Supplementary: Harnessing Frozen Unimodal Encoders for Flexible Multimodal Alignment

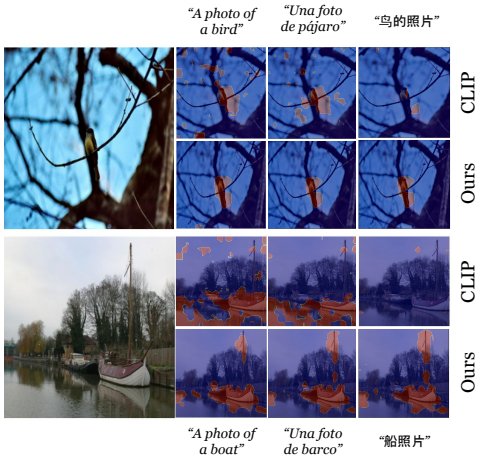

Figure A.1. Compared to CLIP, our approach of aligning DINOv2-MpNet achieves improved segmentation maps focusing on the relevant objects in the multilingual setting.

A. Appendix

A.1. Unlocking parts of text and vision encoders

We evaluated our model with different parts of the vision and text encoders unlocked for DinoV2-ARL, shown in Tab.A.2. Similar to Lit [36] we find that unlocking the vision encoder (*e.g.*, via BitFit [35]) reduced performance, while full unlock resulted in unstable training. In contrast, unlocking the text encoder or applying BitFit$_{text}$ slightly improved performance with increased training costs.

A.2. Training CLIP with same dataset

We compare our approach against CLIP-ViT-L models trained from scratch, and projector-only trained in Tab. A.3. We see that our 20M dataset is not enough to train the CLIP model (427M params) from scratch. Meanwhile, projector-only training of CLIP improves over OpenAI CLIP on COCO I2T and achieves competitive performance on Imagenet. Notably, none of the trained CLIP models outperform DINOv2-ARL.

A.3. Multi-lingual 0-shot Semantic Segmentation

The lower compute and paired data requirements of the framework lead to application flexibility simply by swapping the unimodal encoders. (see Sec. 6.2-6.4 in the main paper). An additional advantage of this flexibility is showcased in Fig. A.1 and Tab. A.1, where we use our aligned DINOv2-MpNet to perform multi-lingual semantic segmentation. Our segmentation scores stay consistent with different languages while CLIP often fails on non-english languages.

Table A.1. Multilingual Segmentation IOU scores.

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

Table A.2. Unlocking Encoders.

Method (15 epochs)	Imagenet	COCO I2T
$BitFit_{all}$	67.67	53.16
$BitFit_{text}$	74.58	56.72
Text unlock	**75.90**	**56.62**
Projectors	75.04	56.32

Table A.3. CLIP on our dataset.

Method (30 epochs)	Imagenet	COCO I2T
$CLIP_{scratch}$	50.30	36.12
$CLIP_{openai}$	75.32	56.31
$CLIP_{projectors}$	72.10	59.04
DINOv2-ARL	**76.45**	**60.14**

A.4. Toy Example using Random Latent Model

Similar to Sec. 3.2 (main paper), here we investigate whether semantically similar encoder embedding spaces can be aligned through a simple projection transformation, using a random latent model.

In our experiment, we generated 10^3 instances of two vector sets, A and B, each containing 32 vectors of 16 dimensions. Following the approach in [12, 19], we modeled the world using a latent distribution Z, with Image and Text representations (A and B) as random independent non-linear transformations from Z with additive noise. For each sampled pair of A and B matrices, we calculated the CKA and the minimum CLIP loss. The non-linear transform was defined as a randomly initialized 2-layer MLP with ReLU non-linearity and hidden dimensions significantly larger than the input dimensions, ensuring it could universally approximate the non-linear transformation [11]. Figure A.3 was used to generate each instance.

Figure A.2 illustrates the results of this experiment, showing a clear negative correlation between CKA and minima of the CLIP loss. As CKA increases, indicating greater similarity between the similarity structures of A and B, the minima of CLIP loss consistently decreases. Despite arising from a simplified experiment, the observed strong in-

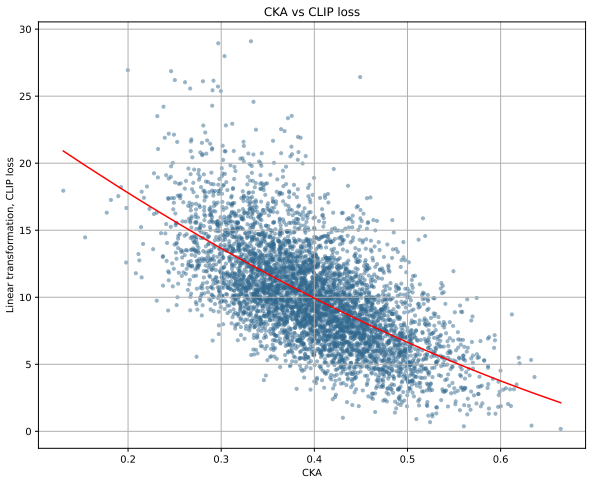

Figure A.2. **CLIP Loss minima are negatively correlated to CKA.** We plot CKA vs CLIP Loss for random instances of A and B.

```
# 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)
```

Figure A.3. **Code for initializing A and B from a latent world model Z.** Random instances of A, B are generated using random non-linear transformations of latent vector Z denoting a representation of the real world.

verse relationship between CKA and CLIP loss provides empirical support for using CKA as a predictor of alignment potential between embedding spaces. Since CLIP loss is lower-bounded by mutual information, and mutual information is correlated with HSIC, higher CKA suggests a stronger alignment between embeddings. This implies that the achievable minima of CLIP loss is lower when the embedding spaces already have a higher CKA, reflecting greater mutual information and ease of alignment.

A.5. Embedding Graph structures visualized

To visually demonstrate how CKA represents similarities in graph structures across different encoder spaces, we conducted an experiment using the MSCOCO validation set. We examined encoder outputs for DINOv2 and All-Roberta-Large-v1, before and after projection, focusing on relationships between formed clusters in both domains. For each cluster, we identify COCO detection class and COCO

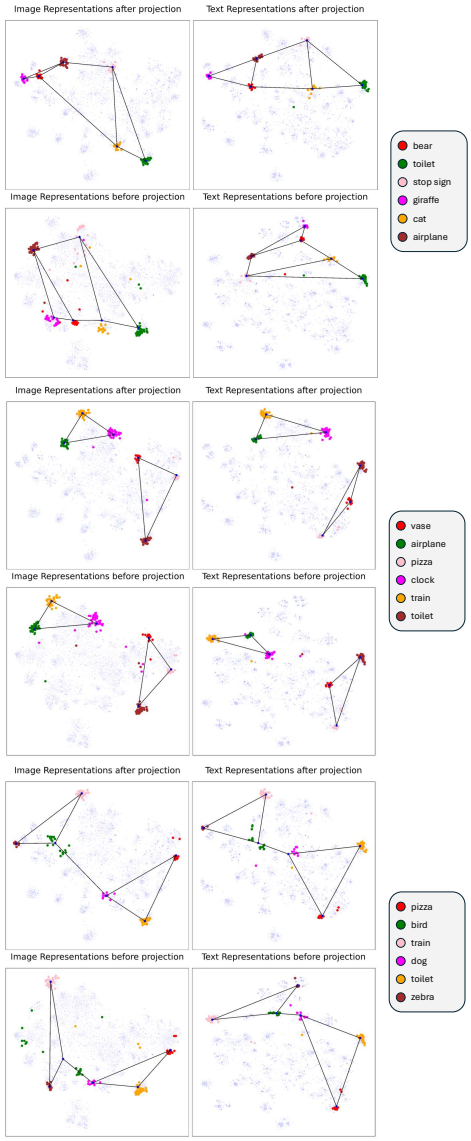

Figure A.4. TSNE visualizations of encoder outputs for six COCO detection classes. Left: DINOv2 (vision), Right: All-Roberta-Large-v1 (text).

image-caption pairs where the image contained only the respective class among its detection annotations. We then extracted encoder outputs for these samples from both vision and text encoders, before and after applying our projection layers, and applied the TSNE algorithm to visualize their structure in a lower-dimensional space. For each visualization, we pick 6 classes to highlight the shape similarities between graphs of encoder spaces.

Figure A.4 shows the resulting TSNE visualizations for the six selected classes across four conditions: vision pre-projection, vision post-projection, text pre-projection, and text post-projection. The visualizations reveal striking

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
LiLT$_{DA}$-base	0.5M	15.9	12.9	37.6	7.2	1.6	1.1	13.3	1.7	25.6	2.3	19.1
LiLT$_{LwA}$-base	0.5M	14.4	12.1	42.3	4.8	1.3	2.1	12.3	1.6	26.5	1.4	26.6
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

Table A.4. **0-shot domain transfer to classification datasets.** We compare the performance of our DINOv2-ARL projector model, trained on a 20M dataset, against CLIP models from OpenAI and LAION across various datasets. Despite the smaller training size, our model achieves a 76.3% accuracy on ImageNet, outperforming comparably sized CLIP models.

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
LiLT$_{DA}$-base	47.6	34.46	41.4	29.1
LiLT$_{LwA}$-base	56.8	41.7	47.0	33.7
DINOv2-MpNet (Ours)	84.6	71.2	58.0	42.6
DINOv2-ARL (Ours)	87.5	**74.1**	**60.1**	**45.1**

Table A.5. **Image, Text Retrieval on COCO/Flickr30k.** Our model shows comparable text retrieval scores and significantly better image retrieval results.

similarities in cluster shapes and relative positions across the different encoder spaces, particularly before projection. This visual similarity aligns with our quantitative CKA results, providing an intuitive illustration of how CKA captures structural similarities between different embedding spaces.

A.6. Comparison to LiLT

Tables A.4 and A.5 report the zero-shot domain classification and retrieval performance of LiLT models [13]. The vision encoder is initialized with the DeiT base model [31], and the text encoder is from SimCSE [9]. The LiLT$_{\text{DA}}$-base model is trained by duplicating and appending the last transformer layer, while only unlocking the last encoder and projector layers. The LiLT$_{\text{LwA}}$-base model introduces trainable layerwise adapters for both the vision and text encoders. LiLT public checkpoints are trained on 500k image-caption pairs from the COCO dataset. However, LiLT's performance lags behind CLIP models and our DINOv2-ARL projector model, primarily due to suboptimal encoder pairs and limited concept coverage in the COCO training set for alignment.

A.7. Encoder Pairs Ablations

Similar to Sec 5.1 (main paper), we train our projector configurations on various combinations of unimodal encoders using the COCO dataset and evaluate image/text retrieval accuracies on the Flickr30k test set, plotting these against CKA scores. In Fig. A.5 both the Image and Text retrieval accuracies shows a strong correlation with CKA suggesting that CKA can effectively predict which encoder pairs will align well with projector training.

A naive approach to choosing the best encoder pair is to chose the unimodal encoders with highest performance in their respective modalities, but it's not straightforward which benchmarks can be more predictive of ease of alignment. To demonstrate this, we consider the same ablation as above, but with DINOv2 and 14 different text encoders from the SentenceTransformers [26] library. We consider 2 types of text model benchmarks. 1. Sentence Embedding task or Semantic Textual Similarity (STS) is the task of evaluating how similar two texts are in terms of meaning. These models take a source sentence and a list of sentences and return a list of similarity scores. The task is evaluated using Spearman's Rank Correlation. We average over 14 datasets reported in [26, 27]. 2. Semantic Search (SS) is the task of retrieving relevant documents or passages based on the semantic content of a query. Rather than relying solely on keyword matching, semantic search models generate embeddings for both the query and the documents, allowing for retrieval based on contextual and conceptual similarity and is evaluated using Normalized Discounted Cumulative Gain (nDCG), which measure the relevance of retrieved documents in ranked lists. We average over 6 datasets reported in [26, 27].

In Fig A.6, we see that there is a clear correlation (pearson corr.=0.81, p=4e-4) between downstream Flickr30k performance and CKA on the COCO val set, suggesting that CKA is a better predictor of ease of alignment. The average unimodal performance (pearson corr.=0.47, p=0.08), as well as the semantic search (SS) performance (pearson corr.=0.13, p=0.65), are not predictive of the ease of alignment. Meanwhile, Sentence Task Similarity (STS) tasks are more predictive of downstream alignment (pearson corr.=0.72, p=0.003) but still worse than CKA and it's not intuitive which unimodal performance is to be considered.

A.8. Data Curation Implementation Details

We streamline our class collection process by precomputing CLIP text embeddings for LAION-400M and CLIP image prototype embeddings for various concepts, allowing us to run different collection methods without needing to recompute embeddings. The embedding process takes just 12 hours on two nodes with 4 A6000 GPUs each. Class-level collection is performed using GPU-accelerated PyTorch code on a single GPU, completing in under an hour. While image-to-image-prototype collection, as in [22], could yield higher-quality results, it demands significantly more GPU resources due to the need to create CLIP embeddings for all LAION-400M images. We find that caption-image-concept

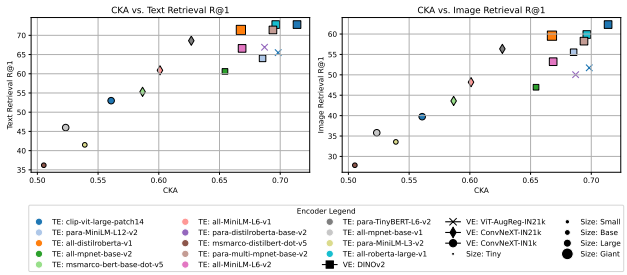

Figure A.5. **Retrieval performance vs. CKA for different encoder pairs.** Text/Image retrieval accuracies on Flickr30k are compared to CKA, calculated on the COCO val set. Models trained on COCO train set. A clear correlation exists between CKA and alignment quality (Pearson correlation = 0.92, p = 2.1e-7), as reflected in retrieval accuracies.

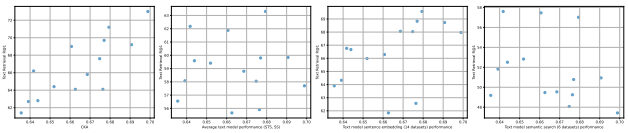

Figure A.6. **Retrieval performance vs. text model performance for DINOv2 and different text encoders.** Text/Image retrieval accuracies on Flickr30k are compared different text encoder tasks performance. CKA is more closely correlated with retrieval performance than text encoder downstream task performance on sentence embedding tasks, semantic search tasks. Models trained on COCO train set.

similarity performs well for image classification accuracy. To support efficient multi-modal model training, we release the LAION-CLASS-Collected parquets for research use.

A.9. Projector training details

We use the standard CLIP loss with a learnable temperature parameter to train the projectors while keeping the vision and text encoders frozen. For our largest experiments on the 20M MIX-CLASS-Collected dataset, we use an effective batch size of 16k and train for 30 epochs. Training is done with a cosine learning rate scheduler, ramping up to 1e-3 in the first epoch. Additional hyperparameters are detailed in the table in the appendix. The training process takes 50 hours on a node with 8 A100 GPUs.

A.10. 0-shot Segmentation Evaluation

In DINOV2-ARL, we perform 0-shot segmentation by computing cosine similarities between each patch and all the ground truth classes and subsequently upscaling to the target size. Each patch is then classified into a corresponding class. Consistent with previous studies, the intersection over union (IoU) is computed solely for the foreground classes. In the zero-shot segmentation process of CLIP models, we employ a technique similar to [37] to alleviate the opposite visualization problem in CLIP models [16]. The patch embeddings from the penultimate layer are passed through the value layer and output MLP of the final self-attention block, followed by projection into the joint embedding space using the vision projector. Meanwhile, our DINOv2-ARL model considers patch embeddings projected into the joint embedding space by the patch projector and augments them with the projected CLS token in a residual manner.

A.11. Multi-Lingual Full Results

Another significant advantage of using only Projectors to align modalities is the ability to swap the text encoder with multi-lingual encoders trained on various languages, thus potentially extending a CLIP model to accommodate any language. This feature is particularly beneficial for low-resource languages. We demonstrate the feasibility of this approach by training projectors to align the DINOv2 visual encoder with the paraphrase-multilingual-v2 text encoder, using a dataset consisting solely of English image-caption pairs. We selected this specific text encoder as it showed the highest compatibility in terms of CKA with DINOv2. Subsequently, we evaluated the performance of our model on multi-lingual image retrieval using the XTD dataset [1] and on multi-lingual image classification using the ImageNet dataset. For multi-lingual classification, we translate our VDT prompts [20] to the languages being considered using the nllb-700M model [6] and then use the same prompts for all the models being considered including ours.

For both multi-lingual classification and retrieval tasks, our comparisons are structured into two categories as delineated in Table A.7 and Table A.6. The lower sections of each of these tables list models trained exclusively with English captions, more specifically the CLIP-VIT-L models from OpenAI and LAION trained on 400 million image caption pairs of WIT dataset and LAION400M dataset respectively. The upper sections of these tables feature models trained with translated captions, including those employing contrastive training with multi-lingual image-caption pairs such as CLIP-models based on the LAION5B multilingual dataset, which contains image-caption pairs in over 100 languages. We also compare against, M-CLIP [4] models that are trained using English and translated captions to align a multi-lingual text encoder with CLIP's original text encoder through contrastive learning, thereby enhancing performance on multi-lingual tasks. Additionally we also compare against the NLLB-CLIP [33] models developed through LiT [36] techniques, coupling a frozen CLIP visual encoder with an unfrozen multi-lingual text encoder using translated captions from the smaller LAION-COCO dataset. We compare against only model sizes of up to ViT-Large for fair comparison.

Retrieval results: Our model DINOv2-MpNet trained only on English image,caption pairs outperforms all other CLIP models trained only on English image caption pairs, by a large margin of over 43 % on average retrieval performance over 10 languages. We also outperform the next best performing English CLIP model trained on LAION400m English caption retrieval by over 6 percent. On Latin script languages the CLIP models have decent performance while it falls significantly for non Latin languages like JP, KO, PL, RU, TR, and ZH. This is mainly because these models were trained using an English only tokenizer which results in unknown token for most characters of these languages. However our DINOv2-MpNet projector model maintains competitive performance on all languages both Latin script and non Latin script even when compared against models specifically trained using multi-lingual data (Upper half of the table). Amongst the multi-lingual trained CLIP models we perform better than laion5b trained xlm-roberta-base-VitB32 by 4.5 percent. It is to be noted here that we only use 20 million Image caption pairs for alignment while LAION5B has over 5B image-caption pairs from over 100 languages and multi-lingual webli has over 30B image-caption pairs from over 100 languages. It is to be noted that our DINOv2-Mpnet is also competitive with M-CLIP model XLM-Roberta-Large-Vit-B-16Plus(56.1 vs 57.7) which has been trained using translated English sentences of over 175 million data points to over 100 languages, and 3M translated image, caption pairs from CC3m.

Classification results: We see a similar trend when we compare our DINOv2-MpNet projector model against CLIP baselines(lower section), and multi-lingual baselines (upper section) on multi-lingual imagenet classification in Table. Our model showcases competitive performance to that of OpenAI-clip model while beating LAION400m trained ViT-Large on english Imagenet, while performing significantly better on all other languages considered (over 24 percent better on 8 language average). When compared with models trained with multi-lingual data, our model outperforms both nllb-clip models as well as M-CLIP models, beating the next best performing model M-CLIP/XLM-Roberta-Large-Vit-L-14 by over 3 percent despite not training using any multi-lingual text data. We believe that training using translated image-caption pairs of our dataset would further improve the performance of our method, and we leave this as a future work. The main advantage of training using our methods is that we can get highly porformant CLIP-like models using much lesser amount of image-caption pairs, (more than 20x lesser) resulting in quick adaptation to low resource languages given that a multilingual text encoder exists for that language.

model	EN	DE	ES	FR	IT	JP	KO	PL	RU	TR	ZH	average
nllb-clip-base@v1	47.2	43.3	44.1	45.0	44.7	37.9	39.4	45.5	40.6	41.2	41.1	42.3
M-CLIP/XLM-Roberta-Large-Vit-B-32	48.5	46.9	46.4	46.1	45.8	35.0	36.9	48.0	43.2	45.7	45.4	43.9
M-CLIP/XLM-Roberta-Large-Vit-L-14	56.3	52.2	52.7	51.8	53.6	41.5	42.5	54.1	48.4	52.7	53.5	50.3
xlm-roberta-base-ViT-B-32@laion5b	63.2	54.5	54.6	55.7	55.7	47.1	43.8	55.5	50.3	48.2	50.8	51.6
nllb-clip-large@v1	59.9	56.5	56.7	56.0	55.5	**49.3**	**51.7**	57.4	50.4	56.0	52.3	54.2
M-CLIP/XLM-Roberta-Large-Vit-B-16Plus	63.2	**61.4**	**59.8**	59.3	**61.0**	48.3	49.8	64.0	54.8	**59.6**	**58.8**	**57.7**
ViT-L-14@laion400m_e31	64.5	26.7	31.4	38.3	26.6	1.4	0.4	4.8	1.7	4.1	1.0	13.6
openai/clip-vit-large-patch14	59.4	19.9	26.6	28.5	19.2	4.1	0.3	3.9	1.3	2.6	0.7	10.7
DINOv2-MpNet (Ours)	**70.7**	60.6	59.0	**60.6**	60.7	45.6	49.8	58.3	52.7	55.8	57.9	56.1

Table A.6. **Multilingual image-caption retrieval** performance on XTD dataset. DINOv2-MpNet outperforms many baselines despite English-only training. Upper: multilingual-trained models; Lower: English-only trained models.

model	EN	AR	ES	FR	DE	JP	ZH	RU	average
nllb-clip-base@v1	25.4	20.4	23.9	23.9	23.3	21.7	20.3	23.0	22.4
nllb-clip-large@v1	39.1	30.1	36.5	36.0	36.2	32.0	29.0	33.9	33.4
M-CLIP/XLM-Roberta-Large-Vit-B-32	46.2	33.4	43.7	43.3	43.3	31.6	29.1	38.8	37.6
M-CLIP/XLM-Roberta-Large-Vit-B-16Plus	48.0	35.1	46.6	45.4	46.1	32.9	31.3	40.3	39.7
xlm-roberta-base-ViT-B-32@laion5b	63.0	29.0	53.4	53.8	55.8	37.3	26.8	40.3	42.3
M-CLIP/XLM-Roberta-Large-Vit-L-14	54.7	**40.0**	51.9	51.6	51.9	37.2	**35.2**	47.4	45.0
ViT-L-14@laion400m_e32	72.3	6.4	44.7	49.9	48.2	2.7	2.3	4.5	22.7
openai/clip-vit-large-patch14	**75.6**	6.7	46.2	49.6	46.7	6.6	2.2	3.5	23.1
DINOv2-MpNet (Ours)	73.4	38.0	56.8	58.3	61.6	**43.2**	33.3	49.3	48.6

Table A.7. **Multi-lingual classification.** Classification performance comparison of DINOv2-MpNet and various CLIP models and multilingual baselines on multilingual ImageNet. Our DINOv2-MpNet model trained only on English data outperforms even models trained on multi-lingual data. The upper half of the table lists models trained on multiple languages, while the lower half lists models trained only on English data. The models are evaluated on translations of the labels and the prompts made using nllb-200-distilled-600M translation model. [6]

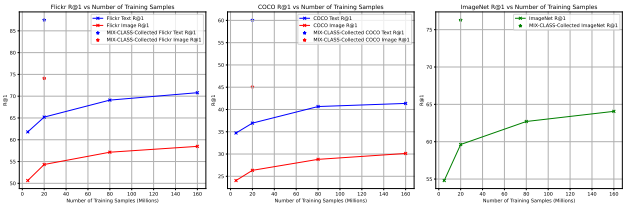

Figure A.7. **Performance scales with higher amounts of randomly sampled LAION data** The performance scales with higher amounts of randomly sample data from LAION400M, but very slowly, highlighting the need for a densely covered and high quality dataset when training projectors only to align modalities.

A.12. Dataset scale

Figure A.7 illustrates that while performance scales with an increasing number of randomly sampled data points from the LAION400M dataset, the rate of improvement diminishes, highlighting the critical need for densely covered and high-quality datasets when training projectors to align modalities. Additionally, the comparative performance of MIX-CLASS-Collected data reveals that datasets curated with more focused criteria can lead to better performance gains than simply increasing the volume of data. This underscores the importance of prioritizing dataset quality over quantity, especially given the observed diminishing returns when using larger data sizes for projector-based alignment.

Model	All SCM	All Neg	All Pick5-SCM	All Pick5-Neg	Base Neg	All Hard-Negs
CLIP Baseline	**40.06**%	60.79%	**11.21**%	**24.06**%	67.56%	41.34%
DINOv2-ARL (Ours)	29.33%	**64.36**%	9.35%	21.39%	**81.94**%	**61.10**%

Table A.8. Performance comparison on DCI dataset benchmarks

A.13. sDCI benchmark results

We evaluate our method on the Densely Captioned Images (DCI) dataset [32], which contains 7,805 images with mask-aligned descriptions averaging over 1,000 words each. To accommodate current models' token limits, the authors also provide sDCI, a summarized version with CLIP-compatible 77-token captions generated by LLMs.

sDCI introduces several benchmarks:

- All SCM (Subcrop-Caption Matching): Matches captions to corresponding image subcrops.
- All Neg: Distinguishes between positive captions and LLM-generated negatives.
- All Pick5-SCM: Similar to All SCM, but uses multiple captions per subcrop.
- All Pick5-Neg: Distinguishes between multiple positive captions and a negative.
- Base Neg: Focuses on caption-negative distinction for full images only.
- All Hard-Negs: Uses the most challenging LLM-generated negatives.

We tested our DINOv2-ARL model on the sDCI dataset benchmarks. Table A.8 presents our results alongside the CLip baseline. Our method demonstrates competitive performance compared to the CLIP baseline across several DCI benchmarks.

In the Subcrop-Caption Matching tasks (All SCM and All Pick5-SCM), our model performs slightly below the CLIP baseline. This suggests that there is room for improvement in our approach when it comes to distinguishing between the different parts that compose an image.

However, our model shows notable improvements in the negative detection tasks. We outperform CLIP on All Neg (64.36% vs. 60.79%), Base Neg (81.94% vs. 67.56%), and All Hard-Negs (61.10% vs. 41.34%). These results demonstrate the potential of our method in aligning vision and language models for a fine-grained understanding of image content, especially in scenarios requiring robust discrimination between relevant and irrelevant captions. Future work could focus on improving the model's performance on subcrop caption matching tasks while maintaining its strong capabilities in negative detection.

A.14. 0-Shot Classification and Retrieval Evaluation Datasets

To evaluate the performance of our DINOv2-ARL projector model and compare it with baseline CLIP models, we utilized a diverse set of datasets for zero-shot classification and retrieval tasks. These datasets span various domains and challenge the models' ability to generalize across different visual concepts.

For zero-shot classification, we employed the following datasets:

- ImageNet [7]: A large-scale dataset with 1000 object categories, widely used as a benchmark for image classification tasks. It contains over 1.2 million training images and 50,000 validation images, with each image labeled with one of 1000 object classes.
- ImageNetV2 [25]: A newer version of ImageNet designed to test the robustness of models trained on the original ImageNet. It features 10,000 new test images collected using the same procedure as the original, but addressing certain biases in the original dataset.
- Caltech101 [15]: A dataset containing pictures of objects belonging to 101 categories, plus a background category. It includes about 40 to 800 images per category, with most categories having about 50 images. The dataset is known for its high intra-class variability.
- Oxford-IIIT Pet [23]: A 37-category pet dataset with roughly 200 images for each class, featuring different breeds of cats and dogs. It includes pixel-level trimap segmentations and breed-level labels for each image.
- Stanford Cars [14]: A dataset of 196 car classes, totaling 16,185 images. Classes are at the level of Make, Model, Year (e.g., 2012 Tesla Model S). It includes 8,144 training images and 8,041 testing images, with bounding box annotations.
- Oxford Flowers102 [21]: A 102 category dataset consisting of 102 flower categories common to the UK. It contains 40 to 258 images per class and provides segmentation data for each image. The dataset is particularly challenging due to the fine-grained nature of the categories.
- Food101 [3]: A large dataset of 101 food categories, with 101,000 images. It features 1000 images per food class, with 250 test images and 750 training images per class. The training images are not manually cleaned, adding a level of noise to the dataset.
- FGVC Aircraft [18]: A fine-grained visual classification dataset with 10,200 images of aircraft, spanning

100 aircraft models. Each model is associated with a specific variant, manufacturer, family, and collection. The dataset includes 6,667 training images and 3,333 test images.

- SUN397 [28]: A scene recognition dataset with 397 categories and 108,754 images, covering a large variety of environmental scenes under various lighting conditions. It provides at least 100 images per class and has been used extensively for scene recognition tasks.

- Caltech-UCSD Birds-200-2011 (CUB) [34]: A dataset for fine-grained image classification with 200 bird species, containing 11,788 images. Each image has detailed annotations including 15 part locations, 312 binary attributes, and 1 bounding box. It's widely used for fine-grained visual categorization research.

- UCF101 [29]: An action recognition dataset with 101 action categories, consisting of realistic action videos collected from YouTube. It contains 13,320 videos from 101 action categories, with videos exhibiting large variations in camera motion, object appearance and pose, illumination conditions, and more.

For zero-shot image-text retrieval, we used:

- Flickr30k [24]: A dataset containing 31,783 images collected from Flickr, each paired with 5 crowd-sourced captions. It focuses on describing the objects and actions in everyday scenes. The dataset is split into 29,783 training images, 1000 validation images, and 1000 test images.

- COCO [17]: A large-scale dataset for object detection, segmentation, and captioning, which we use for its image-caption pairs in the retrieval task. It features over 330,000 images, each with 5 captions. The dataset includes 80 object categories and instance segmentation masks, making it versatile for various computer vision tasks.

These datasets comprehensively evaluate a model's ability to perform zero-shot classification across various domains and its capacity for cross-modal retrieval. By using this diverse set of benchmarks, we can assess the generalization capabilities of our approach compared to existing CLIP models. We use Visually Descriptive Class-Wise prompts from [20] to enable the unimodal-text encoder in our DINOv2-ARL projector model to better identify the zero-shot classes of the downstream datasets.

A.14.1 Concept Coverage Collection datasets

We use a few shot examples from 14 curated computer vision datasets to construct our Concept Image prototypes to curate the images from our uncurated data pool. The 14 curated datasets are described as follows.

- BirdSnap [2]: A fine-grained dataset consisting of 49,829 images of 500 North American bird species. The images are annotated with species labels, and the dataset is primarily used for species classification and fine-grained recognition tasks.

- Caltech101 [15]: A dataset containing pictures of objects belonging to 101 categories, plus a background category. It includes about 40 to 800 images per category, with most categories having about 50 images. The dataset is known for its high intra-class variability.

- EuroSAT [10]: A satellite image dataset with 10 categories related to land use classification (e.g., forests, rivers, residential areas). It contains 27,000 labeled images, with 2700 images per class, widely used in remote sensing and geospatial tasks.

- FGVC Aircraft [18]: A fine-grained classification dataset with 10,000 images of 100 aircraft model variants from 70 manufacturers. It is used for distinguishing between visually similar objects in fine-grained recognition tasks.

- Flowers102 [21]: A dataset containing 102 flower categories, commonly used for fine-grained classification tasks. It has a total of 8,189 images, with 40 to 258 images per category, and is organized into a training, validation, and test set.

- Food101 [3]: A dataset containing 101,000 images of 101 food categories. Each category has 750 training images and 250 test images, commonly used for food classification and recognition tasks.

- GTSRB [30]: The German Traffic Sign Recognition Benchmark dataset, containing over 50,000 images of 43 different traffic sign classes. It is designed for multi-class classification tasks in the context of traffic sign recognition.

- ImageNet [7]: A large-scale dataset with 1,000 object categories, widely used as a benchmark for image classification tasks. It contains over 1.2 million training images and 50,000 validation images, with each image labeled with one of 1,000 object classes.

- Oxford Pets [23]: A dataset of 7,349 images, containing 37 categories of pets (both cats and dogs). Each

image is annotated with species and breed information, commonly used for image classification and segmentation tasks.

- RESISC45 [5]: A dataset of remote sensing images used for scene classification, containing 31,500 images across 45 scene classes. Each class has 700 images with variations in resolution, scale, and orientation.
- Stanford Cars [14]: A dataset with 16,185 images of 196 car models, annotated by make, model, and year. The dataset is designed for fine-grained classification and recognition tasks of vehicles.
- Pascal VOC 2007 [8]: A dataset for object detection, segmentation, and classification, containing 9,963 images of 20 object categories. It is widely used for benchmarking models in computer vision tasks.
- SUN397 [28]: A large-scale scene understanding dataset with 397 categories and 108,754 images. It covers a wide range of environments, from natural to man-made scenes, commonly used for scene classification tasks.
- UCF101 [29]: A video dataset consisting of 13,320 videos across 101 human action categories. It is widely used for action recognition tasks in video analysis and computer vision research.

References

[1] Pranav Aggarwal and Ajinkya Kale. Towards zero-shot cross-lingual image retrieval. *arXiv preprint arXiv:2012.05107*, 2020. 4

[2] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2011–2018, 2014. 7

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 6, 7

[4] Guanhua Chen, Lu Hou, Yun Chen, Wenliang Dai, Lifeng Shang, Xin Jiang, Qun Liu, Jia Pan, and Wenping Wang. mclip: Multilingual clip via cross-lingual transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13028–13043, 2023. 4

[5] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 8

[6] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022. 4, 5

[7] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009. 6, 7

[8] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015. 8

[9] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021. 3

[10] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 7

[11] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989. 1

[12] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024. 1

[13] Zaid Khan and Yun Fu. Contrastive alignment of vision to language through parameter-efficient transfer learning. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[14] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 6, 8

[15] Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, Apr 2022. 6, 7

[16] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023. 4

[17] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 7

[18] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6, 7

[19] Mayug Maniparambil, Raiymbek Akshulakov, Yasser Abdelaziz Dahou Djilali, Mohamed El Amine Seddik, Sanath Narayan, Karttikeya Mangalam, and Noel E. O'Connor. Do vision and language encoders represent the world similarly? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14334–14343, June 2024. 1

[20] Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy, Kevin McGuinness, and Noel E. O'Connor. Enhancing clip with gpt-4: Harnessing visual descriptions as prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 262–271, October 2023. 4, 7

[21] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 6, 7

[22] 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. 3

[23] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 6, 7

[24] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 7

[25] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 6

[26] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. 3

[27] Nils Reimers and Iryna Gurevych. Pretrained models — sentence transformers documentation, 2024. Accessed: 2024-09-24. 3

[28] Vanessa Rouach, Yuliana Pushevsky, Alla Mayboroda, Alina Osherov, and Michal Guindy. Sun-397 the osteosee system measurements, based on parametric electrical impedance tomography, correlate with dual x-ray absorptiometry results for the diagnosis of osteoporosis. *Journal of the Endocrine Society*, 4(Supplement_1):SUN–397, 2020. 7, 8

[29] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 7, 8

[30] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012. 7

[31] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 3

[32] Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26700–26709, 2024. 6

[33] Alexander Visheratin. Nllb-clip–train performant multilingual image retrieval model on a budget. *arXiv preprint arXiv:2309.01859*, 2023. 4

[34] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. Caltech-ucsd birds-200-2011 (cub-200-2011). *Technical Report CNS-TR-2011-001, California Institute of Technology*, 2011. 7

[35] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021. 1

[36] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18123–18133, 2022. 1, 4

[37] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 4