Do Vision and Language Encoders Represent the World Similarly?

Supplementary Material

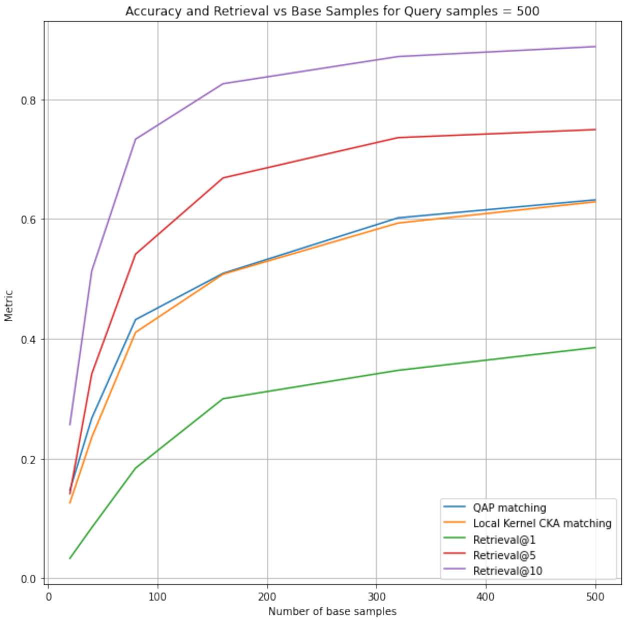


Figure A.1. **Accuracy and Retrieval Scores** of QAP Matching and Local CKA-based retrieval as the number of base samples is varied, keeping the number of query samples fixed at 500.

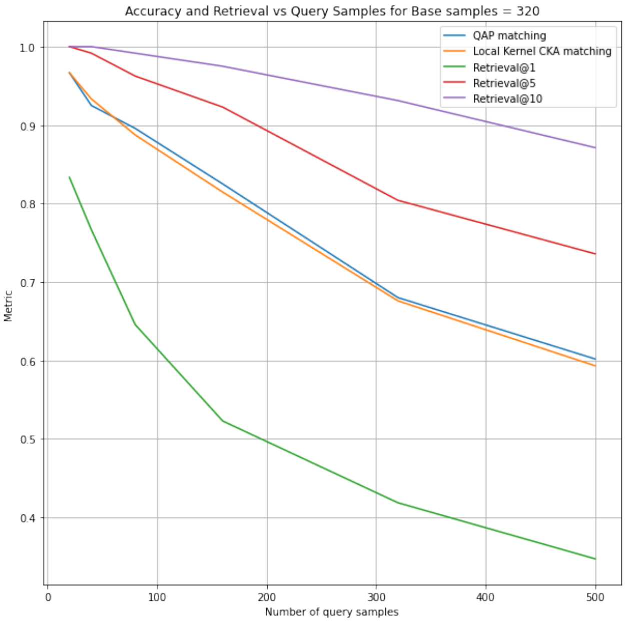


Figure A.2. **Accuracy and Retrieval Scores** of QAP Matching and Local CKA based retrieval as the number of query samples is varied, keeping the number of base samples fixed at 320.

A. Varying the Number of Samples

In Figure A.1, we show QAP and local CKA matching accuracies and retrieval scores for different number of base samples M, keeping the number of query samples N constant at 500. It can be observed that as M increases, accuracy/retrieval scores improve, demonstrating the importance of seed initialization for matching algorithms. Figure A.2 shows the accuracy/retrieval scores as N the number of query samples changes keeping the number of base samples constant at M=320. We see that QAP matching accuracy as local CKA-based retrieval scores decrease with increase in N, but we still get 70% matching accuracy when $\frac{M}{N} = 1$.

B. Vision and Text Encoders

CKA is measured on combinations of a wide variety of vision and text encoders to examine the impact of: model sizes, dataset regimes, and training paradigms on vision-language alignment. This analysis also identifies the optimal pair of unaligned vision and text encoder for caption-matching tasks. Huggingface's transformers library is utilized for vision models, while the sentence transformers library is employed for text encoders. Table A.1 details the vision models, their training data, paradigms, and model types and sizes. Similarly, Table A.2 presents information on various text encoders. The study covers three training paradigms for vision models: supervised, self-supervised, and language-supervised, with training dataset sizes ranging from 1 million to 400 million images. Text encoders predominantly use sentence transformers, trained for semantic search using a contrastive sentence pairs loss, with dataset sizes varying from 500k to 2B.

Kernel CKA of various model combinations is presented in Table A.13. The top-performing text encoder trained exclusively on text information is identified as All-Roberta-large-v1 paired with DINOv2, achieving a CKA of 0.706. Consequently, All-Roberta-large-v1 is selected as the text encoder for all tasks and experiments in the main paper, except for cross-lingual experiments. For these, paraphrase-multilingual-mpnet-base-v2 emerges as the most effective text encoder.

Figure A.3 illustrates the relationship between CKA and text model size across different vision encoder types, training paradigms, and sizes. It is observed that text model size has a limited impact on achieving high CKA with the vision model. Well-trained vision models on large datasets consistently show high kernel CKA with text encoders, regardless of text model size. For instance, language-supervised

Table A.1. **Image Encoders Summary.** List of hugging face vision encoder names and information regarding their train data, paradigm, dataset size, model type, and model sizes for the comparison in Figure A.3 and Table A.13.

Model Name	**Training Data**	**Training Paradigm**	**Model Type**	**Training Data Size**	**Model Size**
facebook\dino-vits8	ImageNet-1k	DinoV1	vit-small	1.2	22
openai\clip-vit-large-patch14-336	CLIP-400M	Language Supervised	vit-large	400	307
facebook\dinov2-base	LVD-142M	DinoV2	vit-base	142	86
facebook\dinov2-small	LVD-142M	DinoV2	vit-small	142	22
facebook\dinov2-large	LVD-142M	DinoV2	vit-large	142	307
facebook\dinov2-giant	LVD-142M	DinoV2	vit-giant	142	1000
openai\clip-vit-base-patch16	CLIP-400M	Language Supervised	vit-base	400	86
facebook\dino-vitb8	ImageNet-1k	DinoV1	vit-base	1.2	86
timm\convnext_base.fb_in1k	ImageNet-1k	Supervised	convnext-base	1.2	89
timm\convnext_tiny.fb_in1k	ImageNet-1k	Supervised	convnext-tiny	1.2	29
facebook\convnext-base-224-22k	ImageNet-21k	Supervised	convnext-base	14.1	89
timm\convnext_base.fb_in22k	ImageNet-21k	Supervised	convnext-base	14.1	89
timm\vit_base_patch16_224.augreg_in21k	ImageNet-21k	Supervised	vit-base	14.1	86
timm\vit_small_patch16_224.augreg_in1k	ImageNet-1k	Supervised	vit-small	1.2	22

Table A.2. **Text Encoders Summary.** List of huggingface text encoder names and information regarding their train data, paradigm, dataset size, and model sizes for the comparison in Figure A.3 and Table A.13

Model Name	**Model Size**	**Train Data**	**Training Paradigm**	**Training Data Size**
all-mpnet-base-v1	109	multiple datasets	contr. sent.	1.12B sent. pairs
gtr-t5-base	110	multiple datasets	contr. sent.	2B sent. pairs
paraphrase-MiniLM-L12-v2	33	multiple datasets	contr. sent.	10M sent. pairs
gtr-t5-large	335	multiple datasets	contr. sent.	2B sent. pairs
all-mpnet-base-v2	109	multiple datasets	contr. sent.	1.12B sent. pairs
average_word_embeddings_komninos	66	Wiki2015	skipgram	2 billion words
average_word_embeddings_glove.6B.300d	120	Wiki2014, GigaWord 5	glove	6 billion tokens
all-MiniLM-L12-v1	33	multiple datasets	contr. sent.	1B sent. pairs
openai_clip-vit-large-patch14	123	CLIP-400M	contr. img-text	400M image-text pairs
all-MiniLM-L12-v2	33	multiple datasets	contr. sent.	1B sent. pairs
all-MiniLM-L6-v2	22	multiple datasets	contr. sent.	1B sent. pairs
sentence-t5-base	110	multiple datasets	contr. sent.	2B sent. pairs
msmarco-distilbert-dot-v5	66	MSMarco	contr. sent.	500k sent. pairs
paraphrase-MiniLM-L3-v2	17	multiple datasets	contr. sent.	10M sent. pairs
paraphrase-albert-small-v2	11	multiple datasets	contr. sent.	10M sent. pairs
all-MiniLM-L6-v1	22	multiple datasets	contr. sent.	1B sent. pairs
all-distilroberta-v1	82	OpenWebTextCorpus	contr. sent.	1B sent. pairs
sentence-t5-large	335	multiple datasets	contr. sent.	2B sent. pairs
All-Roberta-large-v1	355	multiple datasets	contr. sent.	1B sent. pairs
msmarco-bert-base-dot-v5	109	MSMarco	contr. sent.	500k sent. pairs
sentence-t5-xxl	4870	multiple datasets	contr. sent.	2B sent. pairs
paraphrase-TinyBERT-L6-v2	66	multiple datasets	contr. sent.	10M sent. pairs
sentence-t5-xl	1240	multiple datasets	contr. sent.	2B sent. pairs
gtr-t5-xxl	4870	multiple datasets	contr. sent.	2B sent. pairs
paraphrase-distilroberta-base-v2	82	multiple datasets	contr. sent.	10M sent. pairs
gtr-t5-xl	1240	multiple datasets	contr. sent.	2B sent. pairs

models (green) and DINOv2 models, which are trained on datasets with hundreds of millions of instances (such as LVD-142's 142 million images and CLIP-400M's 400 million image-caption pairs), demonstrate high CKA with language encoders of various sizes.

C. Layerwise CKA Analysis

Figure A.4, Table A.3, and Table A.4 show the progression of CKA and QAP matching scores across layers for both text and vision models. We explore two configurations: one involves comparing layers of All-Roberta-large-V1 and DINOv2 VIT-L/14, while the other examines layers

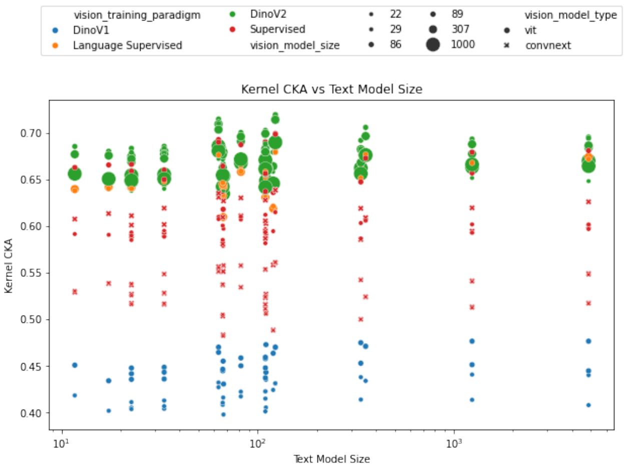


Figure A.3. **CKA *vs.* text model size** for vision encoders of different training paradigms, model types, and model sizes. We see that text model size is not the most important for high semantic similarity with vision models.

of CLIP's vision and text hidden states. For CLIP, the layer $proj$ points to the final image and text embeddings that were passed through the final projection layers. In the first configuration, CKA and QAP scores gradually improve where the image model layer has a far greater effect on the similarity than the text model layer. On the other hand, the second configuration reveals that the QAP matching score in CLIP manifests prominently in the absolute last layers of both the vision/text encoders.

As shown in Table A.3, the CLIP model obtains a significant jump in matching score after the projection head, highlighting the central role of this layer in aligning text and image modalities within a unified representation space. Here, the QAP matching accuracy does not follow a linear increase over the layers for CLIP, but rather suddenly jumps from 0.29 to 0.79 from the last layer to the projection head. This likely suggests that most of the CLIP performance comes from the projection heads ensuring a high statistical similarity. In contrast, Table A.4 shows that DINOv2 and All-Roberta-large-v1 demonstrate a consistent improvement in the matching accuracy across successive layers, suggesting an inherent alignment process within their architectures in a hierarchical way. Here, the QAP matching accuracy linearly increases for the DINOv2 and All-Roberta-large-v1 combination when we fix the last layer of All-Roberta-large-v1 and vary the layers of DINOv2. Inversely, when we fix the last layer of DINOv2 and vary the layers of the text encoder, the QAP starts high at 0.44 and reaches 0.68 at the top layer, thus, we hypothesize that the text encoder representations do not change as much as the image representations.

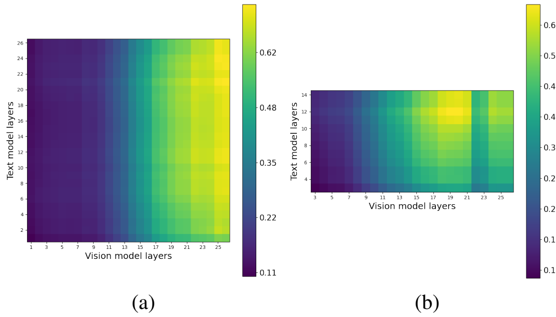


Figure A.4. **Layer-wise CKA heatmap illustration.** The heatmaps depict the CKA scores obtained by varying the layers from which the text and visual embeddings are taken. **On the left:** CKA scores for All-Roberta-large-v1 and DINOv2 unaligned combination. **On the right:** CKA scores for CLIP text and vision encoders. In both cases, we observe that the CKA scores are low for earlier layer embeddings of the vision model and they improve when the embeddings later layers are considered. This illustrates that both aligned and unaligned text-vision encoders behave similarly in terms of the cross-modal similarity w.r.t. CKA.

Table A.3. **QAP accuracy for different layers** of vision and text encoder of CLIP model.

		Vision					
		6th	11th	16th	21st	26th	proj
Text	6th	0.02	0.022	0.022	0.098	0.126	0.118
	11th	0.028	0.038	0.016	0.248	0.278	0.278
	14th	0.026	0.03	0.036	0.238	0.282	0.296
	proj	0.038	0.026	0.034	0.622	0.716	0.792

Table A.4. **QAP accuracy for different layers** of DINOv2 and All-Roberta-large-v1 models.

		Vision				
		6th	11th	16th	21st	26th
Text	6th	0.008	0.020	0.150	0.314	0.448
	11th	0.010	0.022	0.146	0.360	0.498
	16th	0.008	0.016	0.194	0.334	0.500
	21st	0.002	0.004	0.148	0.420	0.538
	26th	0.008	0.016	0.198	0.450	0.672

D. Mathematical Relationship between Local CKA-based Retrieval and Relative Representations

In this section, we provide derivations that show that the relative representations method [3] can be seen as a particular case of our proposed localCKA method. Denote the set of query and base representations samples respectively as $\mathbf{Q}_A = [\boldsymbol{q}_1^A, \ldots, \boldsymbol{q}_N^A] \in \mathbb{R}^{d_A \times N}$ and $\mathbf{B}_A = [\boldsymbol{b}_1^A, \ldots, \boldsymbol{b}_M^A] \in \mathbb{R}^{d_A \times M}$, where $A \in \{I, C\}$ for images

Table A.5. **Impact of adding noise to the embeddings**. Performance comparison, in terms of matching accuracy, between relative representations [3] and our global CKA-based QAP approach is shown for the image-caption matching task with 320 base samples and 500 query samples on COCO validation set. Gaussian noise with std-dev (σ) being a multiple of the embeddings std-dev is added to both image and textual embeddings. Noise level of 0 ($\sigma = 0$) denotes the performance for the original embeddings. The relative performance drop for a noise level from its reference ($\sigma = 0$) is shown in parenthesis. In comparison to relative representations, our QAP approach performance drops at a slower rate as σ increases, illustrating better noise robustness for our approach.

Method	**Noise Level** (σ)					
	0.0	0.1	0.2	0.3	0.4	0.5
Relative representations [3]	47.3	45.3 (↓4.4)	44.2 (↓6.5)	41.3 (↓12.7)	39.0 (↓17.6)	35.6 (↓24.8)
Ours (QAP)	53.9	53.7 (↓0.3)	51.8 (↓3.9)	48.7 (↓9.5)	46.9 (↓13.0)	43.3 (↓19.6)

and captions, the retrieval matrix for the relative representations (RR) method is therefore given by:

$$\mathbf{R}^{\text{RR}} = \mathbf{Q}_I^\top \mathbf{B}_I \mathbf{B}_C^\top \mathbf{Q}_C \in \mathbb{R}^{N \times N}.$$

From which, for instance, the i-th image query is mapped to its corresponding caption via:

$$\arg\max_j R_{ij}^{\text{RR}} = \arg\max_j \, (\boldsymbol{q}_i^I)^\top \mathbf{B}_I \mathbf{B}_C^\top \boldsymbol{q}_j^C. \qquad (1)$$

Whereas, our proposed localCKA method constructs the retrieval matrix $\mathbf{R}^{\text{Ours}}$ having entries $R_{ij}^{\text{Ours}} = \text{localCKA}\left(\boldsymbol{q}_i^I, \boldsymbol{q}_j^C\right)$ with:

$$\text{localCKA}\left(\boldsymbol{q}_i^I, \boldsymbol{q}_j^C\right) = \text{CKA}\left(\mathbf{K}_{[\mathbf{B}_I, \boldsymbol{q}_i^I]}, \mathbf{K}_{[\mathbf{B}_C, \boldsymbol{q}_j^C]}\right). \qquad (2)$$

In particular, taking the particular case of the linear kernel and defining the CKA score as the trace of the product of two kernels, i.e., $\text{CKA}(\mathbf{K}, \mathbf{L}) = \text{tr}\,(\mathbf{KL})$. We first have, for $A \in \{I, C\}$:

$$\mathbf{K}_{[\mathbf{B}_A, \boldsymbol{q}_i^A]} = [\mathbf{B}_A, \boldsymbol{q}_i^A]^\top [\mathbf{B}_A, \boldsymbol{q}_i^A] = \begin{bmatrix} \mathbf{B}_A^\top \mathbf{B}_A & \mathbf{B}_A^\top \boldsymbol{q}_i^A \\ \left(\mathbf{B}_A^\top \boldsymbol{q}_i^A\right)^\top & \|\boldsymbol{q}_i^A\|^2 \end{bmatrix}.$$

Hence, we have:

$$\begin{aligned} \text{tr}\left(\mathbf{K}_{[\mathbf{B}_I, \boldsymbol{q}_i^I]} \mathbf{K}_{[\mathbf{B}_C, \boldsymbol{q}_j^C]}\right) &= \text{tr}\left(\mathbf{B}_I^\top \mathbf{B}_I \mathbf{B}_C^\top \mathbf{B}_C\right) \\ &+ 2 \underbrace{\left(\boldsymbol{q}_i^I\right)^\top \mathbf{B}_I \mathbf{B}_C^\top \boldsymbol{q}_j^C}_{\text{relative representations term}} + \|\boldsymbol{q}_i^I\|^2 \|\boldsymbol{q}_j^C\|^2. \end{aligned}$$

Therefore, in this particular case, there is equivalence between our method and the relative representations method, since $R_{ij}^{\text{Ours}} = R_{ij}^{\text{RR}} + c$ where c is a constant scalar if the representations are normalized. As such, the relative representations method falls within our proposed localCKA method if one considers the linear kernel and takes the trace instead of the HSIC metric. Therefore, our proposed method is more general since it relies on general kernel functions and the HSIC metric, which might explain its performance.

Impact of noise addition: Table A.5 shows the performance comparison between relative representations [3] and our global CKA-based QAP approach for the image-caption matching task with 320 base samples and 500 query samples on COCO validation set. For this experiment, 10 trials were conducted with different seeds and clustering of base samples was employed. Gaussian noise with std-dev (σ) being a multiple of the embeddings std-dev is added to both image and textual embeddings. The performance of original embeddings is also shown for reference (noise level of 0, *i.e.*, $\sigma = 0$). The relative performance drop for a noise level from its reference ($\sigma = 0$) is shown in parenthesis. Compared to relative representations, our QAP approach performance drops at a slower rate as σ increases. *E.g.*, for $\sigma = 0.2$, relative representations matching accuracy drops 6.5% from it maximum of 47.3, while ours is more robust and drops only 3.9% from its maximum of 53.9 when $\sigma = 0$. These results show that our QAP approach is more robust to noise addition, in comparison to relative representations.

E. Other text encoders

Evaluating on COCO with M=320 and N=500, Table A.6 shows that DINOv2-large achieves high QAP accuracy and retrieval performance when combined with different text encoders. This underscores the potential of pairing well-trained sentence and vision encoders for achieving high semantic similarity between image and text embeddings

Table A.6. Comparison of CKA, QAP acc. and local CKA retrieval for different text encoders with DINOv2-large image encoder.

Text Encoder	Kernel CKA	QAP Acc.	Ret @ 5
all-roberta-large-v1	0.690	64.93	77.27
paraphrase-distilroberta-base-v2	0.689	65.07	76.33
paraphrase-mpnet-base-v2	0.695	68.20	81.07
sentence-t5-large	0.660	57.87	69.13
sentence-t5-xxl	0.677	63.40	73.00

F. Simple projection

We trained a 2-layer MLP on frozen DINOv2-large encoder till convergence using CLIP loss and MSE loss. For fair comparison with our setting, we use 320 training and 500 query image-text samples. Results in Table A.7 are averaged over 3 seeds. Notably, QAP matching and local-CKA retrieval excel over projection learning, which demands hyperparameter tuning. In contrast, QAP and local-CKA provide a novel, training-free mechanism to evaluate encoder representational similarity, demonstrating effective latent space communication.

G. Effect of unimodal tasks on alignment

Table A.8 shows using ViT, DETR, DPT, and SegFormer vision encoders for local-CKA and QAP matching on COCO captions (M=320, N=500). ViT is trained on ImageNet-1k (classification), DETR on COCO 2017 (detection), DPT on 1.4M depth images (depth estimation), and SegFormer is fine-tuned on ADE20k (semantic segmentation). Results indicate that classification models exhibit higher semantic similarity to all-roberta-large text encoder in QAP accuracy and local-CKA scores than pixel-level tasks such as object detection, segmentation, and depth estimation.

Table A.7. QAP acc. and Top-5 retrieval scores on COCO.

Method	QAP acc	Ret @ 5
Proj. + MSE	59.8	73.0
Proj. + CLIP	55.4	68.1
QAP	**65.9**	-
Local CKA	64.3	**76.0**

Table A.8. Unimodal tasks' effect on image-text alignment.

Vision model	QAP acc	Ret @ 5
ViT	35.3	56.1
DETR	26.5	39.8
DPT	22.7	34.1
Segformer	16.8	33.4

H. Additional Retrieval Results

While the performance on the image retrieval task was reported in Table 2 of the main manuscript, here in Table A.9, we show the NoCaps and Coco caption retrieval results in the reverse setting. In this configuration, the retrieving objective shifts to finding the correct caption from a pool of N captions when given a single image. The matching objective remains consistent, but, instead of shuffling the captions, the images themselves are shuffled. While the matching accuracies express minimal changes in this setting, the retrieval accuracies display notable discrepancies.

A plausible explanation for the reduced retrieval scores associated with the relative representation method is the heightened semantic variability inherent in the image domain compared to the caption domain. A considerable number of images share very similar captions, leading to a compressed semantic space for the captions. Consequently, caption embeddings become more closer to one another, making the retrieval a lot harder.

I. Additional Cross-Lingual Matching Results

For completeness, we report the results in Table A.10 for the reverse setting of the cross-lingual image caption matching/retrieval task mentioned in the main paper. Given N captions in say, German, and N shuffled images the objective is to match each German caption with the correct image. In retrieval, the goal is to select the most fitting image from the retrieval set given a German caption. We notice that the matching accuracies remain the same as the direction doesn't affect the matching. However, in the case of reverse retrieval, we notice that CLIP's retrieval@5 drops by over 4.5% on average when compared to our local CKA based retrieval of 2.1%.

In Table A.11 we report the results for when we use language-specific BERT Sentence encoders for the cross-lingual caption matching/ retrieval task for 5 languages. For all these cases, the vision encoder is kept fixed as OpenAI's CLIP-VIT-L-14 trained on English image, caption pairs. We notice that the semantic alignment with the vision encoder in terms of CKA as well as matching/retrieval performance drops with language-specific encoders when compared to using a multi-lingual model like multilingual-mpnet-base-v2. We believe this could be due to the multilingual model being trained on a lot more data in comparison to the language-specific ones thus resulting in more meaningful embedding spaces.

J. Qualitative results

In Table A.12, we present instances of retrieval mispredictions where the original image fails to rank within the top five closest images to the given caption, as determined by local Kernel CKA method. Building upon the experimental methodology outlined in the main paper, we selected 320 base samples and conducted local Kernel CKA retrieval using an additional 500 query samples. We used All-Roberta-large-v1 for text embeddings and DINOv2 ViT-L/14 for image embeddings. The results distinctly illustrate that despite the failure to retrieve the exact original image, the alternative images identified in the top five still exhibit a considerable degree of semantic similarity to the provided caption. This underscores the robustness of the local Kernel CKA retrieval approach, revealing its capability to identify images that, while not the precise match, maintain semantic coherence with the specified caption.

Table A.9. **Reverse Caption Retrieval Results for COCO and NoCaps**. In this setting, the retrieval objective is, given one image, to retrieve the correct caption from the overall set of N captions. The matching objective remains quite similar but instead of shuffling the captions, this time, the images are shuffled.

Method	**Vision Model**	**NoCaps** [1]		**COCO** [2]	
		Matching accuracy	Top-5 retrieval	Matching accuracy	Top-5 retrieval
Cosine Similarity*	CLIP [5]	99.5	99.6	97.1	98.5
Linear regression	CLIP-V [5]	63.6	70.1	72.6	83.9
	ConvNeXt [6]	22.8	38.9	43.8	65.7
	DINOv2 [4]	46.8	59.9	56.2	75.9
Relative representations [3]	CLIP-V [5]	61.3	3.0	61.6	2.9
	ConvNeXt [6]	25.5	2.7	38.6	12.9
	DINOv2 [4]	45.9	38.1	47.7	43.7
Ours: QAP	CLIP-V [5]	67.3	-	72.8	-
	ConvNeXt [6]	45.9	-	65.1	-
	DINOv2 [4]	58.5	-	65.9	-
Ours: Local CKA	CLIP-V [5]	65.1	65.9	71.9	80.5
	ConvNeXt [6]	44.8	33.0	63.8	74.3
	DINOv2 [4]	55.7	64.2	64.3	76.0

Table A.10. **Cross-Lingual image matching and retrieval performance comparison. Here we use multilingual captions to retrieve images from the COCO validation set.** Using QAP and local CKA-based methods we are able to do cross-lingual image matching/retrieval using CLIP's ViT-L vision encoder and a multi-lingual sentence transformer paraphrase-multilingual-mpnet-base-v2. While CLIP performs well on the Latin languages, it degrades on non-Latin languages. In comparison, our QAP and Local-CKA-based methods perform comparably in Latin languages while outperforming non-Latin languages, highlighting the efficacy of our training-free transfer approach.

Language		**Kernel CKA**		**Matching Accuracy**				**Retrieval @ 5**	
		CLIP	Ours	CLIP	Relative[3]	Linear	Ours (QAP)	CLIP	Ours (Local)
Latin	de	0.472	0.627	43.5	35.0	19.3	39.7	54.9	57.2
	en	0.567	0.646	80.9	52.5	25.6	51.3	90.4	66.7
	es	0.471	0.634	50.4	37.8	19.7	40.9	63.9	57.9
	fr	0.477	0.624	50.8	37.5	18.8	40.3	65.9	56.9
	it	0.472	0.638	41.9	37.2	19.7	38.7	52.9	57.0
Non-Latin	jp	0.337	0.598	12.9	28.3	15.2	30.2	17.8	48.6
	ko	0.154	0.620	0.9	30.4	15.3	31.3	2.2	48.4
	pl	0.261	0.642	8.1	36.6	21.0	40.0	15.7	55.9
	ru	0.077	0.632	1.7	31.8	16.3	34.8	3.5	53.9
	tr	0.301	0.624	7.8	35.8	18.7	38.9	14.6	53.1
	zh	0.133	0.641	2.4	36.5	19.2	39.9	4.8	53.7
	Avg.	–	–	27.4	36.3	18.9	**38.7**	35.1	**55.4**

Table A.11. **Language-specific encoders for cross-lingual caption matching/retrieval for 5 languages**. Language-specific encoders have less semantic similarity with the vision encoder in terms of CKA as well as poorer matching/accuracy performances when compared to multi-lingual models like multilingual-mpnet-base-v2 which is reported in Table 4.

Language	**Language model**	**CKA**	**Linear**	**Relative**	**QAP**	**Retrieval@5**
es	hiiamsid\sentence_similarity_spanish_es	0.568	15.9	25.1	28.6	50.0
fr	dangvantuan\sentence-camembert-large	0.569	22.5	31.5	35.0	53.1
it	nickprock\sentence-bert-base-italian-uncased	0.543	16.0	22.0	26.4	47.8
jp	colorfulscoop\sbert-base-ja	0.457	9.2	12.1	14.5	33.7
tr	emrecan\bert-base-turkish-cased-mean-nli-stsb-tr	0.564	23.1	34.7	38.3	54.3

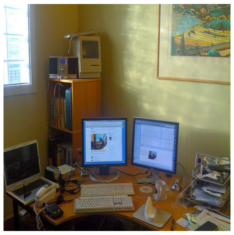
Original Image	Caption	Top-3 Retrieved Images		
	Two desktop computers sitting on top of a desk.			
	A mother and baby elephant walking in green grass in front of a bond.			
	a man is riding a surfboard at the beach			
	The Big Ben clock tower towering over the city of London.			
	A computer mouse is beside a notebook computer.			

Table A.12. **Local Kernel CKA Retrieval Mispredictions.** In accordance with the experimental protocol detailed in the main paper, we selected 320 base samples and conducted local Kernel CKA retrieval using an additional 500 query samples. Presented above are five example prediction retrievals for instances where the original image failed to secure a position within the top-5 retrievals. We observe that although the original image was not in the retrieved top-5, the retrieved images (top-3 shown here) closely resemble the corresponding caption, thereby highlighting the efficacy of our approach.

Table A.13. **CKA for combinations of different vision and text encoders.** V, V_tr, V_tr_size, V_mod_size stand for Vision model name, Vision train set, Vision train set size, and Vision model size respectively. T_mod_size stands for text model size. OpenAI's CLIP text encoder shows highest CKA with facebook dinoV2base closely followed by All-Roberta-large-v1. We make use of All-Roberta-large-v1 as the language encoder for all donwstream tasks and analysis in main text because All-Roberta-large-v1 has been trained using only text data and can be considered a purely textual encoder.

V	T	CKA	V_tr	V_tr_p	V_tr_size	V_mod_size	T_mod_size
facebook_dinov2-base	openai_clip-vit-large-patch14	0.719	LVD-142M	DinoV2	142	86	123
facebook_dinov2-base	All-Roberta-large-v1	0.706	LVD-142M	DinoV2	142	86	355
timm_vit_base_patch16_224.augreg_in21k	openai_clip-vit-large-patch14	0.698	ImageNet-21k	Supervised	14.1	86	123
facebook_dinov2-large	sentence-t5-xxl	0.684	LVD-142M	DinoV2	142	307	4870
openai_clip-vit-large-patch14-336	All-Roberta-large-v1	0.677	CLIP-400M	Lang. Supervised	400	307	355
facebook_dinov2-large	sentence-t5-large	0.668	LVD-142M	DinoV2	142	307	335
facebook_dinov2-small	sentence-t5-xl	0.661	LVD-142M	DinoV2	142	22	1240
facebook_dinov2-small	all-mpnet-base-v2	0.655	LVD-142M	DinoV2	142	22	109
facebook_dinov2-small	all-MiniLM-L6-v1	0.644	LVD-142M	DinoV2	142	22	22
facebook_convnext-base-224-22k	gtr-t5-xxl	0.626	ImageNet-21k	Supervised	14.1	89	4870
timm_vit_small_patch16_224.augreg_in1k	gtr-t5-xl	0.602	ImageNet-1k	Supervised	1.2	22	1240
timm_convnext_base.fb_in22k	all-MiniLM-L6-v2	0.590	ImageNet-21k	Supervised	14.1	89	22
timm_convnext_tiny.fb_in1k	gtr-t5-xl	0.540	ImageNet-1k	Supervised	1.2	29	1240
timm_convnext_base.fb_in1k	msmarco-bert-base-dot-v5	0.512	ImageNet-1k	Supervised	1.2	89	109
facebook_dino-vitb8	msmarco-distilbert-dot-v5	0.445	ImageNet-1k	DinoV1	1.2	86	66
facebook_dino-vits8	all-mpnet-base-v2	0.423	ImageNet-1k	DinoV1	1.2	22	109
facebook_dino-vits8	paraphrase-TinyBERT-L6-v2	0.398	ImageNet-1k	DinoV1	1.2	22	66

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019. 6

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6

[3] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*, 2022. 3, 4, 6

[4] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[6] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023. 6