Assessing and Learning Alignment of Unimodal Vision and Language Models

Le Zhang, Qian Yang, Aishwarya Agrawal

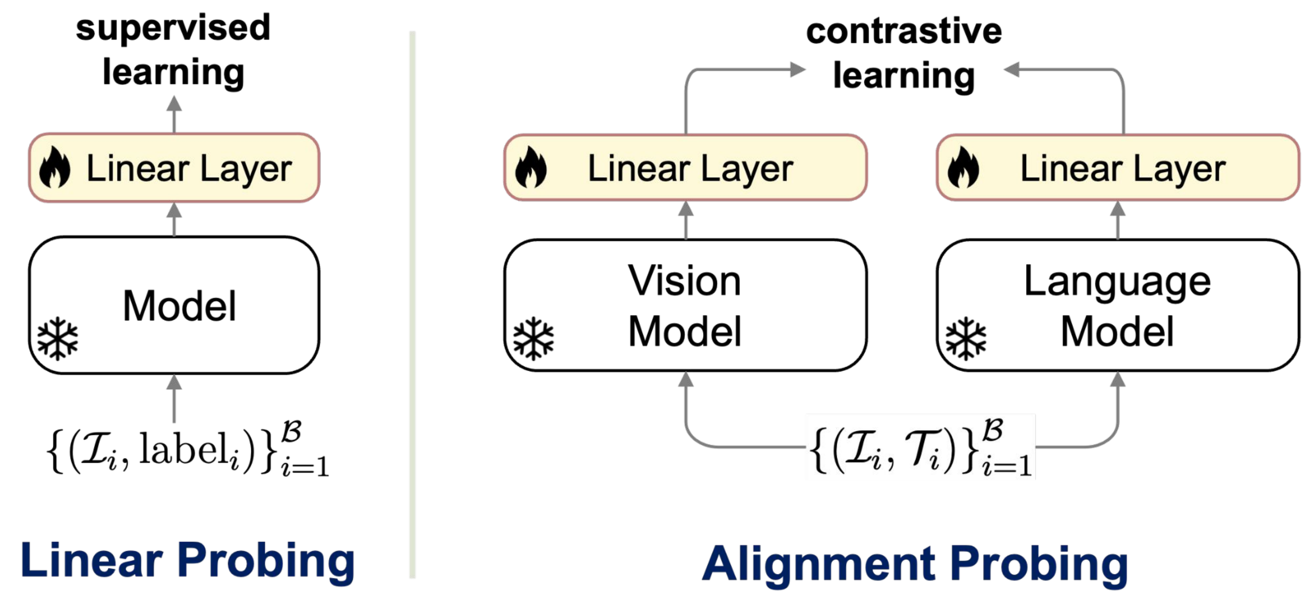
Mila · Université de Montréal

Project page
Scan here !

Part 1: Assessing Alignment

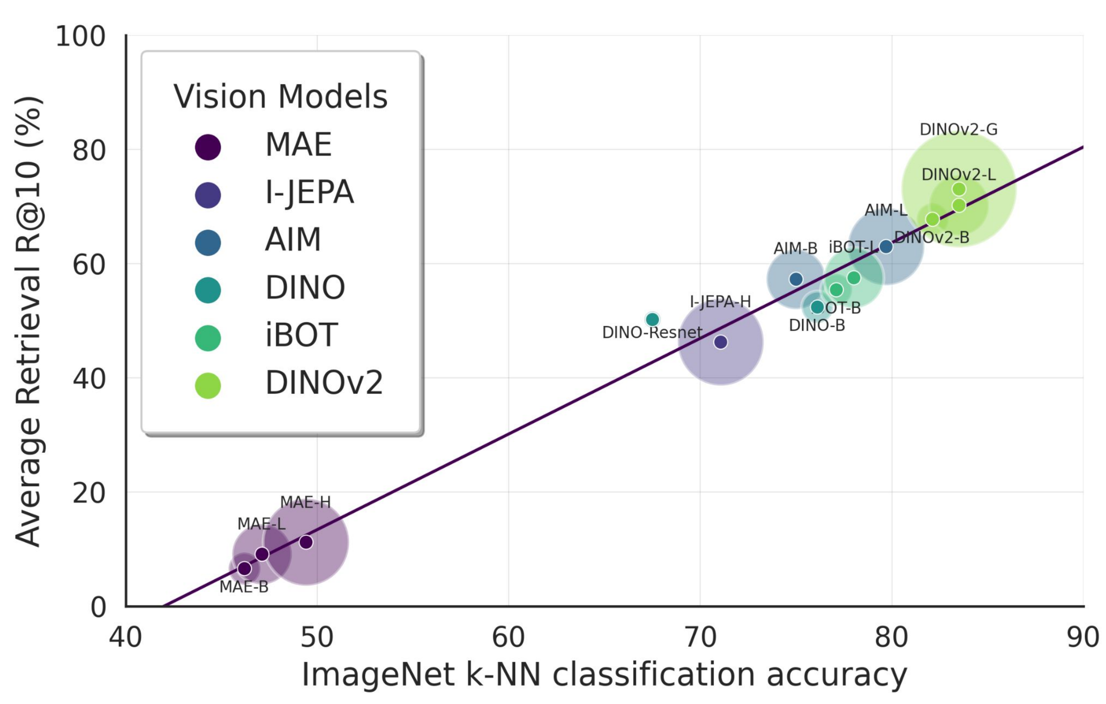
Key Questions

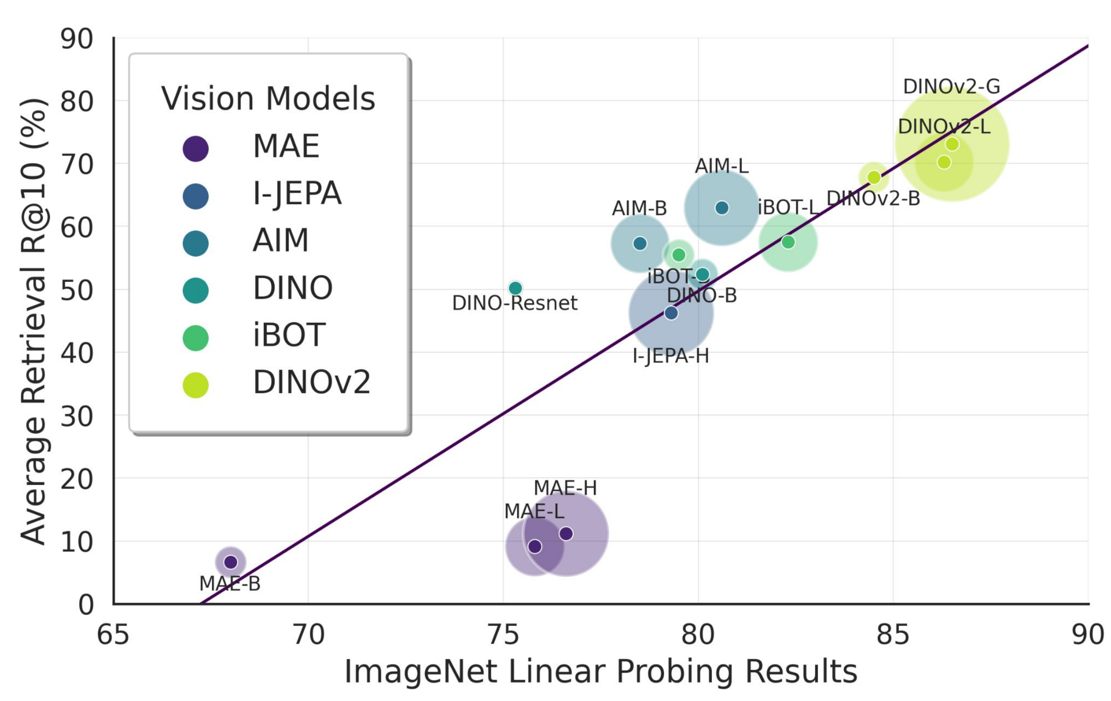
- **Alignment Capability:** How well can unimodal visual and language models align for zero-shot open-vocabulary tasks?
- **Model Architecture Impact:** Do larger models trained on extensive datasets yield better alignment? Does the choice of self-supervised learning (SSL) methods play a more significant role?
- **Representation Properties:** What properties of SSL representations—such as linear separability or clustering quality—drive stronger cross-modal alignment?



Visual-Language Alignment Probing: a direct assessment method inspired by linear probing in SSL evaluation.

Language as Anchor

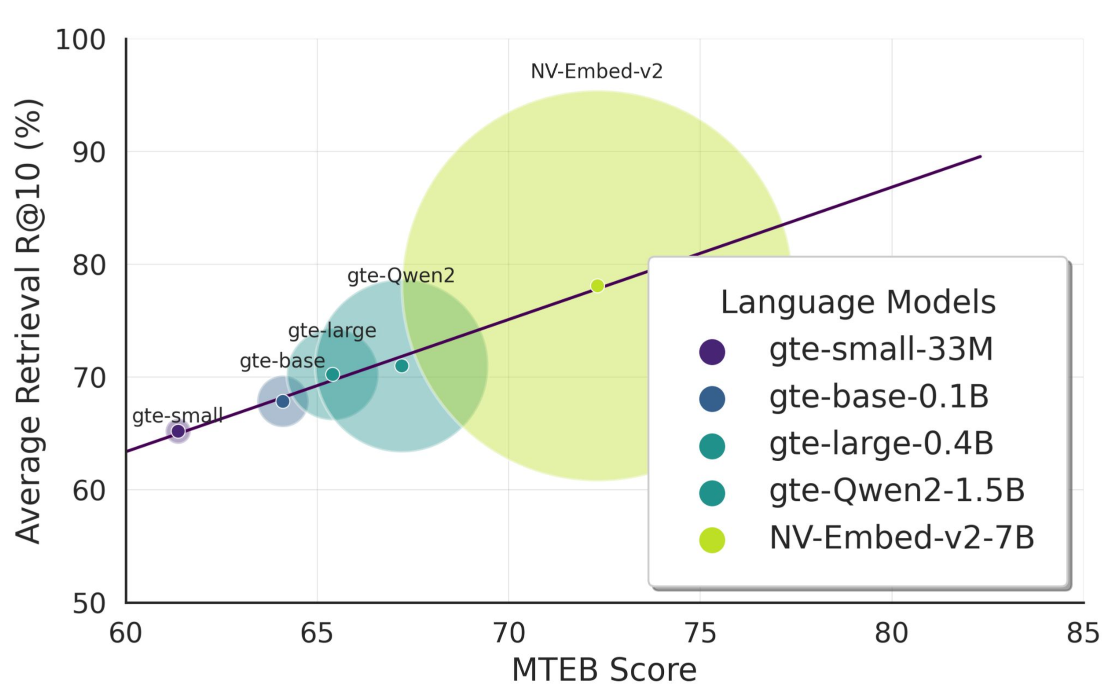


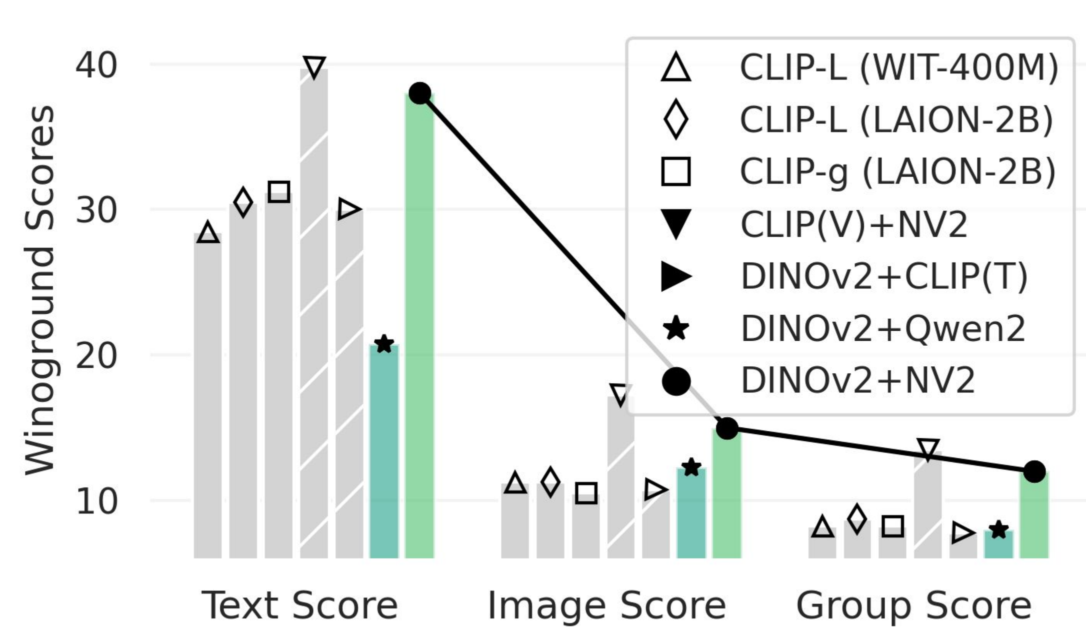


Key Findings

- **SSL Method Matters:** DINOv2-B (86M) > AIM-L (1B parameters)> MAE-familay
- **Representation Properties:** Alignment performance strongly depends on the clustering quality of SSL representation, as reflected by k-NN performancemore than linear separability.

Vision as Anchor





Key Findings

- **Language Understanding Critical:** language understanding capability is essential for vision-language reasoning tasks.
- **CLIP Training Limitations:** Training text encoders solely through CLIP-style contrastive learning proves insufficient for optimal performance.
- **Pretrained LM Advantage:** LLMs as text encoders emerges as a promising strategy for building robust VLMs

Part 2: Learning Alignment

We introduce **S**wift **A**lignment of **I**mage and **L**anguage **(SAIL)**, **aligning pretrained unimodal vision and language models.**

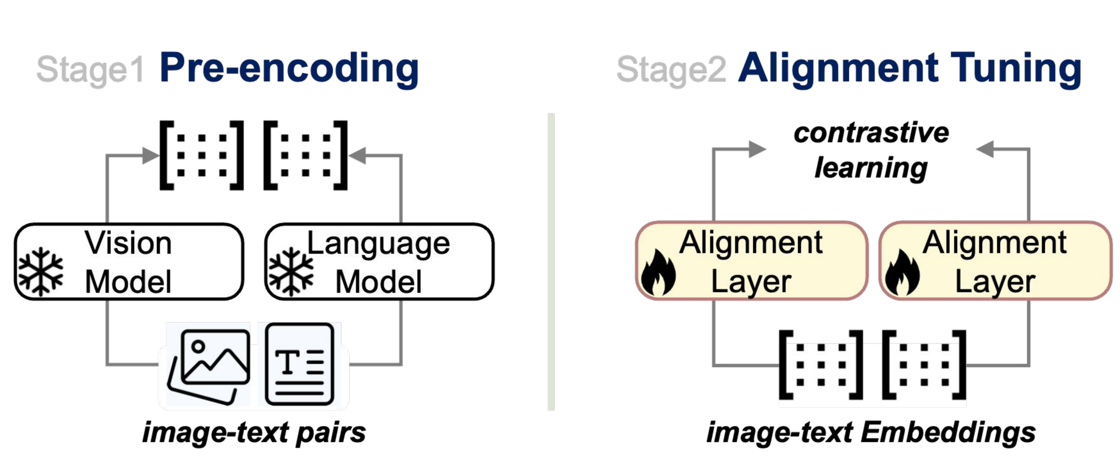
Our efficient two-step training pipeline optimizes both performance and computational costs.

Specifically, SAIL achieves superior alignment through three key optimizations:

Alignment Layer Arch
Advanced non-linear GLU in alignment layers to improve alignment quality

Enhanced Loss Function
Sigmoid binary classification loss with balanced positive/negative contributions

High-Quality Data Selection
MLLM generated captions as additional positives and multiple positive captions contrast loss



Ablation	0	1	2	3	4	5	6	7
Tasks	Baseline	+ MLP × 4	+ GLU × 4	+ GLU × 8	+ Sigmoid	+ |B| → |B|²	+ Long-HQ	+ Multi-Pos
IN-1K 0-shot	33.2	36.8	39.6	45.4	50.7	51.8	48.4	54.0
T2I R@1	11.1	8.0	11.5	16.1	25.4	26.2	31.4	32.9
I2T R@1	13.5	10.7	17.4	22.5	36.0	36.7	44.2	45.4

Table: Ablation results using CC3M on different methods. Baseline refers to aligning unimodal models with only a linear layer using infoNCE loss.

Zero-Shot Vision-Language Tasks

We train **SAIL** a **23M** **M**erged dataset.. The training of SAIL takes **~ 5 hours** on **a single A100 GPU** with batch size up to **32,768**.

SAIL surpasses CLIP with only ~6% of image-text pairs on broad downstream vision-language tasks.

Data	Model	MSCOCO I2T	MSCOCO T2I	Flickr30k I2T	Flickr30k T2I	Winoground T.	Winoground I.	Winoground G.	MMVP 10 Avg.	ImageNet Top1.	10 Classification Avg.
Model Architecture: ViT-B/16											
CC12M	DreamLIP	53.3	41.2	82.3	66.6	26.0	10.00	7.25	24.0	50.3	49.9
	LiT‡	30.0	16.5	54.8	38.5	24.3	6.5	4.8	-	56.2	-
	ShareLock(Llama3)†‡	26.0	13.5	53.9	34.9	26.3	12.8	5.3	-	59.1	-
	ShareLock(NV2)†	39.6	23.1	68.1	49.3	33.25	13	9.75	15.56	61.9	62.0
	SAIL-B (GTE)†	48.2	37.9	76.5	63.9	31.0	11.5	9.5	23.0	58.7	57.7
	SAIL-B (NV2)†	57.3	45.3	84.1	70.1	35.0	17.25	13.0	24.4	68.1	65.4
LAION400M	CLIP-B	55.4	38.3	83.2	65.5	25.7	11.5	7.75	19.3	67	65.5
Model Architecture: ViT-L											
23M Merged	SAIL-L (NV2)†	62.4	48.6	87.6	75.7	40.25	18.75	15.0	28.9	72.1	73.4
LAION400M	CLIP-L	59.7	43.0	87.6	70.2	30.5	11.5	8.75	20.0	75.9	72.7

Table 6. **Results** on standard retrieval, complex reasoning, visual-centric, and classification tasks. We report Recall@1 for MSCOCO and Flickr30k, Text, Image, and Group scores for Winoground, and the average score across 9 visual patterns for MMVP. ‡ indicates cited results, and † denotes a ViT patch size of 14. 10 Classification tasks include: Food101, CIFAR10, CIFAR100, SUN397, Cars, Aircraft, DTD, Pets, Caltech101, and Flowers.

Multimodal LLM tasks

SAIL transforms features from SSL models to be more language-compatible, thus better suited for integration with MLLMs.

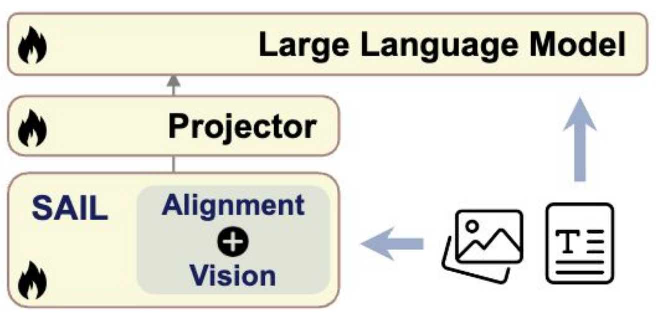


Figure 6. **Using SAIL's vision encoder for MLLMs.**

	Model@224px	VTune	SEED^{IMG}	GQA	VizWiz	PoPE	TextVQA	MMB	VQA^{v2}
0	DINOv2-L	✗	61.47	61.08	44.12	85.5	45.37	56.96	74.4
1	DINOv2-L	✓	62.12	61.53	46.59	85.7	45.92	58.85	74.69
2	SAIL-L	✓	65.43	62.63	50.00	86.16	46.53	60.14	76.77
3	CLIP-L/14*	✗	64.05	61.58	48.87	85.74	54.56	63.06	75.32
4	CLIP-L/14*	✓	64.15	61.54	49.93	85.73	54.18	64.12	76.36

Table 4. **LLaVA-1.5 with various vision models.** *Reproduced using OpenAI CLIP-L@224 [34]. VTune indicates if the vision encoder is fine-tuned during the instruction tuning stage.