The power of VLMs

Advanced LLM Night

Eivind Kjosbakken – Findable

January 2025

Me

Master's in Engineering and ICT from NTNU

- Main focus on AI
- One year exchange at TU Delft

Me

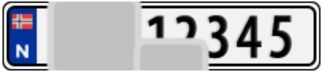
Master's in Engineering and ICT from NTNU

- Main focus on AI
- One year exchange at TU Delft

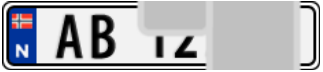
Have been a part of Findable since 2021

- Part-time AI-trainer
- Wrote my master's thesis within ML at Findable
- Full-time data scientist

License Plate Recognition

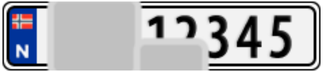


+

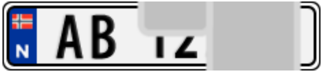


Provide the full license plate number

License Plate Recognition



+



Provide the full license plate number

↓

The full license plate number is AB12345.

Performed with Qwen2-VL-7B-Instruct

Overview

Overview

How vision is integrated in VLMs

Overview

How vision is integrated in VLMs

VLM usecases

Overview

How vision is integrated in VLMs

VLM usecases

- Information extraction

Overview

How vision is integrated in VLMs

VLM usecases

- Information extraction
- Classification

Overview

How vision is integrated in VLMs

VLM usecases

- Information extraction
- Classification

VLM Drawbacks

Overview

How vision is integrated in VLMs

VLM usecases

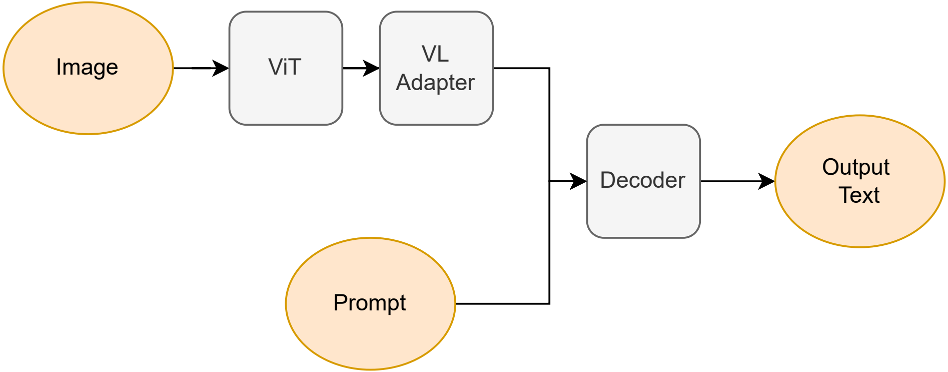
- Information extraction
- Classification

VLM Drawbacks

Staying up to date with VLM advances

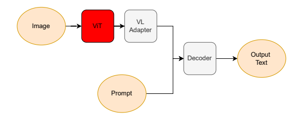
Intro to VLMs

How do they work

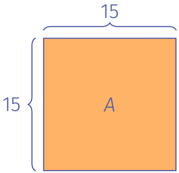


How vision integrates in VLMs - ViT

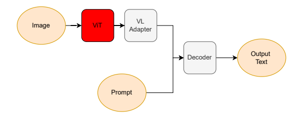
VLM overview



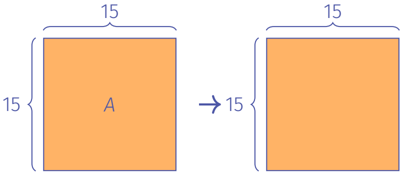
How vision integrates in VLMs - ViT



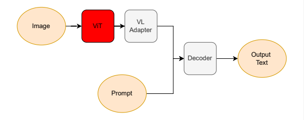
VLM overview



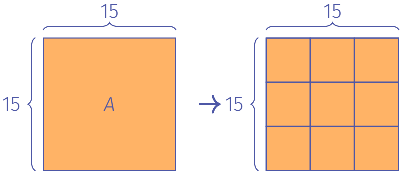
How vision integrates in VLMs - ViT



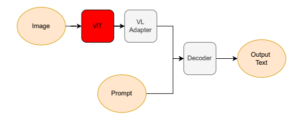
VLM overview



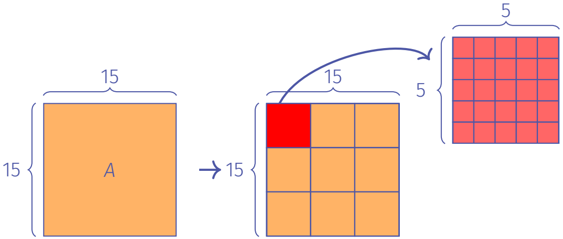
How vision integrates in VLMs - ViT



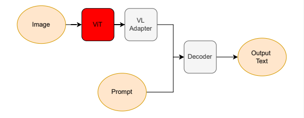
VLM overview



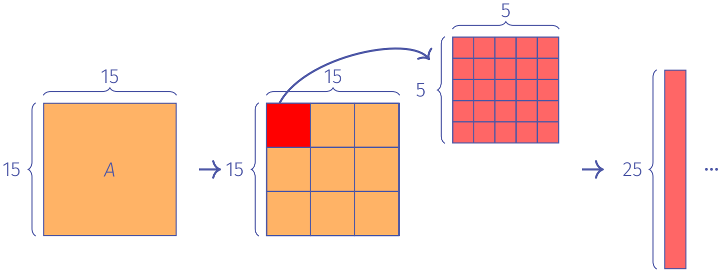
How vision integrates in VLMs - ViT



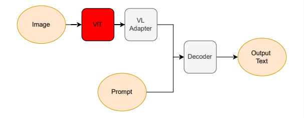
VLM overview



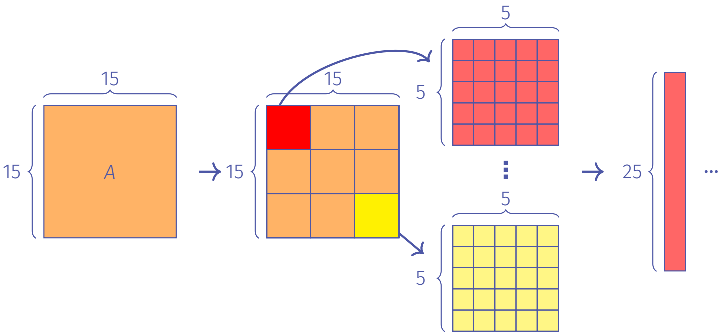
How vision integrates in VLMs - ViT



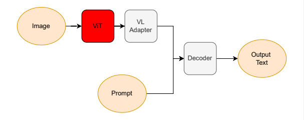
VLM overview



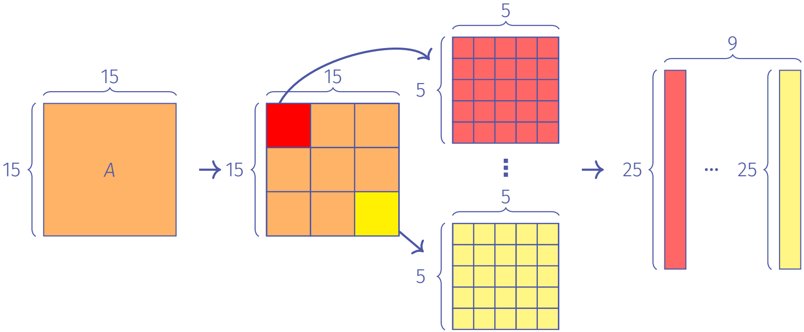
How vision integrates in VLMs - ViT



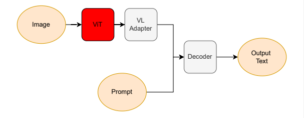
VLM overview



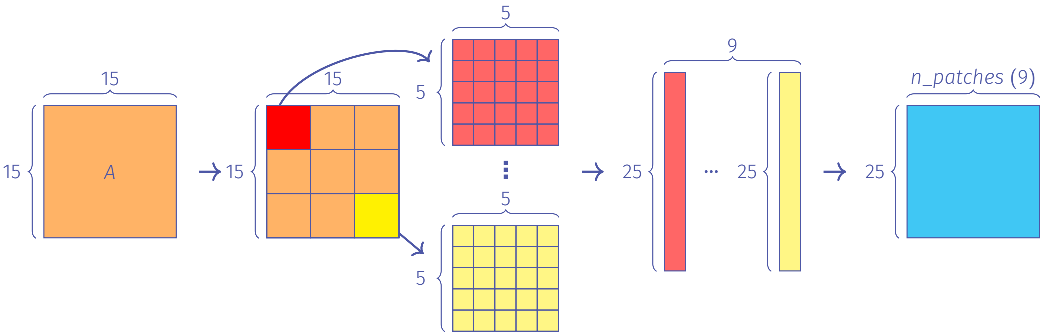
How vision integrates in VLMs - ViT



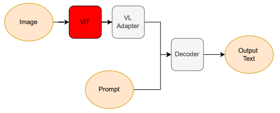
VLM overview



How vision integrates in VLMs - ViT

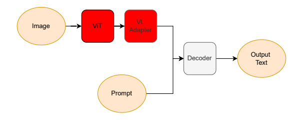


VLM overview

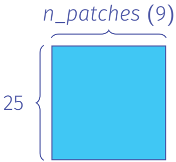


How vision integrates into the LLM - Transformer layers

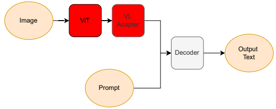
VLM overview



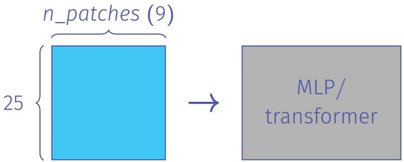
How vision integrates into the LLM - Transformer layers



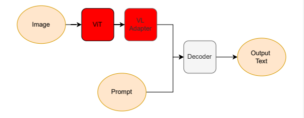
VLM overview



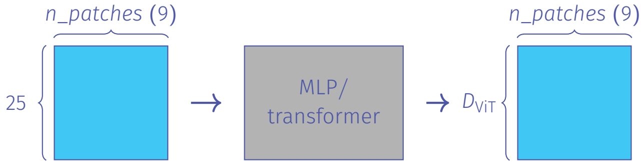
How vision integrates into the LLM - Transformer layers



VLM overview



How vision integrates into the LLM - Transformer layers



$n\_patches$ (9)

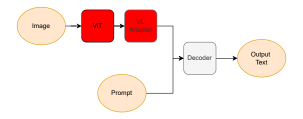
25

MLP/ transformer

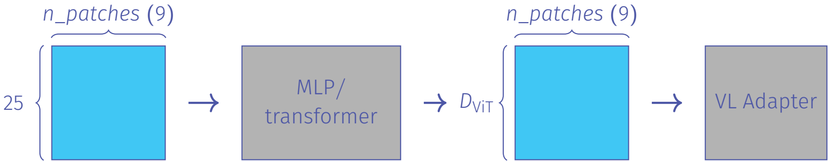
$n\_patches$ (9)

D_{ViT}

VLM overview

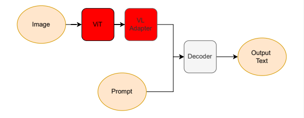


How vision integrates into the LLM - Transformer layers

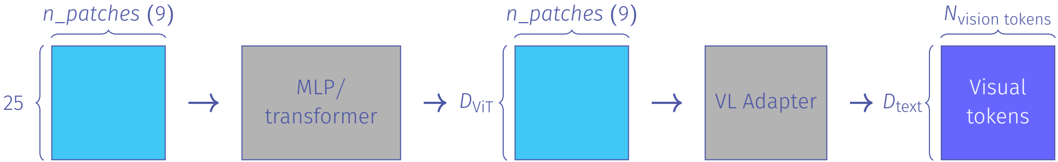


$\overbrace{\quad}^{n\_patches\ (9)}$ 25 → MLP/ transformer → D_{ViT} $\overbrace{\quad}^{n\_patches\ (9)}$ → VL Adapter

VLM overview



How vision integrates into the LLM - Transformer layers



$n\_patches$ (9)

25

MLP/ transformer

D_{ViT}

$n\_patches$ (9)

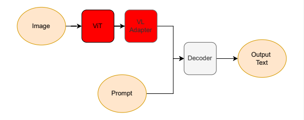
VL Adapter

D_{text}

$N_{\text{vision tokens}}$

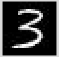
Visual tokens

VLM overview



Decoder input sequence

Prompt

Describe this image 

Decoder input sequence

Prompt

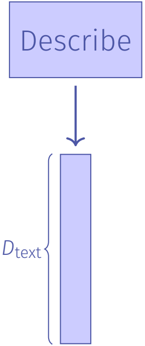
Describe this image

Describe

Decoder input sequence

Prompt

Describe this image



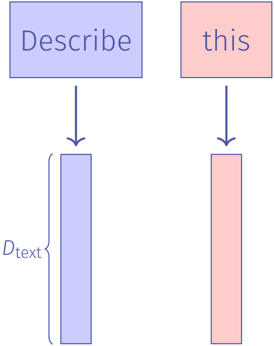
Describe

D_{text}

Decoder input sequence

Prompt

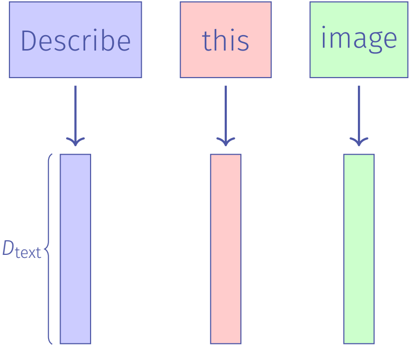
Describe this image



Decoder input sequence

Prompt

Describe this image



Describe

this

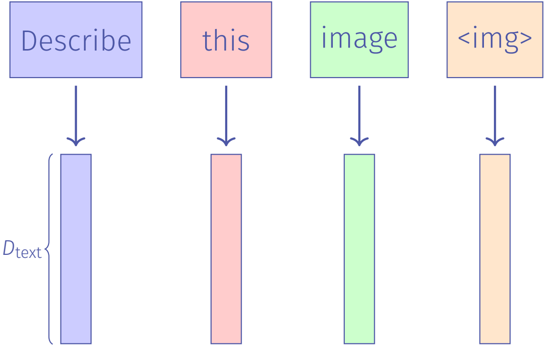
image

D_{text}

DECODER INPUT SEQUENCE

Prompt

Describe this image

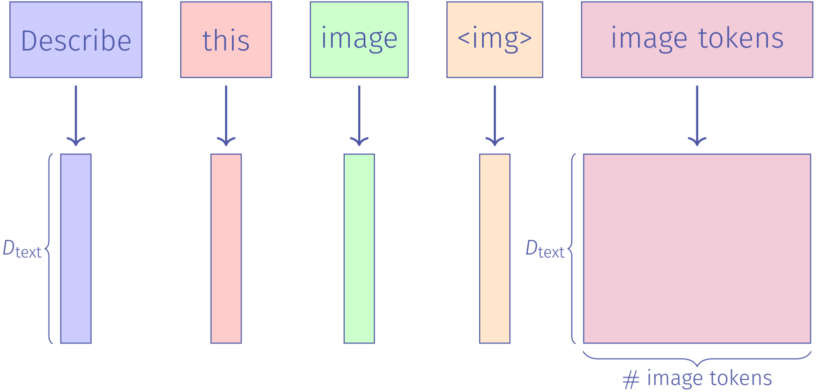

Describe this image <img>

D_{text}

Decoder input sequence

Prompt

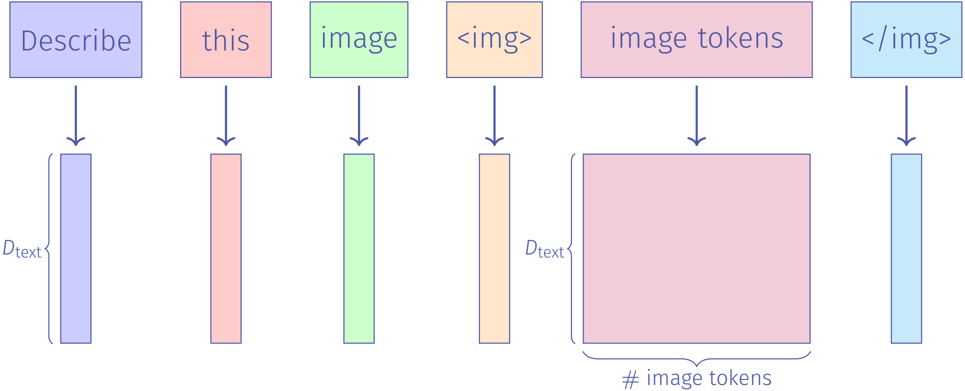
Describe this image



Decoder input sequence

Prompt

Describe this image



Describe

this

image

<img>

image tokens

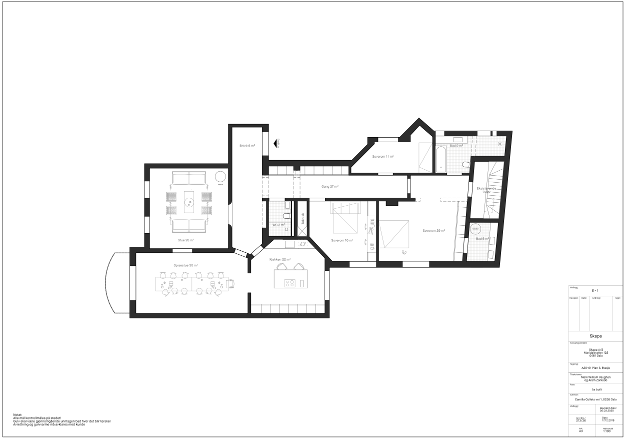
</img>

D_{text}

D_{text}

# image tokens

ARK DRAWING EXAMPLE



Notat:
Alle mål kontrollmåles på stedet!
Gulv skal være gjennomgående unntagen bad hvor det blir terskel
Avrettning og gulvvarme må avklares med kunde

Vedlegg: E - 1			
Revisjon:	Dato:	Endring:	Sign:
Skapa			
Ansvarlig arkitekt: Skapa A/S Maridalsveien 122 0461 Oslo			
Tegning: A20-01 Plan 3. Etasje			
Tiltakshaver: Mark William Vaughan og Aram Zarkoob			
Fase: As built			
Adresse: Camilla Colletts vei 1, 0258 Oslo			
Vedlegg:	Revidert dato: 05.03.2020		
Gnr./Bnr./ 213/36	Dato 17.12.2018		
Ark: A3	Målestokk 1:100		

Not a real customer document. Used with permission City of Oslo Planning and Building Services

VLM INFORMATION EXTRACTION

VLM INFORMATION EXTRACTION

Input

- Single or multiple images (no text necessary)
- Prompt

VLM INFORMATION EXTRACTION

Input

- Single or multiple images (no text necessary)
- Prompt

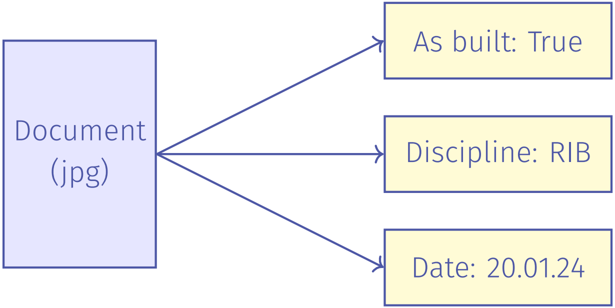
Output is a series of filled out fields

VLM INFORMATION EXTRACTION

Input

- Single or multiple images (no text necessary)
- Prompt

Output is a series of filled out fields

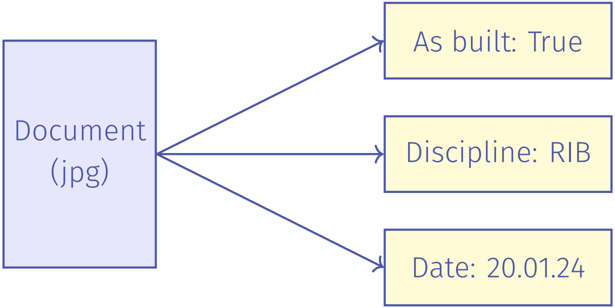


VLM INFORMATION EXTRACTION

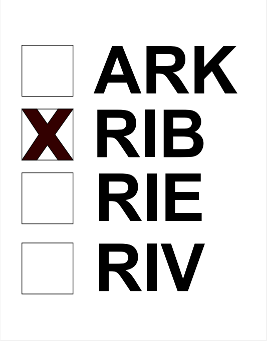
Input

- Single or multiple images (no text necessary)
- Prompt

Output is a series of filled out fields



Textual task requiring visual info


☐ ARK
☒ RIB
☐ RIE
☐ RIV

VLM INFORMATION EXTRACTION

Example Information Extraction prompt

```
prompt = ”””
Given the document,
fill out the following JSON object.
Enter None if the field is not available
{
    ”As_built” : ””,
    ”Discipline” : ””,
    ”Drawing_Date”: ””
}
”””
```

VLM CLASSIFICATION

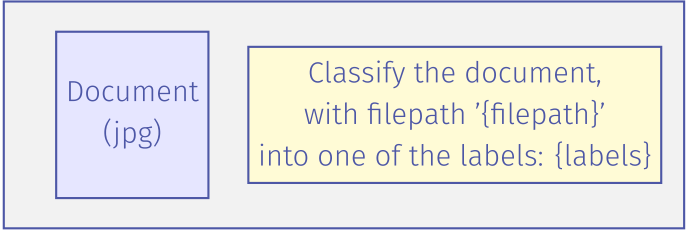
VLM CLASSIFICATION

SFT

VLM CLASSIFICATION

SFT

Input

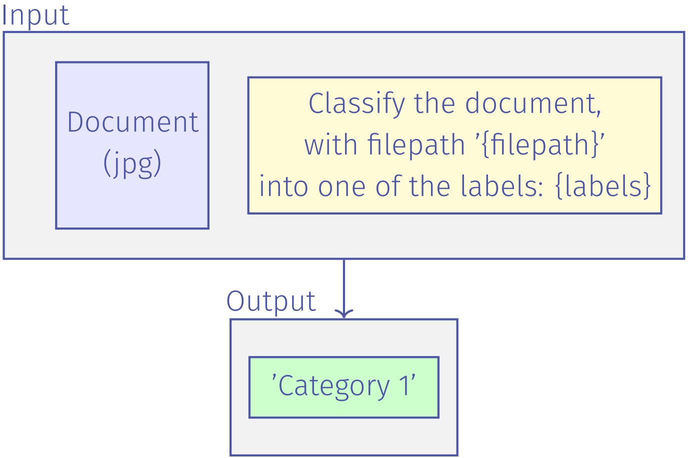


Document (jpg)

Classify the document, with filepath '{filepath}' into one of the labels: {labels}

VLM CLASSIFICATION

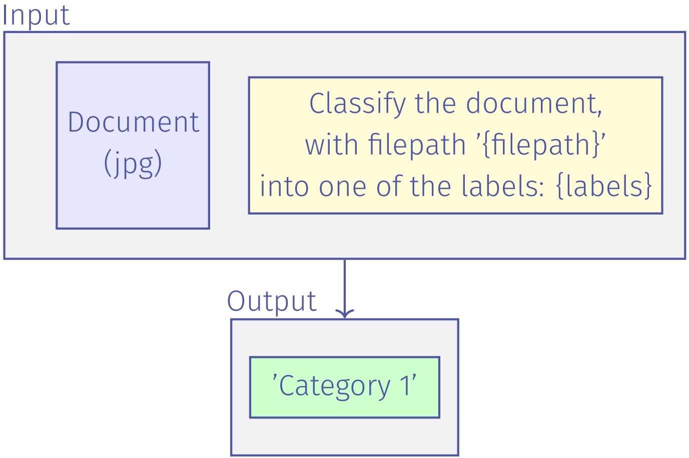
SFT


Input
Document (jpg)
Classify the document, with filepath '{filepath}' into one of the labels: {labels}
Output
'Category 1'

VLM CLASSIFICATION

SFT

Open-source VLMs



Input

Document (jpg)

Classify the document, with filepath '{filepath}' into one of the labels: {labels}

Output

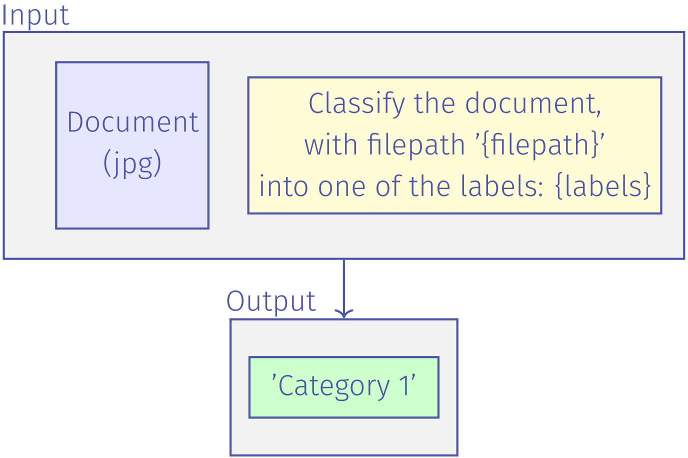
'Category 1'

VLM CLASSIFICATION

SFT

Open-source VLMs

Improved performance



Classify the document, with filepath '{filepath}' into one of the labels: {labels}

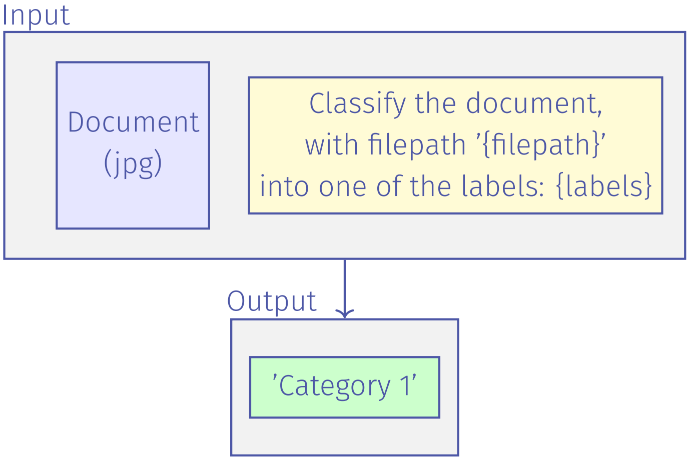
VLM CLASSIFICATION

SFT

Open-source VLMs

Improved performance

How to avoid alignment problems



Input

Document (jpg)

Classify the document, with filepath ’{filepath}’ into one of the labels: {labels}

Output

’Category 1’

VLM DRAWBACKS

VLM DRAWBACKS

Time consuming to set up open-source VLMs

VLM DRAWBACKS

Time consuming to set up open-source VLMs

- Model should be available on platforms like HF

VLM DRAWBACKS

Time consuming to set up open-source VLMs

- Model should be available on platforms like HF
- Often niche package criteria

VLM DRAWBACKS

Time consuming to set up open-source VLMs

- Model should be available on platforms like HF
- Often niche package criteria
- Configure dataset into correct format

VLM DRAWBACKS

Time consuming to set up open-source VLMs

- Model should be available on platforms like HF
- Often niche package criteria
- Configure dataset into correct format

This makes it costly to test different open-source VLMs

VLM DRAWBACKS

Time consuming to set up open-source VLMs

- Model should be available on platforms like HF
- Often niche package criteria
- Configure dataset into correct format

This makes it costly to test different open-source VLMs

Compute heavy

VLM DRAWBACKS

Time consuming to set up open-source VLMs

- Model should be available on platforms like HF
- Often niche package criteria
- Configure dataset into correct format

This makes it costly to test different open-source VLMs

Compute heavy

- High experimentation time for training

VLM DRAWBACKS

Time consuming to set up open-source VLMs

- Model should be available on platforms like HF
- Often niche package criteria
- Configure dataset into correct format

This makes it costly to test different open-source VLMs

Compute heavy

- High experimentation time for training
- Expensive inference

VLM DRAWBACKS

Time consuming to set up open-source VLMs

- Model should be available on platforms like HF
- Often niche package criteria
- Configure dataset into correct format

This makes it costly to test different open-source VLMs

Compute heavy

- High experimentation time for training
- Expensive inference
- Limited number of pages due to compute requirements

Staying up to date

Staying up to date

Newsletters

- TLDR AI
- AlphaSignal

Staying up to date

Newsletters

- TLDR AI
- AlphaSignal

GitHub stars

STAYING UP TO DATE

Newsletters

- TLDR AI
- AlphaSignal

GitHub stars

PapersWithCode

Staying up to date

Newsletters

- TLDR AI
- AlphaSignal

GitHub stars

PapersWithCode

Specific blogs

- OpenAI
- Anthropic

STAYING UP TO DATE

Newsletters

- TLDR AI
- AlphaSignal

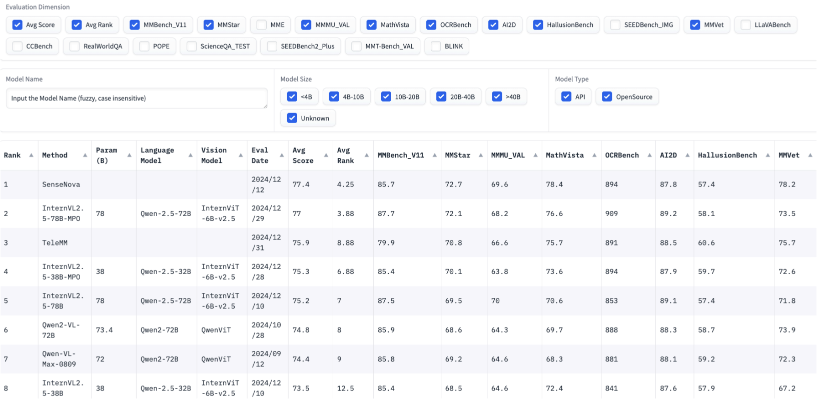
GitHub stars

PapersWithCode

Specific blogs

- OpenAI
- Anthropic

VLM leaderboards (HF)



Evaluation Dimension

☑ Avg Score ☑ Avg Rank ☑ MMBench_V11 ☑ MMStar □ MME ☑ MMMU_VAL ☑ MathVista ☑ OCRBench ☑ AI2D ☑ HallusionBench □ SEEDBench_IMG ☑ MMVet □ LLaVABench □ CCBench □ RealWorldQA □ POPE □ ScienceQA_TEST □ SEEDBench2_Plus □ MMT-Bench_VAL □ BLINK

Model Name

Input the Model Name (fuzzy, case insensitive)

Model Size

☑ <4B ☑ 4B-10B ☑ 10B-20B ☑ 20B-40B ☑ >40B ☑ Unknown

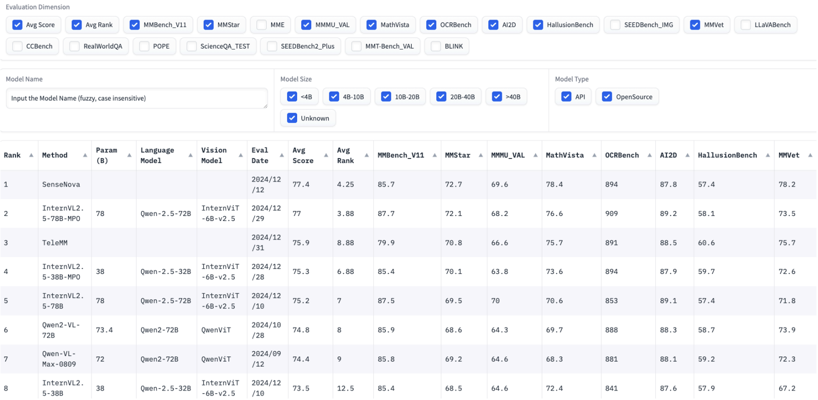
Model Type

☑ API ☑ OpenSource

Rank ▲	Method ▲	Param (B) ▲	Language Model ▲	Vision Model ▲	Eval Date ▲	Avg Score ▲	Avg Rank ▲	MMBench_V11 ▲	MMStar ▲	MMMU_VAL ▲	MathVista ▲	OCRBench ▲	AI2D ▲	HallusionBench ▲	MMVet ▲
1	SenseNova				2024/12/12	77.4	4.25	85.7	72.7	69.6	78.4	894	87.8	57.4	78.2
2	InternVL2.5-78B-MPO	78	Qwen-2.5-72B	InternViT-6B-v2.5	2024/12/29	77	3.88	87.7	72.1	68.2	76.6	909	89.2	58.1	73.5
3	TeleMM				2024/12/31	75.9	8.88	79.9	70.8	66.6	75.7	891	88.5	60.6	75.7
4	InternVL2.5-38B-MPO	38	Qwen-2.5-32B	InternViT-6B-v2.5	2024/12/28	75.3	6.88	85.4	70.1	63.8	73.6	894	87.9	59.7	72.6
5	InternVL2.5-78B	78	Qwen-2.5-72B	InternViT-6B-v2.5	2024/12/10	75.2	7	87.5	69.5	70	70.6	853	89.1	57.4	71.8
6	Qwen2-VL-72B	73.4	Qwen2-72B	QwenViT	2024/10/28	74.8	8	85.9	68.6	64.3	69.7	888	88.3	58.7	73.9
7	Qwen-VL-Max-0809	72	Qwen2-72B	QwenViT	2024/09/12	74.4	9	85.8	69.2	64.6	68.3	881	88.1	59.2	72.3
8	InternVL2.5-38B	38	Qwen-2.5-32B	InternViT-6B-v2.5	2024/12/10	73.5	12.5	85.4	68.5	64.6	72.4	841	87.6	57.9	67.2

STAYING UP TO DATE

- Newsletters
  - TLDR AI
  - AlphaSignal
- GitHub stars
- PapersWithCode
- Specific blogs
  - OpenAI
  - Anthropic
- VLM leaderboards (HF)
- LLM Night



Evaluation Dimension

☑ Avg Score ☑ Avg Rank ☑ MMBench_V11 ☑ MMStar ☐ MME ☑ MMMU_VAL ☑ MathVista ☑ OCRBench ☑ AI2D ☑ HallusionBench ☐ SEEDBench_IMG ☑ MMVet ☐ LLaVABench ☐ CCBench ☐ RealWorldQA ☐ POPE ☐ ScienceQA_TEST ☐ SEEDBench2_Plus ☐ MMT-Bench_VAL ☐ BLINK

Model Name

Input the Model Name (fuzzy, case insensitive)

Model Size

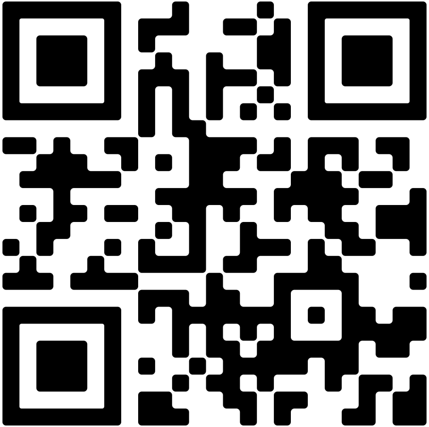
☑ <4B ☑ 4B-10B ☑ 10B-20B ☑ 20B-40B ☑ >40B ☑ Unknown

Model Type

☑ API ☑ OpenSource

Rank ▲	Method ▲	Param (B) ▲	Language Model ▲	Vision Model ▲	Eval Date ▲	Avg Score ▲	Avg Rank ▲	MMBench_V11 ▲	MMStar ▲	MMMU_VAL ▲	MathVista ▲	OCRBench ▲	AI2D ▲	HallusionBench ▲	MMVet ▲
1	SenseNova				2024/12/12	77.4	4.25	85.7	72.7	69.6	78.4	894	87.8	57.4	78.2
2	InternVL2.5-78B-MPO	78	Qwen-2.5-72B	InternViT-6B-v2.5	2024/12/29	77	3.88	87.7	72.1	68.2	76.6	909	89.2	58.1	73.5
3	TeleMM				2024/12/31	75.9	8.88	79.9	70.8	66.6	75.7	891	88.5	60.6	75.7
4	InternVL2.5-38B-MPO	38	Qwen-2.5-32B	InternViT-6B-v2.5	2024/12/28	75.3	6.88	85.4	70.1	63.8	73.6	894	87.9	59.7	72.6
5	InternVL2.5-78B	78	Qwen-2.5-72B	InternViT-6B-v2.5	2024/12/10	75.2	7	87.5	69.5	70	70.6	853	89.1	57.4	71.8
6	Qwen2-VL-72B	73.4	Qwen2-72B	QwenViT	2024/10/28	74.8	8	85.9	68.6	64.3	69.7	888	88.3	58.7	73.9
7	Qwen-VL-Max-0809	72	Qwen2-72B	QwenViT	2024/09/12	74.4	9	85.8	69.2	64.6	68.3	881	88.1	59.2	72.3
8	InternVL2.5-38B	38	Qwen-2.5-32B	InternViT-6B-v2.5	2024/12/10	73.5	12.5	85.4	68.5	64.6	72.4	841	87.6	57.9	67.2

Link to presentations and code



Sources

- Vision Transformer (ViT):
  An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
- Qwen-VL:
  Qwen-VL: A Versatile Vision-Language Model for Understanding, Generation, and Retrieval
- Qwen-VL-2:
  Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution
- VisionLLaMA:
  VisionLLaMA: A Unified LLaMA Backbone for Vision Tasks
- Pixtral:
  Pixtral 12B

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was som unprocessed data that should have been added to the final page this extra pag has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because LaTeX now knows how many pages to expect for this document.