Supplemantary Material for "MoDE: CLIP Data Experts via Clustering"

Contents

A. Full Results

Firstly, we provide the complete results of MoDE when scaling up the number of coarse-grained clusters. As shown in Table A, when more data experts are learned, the average accuracy on CLIP benchmark keeps improving.

	Average	ImageNet	Food-101	CIFAR10	CIFAR100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	MNIST	FER-2013	STL-10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	HatefulMemes	SST2
MetaCLIP	59.8	67.6	82.6	95.2	77.7	67.8	66.8	77.2	26.9	58.9	90.9	92.5	69.7	42.7	48.3	96.3	49.9	66.5	39.2	29.3	17.7	50.0	68.0	47.6	19.4	53.5	53.1
MoDE-2	61.2	68.7	84.1	95.3	78.6	69.5	67.0	80.8	30.9	60.6	91.0	92.9	71.9	40.8	50.4	96.3	51.3	67.9	44.2	31.4	18.3	51.3	69.0	47.4	23.2	52.6	54.4
MoDE-4	61.7	68.8	85.8	95.2	79.0	74.4	67.5	83.3	29.5	60.3	91.9	92.9	72.1	49.7	46.9	96.4	50.3	66.8	51.6	28.5	19.6	50.1	68.4	48.3	21.6	52.6	52.2
MoDE-8	63.4	69.3	88.1	95.6	80.1	76.0	68.2	87.7	46.7	60.9	91.2	93.4	77.1	46.5	47.2	97.1	58.3	67.7	52.7	27.4	18.5	50.1	68.6	48.2	25.2	53.3	52.1
MoDE-16	64.0	70.7	88.4	96.1	80.5	80.8	67.9	87.1	44.6	59.9	92.2	93.2	79.4	50.1	49.8	97.1	60.3	67.7	48.5	26.1	18.9	55.3	68.1	48.3	25.7	54.1	51.9
MoDE-32	64.0	69.6	88.2	95.9	80.8	80.1	68.3	88.9	44.1	59.9	92.6	93.5	83.0	42.9	46.9	97.4	56.2	67.3	48.8	30.7	19.2	58.0	68.2	48.1	30.6	53.5	50.2

Table A. Performance details of Fig. 3 when scaling MoDE-n based on ViT-B/32 on 2.5B image-caption pairs.

We noticed the work LiMoE [5] which follows conventional Deep Mixture of Expert models and trains a stack of Transformer MoE layers on all 3.6B image-caption pairs [7]. However, the number of parameters in a single LiMoE network is much larger than a single dense baseline. As all of the network parameters are trained synchronously, it will cause huge memory usage. Meanwhile, comparing with MoDE-4 trained on different data clusters while the total pre-train set has only about 2.5B image-caption pairs, our system is more flexible and also achieve better results consistently.

Task	Dataset	ViT-B/32		ViT-B/16		ViT-L	
		MoDE	LiMoE	MoDE	LiMoE	MoDE (L14)	LiMoE (L16)
zero-shot classification	ImageNet	68.9	67.5	74.3	73.7	79.4	78.6
zero-shot text retrieval	COCO	57.4	45.7	62.7	51.3	65.6	55.7
zero-shot image retrieval	COCO	39.9	31.0	44.1	36.2	48.2	39.6

Table B. Performance comparison between MoDE and LiMoE [5]

Secondly, we summarize the results for robustness evaluation in Table C and zero-shot retrieval in Table D. The results in each table are separated by the scale of pre-train dataset. Consistently, our approach can outperform the MetaCLIP Baseline in all cases. MoDE also achieves the best score in most cases.

Approach	ViT	Avg.	IN-Sketch	IN-V2	IN-A	IN-O	IN-R	Avg.	IN-Sketch	IN-V2	IN-A	IN-O	IN-R
OpenAI CLIP	B/32	49.4	42.3	56.0	31.5	47.8	69.4	-	-	-	-	-	-
OpenCLIP		50.6	49.4	55.1	21.7	53.5	73.4	52.9	53.7	58.1	26.3	50.0	76.4
MetaCLIP		52.2	53.3	57.6	28.6	46.8	74.8	54.4	56.0	59.6	29.9	48.3	78.1
MoDE-2		53.0	53.9	57.9	29.4	48.0	75.7	55.2	57.1	60.5	31.2	48.4	79.0
MoDE-4		53.4	54.4	58.5	30.8	47.6	76.0	56.5	57.6	61.6	34.2	49.2	80.0
OpenAI CLIP	B/16	56.0	48.3	61.9	50.0	42.3	77.7	-	-	-	-	-	-
OpenCLIP		54.8	52.4	59.7	33.2	50.7	77.9	56.7	56.1	62.3	38.2	46.3	80.6
MetaCLIP		57.7	57.9	62.6	47.0	39.2	81.8	60.1	60.2	65.0	49.5	41.6	84.2
MoDE-2		58.4	58.5	63.2	47.9	39.9	82.3	62.3	62.4	66.5	52.0	45.2	85.5
MoDE-4		59.0	58.8	63.7	49.2	40.4	82.9	63.3	62.8	67.1	55.7	44.5	86.6
OpenAI CLIP	L/14	64.1	59.6	69.8	70.7	32.3	87.9	-	-	-	-	-	-
OpenCLIP		59.6	59.6	65.5	46.5	42.0	84.7	62.2	63.3	67.8	53.9	38.7	87.4
MetaCLIP		63.8	65.0	69.8	66.4	28.9	88.9	67.2	68.9	72.6	72.3	30.2	92.1
MoDE-2		64.0	65.2	70.0	66.9	28.9	89.0	67.6	69.3	72.8	73.0	30.6	92.3
MoDE-4		64.1	65.3	70.1	66.8	29.4	89.0	68.2	69.9	73.3	74.0	31.3	92.7
Pre-Train Dataset:		400M Image-Caption Pairs						OpenCLIP:2.3B, MetaCLIP / MetaCLIP: 2.5B					

Table C. **Zero-Shot Robustness Evaluation**. The results are separated by the scale of pre-train set. Entries in blue are the best ones.

B. Ablation Study Details for Clustering

Firstly, for ablation details on Clustering Strategy in Sec. 5.2, we show details in Table E for Table 6 and Table F for Fig. 5.

Approach	Text Retrieval						Image Retrieval						Text Retrieval						Image Retrieval					
	COCO			Flickr30k			COCO			Flickr30k			COCO			Flickr30k			COCO			Flickr30k		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ViT-B/32																								
OpenAI CLIP	50.2	75.0	83.5	78.9	94.9	98.2	30.4	56.0	66.9	58.8	83.6	90.0	-	-	-	-	-	-	-	-	-	-	-	-
OpenCLIP	52.5	76.8	84.7	78.8	94.1	97.0	35.3	60.9	71.7	61.7	85.5	90.9	56.3	79.8	87.1	84.1	96.2	98.3	39.3	65.4	75.6	66.7	88.4	93.1
MetaCLIP	51.8	76.4	84.7	77.8	93.5	97.1	35.9	61.8	72.1	62.3	85.5	91.5	55.2	78.9	86.5	80.7	95.2	97.3	38.1	64.1	74.3	65.1	87.7	92.7
MoDE-2	53.3	76.7	84.8	78.6	94.3	96.9	36.4	62.1	72.6	63.0	86.1	91.8	56.7	80.2	87.5	82.8	95.1	98.2	39.5	65.3	75.3	66.4	89.0	93.6
MoDE-4	53.7	77.2	85.1	78.5	94.9	96.8	36.7	62.5	73.0	63.6	86.4	91.7	57.4	80.1	87.3	82.9	95.6	97.7	39.9	66.1	75.7	66.7	88.4	93.3
ViT-B/16																								
OpenAI CLIP	52.4	76.7	84.6	86.2	98.0	99.5	33.1	58.4	69.0	69.8	90.4	94.6	-	-	-	-	-	-	-	-	-	-	-	-
OpenCLIP	55.4	79.7	86.9	83.4	96.8	98.5	38.4	63.6	73.9	65.7	88.3	93.0	59.5	81.8	88.6	86.2	98.0	99.5	42.3	67.7	77.1	69.8	90.4	94.6
MetaCLIP	56.4	79.9	87.1	85.7	97.2	98.7	40.0	65.3	75.3	67.6	89.6	94.2	59.4	80.6	87.8	85.5	97.4	98.9	41.4	67.2	76.9	70.7	90.8	94.5
MoDE-2	57.5	80.3	87.6	86.5	97.0	98.8	40.4	65.6	75.6	68.7	89.4	94.2	60.7	82.6	89.0	87.3	97.6	99.2	43.1	68.6	77.8	72.1	91.8	95.3
MoDE-4	57.7	81.1	88.1	86.6	97.5	98.8	41.0	66.2	75.8	68.7	90.0	94.2	62.7	82.9	89.8	89.4	98.0	99.6	44.1	69.5	78.7	72.6	91.8	95.4
ViT-L/14																								
OpenAI CLIP	56.3	79.4	86.6	85.2	97.4	99.2	36.5	61.0	71.1	64.9	87.2	92.0	-	-	-	-	-	-	-	-	-	-	-	-
OpenCLIP	59.7	82.2	89.4	87.6	97.8	99.5	43.0	68.0	77.4	70.2	90.9	94.6	63.3	83.9	90.8	89.5	98.7	99.4	46.5	71.1	79.8	75.5	92.9	95.9
MetaCLIP	60.0	82.9	89.4	86.2	98.1	99.6	43.8	68.7	77.8	73.4	92.3	95.7	64.4	85.0	91.3	90.1	98.6	99.3	47.1	71.4	80.3	76.5	93.6	96.5
MoDE-2	60.8	82.6	89.4	87.7	98.1	99.5	44.2	68.6	77.8	73.4	92.5	95.8	65.2	85.3	91.6	90.9	98.9	99.6	47.9	72.1	80.6	77.2	93.7	96.6
MoDE-4	60.6	82.9	89.1	86.7	97.9	99.7	44.2	68.6	77.8	73.5	92.1	95.9	65.5	85.4	91.8	91.2	99.0	99.7	48.2	72.4	80.7	77.6	93.7	96.7
Pre-Train Dataset:	400M Image-Caption Pairs												OpenCLIP:2.3B, MetaCLIP / MetaCLIP: 2.5B											

Table D. **Zero-shot Retrieval**. The results are separated by the scale of pre-train set. Entries in blue are the best ones.

	Average	ImageNet	Food-101	CIFAR10	CIFAR100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	MNIST	FER-2013	STL-10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	HatefulMemes	SST2
400M Image-Caption Pairs																											
MetaCLIP	58.2	65.5	80.6	91.3	70.2	63.4	63.0	70.7	26.8	52.8	88.7	91.9	68.5	41.5	35.9	95.4	52.6	64.2	35.8	30.7	17.2	55.5	66.1	45.4	30.6	56.4	53.4
OneStep-2	58.0	65.0	80.4	91.3	69.9	62.2	62.5	69.0	27.1	52.7	88.5	91.7	67.3	40.2	32.3	95.0	54.8	63.9	36.2	36.6	16.7	54.5	66.4	45.1	26.4	57.9	54.0
CoarseCluster-2	58.5	66.1	81.1	91.0	70.6	65.3	63.1	71.8	27.1	53.5	89.0	92.2	68.7	45.2	33.5	95.4	52.0	63.7	34.9	34.2	17.3	54.3	66.1	45.5	29.3	56.6	54.6
MoDE-2	58.6	66.1	81.2	90.9	70.5	65.2	63.0	72.0	28.3	53.5	89.4	92.3	68.2	45.2	33.5	95.4	51.9	63.7	34.9	34.2	17.3	54.3	65.9	45.5	29.3	56.6	54.6
CoarseCluster-4	58.7	66.2	82.2	91.2	70.8	67.4	63.2	73.7	28.2	54.0	89.7	92.2	69.8	38.1	33.2	95.6	53.5	64.0	35.2	33.8	17.8	53.2	66.4	45.7	29.9	57.1	53.3
MoDE-4	59.0	66.4	82.3	91.3	70.9	67.0	63.7	73.8	30.1	52.6	89.9	92.1	69.2	37.9	33.2	95.7	53.5	64.1	35.2	33.9	17.1	58.4	66.6	45.9	30.0	58.0	54.5
2.5B Image-Caption Pairs																											
MetaCLIP	59.8	67.7	82.6	95.2	77.7	67.8	66.8	77.2	26.9	58.9	90.9	92.5	69.7	42.7	48.3	96.3	49.9	66.5	39.2	29.3	17.7	50.0	68.0	47.6	19.4	53.5	53.1
OneStep-2	59.8	67.6	82.3	94.8	77.5	67.8	66.3	76.8	26.4	58.1	90.9	92.0	68.7	45.1	47.6	96.1	50.1	65.9	43.6	29.8	17.6	49.7	67.7	47.2	20.2	53.6	51.8
CoarseCluster-2	60.6	68.6	81.9	95.1	77.8	68.7	68.0	77.8	27.3	57.2	90.3	92.6	68.4	44.7	50.3	96.3	50.7	67.2	47.1	33.2	18.4	50.6	69.6	48.2	19.1	52.5	53.2
MoDE-2	61.2	68.7	84.1	95.3	78.6	69.5	67.0	80.8	30.9	60.6	91.0	92.9	71.9	40.8	50.4	96.3	51.3	67.9	44.2	31.4	18.3	51.3	69.0	47.4	23.2	52.6	54.4
CoarseCluster-4	61.3	69.1	83.9	95.1	78.1	73.1	67.5	82.2	27.8	60.4	90.9	92.9	70.1	49.7	46.9	96.2	50.4	66.4	50.4	28.4	18.8	50.0	68.8	48.1	21.6	52.7	52.9
MoDE-4	61.7	68.8	85.8	95.2	79.0	74.4	67.5	83.3	29.5	60.3	91.9	92.9	72.1	49.7	46.9	96.4	50.3	66.8	51.6	28.5	19.6	50.1	68.4	48.3	21.6	52.6	52.2

Table E. Performance details for ablation study on clustering strategy in Table 6 (Sec. 5.2). The experiments are performed on ViT-B/32. The results are separated by the scale of pre-train set.

	Average	ImageNet	Food-101	CIFAR10	CIFAR100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	MNIST	FER-2013	STL-10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	HatefulMemes	SST2
MetaCLIP	58.2	65.5	80.6	91.3	70.2	63.4	63.0	70.7	26.8	52.8	88.7	91.9	68.5	41.5	35.9	95.4	52.6	64.2	35.8	30.7	17.2	55.5	66.1	45.4	30.6	56.4	53.4
$m = 2$	58.0	65.0	80.4	91.3	69.9	62.2	62.5	69.0	27.1	52.7	88.5	91.7	67.3	40.2	32.3	95.0	54.8	63.9	36.2	36.6	16.7	54.5	66.4	45.1	26.4	57.9	54.0
$m = 4$	58.2	65.2	81.2	91.1	70.1	62.6	63.0	72.1	28.2	53.5	89.1	92.1	69.0	37.7	33.5	95.3	53.0	63.6	36.0	35.5	17.6	54.2	65.8	45.0	28.2	56.4	55.0
$m = 32$	58.4	66.0	81.2	91.4	70.7	64.9	63.1	71.7	28.0	53.2	88.8	92.1	69.0	38.1	32.9	95.4	53.1	64.2	36.3	34.0	17.4	54.1	66.0	45.6	29.0	56.3	55.2
$m = 256$	58.4	65.9	80.6	91.2	70.2	64.2	63.5	70.5	27.2	52.6	88.8	92.2	68.5	40.0	35.2	95.3	53.5	64.0	39.5	34.9	17.2	53.7	66.0	45.6	28.3	56.0	54.4
$m = 512$	58.5	65.9	81.2	91.2	70.3	64.6	63.6	72.0	29.0	52.5	89.2	92.0	69.7	40.0	35.3	95.5	52.9	63.2	39.1	34.9	17.1	53.7	66.1	45.2	27.4	56.0	54.6
$m = 1024$	58.6	66.1	81.2	90.9	70.5	65.2	63.0	72.0	28.3	53.5	89.4	92.3	68.2	45.2	33.5	95.4	51.9	63.7	34.9	34.2	17.3	54.3	65.9	45.5	29.3	56.6	54.6
$m = 2048$	58.7	65.8	81.4	91.2	70.4	66.1	63.3	72.1	29.6	51.5	89.1	92.4	70.2	43.0	33.2	95.1	53.1	63.8	32.9	32.7	17.1	57.9	66.7	45.2	31.5	56.1	54.9

Table F. Performance details of MoDE-2 when ablating the number of finegrained clusters in Step 1 (Fig. 5 in Sec 5.2). Experiments are performed on ViT-B/32 on 400M image-caption pairs.

Then, for the embedding types, we provide the details of MoDE-2 in **??**. We note that the SimCSE [2] can be trained via unsupervised or supervised ways. The unsupervised training strategy utilizes dropout masks to generate two views from the same sentence to build positive pair while the latter one uses two sentences which are of similar semantic meaning as positive samples to each other. Regardless the training strategy, we found the average score on CLIP benchmark is the same.

Meanwhile, when both image and language embeddings are used for clustering, we concatenate their embeddings and we experimentally found that adding the language and image embeddings pair-wisely cannot result in meaningful cluster. How-

	Average	ImageNet	Food-101	CIFAR10	CIFAR100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	MNIST	FER-2013	STL-10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	HatefulMemes	SST2
MetaCLIP	58.2	65.5	80.6	91.3	70.2	63.4	63.0	70.7	26.8	52.8	88.7	91.9	68.5	41.5	35.9	95.4	52.6	64.2	35.8	30.7	17.2	55.5	66.1	45.4	30.6	56.4	53.4
DINOv2	58.1	65.2	80.5	91.2	70.3	63.4	63.1	69.8	26.5	51.6	89.0	91.8	68.1	41.0	36.4	95.2	53.4	63.0	37.3	35.0	16.7	53.7	65.6	45.4	26.8	56.0	53.5
Image (CLIP Seed)	58.3	64.7	80.6	91.3	70.7	63.0	63.0	70.8	27.4	53.4	87.8	92.1	68.9	42.2	33.2	95.2	53.6	62.4	38.8	34.4	16.9	61.6	65.9	45.2	20.3	57.8	55.6
Image & Lang. (CLIP Seed)	58.4	65.5	80.3	91.3	70.2	63.4	63.0	70.3	27.7	52.0	88.7	91.8	68.3	40.0	35.3	95.1	54.4	64.4	38.9	36.0	16.7	54.0	66.2	45.7	27.4	56.6	54.6
Lang. (CLIP Seed)	58.3	65.2	80.7	91.3	69.8	64.8	62.6	71.9	26.9	52.3	88.8	91.7	68.6	39.0	34.1	95.2	54.1	63.1	38.1	33.8	16.8	54.8	66.1	45.2	27.6	57.5	55.8
SimCSE-UnSup	58.6	65.7	80.3	91.4	69.6	64.4	63.0	71.8	26.6	52.0	88.9	92.1	69.2	41.0	37.7	95.4	54.4	64.2	39.0	35.1	17.3	53.5	66.3	45.6	26.8	56.8	55.5
SimCSE-Sup	58.6	66.1	81.2	90.9	70.5	65.2	63.0	72.0	28.3	53.5	89.4	92.3	68.2	45.2	33.5	95.4	51.9	63.7	34.9	34.2	17.3	54.3	65.9	45.5	29.3	56.6	54.6

Table G. Performance details on CLIP evaluation benchmark for ablating the embedding types for clustering (Table 7 in Sec. 5.3). The experiments evalutes MoDE-2 based on ViT-B/32 on 400M image-caption pairs.

ever, at inference time, the ensembling weights should be calculated for all image-class pairs in the zero-shot classification task, which is computational heavy but provides very limited gain over the baseline.

C. Application in Retrieval-Enhanced Setup

The retrieval-enhanced setup [3] is to select & retrieve a subset of training data from a large corpus and only improve the performance on tasks of interest. Through data clustering, we can also select the clusters given the task metadata as prior. We use the SimCSE [2] to extract their embeddings and retrieve the nearest fine-grained clusters for each of them. Then, only a single data expert trained on the selected clusters is used for evaluation. We take ImageNet as an example where the 1000 class names are used to retrieve clusters. As shown in Table H, efficiency of network training can be improved significantly and the performance along the model scale can even be escalated.

Approach	B/32	B/16	L/14	G/14
OpenAI CLIP	63.3	68.4	75.6	-
OpenCLIP	66.6	70.2	75.3	80.1
MetaCLIP	67.6	72.1	79.2	-
Ours	71.4	75.3	80.3	-

Table H. Performance comparison on ImageNet in retrieval-enhanced setup.

	Average	ImageNet	Food-101	CIFAR10	CIFAR100	CUB	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	MNIST	FER-2013	STL-10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	HatefulMemes	SST2
ViT-B/32																											
OpenAI CLIP	56.6	63.4	83.7	89.8	65.1	53.7	62.0	59.7	19.6	44.0	87.2	87.4	66.9	48.2	46.6	97.1	44.9	61.0	32.6	28.7	17.2	62.5	63.9	48.0	23.6	56.4	58.6
OpenCLIP	61.5	66.6	82.0	93.6	75.8	66.0	68.3	86.0	23.9	56.1	90.5	91.9	70.5	70.0	50.4	96.6	49.3	65.7	49.3	32.7	16.7	51.7	64.9	45.6	24.2	52.4	57.2
MetaCLIP	59.8	67.6	82.6	95.2	77.7	67.8	66.8	77.2	26.9	58.9	90.9	92.5	69.7	42.7	48.3	96.3	49.9	66.5	39.2	29.3	17.7	50.0	68.0	47.6	19.4	53.5	53.1
Ours	61.9	70.1	85.4	95.7	80.1	74.4	67.0	81.2	36.4	58.5	91.4	93.5	72.7	44.7	42.2	96.8	53.0	69.1	41.8	35.8	18.6	58.7	69.8	48.9	21.7	49.7	51.3
ViT-B/16																											
OpenAI CLIP	59.6	68.3	88.8	90.8	68.2	55.6	64.0	64.6	24.0	45.1	88.9	89.1	69.4	51.8	53.0	98.2	54.8	65.5	43.3	21.7	22.8	56.3	68.5	52.3	25.5	58.7	60.5
OpenCLIP	62.4	70.2	86.2	94.9	76.9	70.5	70.6	88.2	26.6	56.3	90.4	93.1	71.0	65.8	53.3	97.9	55.2	68.3	48.3	11.9	20.3	51.2	68.1	48.9	24.8	53.0	59.5
MetaCLIP	63.5	72.1	88.3	95.7	79.0	71.4	68.5	82.9	30.3	62.1	91.7	93.3	73.9	66.1	47.0	98.4	51.1	71.1	46.6	16.6	22.7	50.5	73.0	52.5	30.8	57.4	59.0
Ours	64.8	74.0	89.8	96.3	81.2	76.2	69.4	85.3	39.1	58.4	92.8	93.8	75.9	57.4	48.3	98.6	54.8	72.3	46.5	28.0	23.3	50.0	74.3	53.4	29.2	57.8	58.4
ViT-L/14																											
OpenAI CLIP	65.7	75.5	93.0	95.6	78.3	63.3	66.8	77.8	31.3	55.3	93.6	93.3	79.3	76.4	56.9	99.4	61.9	70.9	50.6	19.2	31.9	50.1	75.7	60.2	22.3	59.7	68.9
OpenCLIP	65.7	74.0	88.6	95.8	78.3	73.5	73.5	91.4	34.6	61.2	92.7	93.3	74.4	64.4	53.9	98.5	58.6	71.9	51.6	26.1	24.4	58.0	73.3	52.0	27.4	55.1	60.4
MetaCLIP	69.8	79.2	93.4	97.6	84.2	80.1	73.8	88.7	44.6	68.1	94.7	95.4	81.8	64.4	55.1	99.3	59.2	74.6	56.3	29.7	34.0	67.3	81.6	62.0	25.9	58.0	66.7
Ours	70.0	79.4	93.7	97.7	85.0	81.6	73.8	89.2	47.5	68.3	95.7	95.4	83.8	69.5	52.9	99.4	62.4	74.1	59.1	29.3	34.3	58.4	81.8	62.2	23.9	57.1	65.1

Table I. Performance on CLIP evaluation benchmark via in Retrieval-Enhanced setup. The class names of all 26 tasks are jointly used to determine the data clusters. OpenCLIP is trained on LAION-2B with 2.3B image-caption pairs. OpenAI CLIP is trained on WIT400M and its results are included here for complete result summary purpose only.

Besides using the class names of a single dataset to retrieve the most important finegrained data clusters, we can also use the class names of all tasks in CLIP benchmark. The detailed results are summarized in Table I.

D. Downstream Evaluation with Vision Encoders

Besides zero-shot generalization, the set of vision encoders can also be directly ensembled in downstream application. We use ImageNet for evaluation and assume the language metadata such as class names is not available. As such, all vision encoders are equally weighted by default.

Firstly, we evaluate the robustness by ensembling the encoder outputs. Specifically, for each image, we concatenate the outputs from all (n) vision encoders as the image feature and feed it into a linear layer for classification. To maintain reasonable training cost, only linear probing is considered where we exclusively train the linear classifier from scratch and fix all vision encoders. As shown in Table J, our MoDE achieves consistent and clear performance gain over MetaCLIP Baseline.

Model	Linear Probe*			Linear Probe		
	ViT-B/32	ViT-B/16	ViT-L/14	ViT-B/32	ViT-B/16	ViT-L/14
MetaCLIP	69.3	73.3	80.3	67.5	73.8	82.3
MoDE-2	68.9	73.8	80.6	71.3	76.9	83.9
MoDE-4	69.1	74.5	80.6	74.1	79.6	84.7

*: Initialize classifier with language embeddings as in OpenCLIP [6].

Table J. Performance comparison on ImageNet via linear probing on concatenated features.

For comparison, we take MoDE-4 with ViT-B/16 vision encoders as an example and summarize the accuracy, for each vision encoder, via linear probing and finetuning (*i.e.*, all parameters are trained). We can find that linear probing on the concatenated features achieves higher score than finetuning a single model but with much less training cost, which further indicates the great potential of efficiency by our framework.

Data Experts	Zero-Shot	Linear Probe*	Linear Probe	Finetune
MetaCLIP	72.1	73.3	73.8	76.7
0	63.3	66.4	67.3	75.7
1	68.5	71.3	72.0	76.9
2	65.2	68.2	68.8	76.3
3	72.9	74.9	74.2	77.2

*: Initialize classifier with language embeddings as in OpenCLIP [6].

Table K. Accuracy on ImageNet for each ViT-B16 vision encoder of data experts in MoDE-4.

Then, as shown in Table K, a strong correlation between the zero-shot classification and linear probing & finetuning application. The expert model achieving higher zero-shot accuracy also hits the best score in both linear probing and finetuning. In this way, by training data expert on each coarse-grained cluster, we increase the quality negative within each mini-batch to learn stronger vision encoders effectively.

Meanwhile, as all vision encoders are separately trained, the learned embedding spaces are not necessarily aligned with each other. As a result, we experimentally found that adding the model outputs element-wisely in model ensembling does not introduce gain, *e.g.*, for MoDE-4 with ViT-B/16 encoders, the accuracy is only 74.5 compared with 79.6 in Table J.

Approach	MetaCLIP	MoDE-2	MoDE-4	MoDE-8	MoDE-16	MoDE-32
Acc.	73.7	74.0	74.2	73.9	74.1	74.1

Table L. Accuracy on ImageNet via parameter averaging.

Finally, in addition to directly aggregate the feature outputs by all data experts, the parameters learned in MoDE can also be ensembled via averaging and then used as initialization of a single network for finetuning. As shown in Table L, we use ViT-B/32 vision encoder, and achieve consistent gain over MetaCLIP Baseline.

E. Implementation Detail

Clustering. We first sample 100M captions from the 400M image-caption pairs to learn the cluster centers in an unsupervised manner. Then, we use nearest neighbor to determine the cluster assignment for all other samples in the 400M as well as 2.5B dataset. We also observed that the cluster centers learned by using less than 2M samples can also result in similar clustering assignments using spherical K-means clustering [1] via FAISS [4]. In practice, we observed that the balanced K-means clustering algorithm does not necessarily enforce strict balance regarding the distribution of the clusters. For example, for the two coarse-grained clusters on 400M dataset used to train MoDE-2, the number of samples for each cluster are around 170M and 230M respectively. Consequently, as mentioned for Random-2 in Sec. 5.1, mimic the size of subsets by MoDE-2 in the random splitting for fair comparison.

Similarity matrix. For task-level adaptation, as mentioned in Sec. 3.4, we use the nearest neighbor fine-grained cluster ($\arg\max_{s\in S} \mathbf{A}_{l,s}$) for each class $l \in L$. In other words, we apply a maximum filter for each row, *i.e.*, $\mathbf{A}_l$, where the non-maximum values are reset as 0, *i.e.*, $\mathbf{A}_{l,s'} = 0$ if $s' \neq \hat{s}$ where $\hat{s} = \arg\max_{s\in S} \mathbf{A}_{l,s}$. Then, we set $\lambda = 5$ according to our experimental cross validation.

Routing Weights. As described in Eq. (6), the routing weight $p(c|\mathbf{T})$ of a data expert $f(\cdot|c)$ is essentially obtained via softmax normalization. At inference time, we note the routing weights should be reasonably distant from each other. Consequently, given the classification task with the class names L, we use the number of classes $|L|$ to roughly adjust the weights. Firstly, when $|L|$ is small, *e.g.*, $|L| < 10$, though only one data expert can be activated, the selection could be sensitive to noisy routing. Then, we soften the values in $\mathbf{A}$ by multiplying $\exp(0.5 - \sqrt{|L|})$ to ensemble two data experts in most cases. Then, when $|L|$ is large, *e.g.*, $|L| > 200$, the normalized weights tend to be over-smooth, we thus use a much smaller temperature by dividing the λ by $\log(|L|)$. Then, we can only select a few data experts and have low-entropy routing weights.

References

[1] Inderjit S Dhillon and Dharmendra S Modha. Concept decompositions for large sparse text data using clustering. *Machine learning*, 42:143–175, 2001. 6

[2] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 6894–6910. Association for Computational Linguistics (ACL), 2021. 3, 4

[3] Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. Retrieval-enhanced contrastive vision-text models. *arXiv preprint arXiv:2306.07196*, 2023. 4

[4] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 6

[5] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022. 2

[6] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 5

[7] 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. 2