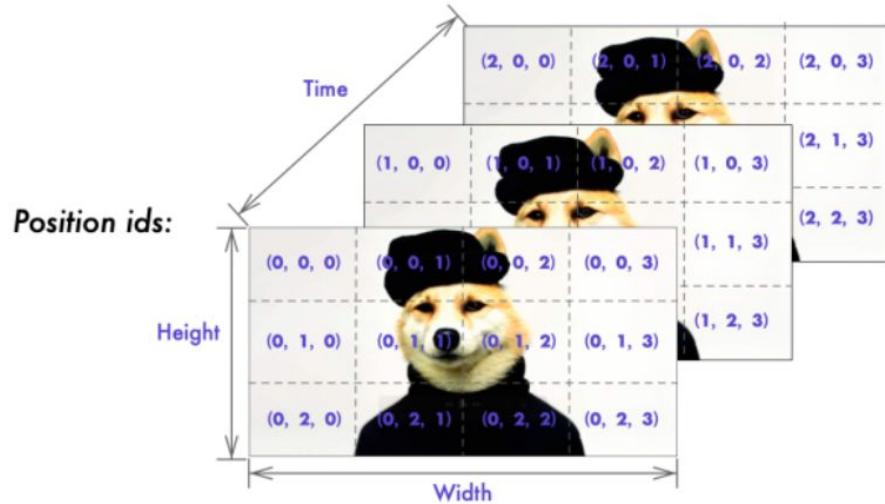
Figure 2: Qwen2-VL is capable of accurately identifying and comprehending the content within images, regardless of their clarity, resolution, or extreme aspect ratios.



Qwen2-VL

Dynamic resolution/ratio + support vidéo

- Remplace les positional embeddings du ViT vanilla par RoPE-2D pour supporter les différentes résolutions/ratio d'image
- Remplace les positional embeddings du LLM vanilla par un RoPE-3D pour supporter les inputs vidéo.
- Compression des tokens visuels par groupe de 2x2 patch disjoints avec un MLP pour réduire le nombre de visual tokens pour permettre de traiter des inputs vidéos plus longs.



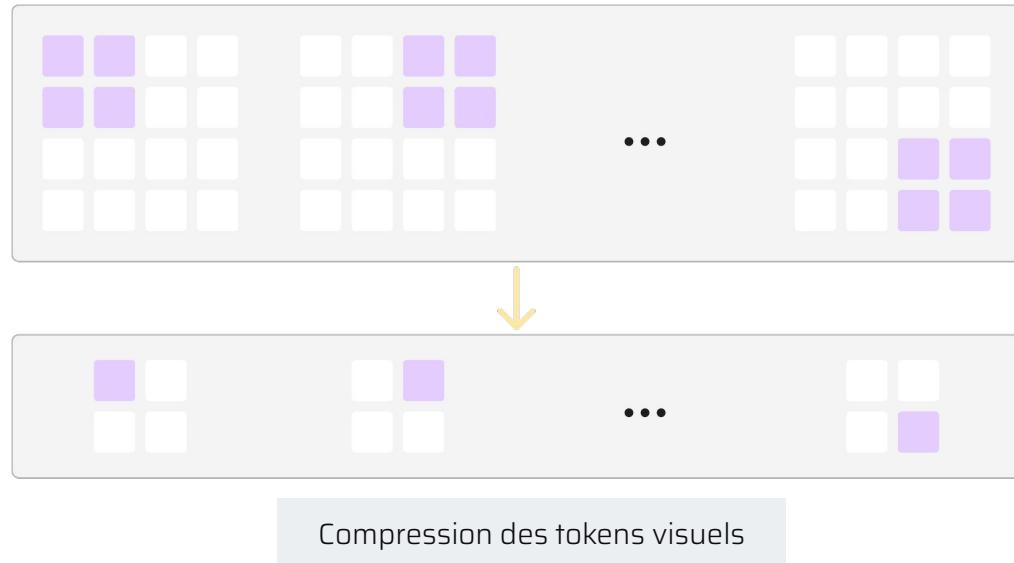
RoPE-3D : 2 dimensions pour hauteur/largeur et 1 dimension pour l'indice de la frame vidéo



Qwen2-VL

Dynamic resolution/ratio + support vidéo

- Remplace les positional embeddings du ViT vanilla par RoPE-2D pour supporter les différentes résolutions/ratio d'image
- Remplace les positional embeddings du LLM vanilla par un RoPE-3D pour supporter les inputs vidéo.
- **Compression des tokens visuels** par groupe de 2x2 patch disjoints avec un MLP pour réduire le nombre de visual tokens pour permettre de traiter des inputs vidéos plus longs.



Pixtral

Pixtral ViT : un vision encoder custom capable de gérer de multiples résolutions et aspect ratios

- Reprend le full attention de PaliGemma mais ajoute le support multi-image grâce à un block-diagonal mask
- Ajoute des break tokens [b] pour délimiter chaque ligne de patches
- Remplace les positional embeddings du ViT vanilla par RoPE-2D pour supporter les différentes résolutions/ratio d'image

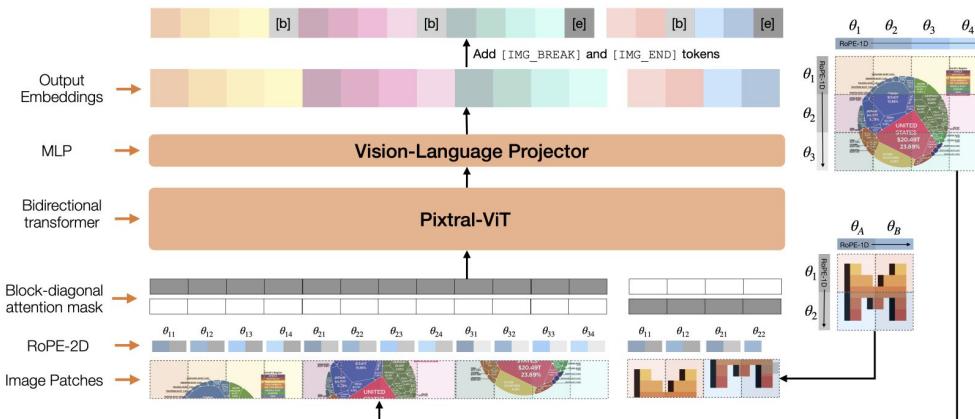


Figure 2: Pixtral Vision Encoder. Pixtral uses a new vision encoder, which is trained from scratch to natively support variable image sizes and aspect ratios. Block-diagonal attention masks enable sequence packing for batching, while RoPE-2D encodings facilitate variable image sizes. Note that the attention mask and position encodings are fed to the vision transformer as additional input, and utilized only in the self-attention layers.



Pixtral

Pixtral ViT : un vision encoder custom capable de gérer de multiples résolutions et aspect ratios

- Reprend le full attention de PaliGemma mais ajoute le support multi-image grâce à un **block-diagonal mask**
- Ajoute des break tokens **[b]** pour délimiter chaque ligne de patches
- Remplace les positional embeddings du ViT vanilla par RoPE-2D pour supporter les différentes résolutions/ratio d'image

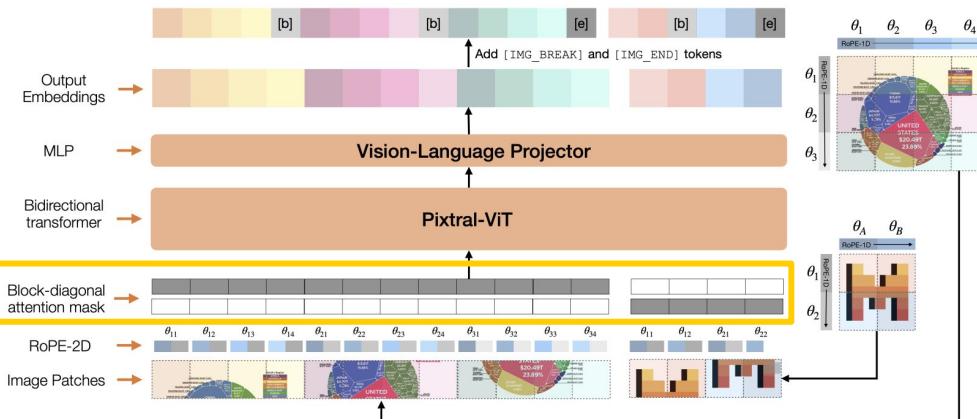


Figure 2: Pixtral Vision Encoder. Pixtral uses a new vision encoder, which is trained from scratch to natively support variable image sizes and aspect ratios. Block-diagonal attention masks enable sequence packing for batching, while RoPE-2D encodings facilitate variable image sizes. Note that the attention mask and position encodings are fed to the vision transformer as additional input, and utilized only in the self-attention layers.

Pixtral

Pixtral ViT : un vision encoder custom capable de gérer de multiples résolutions et aspect ratios

- Reprend le full attention de PaliGemma mais ajoute le support multi-image grâce à un block-diagonal mask
- Ajoute des **break tokens [b]** pour délimiter chaque ligne de patches
- Remplace les positional embeddings du ViT vanilla par RoPE-2D pour supporter les différentes résolutions/ratio d'image

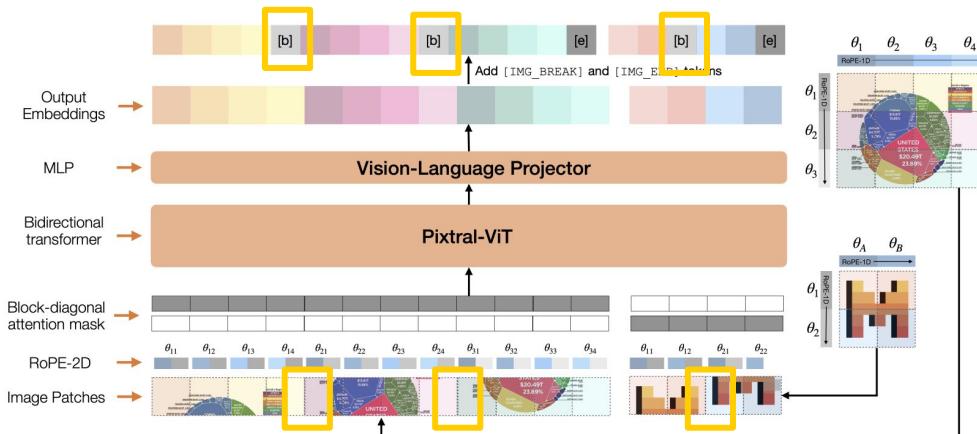


Figure 2: Pixtral Vision Encoder. Pixtral uses a new vision encoder, which is trained from scratch to natively support variable image sizes and aspect ratios. Block-diagonal attention masks enable sequence packing for batching, while RoPE-2D encodings facilitate variable image sizes. Note that the attention mask and position encodings are fed to the vision transformer as additional input, and utilized only in the self-attention layers.

Pixtral

Pixtral ViT : un vision encoder custom capable de gérer de multiples résolutions et aspect ratios

- Reprend le full attention de PaliGemma mais ajoute le support multi-image grâce à un block-diagonal mask
- Ajoute des break tokens [b] pour délimiter chaque ligne de patches
- Remplace les positional embeddings du ViT vanilla par RoPE-2D pour supporter les différentes résolutions/ratio d'image

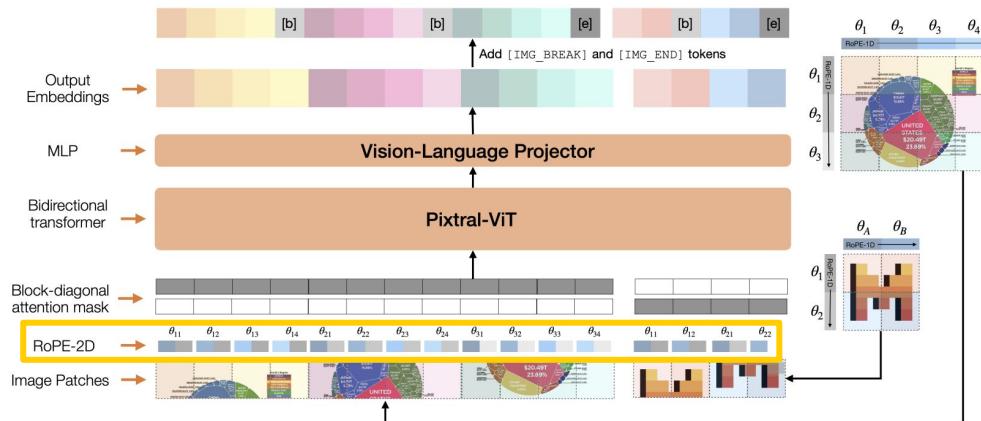


Figure 2: Pixtral Vision Encoder. Pixtral uses a new vision encoder, which is trained from scratch to natively support variable image sizes and aspect ratios. Block-diagonal attention masks enable sequence packing for batching, while RoPE-2D encodings facilitate variable image sizes. Note that the attention mask and position encodings are fed to the vision transformer as additional input, and utilized only in the self-attention layers.

Molmo : Un nouveau process de génération de data simple et riche

Annotation hybride à partir de descriptions audio humaines 🎙️

La lacune des VLMs dans le captioning vient des captions trop courtes utilisées à l'entraînement. Les auteurs de Molmo ont donc eu l'idée de faciliter l'annotation humaine en leur demandant de **décrire à voix haute** les images du dataset.



Pipeline de génération de données pour le captioning pour PixMo/Molmo





ILLUIN
TECHNOLOGY

Change the code

SOMMAIRE KEYNOTE VLM

- 01. Préliminaires**
- 02. Concevoir son VLM**
- 03. Exemples de VLMs récents**
- 04. Limitations**
- 05. Next steps**

Multimodal Visual Patterns (MMVP) : là où les VLMs voient mal



Q: Is the dog facing left or right from the camera's perspective?

The dog is **facing to the right** from the camera's perspective. You can tell by the position of its tail and the orientation of its body.



Q: Do you see any window in this image?

No, there are **no windows visible** in this image. The image shows a traditional red pillar box, which is a type of mailbox found in the United Kingdom.



Q: Is the door of the truck cab open?

No, the door of the truck is **not open** in the image provided.



Q: In this image, how many eyes can you see on the animal?

The image shows **one eye** of the animal. It's a close-up of a bald eagle facing slightly to its right, which presents only one side of its face to the viewer.



Q: Does the keyboard have a backlight?

Based on the image, it appears that the keyboard **does have a backlight**. This is indicated by the illumination seen around the keys, which suggests that there is a light source beneath them to enhance visibility in low-light conditions.



Q: From the camera's perspective, is the piano's back panel on the left or right side?

From the camera's perspective, the piano's back panel is **on the left side**. The keys visible in the foreground suggest that we are looking at the piano from front-right side, thus placing the back panel on the left from this angle.



Q: Is the edge of the heart in the picture white or dark-colored?

The image you've provided is small and of low resolution, but it appears that the hearts have a **dark-colored edge** or outline. This outline varies in visibility, likely due to the quality of the image.



Q: Is the school bus parked facing the camera or away from the camera?

The school bus is parked **facing away from the camera**. You can tell because the rear lights and the back of the exterior rearview mirror are visible. The text "SCHOOL BUS" is also reversed, which is typically seen on the front of a bus, designed so that it reads correctly in a rearview mirror.

Les erreurs des VLMs sur MMVP ont été regroupées en 8 catégories



Orientation and Direction



Presence of Specific Features



State and Condition



Quantity and Count



Positional and Relational Context



Color and Appearance



Structural and Physical Characteristics



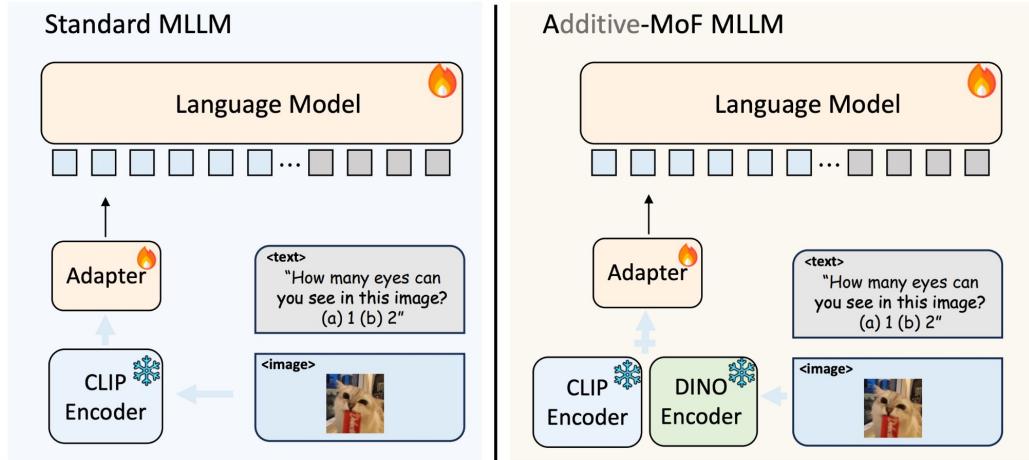
Text



Viewpoint and Perspective

Exemples du MMVP sur lesquels les VLMs répondent de manière incorrecte

Multimodal Visual Patterns (MMVP) : là où les VLMs voient mal



Solution proposée : combiner la représentation de CLIP avec une représentation purement visuelle self-supervised de DINOv2

method	SSL ratio	MMVP	LLaVA
LLaVA	0.0	5.5	81.8
	0.25	7.9 (+2.4)	79.4 (-2.4)
	0.5	12.0 (+6.5)	78.6 (-3.2)
LLaVA + A-MoF	0.625	15.0 (+9.5)	76.4 (-5.4)
	0.75	18.7 (+13.2)	75.8 (-6.0)
	0.875	16.5 (+11.0)	69.3 (-12.5)
	1.0	13.4 (+7.9)	68.5 (-13.3)

Résultats VLM vanilla vs VLM A-MoF avec différents ratios de self-supervised learning (SSL) avec DINOv2

Vers du raisonnement spatial 3D ?

User



You are a cleaning robot that is 1 meter wide, now you are standing in a living room and see the image, you want to move from here to the door that leads to the backyard, do you think I can go through the path between the sofa and the table and chairs.



User



Which can is the closest to the yellow finger that is on the right most side of the picture, the blue, orange or the silver one?



GPT-4V*



As an AI, I'm unable to physically interact with environments, but I can provide some insights based on the image provided. ... From the image, we can make some visual estimations. Judging from the image, it appears that the sofa and the table are quite close to each other, and there may not be enough space to accommodate a 1-meter wide robot....

Spatial-VLM



[VLM Reasoning] How wide is the path between the sofa and the table and chairs?
[VLM Answer] 1.56m [Answer] Yes, the robot can go through the path between the sofa and the table and chairs since it is wider than the robot's width.

GPT-4V



After estimating the distances by measuring the Euclidean distances between the yellow finger and each of the cans, it is clear that the silver can is the closest to the yellow finger with an approximate distance of 110 units. In comparison, the blue can is about 262 units away, and the orange can is approximately 233 units away from the yellow finger.

Spatial-VLM

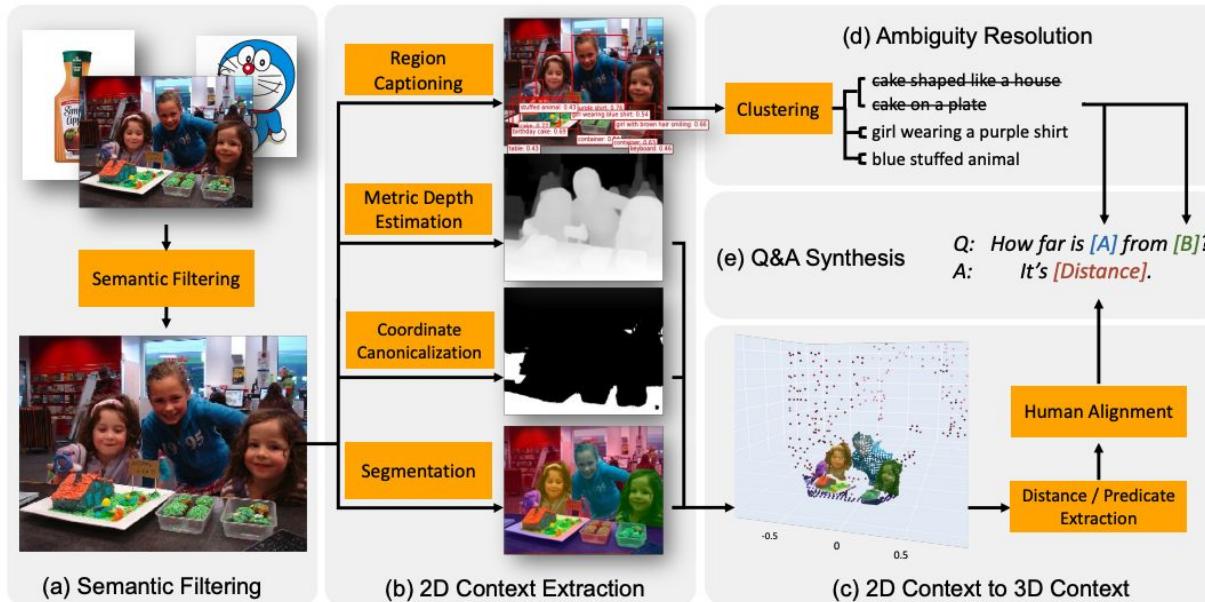


[VLM Reasoning] How far is the yellow finger from the blue can? [VLM Answer] 0.5m [VLM Reasoning] How far is the yellow finger from the orange can? [VLM Answer] 0.59m [VLM Reasoning] How far is the yellow finger from the silver can? [VLM Answer] 0.21m [Answer] The silver can is the closest to the yellow finger that is on the right most side of the picture.

GPT-4V vs Spatial-VLM sur des tâches de raisonnement sur des distances dans l'espace



Vers du raisonnement spatial 3D ?



Distillation des capacités de modèles experts dans le VLM dans Spatial-VLM

Vers du raisonnement spatial ?



Introducing Gemini 2.0: our new AI model for the agentic era,

<https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#ceo-message> (2024)

Autres limitations observées en pratique



Projet DTY OOrion : Description de scènes pour personnes malvoyantes

Florence-2 VL **hallucine souvent sur les couleurs** (comportement aussi identifié dans [ce papier](#)).



Projet DTY SACEM : Extraction d'information documentaire

Claude 3.5 Sonnet a tendance à **rajouter des informations** dans les informations extraites. Comme la KPI client privilégié l'exact match, la solution des VLM a été abandonnée pour une solution OCR+LLM.





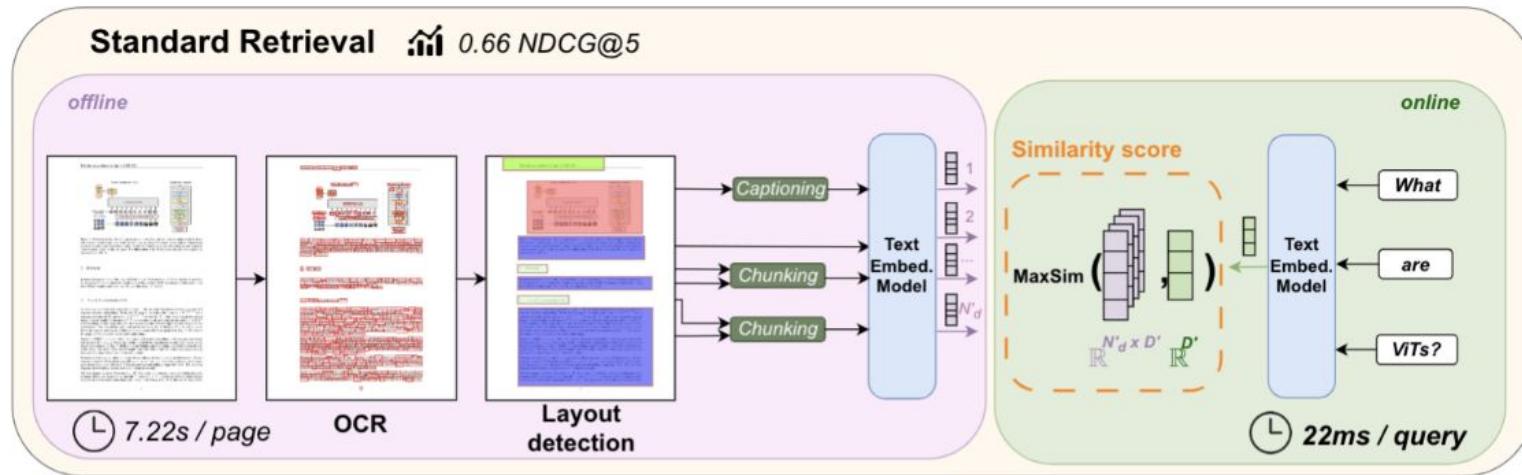
ILLUIN
TECHNOLOGY

Change the code

SOMMAIRE KEYNOTE VLM

- 01. Préliminaires**
- 02. Concevoir son VLM**
- 03. Exemples de VLMs récents**
- 04. Limitations**
- 05. Next steps**

A Standard Document Retrieval Indexing Pipeline



Indexing pipelines are complex and slow

To index documents, pipelines may include **OCR** models, **document layout detection** methods, image and figure **captioning** approaches, **chunking** strategies, and finally **textual embedding models**. In practice, this is complex, slow, and visual elements are not handled well.

ColPali : utiliser les VLM pour du retrieval de document

ColPali

for information retrieval from PDFs

Fig. 1: Common Retrieval

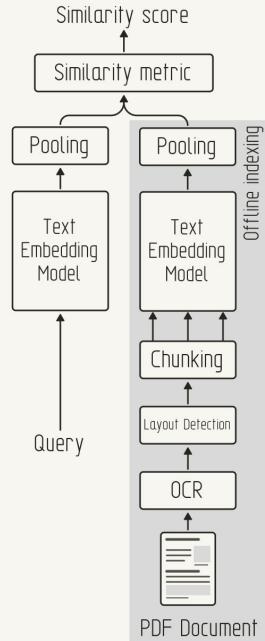
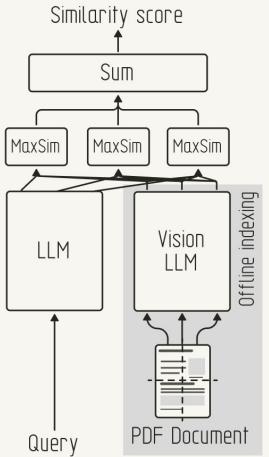


Fig. 2: ColPali



Méthodes de retrieval de documents

- Avant ColPali, indexer des documents nécessitait l'orchestration d'une pipeline de plusieurs modèles.
- ColPali combine les capacités de compréhension de document des VLM et une représentation contextualisée multi-vecteur pour améliorer le retrieval de documents selon les axes donnés dans le tableau suivant.

OCR-BASED PIPELINES

COLPALI

SIMPLE

1 seule modèle



RAPIDE

Embedding rapide



HAUTE PERFORMANCE

Pertinence des documents récupérés



ENTRAINABLE E2E

Facilement adaptable à de nouveaux types de document



ColPali : utiliser les VLM pour du retrieval de document

ColPali est rapide

ColPali est plus un ordre de grandeur plus rapide que la meilleure solution open-source de pipeline d'embedding textuel de document ([unstructured](#)).

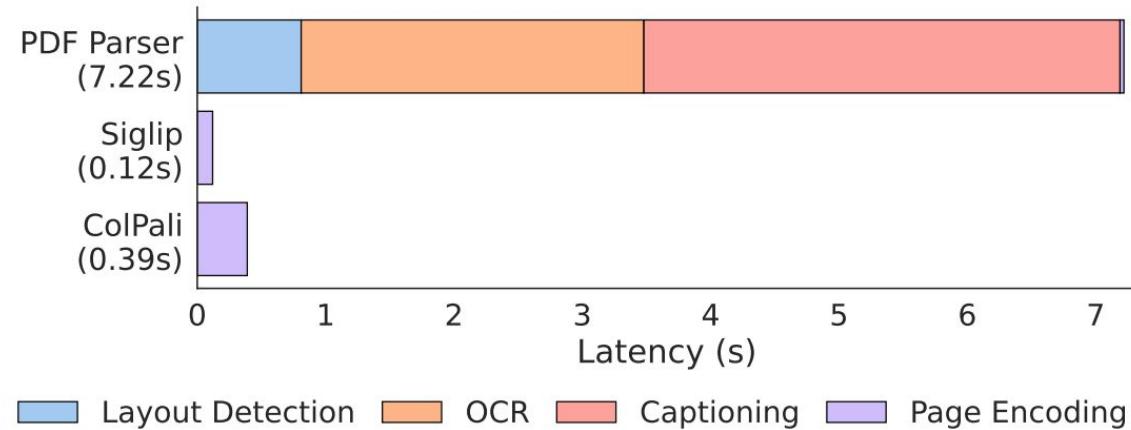


Figure 3: Offline indexing with *ColPali* is much simpler and faster compared to standard retrieval methods. Indexing speeds reported are computed on Nvidia L4 GPUs and detailed in subsection B.5.

ColPali : utiliser les VLM pour du retrieval de document

ColPali retourne des contextes pertinents

Sur le ViDoRe benchmark (représentatif des documents industriels), ColPali est **+14.1** points au-dessus du meilleur retriever textuel.

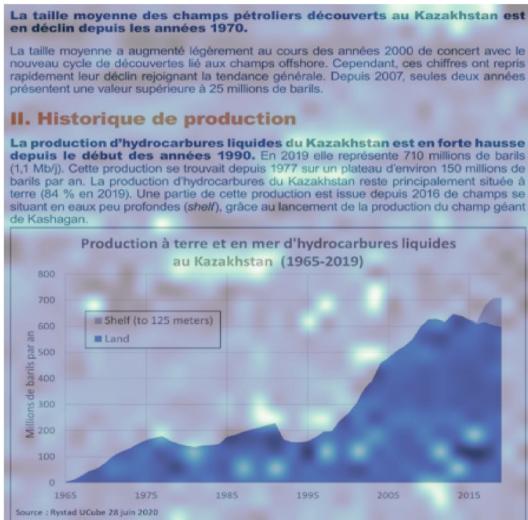
	ArxivQ	DocQ	InfoQ	TabF	TATQ	Shift	AI	Energy	Gov.	Health.	Avg.
Unstructured Text only											
- BM25	-	34.1	-	-	44.0	59.6	90.4	78.3	78.8	82.6	-
- BGE-M3	-	28.4 _{↓5.7}	-	-	36.1 _{↓7.9}	68.5 _{↑8.9}	88.4 _{↓2.0}	76.8 _{↓1.5}	77.7 _{↓1.1}	84.6 _{↑2.0}	-
Unstructured + OCR											
- BM25	31.6	36.8	62.9	46.5	62.7	64.3	92.8	85.9	83.9	87.2	65.5
- BGE-M3	31.4 _{↓0.2}	25.7 _{↓11.1}	60.1 _{↓2.8}	70.8 _{↑24.3}	50.5 _{↓12.2}	73.2 _{↑8.9}	90.2 _{↓2.6}	83.6 _{↓2.3}	84.9 _{↑1.0}	91.1 _{↑3.9}	66.1 _{↑0.6}
Unstructured + Captioning											
- BM25	40.1	38.4	70.0	35.4	61.5	60.9	88.0	84.7	82.7	89.2	65.1
- BGE-M3	35.7 _{↓4.4}	32.9 _{↓5.4}	71.9 _{↑1.9}	69.1 _{↑33.7}	43.8 _{↓17.7}	73.1 _{↑12.2}	88.8 _{↑0.8}	83.3 _{↓1.4}	80.4 _{↓2.3}	91.3 _{↑2.1}	67.0 _{↑1.9}
Contrastive VLMs											
Jina-CLIP	25.4	11.9	35.5	20.2	3.3	3.8	15.2	19.7	21.4	20.8	17.7
Nomic-vision	17.1	10.7	30.1	16.3	2.7	1.1	12.9	10.9	11.4	15.7	12.9
SigLIP (Vanilla)	43.2	30.3	64.1	58.1	26.2	18.7	62.5	65.7	66.1	79.1	51.4
Ours											
SigLIP (Vanilla)	43.2	30.3	64.1	58.1	26.2	18.7	62.5	65.7	66.1	79.1	51.4
BiSigLIP (+fine-tuning)	58.5 _{↑15.3}	32.9 _{↑2.6}	70.5 _{↑6.4}	62.7 _{↑4.6}	30.5 _{↑4.3}	26.5 _{↑7.8}	74.3 _{↑11.8}	73.7 _{↑8.0}	74.2 _{↑8.1}	82.3 _{↑3.2}	58.6 _{↑7.2}
BiPali (+LLM)	56.5 _{↓1.2}	30.0 _{↓1.2}	67.4 _{↓1.3}	76.9 _{↑14.2}	33.4 _{↑2.9}	43.7 _{↑17.7}	71.2 _{↓1.1}	61.9 _{↓1.7}	73.8 _{↓0.4}	73.6 _{↓8.8}	58.8 _{↑0.2}
ColPali (+Late Inter.)	79.1_{↑22.6}	54.4_{↑24.5}	81.8_{↑14.4}	83.9_{↑7.0}	65.8_{↑32.4}	73.2_{↑29.5}	96.2_{↑25.0}	91.0_{↑29.1}	92.7_{↑18.9}	94.4_{↑20.8}	81.3_{↑22.5}

Table 2: Comprehensive evaluation of baseline models and our proposed method on ViDoRe. Results are presented using NDCG@5 metrics, and illustrate the impact of different components. Text-only metrics are not computed for benchmarks with only visual elements.

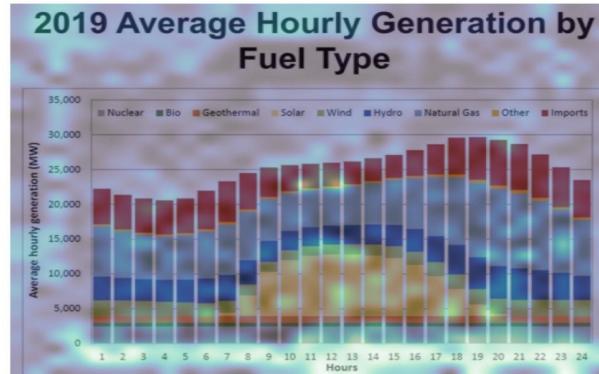
ColPali : utiliser les VLM pour du retrieval de document

ColPali est interprétable

En regardant la similarité des vecteurs représentant patch d'image du Vision Transformer (ViT) et token, on peut tracer une similarity map montrant la contribution de chaque patch au score de retrieval.



Query: "Quelle partie de la production pétrolière du Kazakhstan provient de champs en mer ?"



Query: "Which hour of the day had the highest overall electricity generation in 2019?"



ColPali : Vers du RAG visuel sans OCR

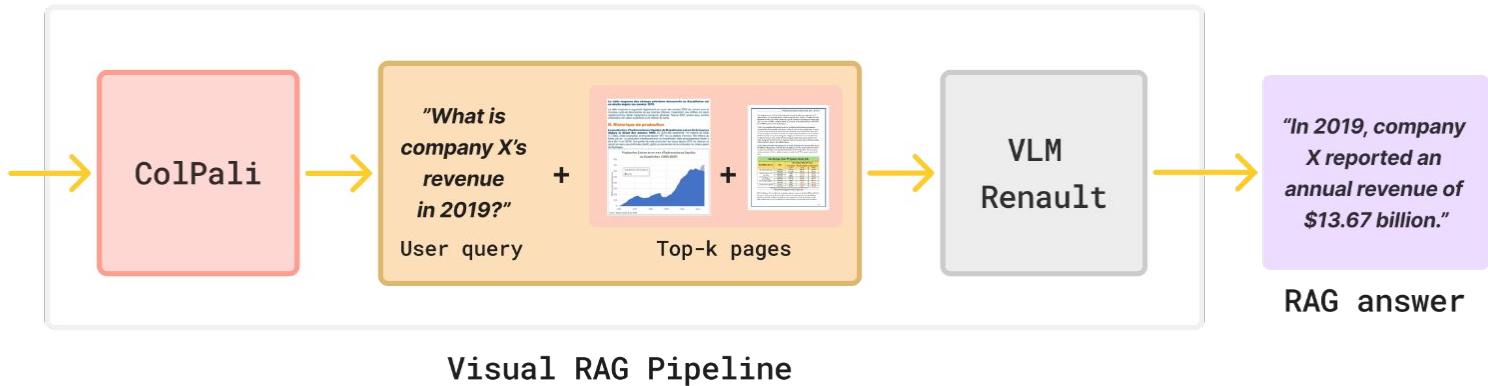
ColPali + VLM = Pipeline de RAG Visuel

En donnant au VLM les top-k images retournées par ColPali, il est possible de créer une pipeline de RAG visuelle simple et capturant tout le signal du document d'origine.

User query:

"What is company X's
revenue in 2019?"

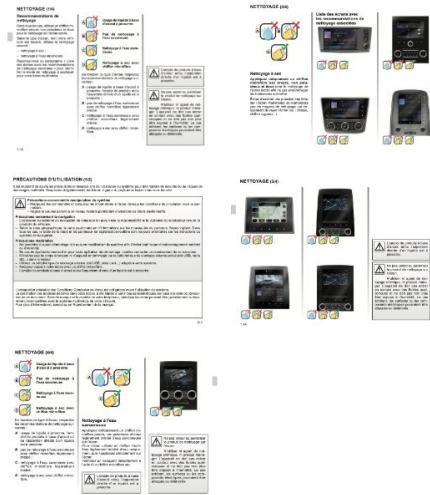
+



Exemple de RAG visuel

User

Quel nettoyage est recommandé pour les différents écrans de la voiture ?



CE QUE VOIT LE VLM

Query utilisateur + top-k pages

Assistant

Pour nettoyer les différents écrans de votre véhicule, voici les recommandations :

1. **Écran tactile** :

- Utiliser un chiffon microfibre propre, non pelucheux et doux.
- **Nettoyage à sec** (D) : Appliquer uniquement un chiffon microfibre sec.
- Évitez d'exercer une pression trop forte sur l'écran.

2. **Écrans nécessitant nettoyage à l'eau savonneuse** (C) :

- Utiliser un chiffon microfibre légèrement imbibé d'eau savonneuse.
- Pour rincer, utiliser un chiffon microfibre légèrement imbibé d'eau uniquement, puis passer délicatement sur l'écran.
- Terminer en essuyant délicatement avec un chiffon microfibre sec.

À proscrire :

- **Usage de liquides à base d'alcool** (A).
- **Aspersion directe de liquide** sur l'écran.
- **Agents chimiques ou produits ménagers**.
- Ne pas verser ou pulvériser de produit de nettoyage.

Respectez ces méthodes selon le type d'écran pour éviter d'endommager votre système multimédia.

RÉPONSE DU VLM

Exemple avec GPT-4o





ColPali : Efficient Document Retrieval with Vision Language Models

An Industrial and Academic Partnership



Manuel Faysse

CentraleSupélec, MICS

Illuin Technology

Doctorant CIFRE



Hugues Sibille

ETH Zurich

Illuin Technology

Research Scientist



Tony Wu

Illuin Technology

Research Scientist



Bilel Omrani

Illuin Technology

Research Scientist



Gautier Viaud

Illuin Technology

Head of R&D



Céline Hudelot

CentraleSupélec, MICS

Directrice du MICS



Pierre Colombo

CentraleSupélec, MICS

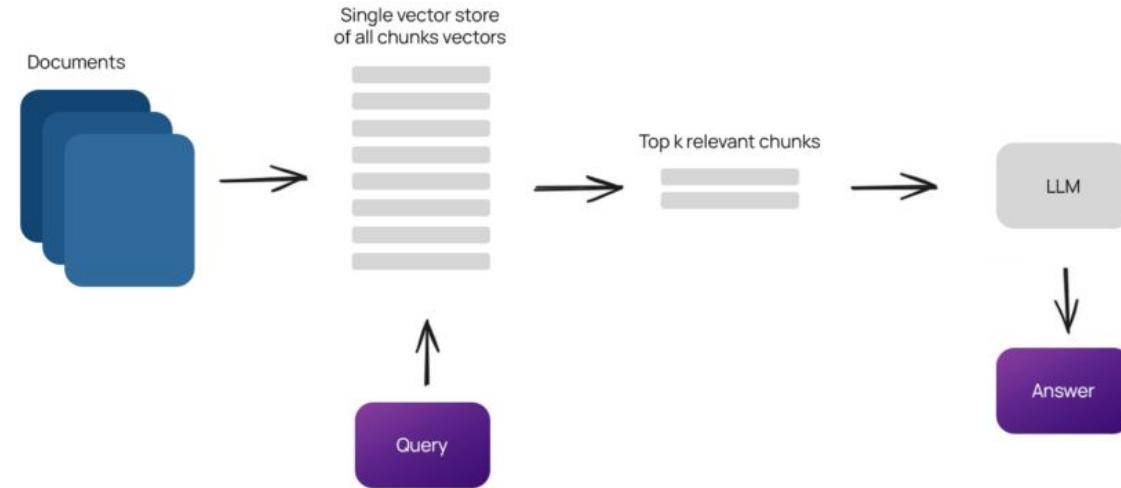
Maître de conférence

RAG: The #1 business use case of Generative AI

RAG is powerful...

Retrieval Augmented Generation

generates responses to user queries by first searching for relevant information in a given document corpus.



...but challenging

The **retrieval** phase is essential to performance but current systems are limited in multiple ways...

Text Embedding Models: A Primer

Text embedding models attempt to represent a text passage in vector space, in a way it can be stored and easily matched to incoming queries.

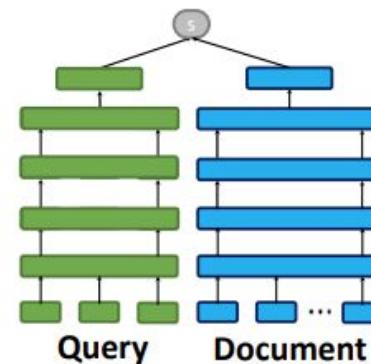
Statistical

Approaches based on **word frequency** statistics. Most common is **BM25**, a simple statistical heuristic many modern methods struggle to beat!

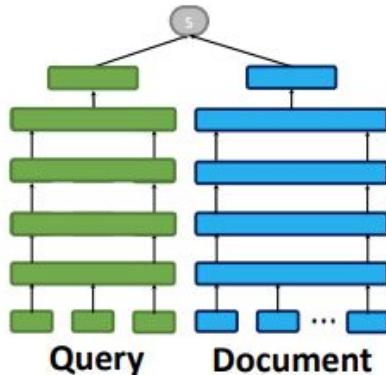
$$\text{score}(D, Q) = \sum_{i=1}^n \text{IDF}(q_i) \cdot \left[\frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)} + \delta \right]$$

Neural

Language Models are used to create a **dense representation** of a text passage. They are often trained contrastively.

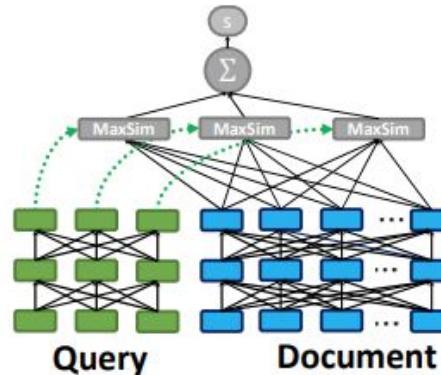


Neural Retrieval can take multiple forms



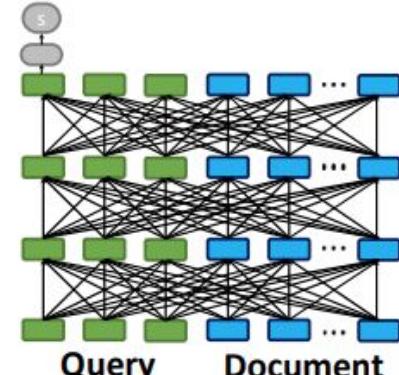
Bi-encoders

- Fast and simple
- Indexing can be offline



Late-interaction

- Richer signal (multi-vector)
- Indexing can be offline



Cross-encoders

- Very rich signal
- Indexing is online and expensive

Late Interaction: A Primer

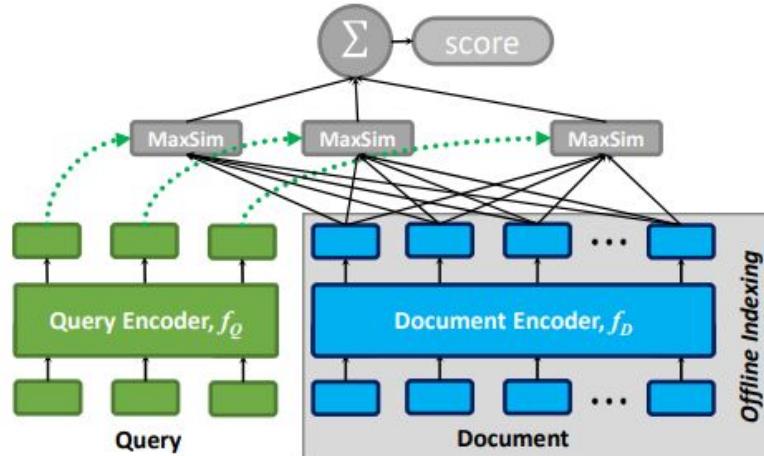


Figure 3: The general architecture of ColBERT given a query q and a document d .

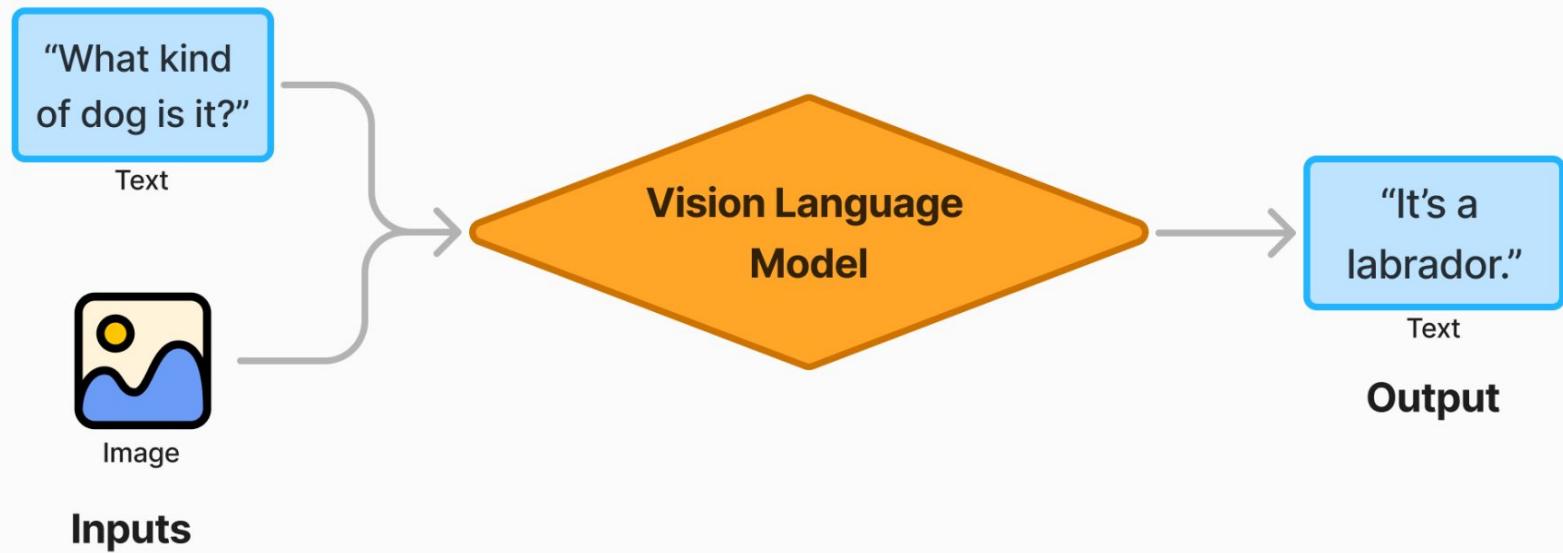
ColBERT

ColBERT stores multiple representations of reduced size per document. During querying, each query token is matched to the most relevant document vector and scores are aggregated per document.

This enables **fast matching**, but also **rich interaction** between query and document tokens.

Vision Language Models: A Primer

Inputs/outputs of a Vision Language Model



Vision Language Models: A Primer

Prompt: Caption this image.



“In this image we can see birds on the wall. In the background, we can see a building with windows and plants in the pots.”

Captioning

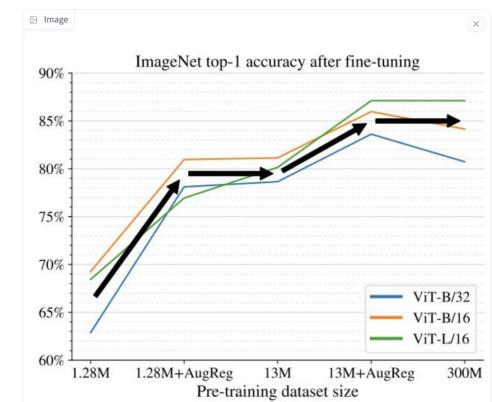
Prompt: Parse this receipt as JSON.



```
{
  "items": [
    {
      "name": "Bintang Bremer",
      "Quantity": 1,
      "Price": 59000
    },
    {
      "name": "Chicken H-H",
      "Quantity": 1,
      "Price": 190000
    },
    {
      "name": "Ades",
      "Quantity": 1,
      "Price": 10000
    }
  ],
  "Subtotal": 259000,
  "Service": 9600,
  "Tax": 52416,
  "Discount": 19000,
  "Total": 302016
}
```

**OCR + Layout
Understanding**

Prompt: What does this image show?



“vit-b/32, vit-1/16, and vit-b/16+augreg are trained on different sizes of training data.”

**Vision Question
Answering & Chart
Understanding**

Vision Language Models: A Primer

The standard VLM architecture

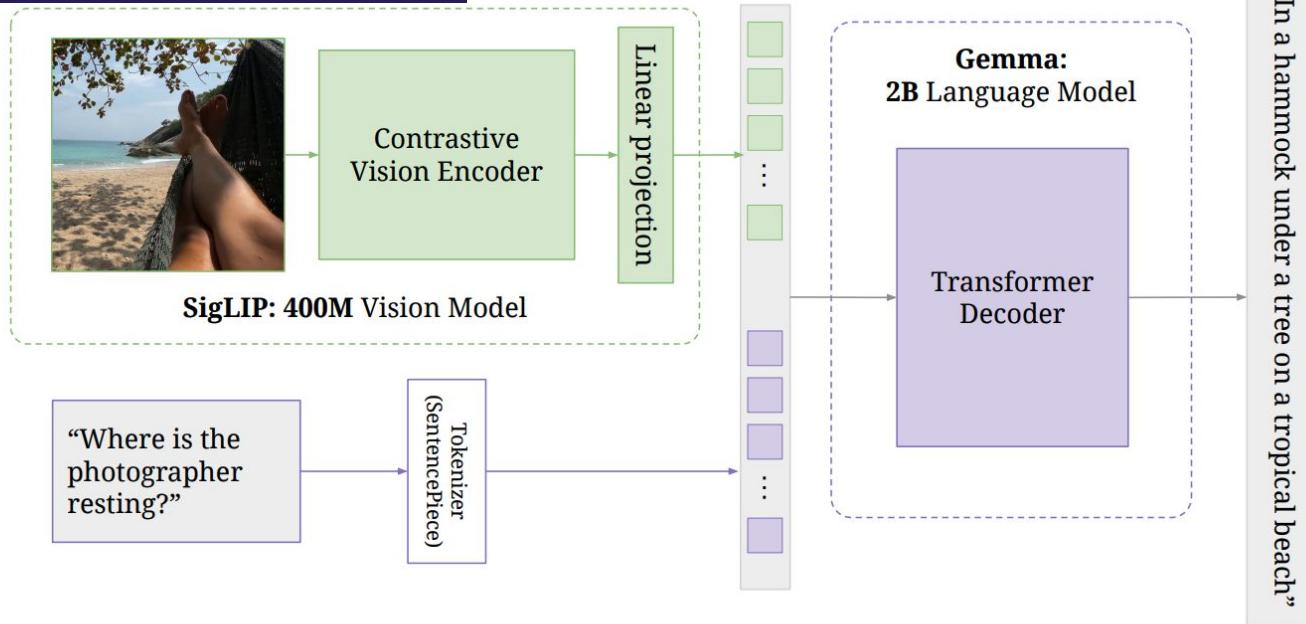


Figure 1 | PaliGemma's architecture: a SigLIP image encoder feeds into a Gemma decoder LM.

Our Approach: Retrieving in Vision Space

We propose to skip document parsing entirely, by working with document “screenshots”.

Standard Pipelines work with extracted textual content

Complex pipelines are necessary to extract text, detect the document layout, caption visual elements, embed the textual content with text embedding models...

vs.

Can we train retrievers that work using the document image ?

Vision Language Models are used to create representations of the image of each document page directly !

Document Retrieval from images is a completely novel concept... does it work well ?



Capable



Faster



Trainable

Evaluation : The ViDoRe Benchmark

We introduce **ViDoRe**, the Visual Document Retrieval Benchmark, to assess retrievers on their capacity to **retrieve visually rich information** in docs, with tasks spanning various **topics, modalities** (figures, tables, text), and **languages!**

Dataset	# Queries	Domain
Academic Tasks		
DocVQA (eng)	500 (500)	Industrial
InfoVQA (eng)	500 (500)	Infographics
TAT-DQA (eng)	1600 (1600)	Varied Modalities
arXiVQA (eng)	500 (500)	Scientific Figures
TabFQuAD (fra)	210 (210)	Tables
Practical Tasks		
Energy (eng)	100 (1000)	Scientific
Government (eng)	100 (1000)	Administrative
Healthcare (eng)	100 (1000)	Medical
AI (eng)	100 (1000)	Scientific
Shift Project (fra)	100 (1000)	Environment

Table 1: *ViDoRe* comprehensively evaluates multimodal retrieval methods. The size of the document corpus is indicated in parentheses.

Query-page pair examples

TAX EFFICIENCY



Tax-efficient investing

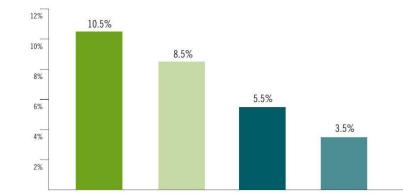
Taxes have the potential to significantly affect your investment returns.

One way to help reach your financial goals is to be tax smart with your investments. You can affect your tax bill by paying attention to how and where you generate investment income, dividends, interest, and capital gains and losses. There are three strategies you can use to try to manage the potential impact on your federal income taxes:

- **Defer:** Retirement savings accounts—including 401(k) and 403(b) plans, IRAs, health savings accounts (HSA), and other tax-deferred products such as deferred annuities—all allow you to put off paying taxes.
- **Manage:** Using asset location strategies, investing in lower turnover funds, understanding mutual fund distributions, and taking advantage of charitable gifts and capital loss deductions can all help you manage your tax burden.
- **Reduce:** Consider tax-free investments, municipal bonds, HSAs, and college savings accounts to help reduce your taxes.

TAXES CAN SIGNIFICANTLY REDUCE RETURNS

Average annual return %



Investment Type	Average Annual Return (%)
Stocks	10.5%
Stocks after taxes	8.5%
Bonds	5.5%
Bonds after taxes	3.5%

*Past performance is no guarantee of future results. This chart is for illustrative purposes only and does not represent actual investment performance. The chart shows the effect of taxes on investment returns. The chart does not reflect the fact that stocks assume that the stocks purchased were held for five years, then sold, and the capital gains realized. The net proceeds from the sale were reinvested. Dividends were taxed when earned and reinvested. From 1926 to 2021, the average return on stocks was 10.5% before taxes. After taxes, the average return was 8.5%. The average return on bonds was 5.5% before taxes. After taxes, the average return was 3.5%. Comparing the after-tax return to the rate of return, you can see that if you invested solely in cash equivalents, you actually lost money in terms of purchasing power. Government bonds and interest, while stock are not guaranteed and have been more volatile than the other asset classes.

© 2022 Morningstar and Precidian Information, dba Financial Fitness Group 2022. All rights reserved.

INVESTMENT STRATEGY

Name of EBP

Technology-Aided Instruction and Intervention (TAII)

Technology-Aided Instruction and Interventions (TAII) are those in which technology is the central feature of an intervention. Given the rapid rise in the inclusion of technology in interventions, this evidence base is intended to include technology that is specifically designed or employed to support learning or the delivery of a skill or learning intervention. Interventions that use a more general form of technology to deliver an alternative EBP (e.g., displaying a visual support on a mobile device, video modeling, alarm on a phone as part of self-management) are not included in this evidence base. TAII includes technologies such as robots, computers, web-based software, applications for devices, and virtual networks. The common features of these interventions are the technology itself (as noted) and instructional procedures for learning to use the technology or supporting its use in appropriate contexts.

- **Technology-Aided Interventions Meeting Criteria:** MindReading software, FaceSay™ (Symbionica, LLC) software
- **Note:** Augmentative and alternative communication (AAC) that incorporates technology is part of the evidence base for AAC and not TAII.

Age Ranges

Age Range	0-3	3-5	6-11	12-14	15-18	19-22
Toddlers	✓	✓	✓	✓	✓	✓
Preschoolers		✓	✓	✓	✓	✓
Elementary School		✓	✓	✓	✓	✓
Middle School			✓	✓	✓	✓
High School				✓	✓	✓
Young Adults					✓	✓

Outcome Areas

Outcome Area	Communication	Social	Joint attention	Play	Cognitive	School readiness	Academic/ Pre-academic	Adaptive/ self-help	Challenging/ Interfering behavior	Vocational	Motor	Mental health	Self-determination
Communication	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Social		✓	✓	✓		✓	✓		✓	✓			✓
Joint attention		✓	✓	✓			✓						
Play	✓	✓	✓	✓									
Cognitive	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
School readiness		✓	✓	✓	✓	✓	✓						
Academic/ Pre-academic		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Adaptive/ self-help	✓	✓	✓	✓									
Challenging/ Interfering behavior			✓	✓					✓	✓			
Vocational													
Motor	✓	✓	✓	✓									
Mental health			✓	✓	✓	✓	✓						
Self-determination													

References

1. Bailey, B., Arculli, J., & Stancliffe, R. J. (2017a). Effects of ABACADABRA instruction on spelling in children with autism spectrum disorder. *Scientific Studies of Reading*, 21(2), 146-164. <https://doi.org/10.1080/10888438.2016.1276618>
2. Bailey, B., Arculli, J., & Stancliffe, R. J. (2017b). Effects of ABACADABRA literacy instruction on children with autism spectrum disorder. *Journal of Educational Psychology*, 109(2), 257-268. <https://doi.org/10.1037/a0040318>
3. Beaumont, R., & Schonoff, K. (2008). A multi-component social skills intervention for children with Asperger syndrome: The Junior Detective Training Program. *Journal of Child Psychology and Psychiatry*, 49(7), 743-753. <https://doi.org/10.1111/j.1469-7610.2008.01920.x>

129 | Evidence-Based Practices for Children, Youth, and Young Adults with Autism

What types of accounts or products allow investors to defer paying taxes?

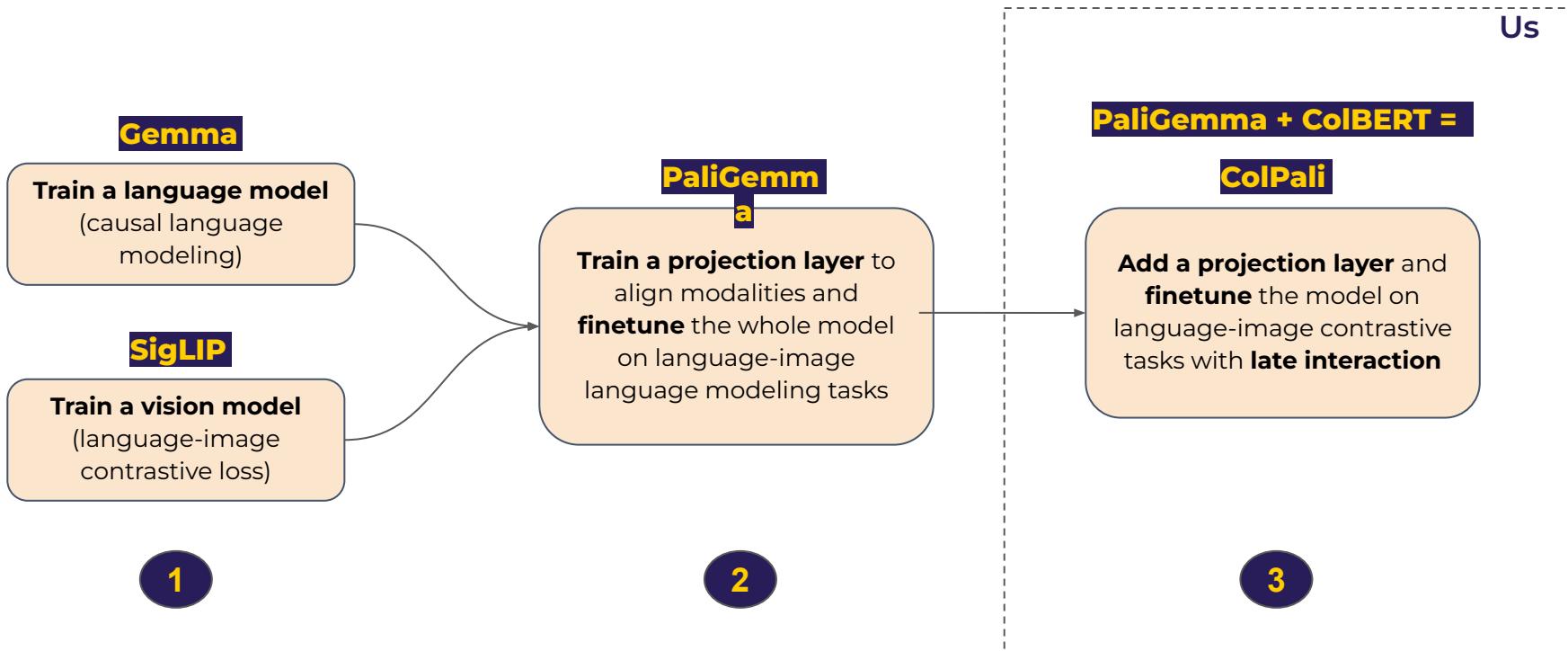
What are some common outcome areas targeted by TAII for different age groups?

Existing Methods on ViDoRe

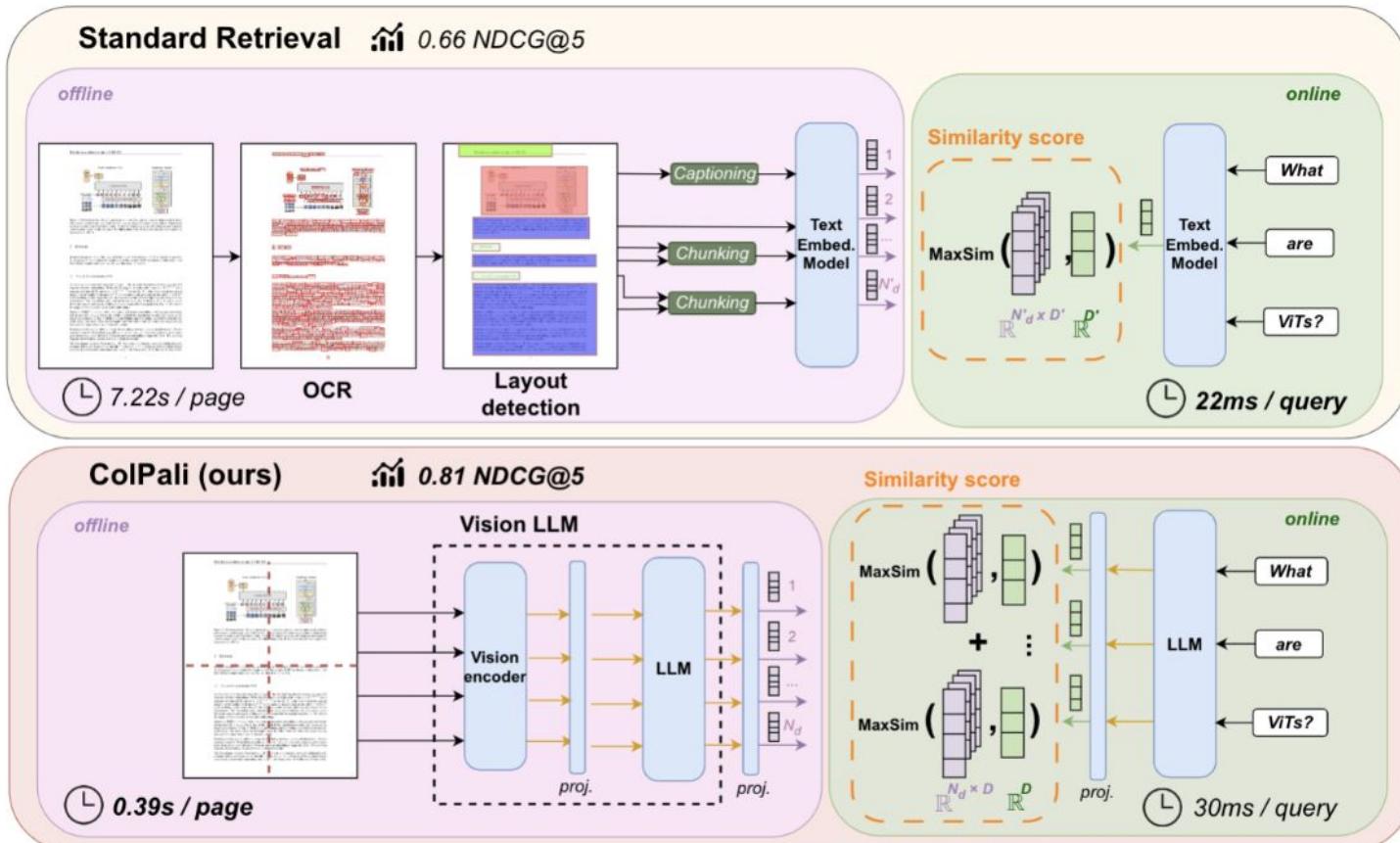
	<u>ArxivQ</u>	<u>DocQ</u>	<u>InfoQ</u>	<u>TabF</u>	<u>TATQ</u>	<u>Shift</u>	<u>AI</u>	<u>Energy</u>	<u>Gov.</u>	<u>Health.</u>	<u>Avg.</u>
Unstructured <small>Text only</small>											
- BM25	-	34.1	-	-	44.0	59.6	90.4	78.3	78.8	82.6	-
- BGE-M3	-	28.4 _{↓5.7}	-	-	36.1 _{↓7.9}	68.5 _{↑8.9}	88.4 _{↓2.0}	76.8 _{↓1.5}	77.7 _{↓1.1}	84.6 _{↑2.0}	-
Unstructured <small>+ OCR</small>											
- BM25	31.6	36.8	62.9	46.5	62.7	64.3	92.8	85.9	83.9	87.2	65.5
- BGE-M3	31.4 _{↓0.2}	25.7 _{↓11.1}	60.1 _{↓2.8}	70.8 _{↑24.3}	50.5 _{↓12.2}	73.2 _{↑8.9}	90.2 _{↓2.6}	83.6 _{↓2.3}	84.9 _{↑1.0}	91.1 _{↑3.9}	66.1 _{↑0.6}
Unstructured <small>+ Captioning</small>											
- BM25	40.1	38.4	70.0	35.4	61.5	60.9	88.0	84.7	82.7	89.2	65.1
- BGE-M3	35.7 _{↓4.4}	32.9 _{↓5.4}	71.9 _{↑1.9}	69.1 _{↑33.7}	43.8 _{↓17.7}	73.1 _{↓2.2}	88.8 _{↑0.8}	83.3 _{↓1.4}	80.4 _{↓2.3}	91.3 _{↑2.1}	67.0 _{↑1.9}
Contrastive VLMs											
Jina-CLIP	25.4	11.9	35.5	20.2	3.3	3.8	15.2	19.7	21.4	20.8	17.7
Nomic-vision	17.1	10.7	30.1	16.3	2.7	1.1	12.9	10.9	11.4	15.7	12.9
SigLIP <small>(Vanilla)</small>	43.2	30.3	64.1	58.1	26.2	18.7	62.5	65.7	66.1	79.1	51.4

VLMs have a lot of benefits but still lag behind more complex pipelines on document retrieval tasks... Can we do better ?

Stages of Training



Zoom on ColPali



Speed benchmarking at indexing time

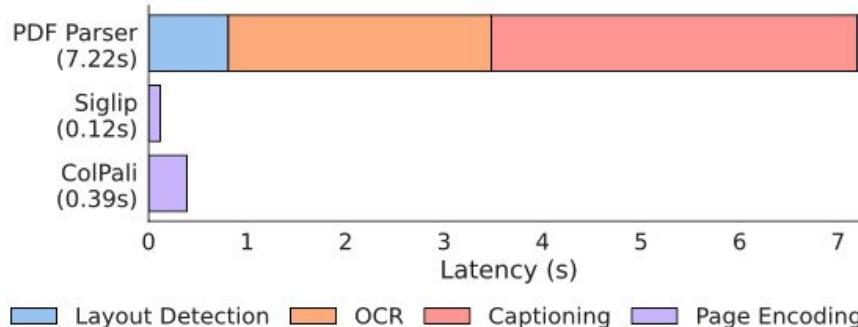


Figure 3: Offline indexing with *ColPali* is much simpler and faster compared to standard retrieval methods. Indexing speeds reported are computed on Nvidia L4 GPUs and detailed in subsection B.5.

ColPali is simple

Encoding document is done with a simple **forward** pass.

ColPali is fast

Encoding a document can be up to **18x faster** than a complex PDF parsing pipeline with OCR & Captioning.

Benefits of the ColPali model



Performance

ColPali leverages the **visual elements** of a document page **organically**, yielding **better retrieval** performance than standard systems on visually rich documents!



Speed

During **indexing**, page embeddings are created in a **single forward pass**.

During **querying**, the late interaction mechanisms enable **immediate matching** using simple dot product and sum operations.



Trainable

ColPali is an **end-to-end** retrieval method, and can be **finetuned** very easily to **adapt** to new languages, document types, modalities...

Retriever Fine-tuning

Corpus



Documents

Repurposed VQA datasets

Charts
Figures
Tables
Infographics
Text - centric

Queries correspond to the original questions



Internet

Public scrapped PDFs with large coverage

Scientific
Medical
Healthcare
Industrial
Tables

Queries are synthesized with Claude Sonnet

Training

We train ColPali for industry-oriented retrieval use cases with a dataset of **127 460** documents.

Models are trained using an **in-batch contrastive objective** for 1 epoch, on 8xMi250x AMD GPUs on Adastra, a French government owned supercomputer. A training run roughly lasts 5 hours.

The model outperforms very strong baselines

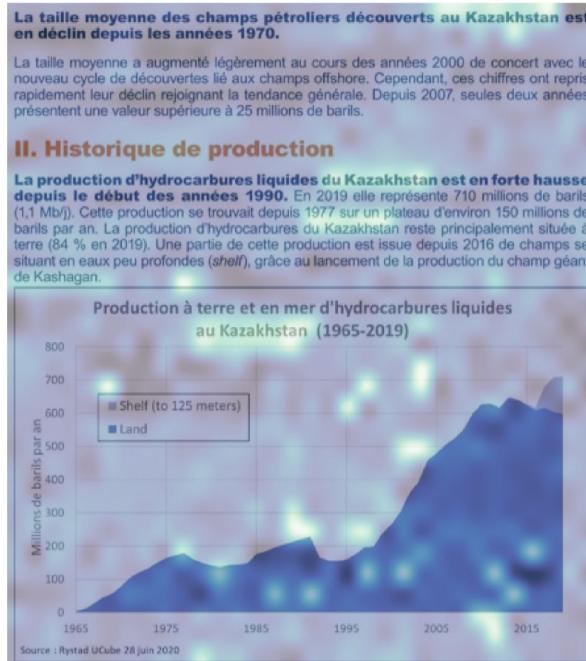
	ArxivQ	DocQ	InfoQ	TabF	TATQ	Shift	AI	Energy	Gov.	Health.	Avg.
Unstructured Text only											
- BM25	-	34.1	-	-	44.0	59.6	90.4	78.3	78.8	82.6	-
- BGE-M3	-	28.4 _{↓5.7}	-	-	36.1 _{↓7.9}	68.5 _{↑8.9}	88.4 _{↓2.0}	76.8 _{↓1.5}	77.7 _{↓1.1}	84.6 _{↑2.0}	-
Unstructured + OCR											
- BM25	31.6	36.8	62.9	46.5	62.7	64.3	92.8	85.9	83.9	87.2	65.5
- BGE-M3	31.4 _{↓0.2}	25.7 _{↓11.1}	60.1 _{↓2.8}	70.8 _{↑24.3}	50.5 _{↓12.2}	73.2 _{↑8.9}	90.2 _{↓2.6}	83.6 _{↓2.3}	84.9 _{↑1.0}	91.1 _{↑3.9}	66.1 _{↑0.6}
Unstructured + Captioning											
- BM25	40.1	38.4	70.0	35.4	61.5	60.9	88.0	84.7	82.7	89.2	65.1
→ - BGE-M3	35.7 _{↓4.4}	32.9 _{↓5.4}	71.9 _{↑1.9}	69.1 _{↑33.7}	43.8 _{↓17.7}	73.1 _{↑12.2}	88.8 _{↑0.8}	83.3 _{↓1.4}	80.4 _{↓2.3}	91.3 _{↑2.1}	67.0 _{↑1.9}
Contrastive VLMs											
Jina-CLIP	25.4	11.9	35.5	20.2	3.3	3.8	15.2	19.7	21.4	20.8	17.7
Nomic-vision	17.1	10.7	30.1	16.3	2.7	1.1	12.9	10.9	11.4	15.7	12.9
SigLIP _(Vanilla)	43.2	30.3	64.1	58.1	26.2	18.7	62.5	65.7	66.1	79.1	51.4
Ours											
SigLIP _(Vanilla)	43.2	30.3	64.1	58.1	26.2	18.7	62.5	65.7	66.1	79.1	51.4
BiSigLIP _(+fine-tuning)	58.5 _{↑15.3}	32.9 _{↑2.6}	70.5 _{↑6.4}	62.7 _{↑4.6}	30.5 _{↑4.3}	26.5 _{↑7.8}	74.3 _{↑11.8}	73.7 _{↑8.0}	74.2 _{↑8.1}	82.3 _{↑3.2}	58.6 _{↑7.2}
BiPali _(+LLM)	56.5 _{↓2.0}	30.0 _{↓2.9}	67.4 _{↓3.1}	76.9 _{↑14.2}	33.4 _{↑2.9}	43.7 _{↑17.2}	71.2 _{↓3.1}	61.9 _{↓11.7}	73.8 _{↓0.4}	73.6 _{↓8.8}	58.8 _{↑0.2}
→ ColPali _(+Late Inter.)	79.1 _{↑22.6}	54.4 _{↑24.5}	81.8 _{↑14.4}	83.9 _{↑7.0}	65.8 _{↑32.4}	73.2 _{↑29.5}	96.2 _{↑25.0}	91.0 _{↑29.1}	92.7 _{↑18.9}	94.4 _{↑20.8}	81.3 _{↑22.5}

Table 2: **Comprehensive evaluation of baseline models and our proposed method on ViDoRe.** Results are presented using NDCG@5 metrics, and illustrate the impact of different components. Text-only metrics are not computed for benchmarks with only visual elements.

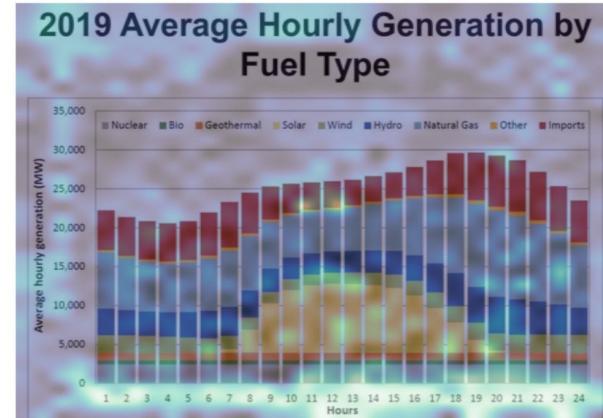
An interpretable mechanism

With the **late interaction** matching mechanism, we can observe which parts of the documents were most relevant with respect to the each token of the query.

This enables **explaining model decisions** and helps model improvement workflows.



Query: "Quelle partie de la production pétrolière du Kazakhstan provient de champs en mer ?"



Query: "Which hour of the day had the highest overall electricity generation in 2019?"

An Open-Source Project

Project rooted in transparency, to serve as a useful resource for industrial practitioners and researchers!



Documented training process from beginning to end



Openly available:

- Model checkpoints
- Leaderboard
- Evaluation Benchmarks
- Code bases



No usage restrictions (MIT)

[Leaderboard](#) [Submit your model](#)

ViDoRe: The Visual Document Retrieval Benchmark

From the paper - ColPali: Efficient Document Retrieval with Vision Language Models [++](#)

Visual Document Retrieval Benchmark leaderboard. To submit results, refer to the corresponding tab.

Refer to the [ColPali paper](#) for details on metrics, tasks and models.

Select Metric		Filter Models by Name		Select Columns to Display											
<input type="text" value="ndcg_at_5"/>		<input type="text" value="Search Models... [press enter]"/>		<input checked="" type="checkbox"/> Energy <input checked="" type="checkbox"/> Healthcare Industry <input checked="" type="checkbox"/> Artificial Intelligence <input checked="" type="checkbox"/> Government Reports <input checked="" type="checkbox"/> InfoVQA <input checked="" type="checkbox"/> DocVQA <input checked="" type="checkbox"/> ArxivQA <input checked="" type="checkbox"/> TabQuad <input checked="" type="checkbox"/> TAT-DQA <input checked="" type="checkbox"/> Shift Project											
Rank	Model	Average	Energy	Healthcare Industry	Artificial Intelligence	Government Reports	InfoVQA	DocVQA	ArxivQA	TabQuad	TAT-DQA	Shift Project			
1	vidore_colpali	81.3	91	94.4	96.2	92.7	81.8	54.4	79.1	83.9	65.8	73.2			
2	vidore_colidefics	76.6	86.9	91.2	95.5	86.6	82.4	48	73.6	81.6	63	57.2			
3	Tevatron-DSE-Phis3-V1(ZeroShot	68.8	88.6	88.5	89	89.2	77.5	39.5	70.4	64.2	40.7	39.8			
4	BAAT_bgg-m3_captioning	67	83.3	91.3	88.8	80.4	71.9	32.9	35.7	69.1	43.8	73.1			
5	BAAT_bgg-m3_ocr	66.1	83.6	91.1	90.2	84.9	60.1	25.7	31.4	70.8	50.5	73.2			
6	bm25_ocr	65.5	85.9	87.2	92.8	83.9	62.9	36.8	31.6	46.5	62.7	64.3			
7	bm25_captioning	65.1	84.7	89.2	88	82.7	70	38.4	40.1	35.4	61.5	60.9			
8	vidore_binali	58.8	61.9	73.6	71.2	73.8	67.4	30	56.5	76.9	33.4	43.7			
9	vidore_bisiglip	58.6	73.7	82.3	74.3	74.2	70.5	32.9	58.5	62.7	30.5	26.5			
10	google_siglip-so400m-patch14-	51.4	65.7	79.1	62.5	66.1	64.1	30.3	43.2	58.1	26.2	18.7			
11	jinaai_jina-clip-v1	17.7	19.7	20.8	15.2	21.4	35.5	11.9	25.4	20.2	3.3	3.8			



Use cases: Where does the model shine ?



Complex Documents

ColPali particularly shines when retrieving complex documents (forms, figures, tables, complex formats, etc...)



Online Retrieval

ColPali's indexing speed enables online RAG applications in which users upload their own documents and can query them moments later.



OCR

By not necessitating upstream OCR, ColPali is perfect for scanned document retrieval and is robust to old & damaged documents.



New Settings

ColPali is end-to-end trainable. As such, it can easily be tuned for new languages, tasks or document formats !

Towards end-to-end vision systems

MMLONGBENCH-DOC: Benchmarking Long-context Document Understanding with Visualizations

Yubo Ma¹, Yuhang Zang^{2*}, Liangyu Chen¹, Meiqi Chen³, Yizhu Jiao⁴,
 Xinze Li¹, Xinyuan Lu⁵, Ziyu Liu⁶, Yan Ma⁷, Xiaoyi Dong², Pan Zhang²,
 Liangming Pan⁸, Yu-Gang Jiang⁹, Jiaqi Wang², Yixin Cao^{9,*}, Aixin Sun²

¹ S-Lab, Nanyang Technological University, ² Shanghai AI Laboratory, ³ Peking University, ⁴ University of Illinois Urbana-Champaign, ⁵ National University of Singapore, ⁶ Wuhan University, ⁷ Singapore Management University, ⁸ University of California, Santa Barbara, ⁹ Fudan University

Abstract

Understanding documents with rich layouts and multi-modal components is a long-standing and practical task. Recent Large Vision-Language Models (LVLMs) have made remarkable strides in various tasks, particularly in single-page document understanding (DU). However, their abilities on long-context DU remain an open problem. This work presents **MMLONGBENCH-DOC**, a long-context, multi-modal benchmark comprising 1,091 expert-annotated questions. Distinct from previous datasets, it is constructed upon 135 lengthy PDF-formatted documents with an average of 47.5 pages and 21,214 textual tokens. Towards comprehensive evaluation, answers to these questions rely on pieces of evidence from (1) different sources (text, image, chart, table, and layout structure) and (2) various locations (*i.e.*, page number). Moreover, 33.0% of the questions are *cross-page questions* requiring evidence across multiple pages. 22.5% of the questions are designed to be *unanswerable* for detecting potential hallucinations. Experiments on 14 LVLMs demonstrate that long-context DU greatly challenges current models. Notably, the best-performing model, GPT-4o, achieves an F1 score of only 44.9%, while the second-best, GPT-4V, scores 30.5%. Furthermore, 12 LVLMs (all except GPT-4o and GPT-4V) even present worse performance than their LLM counterparts which are fed with lossy-parsed OCR documents. These results validate the necessity of future research toward more capable long-context LVLMs.

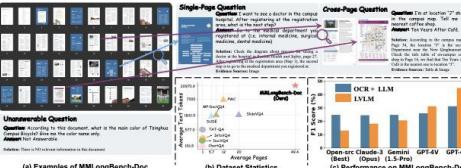


Figure 1: MMLONGBENCH-DOC evaluates understanding abilities of LVLMs on lengthy documents that span tens of pages and incorporate multi-modal elements. Experiments (bottom-right) indicate that most LVLMs struggle, even falling behind LLMs that are fed with only OCR-parsed documents.

If retrieval can be done in **Vision Space**, so can question answering.

MMLongBench-Doc shows GPT4o with document screenshots outperforms GPT4o with text input, confirming our vision for a **fully vision-based RAG system!**

Towards production optimized versions

Token Pooling

Token pooling by clustering similar tokens within a given document and averaging their representation can help **reduce memory** requirements and **speed up retrieving**.



Faster



Memory Requirements

Reranking

To reduce the retrieval latency, we can opt for a 2-step pipeline:

1. **Pre-ranking:** Use a single-vector embedding per document to get a reduced number of candidates (BiPali).
2. **Reranking:** Use ColPali's multi-vector embeddings to rerank top candidates.



Faster