Harnessing Frozen Unimodal Encoders for Flexible Multimodal Alignment

Mayug Maniparambil*, Raiymbek Akshulakov*, Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Ankit Singh, Noel E. O'Connor
(* equal contribution)

ML Labs
Berkeley UNIVERSITY OF CALIFORNIA
TII Technology Innovation Institute

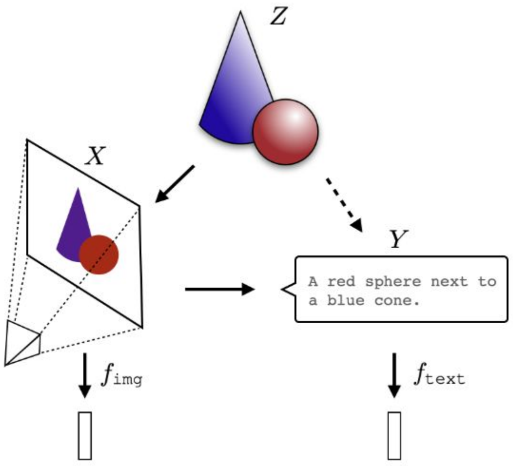
CVPR Nashville JUNE 11-15, 2025

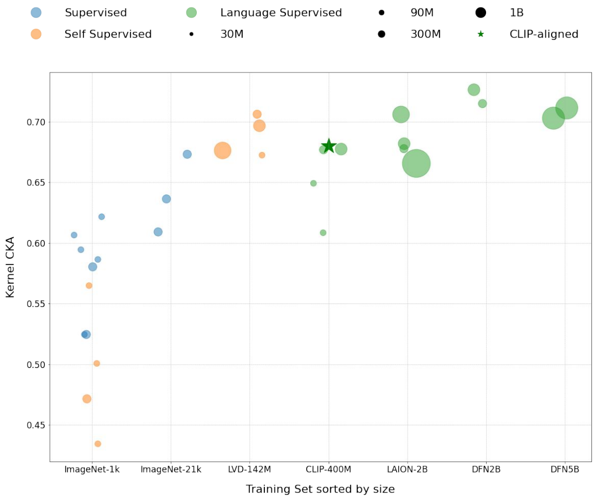
Introduction

- Given well trained unimodal vision and language models have **high semantic similarity (Platonic Rep. Hypothesis)** are they **separated by simple projection transformations?**
- Using toy examples we find that embedding spaces with **high semantic similarity** can be aligned using **Projectors**
- We introduce a simple framework for aligning frozen unimodal vision and language embeddings to get CLIP models
- **65x less compute, 20x less data** compared to CLIP models
- **Flexible adaptation** to 0-shot classification, retrieval, long context, multilingual and localization tasks by swapping out text encoders.

Platonic representation hypothesis

- The Platonic Representation Hypothesis [1] and our earlier work DoVisLang [2] suggest that vision and language encoders are converging toward a shared latent reality.
- Specifically, [2] shows that well-trained unimodal vision and language models exhibit high semantic similarity.
- This raises the “so what?” question: Can we use this convergence to build better multimodal models?
- In this work, we investigate whether semantically similar unimodal spaces can be bridged by simple transformations, like a lightweight 2-layer MLP.

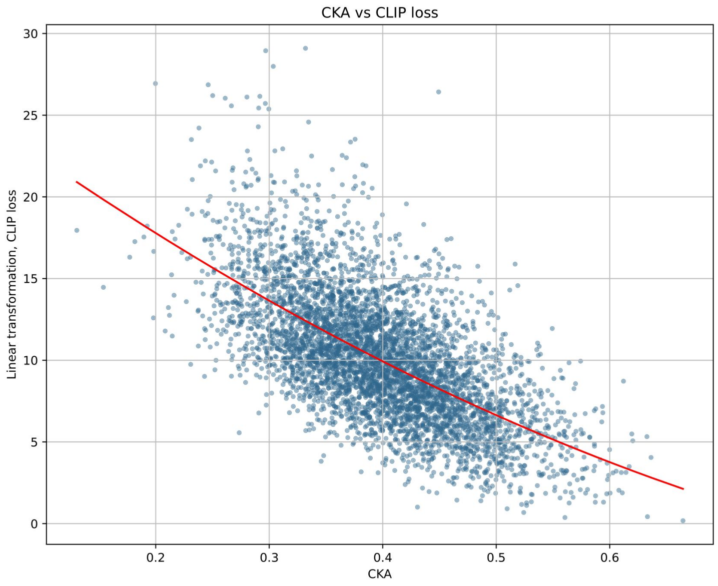




[1] Maniparambil, Mayug, et al. "Do Vision and Language Encoders Represent the World Similarly?." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

[2] Huh, Minyoung, et al. "The platonic representation hypothesis." *arXiv preprint arXiv:2405.07987* (2024).

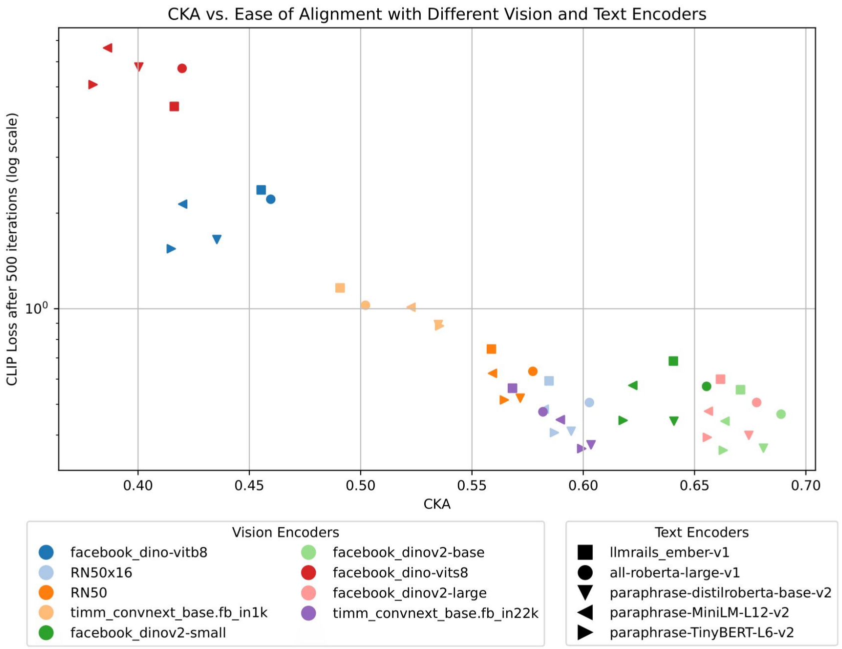
Toy Experiments - Synthetic embeddings




```
# Init Z with random values scaled to [-1, 1]
Z = 2 * rand(n, d) - 1

# Define non-linear transforms T1 and T2
T1, T2 = NLTransform(d, d), NLTransform(d, d)

# Sample random weights w1 and w2
w1, w2 = rand(1), rand(1)

# Compute A and B using transforms
A = T1(Z) + w1 * rand(n, d)
B = T2(Z) + w2 * rand(n, d)
```

- We study this on 2 toy experiments- With synthetic and real embeddings
- Semantic similarity measured using Centered Kernel Alignment
- Ease of Alignment measured as CLIP loss value after 500 iterations
- Ease of Alignment increases with higher CKA b/w embedding spaces

Toy experiment- Real embeddings

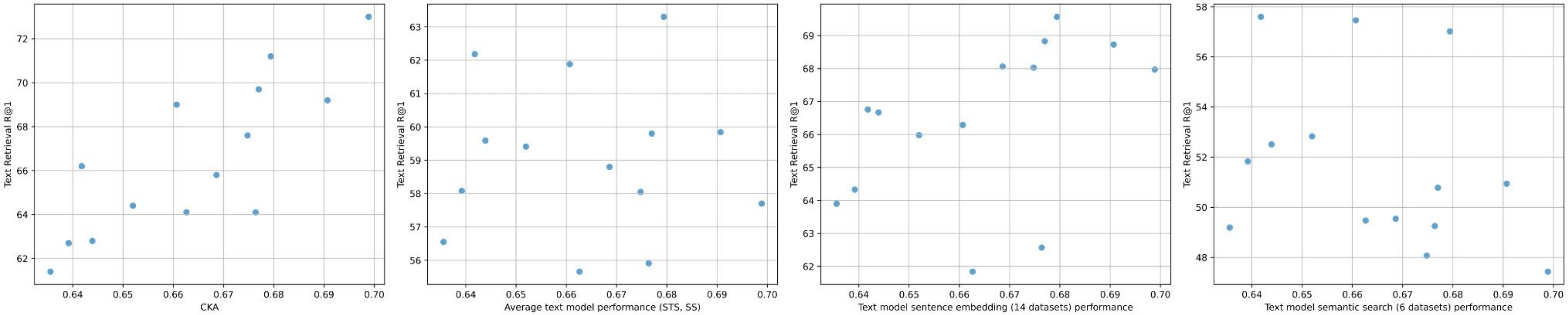


- Ease of alignment increases with higher semantic similarity b/w unimodal embedding spaces.
- Real embeddings from 9 vision encoders and 5 language encoders
- Small Scale experiment on a toy subset of the coco dataset.

Framework

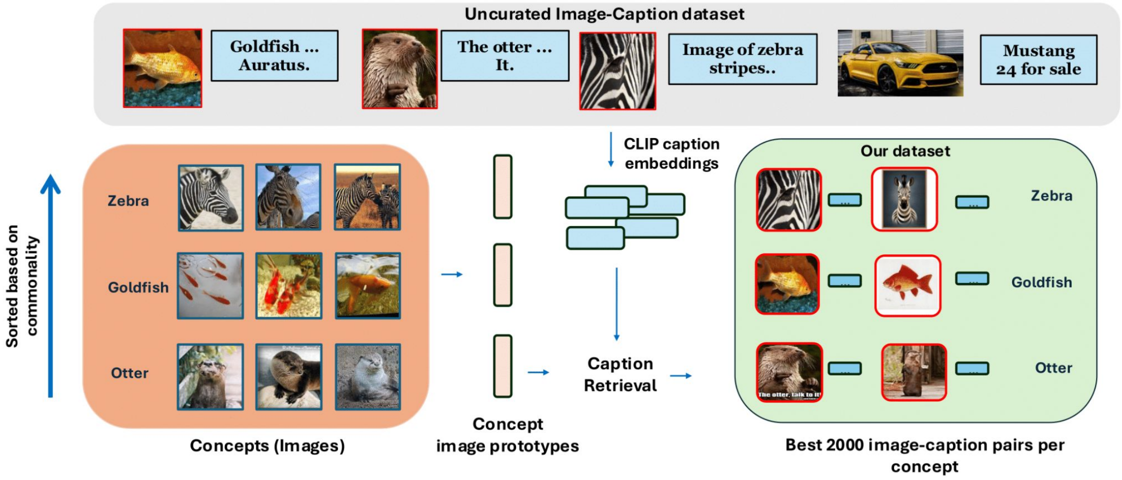
- Choose semantically similar encoder pairs;
- Curate dataset - **high concept coverage** and **high alignment**
- **Train Projectors** using vanilla contrastive loss

Why CKA? Unimodal performance is not predictive of alignment



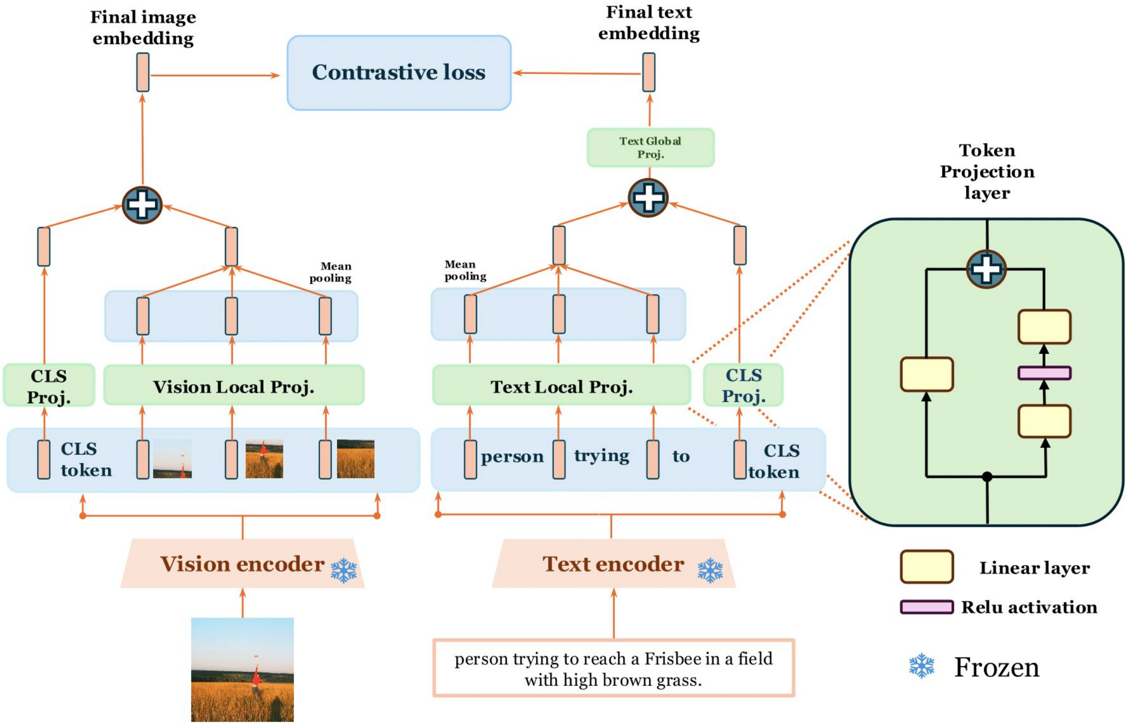
- We measure the performance after alignment for several text encoders to DinoV2 and compare the Text Retrieval Scores with the unimodal scores of pure text tasks.
  - Considered text tasks - Sentence Text Similarity and Semantic Search
- CKA to DinoV2 embeddings is more predictive of downstream performance compared to unimodal performance
- Not clear which unimodal performance to consider while CKA measures semantic similarity to the vision encoder space– hence more intuitive.

Concept Rich Data Curation



- We train only **11M parameters**, so a **small, high-quality dataset** is sufficient.
- Use **CC3M + CC12M** (15M pairs) for strong semantic alignment.
- Add **3,000 curated concepts** using few-shot image prototypes.
- For each concept, retrieve **2,000 image-caption pairs** → **6M extra samples**.
- **Total dataset: 21M pairs** with high alignment and broad concept coverage.

Projector Training on Frozen embeddings



- Train lightweight projectors on top of frozen vision and language encoders.
- Use separate projectors for CLS and local tokens to capture both global and fine-grained information.

Results

- CLIP-level performance with much lower data and compute.
- Supports a wide range of zero-shot tasks:
  - Multilingual classification & retrieval
  - Semantic & multilingual segmentation
  - Long-context retrieval
- Achieved without any specialized alignment data—no multilingual, localization, or long-caption supervision needed.

0-shot classification and Retrieval

Model	N	ImageNet	ImageNetv2	Caltech	Pets	Cars	Flowers	Food	Aircrafts	SUN	CUB	UCF101
LAION-CLIP VIT-L	400M	72.7	65.4	92.5	91.5	**89.6**	73.0	90.0	24.6	70.9	**71.4**	71.6
OpenAI-CLIP VIT-L	400M	75.3	**69.8**	92.6	**93.5**	77.3	**78.7**	**92.9**	**36.1**	67.7	61.4	**75.0**
LiT L16L	112M	75.7	66.6	89.1	83.3	24.3	76.3	81.1	15.2	62.5	58.7	60.0
DINOv2-MpNet (Ours)	20M	74.8	68.0	91.8	91.7	71.0	75.8	87.5	23.0	71.9	63.2	71.0
DINOv2-ARL(Ours)	20M	**76.3**	69.2	**92.8**	92.1	73.9	78.4	89.1	28.1	**72.6**	66.1	73.2

- **Outperforms CLIP** models from **OpenAI** and **LAION** on zero-shot classification and retrieval.
- Uses **DINOv2-Large** for vision and **All-Roberta-Large-v1** for text.

Model	Flickr		COCO	
	I2T	T2I	I2T	T2I
LAION-CLIP VIT-L	**87.6**	70.2	59.7	43.0
OpenAI-CLIP VIT-L	85.2	64.9	56.3	36.5
LiT L16L	73.0	53.4	48.5	31.2
DINOv2-MpNet (Ours)	84.6	71.2	58.0	42.6
DINOv2-ARL (Ours)	87.5	**74.1**	**60.1**	**45.1**

Data and Compute for Alignment

Model	Data	SS	Trainable / Total	Compute	IN 0-shot
OpenAI CLIP	400M	12.8B	427M / 427M	21,845	72.7%
LAION400M CLIP	400M	12.8B	427M / 427M	25,400	75.3%
DINOv2-ARL	20M	0.6B	11.5M / 670M	400	76.3%

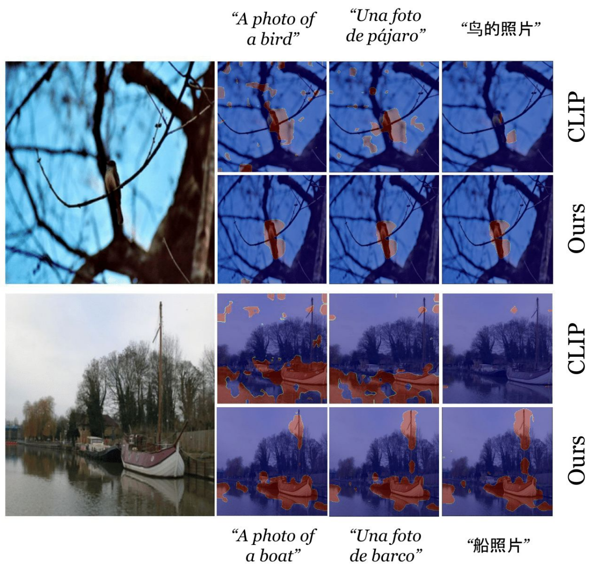
- Only 11.5M parameters are trained, while the encoders remain frozen
- The lower parameter count means we can train with a high quality concept rich dataset of just 20M image-caption pairs.
- The alignment data requirement is 20X lower than CLIP while compute requirement is 65X lower.

Flexibility: Multilingual Classification/ Retrieval

model	classification						retrieval					
	EN	DE	FR	JP	RU	average	EN	DE	FR	JP	RU	average
nllb-clip-base@v1	25.4	23.3	23.9	21.7	23.0	23.5	47.2	43.3	45.0	37.9	40.6	42.8
M-CLIP/XLM-Roberta-Large-Vit-B-32	46.2	43.3	43.3	31.6	38.8	40.6	48.5	46.9	46.1	35.0	43.2	43.9
M-CLIP/XLM-Roberta-Large-Vit-L-14	54.7	51.9	51.6	37.2	47.4	48.6	56.3	52.2	51.8	41.5	48.4	50.0
xlm-roberta-base-ViT-B-32@laion5b	63.0	55.8	53.8	37.3	40.3	50.0	63.2	54.5	55.7	47.1	50.3	54.2
nllb-clip-large@v1	39.1	36.2	36.0	32.0	33.9	35.4	59.9	56.5	56.0	**49.3**	50.4	54.4
M-CLIP/XLM-Roberta-Large-Vit-B-16Plus	48.0	46.1	45.4	32.9	40.3	42.5	63.2	**61.4**	59.3	48.3	**54.8**	57.4
ViT-L-14@laion400m	72.3	48.2	49.9	2.7	4.5	35.5	64.5	26.7	38.3	1.4	1.7	26.5
openai/clip-vit-large-patch14	**75.6**	46.7	49.6	6.6	3.5	36.4	59.4	19.9	28.5	4.1	1.3	22.6
DINOv2-MpNet (Ours)	73.4	**61.6**	**58.3**	**43.2**	**49.3**	**57.1**	**70.7**	60.6	**60.6**	45.6	52.7	**58.0**

- Swap out encoders for flexibility.
- Using **Multilingual-MpNet** with DinoV2 and **trained only on English captions**
- We outperform models trained with muti-lingual data like M-CLIP, nllb-CLIP and CLIP models trained on LAION-5B
- Shows that strong multilingual features of MpNet is retained in the joint embedding space after alignment.

0-shot segmentation; English / Multilingual

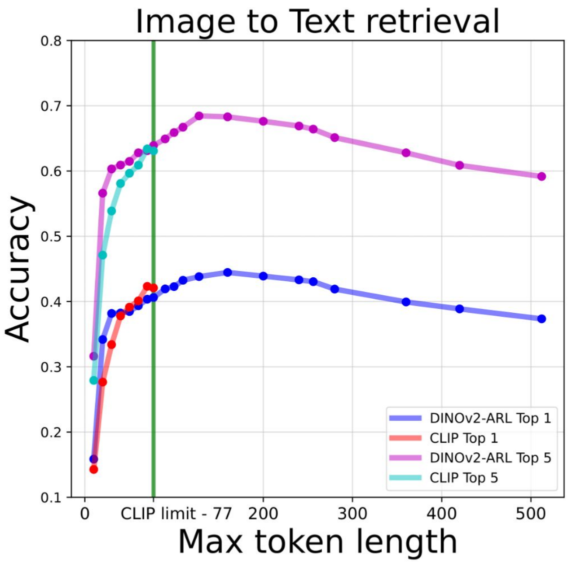


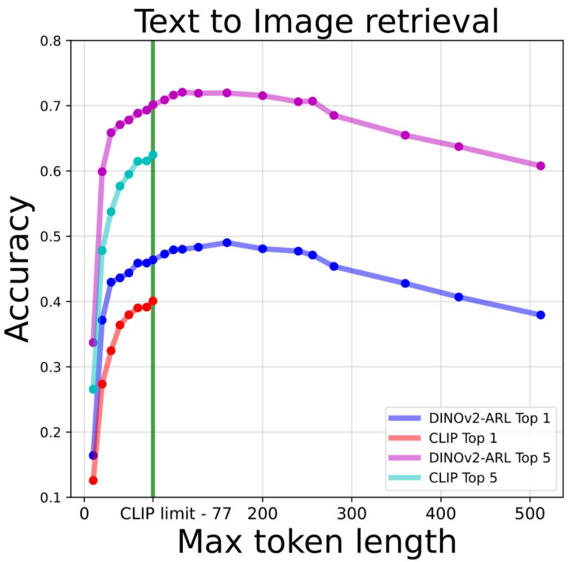
Model	Pascal VOC	Pascal Context
OpenAI-CLIP-VIT-L*	23.46	14.25
SPARC	27.36	21.65
DINOv2-ARL	**31.37**	**24.61**

Language	CLIP	*DINOv2-MpNet*
EN	23.46	29.07
ES	18.86	28.69
ZH	8.46	28.06
FR	15.12	28.48
DE	21.30	27.91
RU	5.72	26.85

- DINOv2's localization strength is preserved after alignment.
- On English Pascal VOC, we outperform SPARC, despite using no fine-grained supervision.
- On multilingual VOC, CLIP fails on non-English prompts—our model segments accurately using multilingual text.

Long context retrieval





- Our DINOv2-ARL model handles queries longer than 77 tokens.
- Powered by All-Roberta-Large, results improve up to 200 tokens.
- Trained only on normal-length captions—yet retains long-context understanding in the joint space.

Conclusion

- **Semantically similar vision and language encoders can be aligned using simple projection layers.**
- **This is the first practical application of the Platonic Representation Hypothesis**, showing that meaningful alignment is possible with minimal data and compute.
- Our **framework achieves CLIP-level zero-shot performance** using **orders of magnitude less data and compute**.
- The **modular design** allows flexible adaptation—just **swap in new frozen vision or text encoders** for new tasks or domains.
- This makes **multimodal research accessible to researchers with limited compute resources**.
- All training code, data curation scripts, and the **concept-rich dataset** are available on our GitHub.
- We envision a future where multimodal alignment can be done for any modality—**audio, 3D, text—by simply aligning frozen unimodal encoders**. Think of it as **Sentence Transformers—but for multimodal alignment. Freeze-align** it!