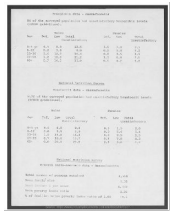
AlignVLM: Bridging Vision and Language Latent Spaces for Multimodal Document Understanding

NeurIPS 2025

(5) AlignVLM: Motivation

Text-rich Visual Understanding: Essential for tasks in enterprise such as document understanding, trends analysis, table extraction where both **language** and **visuals** are critical.



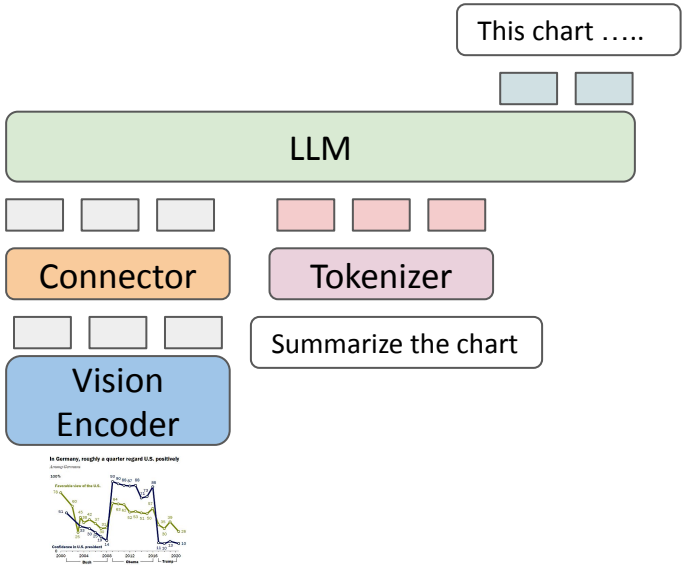
PDF Analysis

Infographics & Charts

Table Extraction

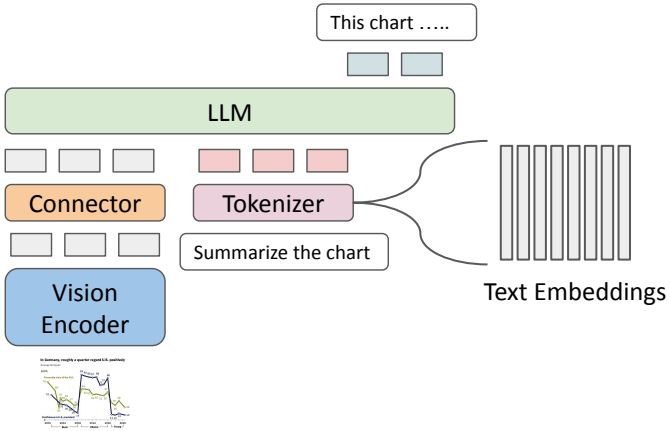
Vision Language Models (VLMs) typically consist of ***three*** components:

- Vision Encoder, pretrained on images.
- LLM, pretrained on text
- Connector that maps visual features into the LLM's text space.

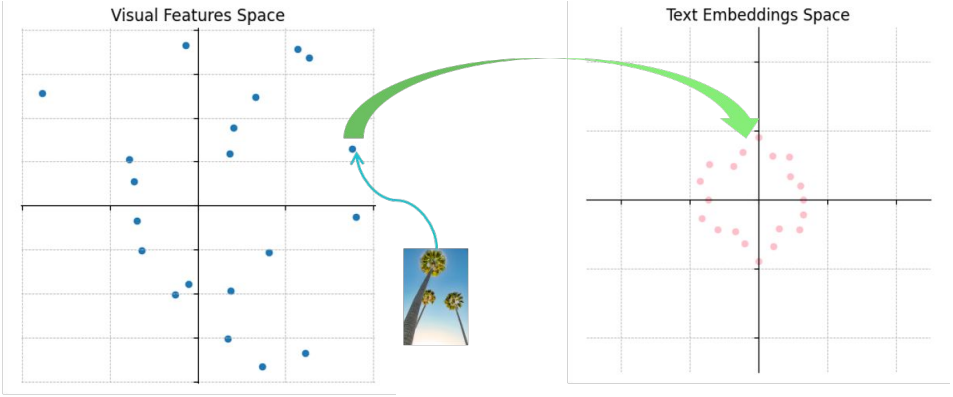


(5) AlignVLM: Motivation

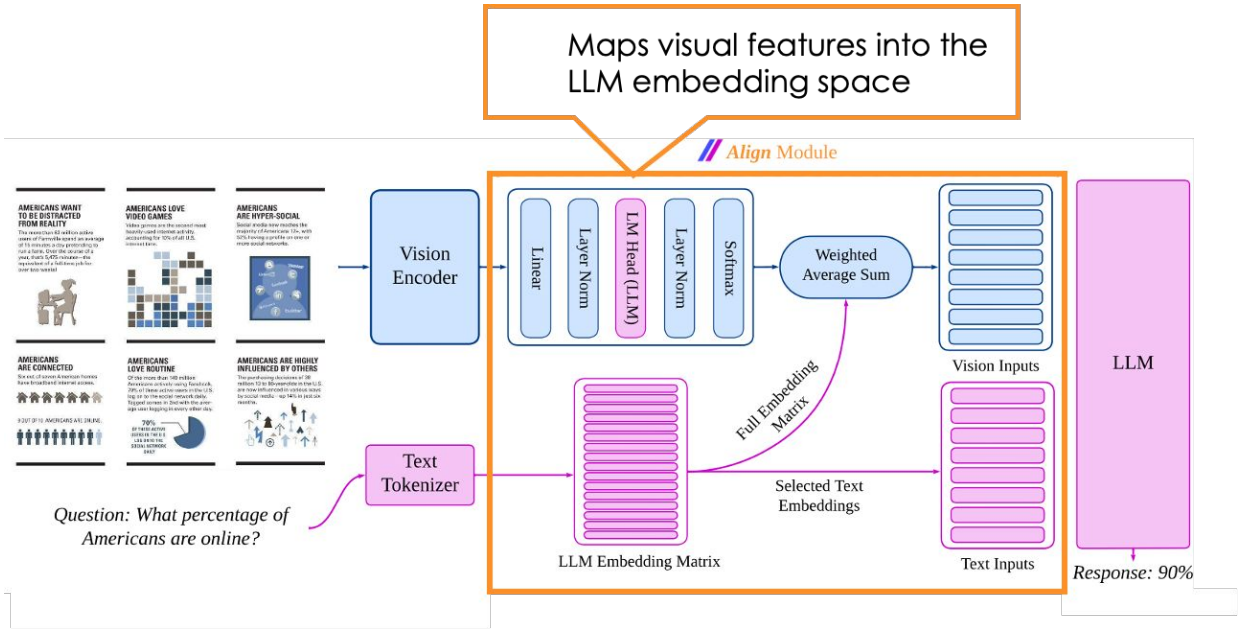
- LLM is pretrained to process a discrete set of text embeddings
- Existing connector (e.g., MLP) produce continuous visual features
    - **Out-of-distribution (OOD):** making the connector data hungry!
    - **Unconstrained Mapping:** They do not enforce any hard constraints which makes them *prone to noise.*



Can we exploit the LLM's inductive bias by aligning visual features directly with text embeddings?



(5) AlignVLM: Architecture



- Map visual features to a distribution over LLM token embeddings.

$$\mathbf{P}_{\text{vocab}} = \text{softmax}(\text{LayerNorm}(\mathbf{W}_2\,\text{LayerNorm}(\mathbf{W}_1\mathbf{F}))) \quad (1)$$

- Computes final features as a weighted average of text embeddings.

$$\mathbf{F}'_{\text{align}} = \mathbf{P}_{\text{vocab}}^{\top}\mathbf{E}_{\text{text}}.$$

- Constrains visual inputs to the convex hull of the LLM's embedding space, making them familiar to the LLM.

(5) AlignVLM: Training Setup

Training Stages

Stage 1: Natural Image Understanding
Data: CC-12M (Image-Caption)

Stage 2: Document Understanding
Data: BigDocs-7.5M

Stage 3: Instruction Tuning for downstream tasks
Data: BigDocs-Docdownstream

Model components

- **LLM:** Llama 3.2 Family (1B, 3B, 8B)
- **Vision Encoder:** SigLip-400m

Evaluation Benchmarks

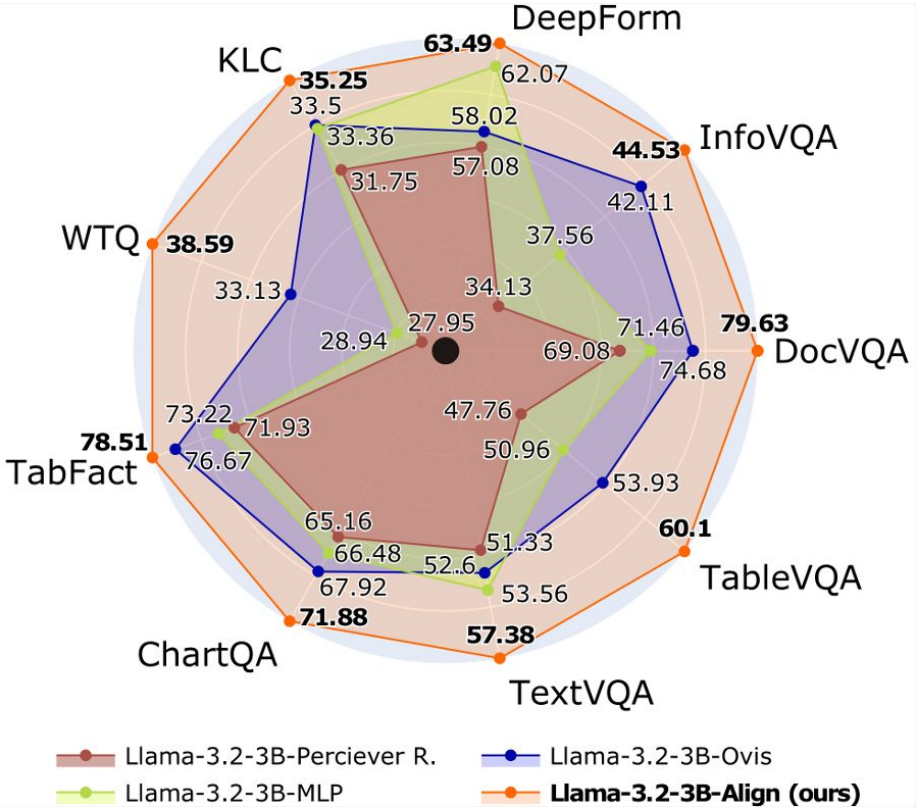
Nine document benchmarks, including:

- DocVQA, InfoVQA, ChartQA, TableVQA, etc.

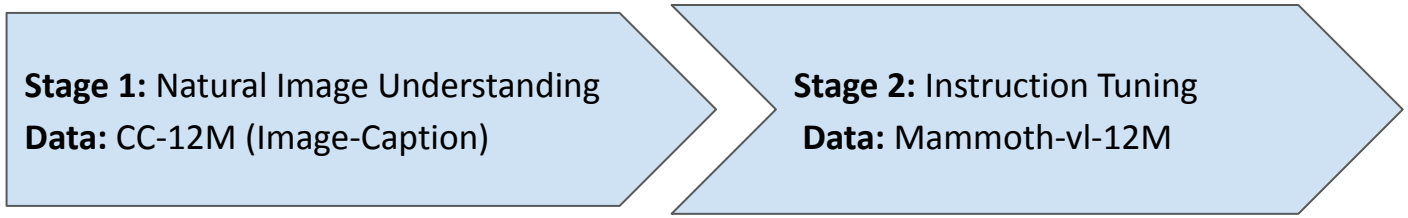
(5) AlignVLM: Results

- We compare **our Align Module** against common connectors:
  - MLP, Perceiver Resampler, Ovis
- Trained under similar configurations to ensure a fair comparison.

The **Align Module** outperforms them all and achieves better accuracy on diverse document understanding tasks.

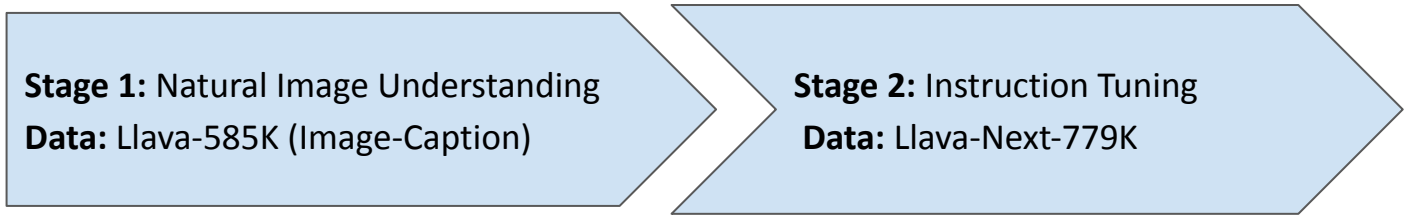


(5) AlignVLM: General Vision Tasks



Model	MMMU (dev)	SeedBench	MMVet	POPE	GQA
Llama-3.2-3B-MLP	35.66	71.68	44.95	84.11	37.07
Llama-3.2-3B-ALIGN (ours)	**38.66**	**72.87**	**47.75**	**84.73**	**42.77**

(5) AlignVLM: Low Resource Setup



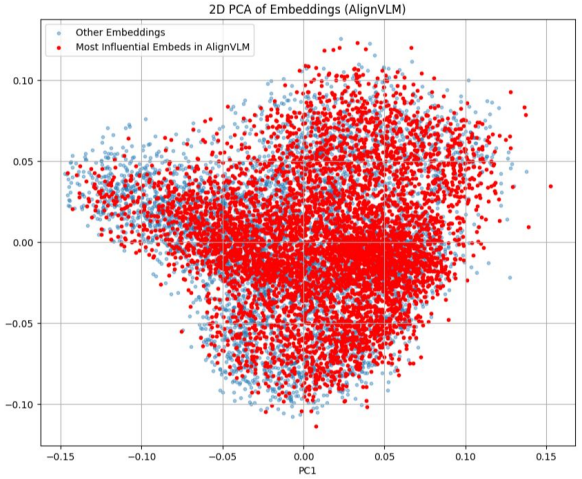
Model	DocVQA	InfoVQA	ChartQA	TextVQA	Average	Δ
LLama-3.2-3B-MLP (Llava Next)	42.11	19.93	48.44	51.97	40.61	
LLama-3.2-3B-Align (Llava Next)	71.43	30.50	69.72	65.63	59.32	+18.71
LLama-3.2-3B-MLP (BigDocs)	71.46	37.56	66.48	53.56	57.26	
LLama-3.2-3B-Align (BigDocs)	79.63	44.53	71.88	57.38	**63.35**	+6.09

(5) AlignVLM: Pruning

Token utilization: Align focuses on ~3.4K out of 128K tokens, ignoring the rest.

Latent coverage: These tokens span the LLM's embedding space.

Efficiency: Pruning to these tokens preserves performance and boosts efficiency.



Model	DocVQA	InfoVQA	Deepform	KLC	WTQ	TabFact	ChartQA	TextVQA	TableVQA
AlignVLM-3B (3.4K tokens)	79.40	44.13	63.64	35.02	38.26	78.83	71.72	57.48	59.80
AlignVLM-3B (full)	79.60	44.53	63.49	35.25	38.59	78.51	71.88	57.38	60.10