Appendix

This appendix is organized as follows.

- In Section A, we present more results on ImageNet (Sec. A.1) and ELEVATER (Sec. A.2), with additional studies on a broader selection of checkpoints and more visualizations for a better understanding of our approach.
- In Section B, we further present more analysis and discussion for a more comprehensive analysis and understanding of our approach.
- In Section C, we provide implementation details (Sec. C.1,C.2) and cost analysis (Sec. C.3) of our retrieval system and model customization pipelines.

A. More Results on Image-level Tasks

A.1. ImageNet

Comparison of a broader selection of checkpoints. To further study the improvement of REACT over the pretrained vision-language models, we present more results with different pretraining data (WIT-400M, LAION-400M, LAION-2B) and different vision Transformer model sizes (B/32, B/16, L/14). For a fair study, we use the same set of 10M retrieved image-text pairs from LAION-400M dataset for all configurations. Results are presented in Table 1 (first column).

From the table, we can see that REACT with both tuning strategies consistently improves over the base checkpoints across different pretrained data and different vision backbones, and locked-text-gated-image tuning consistently performs better than the locked-text tuning only. Though both benefiting from REACT customization, CLIP checkpoints that are trained on WIT-400M data benefit slightly more than OpenCLIP checkpoints. This suggests that during the model customization stage, leveraging unseen data can potentially give the model a larger gain compared with the seen data during pretraining.

We further study the case when all retrieval data is already observed by models during their pretraining stage. Specifically, we study OpenCLIP checkpoints pretrained on LAION-400M [22], and LAION-2B [21] (a super-set of LAION-400M). From the results, we see that by revisiting the already observed LAION-400M data, REACT (locked-text) shows +2.8/+3.0 improvements on B32/B16 checkpoints, respectively, which purely comes from the model customization stage, with neither additional model parameters, nor additional training data. Interestingly, even on OpenCLIP checkpoints that is pretrained with a much larger LAION-2B, REACT can still improves over OpenCLIP by +0.9/+2.9 with B32 backbone with locked-text and locked-text-gated-image tuning strategy, respectively. These findings suggest that leveraging the original pretraining dataset only at the pre-training stage is sub-optimal, it is of much larger potential to explore the web-scale data using the proposed model customization stage.

Robustness. We also conduct zero-shot evaluation on other ImageNet variants: ImageNet-V2 [15], ImageNet-R [9], ImageNet-A [10], ImageNet-Sketch [25] in Table 1. With the REACT customization, model robustness towards different ImageNet variants consistently improves on ImageNet-V2, ImageNet-R, and ImageNet-Sketch. We notice that for some checkpoints, the accuracy drops after model customization on ImageNet-R dataset: an adversarial dataset with a collection of selected images from the web that can "fool" common classifiers. We find that classifiers trained on the LAION dataset are more prone to such adversarial attacks, while REACT customization helps it recover from such attacks to some extent: accuracy improves for OpenCLIP checkpoints that are trained on LAION, especially when locked-text gated-image strategy is used.

Linear Probe. We further study the full-shot performance on ImageNet-1K of REACT using the linear probing protocol. ImageNet-1K contains around 1.28M training images, and it represents one of the most standard data-rich settings. We use the DINO [5] code base for the linear probe experiments. As shown in Table 2, REACT improves over CLIP by +0.6/+1.9 with the locked-text and locked-text-gated-image tuning, respectively. This suggests that the REACT customization adequately adapts the visual encoder to the ImageNet domain, resulting in better feature representations.

Low-Shot Adaptation. We extend the scope of ImageNet-1K experiments to low-shot settings: 1% and 10% labelled data settings, and provide the first strong baselines using CLIP checkpoints. We present results in Table 3. First, we find that when the linear head is randomly initialized, it often results in sub-optimal low-shot performance, as the knowledge from the CLIP's language encoder is completely discarded. We advocate using language-augmented initialization of the linear head [16], which improves the 1% label adaptation performance of CLIP ViT-B/16 from 70.9% to 74.3% (+3.4%). With REACT customization stage, it further improves by 3.1% to 77.4%, outperforming prior arts with similar model sizes. Furthermore, when we further scale up the model size to ViT-L/14, CLIP achieves 80.5% accuracy, which is on par with the previous SoTA. REACT further improves the accuracy by 1.1%, setting a new state-of-the-art of 81.6% accuracy on 1% label settings. Similar trend is observed in 10% label setting: CLIP is on-par with the prior art, and our REACT customization pushes the new SoTA towards 85.1% accuracy.

Sample Overlap. There is a chance that the LAION-400M dataset contains *some* of downstream ImageNet images, and our retrieval system *may* retrieve these image-text pairs. One

f_θ	Pretrain Data	Method	ImageNet	ImageNet-V2	ImageNet-R	ImageNet-A	IN-Sketch
B/32	WIT-400M	CLIP	63.2	55.9	69.3	31.4	42.3
		REACT (Locked-Text)	**66.9** (+3.7)	**58.6** (+2.7)	**77.9** (+8.6)	**23.0** (-8.4)	**54.2** (+11.8)
		REACT (Locked-Text Gated-Image)	**68.6** (+5.4)	**61.0** (+5.1)	**78.2** (+8.9)	**30.8** (-0.6)	**53.9** (+11.6)
	LAION-400M	OpenCLIP	62.9	55.2	73.4	21.8	49.4
		REACT (Locked-Text)	**65.7** (+2.8)	**57.3** (+2.1)	**77.5** (+4.1)	**20.2** (-1.6)	**54.8** (+5.5)
		REACT (Locked-Text Gated-Image)	**66.4** (+3.5)	**58.7** (+3.5)	**77.8** (+4.4)	**22.7** (+0.9)	**54.8** (+5.5)
	LAION-2B	OpenCLIP	66.6	58.2	76.5	26.2	53.5
		REACT (Locked-Text)	**67.5** (+0.9)	**59.5** (+1.3)	**79.1** (+2.6)	**23.8** (-2.5)	**57.1** (+3.6)
		REACT (Locked-Text Gated-Image)	**69.5** (+2.9)	**61.6** (+3.5)	**80.2** (+3.7)	**27.9** (+1.6)	**58.4** (+4.8)
B/16	WIT-400M	CLIP	68.6	61.8	77.6	49.7	48.3
		REACT (Locked-Text)	**71.6** (+3.0)	**64.4** (+2.6)	**83.4** (+5.8)	**38.8** (-10.9)	**58.3** (+10.0)
		REACT (Locked-Text Gated-Image)	**73.4** (+4.8)	**66.8** (+5.0)	**84.0** (+6.4)	**48.5** (-1.2)	**58.3** (+10.1)
	LAION-400M	OpenCLIP	67.1	59.4	77.9	33.0	52.4
		REACT (Locked-Text)	**69.9** (+2.8)	**62.4** (+3.0)	**81.8** (+3.9)	**33.7** (+0.7)	**58.1** (+5.7)
		REACT (Locked-Text Gated-Image)	**70.5** (+3.4)	**63.0** (+3.6)	**82.3** (+4.3)	**37.8** (+4.9)	**57.4** (+5.1)
L/14	WIT-400M	CLIP	75.3	69.6	87.8	70.5	59.6
		REACT (Locked-Text Gated-Image)	**78.1** (+2.8)	**71.5** (+1.9)	**89.9** (+2.1)	**68.6** (-2.0)	**64.8** (+5.2)
	LAION-2B	OpenCLIP	75.3	67.9	84.1	42.0	63.3
		REACT (Locked-Text Gated-Image)	**76.4** (+1.1)	**68.9** (+1.0)	**89.0** (+4.9)	**55.2** (+13.2)	**65.4** (+2.0)

Table 1. Comparison with public checkpoints from CLIP [20] and OpenCLIP [12]. All REACT checkpoints use 10M retrieved samples from LAION-400M [22] dataset during model customization stage. It consistently outperforms base CLIP and OpenCLIP checkpoints.

Method	Accuracy
CLIP [20]	80.2
CLIP†	79.5
REACT (Locked-Text)	**80.1** (+0.6)
REACT (Locked-Text Gated-Image)	**81.4** (+1.9)

Table 2. Linear Probe on ImageNet-1K. CLIP†: reproduced by our implementation.

may question that if the performance gain of REACT model customization actually comes from these samples.

We carefully study the de-duplication experiments. We compute the pairwise distance of the visual features between the images from the retrieved set and the ImageNet train/val set, and set the cutoff threshold to 0.95 (Fig. 1, Bottom). Note that 0.95 is a high threshold, as ∼85K (∼1% of 10M total retrieved images) images are removed, among which, only a few of them overlap with ImageNet train/val. This suggests that the LAION data contain ImageNet images, making the publicly available OpenCLIP checkpoints less rigorous when reporting the zero-shot task transfer performance. As for CLIP, as its pre-training data is not publicly available, it remains unknown if any ImageNet images are observed in its pre-training. We set it to 0.95 mainly to ensure that the overlapping images are removed from the retrieved sets so as to carefully study its effect.

As shown in Table 4, even after aggressively removing 85K images, the final model's performance is similar (-0.2%) to the checkpoint trained on the unfiltered retrieved set. We

Method	Backbone	# Params	1%	10%
Self-supervised or semi-supervised methods				
iBOT [30]	ViT-B/16	86M	69.7	–
DINO [3]	ViT-B/8	86M	70.0	–
MSN [2]	ViT-B/4	86M	75.7	–
MSN [2]	ViT-L/7	304M	75.1	–
PAWS [3]	RN50x4	375M	69.9	79.0
SimCLRv2 [6]	RN152x3	795M	74.9	80.1
SimCLRv2 (self-distilled) [6]	RN152x3	795M	76.6	80.9
Semi-ViT [4]	ViT-H/14	632M	80.0	84.3
SEER [7]	RegNetY	1.3B	60.5	77.9
Language-image learning methods				
CLIP [20] (Zero-Shot)	ViT-B/16	86M	68.6	
CLIP (Random Init.)	ViT-B/16	86M	70.9	80.1
CLIP (Language Init. [16])	ViT-B/16	86M	74.3	80.4
REACT (Locked-Text)	ViT-B/16	86M	76.1	80.8
REACT (Locked-Text Gated-Image)	ViT-B/16	129M	77.4	81.8
CLIP [20] (Zero-Shot)	ViT-L/14	304M	75.3	
CLIP (Language Init. [16])	ViT-L/14	304M	80.5	84.7
REACT (Locked-Text)	ViT-L/14	304M	**81.6**	**85.1**
REACT (Locked-Text Gated-Image)	ViT-L/14	380M	**81.6**	**85.0**

Table 3. Low-shot (1% and 10% labels) on ImageNet-1K. For CLIP and REACT experiments, unless noted, we use language initialization [16] by default for the optimal low-shot performance.

further visualize the validation accuracy curve as training proceeds in Fig. 1 (Top). The model behaviors during training are very similar between filtered (dashed curve) and unfiltered (solid curve) retrieved sets, for both tuning strategies. This ensures that the gains are not due to the overlapping samples, and REACT effectively learns and adapts to the ImageNet domain during the model customization stage.

Quality Control. The quality of retrieved data matters for vision-language pre-training. As an initial attempt of the

Method	Accuracy	
	Filtered	Unfiltered
OpenCLIP [12]	–	62.9
REACT (Locked-Text)	66.7	66.9
REACT (Locked-Text Gated-Image)	68.4	68.6

Table 4. The study of sample overlap. Comparison between checkpoints trained on filtered and unfiltered retrieved set. REACT is customized based on CLIP ViT-B/32 checkpoint, whose ImageNet zeroshot accuracy is 63.2.

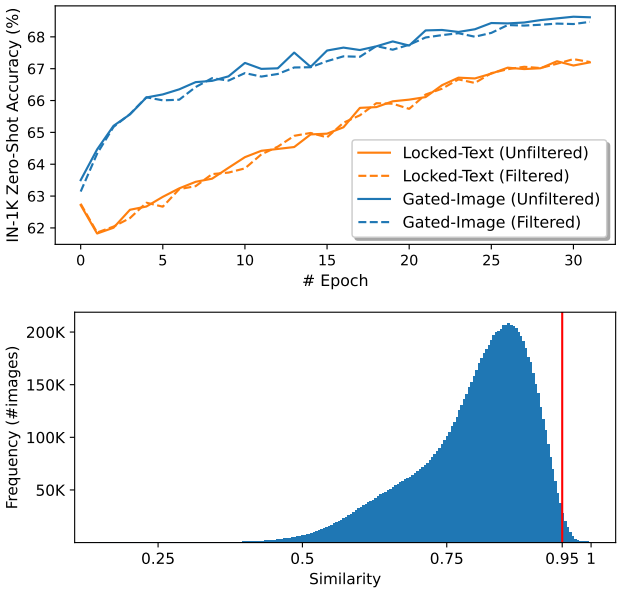

Figure 1. The study of sample overlap. Top: comparison of validation accuracy curve between checkpoints using filtered (dashed line) and unfiltered (solid line) retrieval set during the model customization stage. Both the training behavior and the final model performance are similar. Bottom: histogram of the sample similarity between the retrieved samples and ImageNet training set (nbins=200). The cutoff threshold is set to 0.95 (red line).

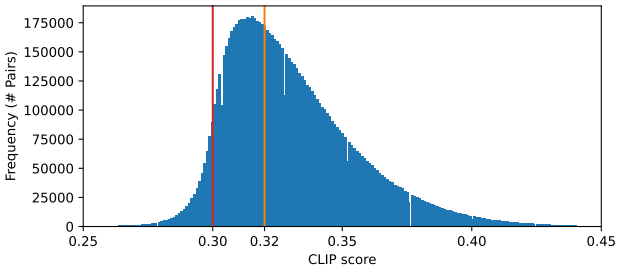

Figure 2. Quality control. Sample frequency under different CLIP scores.

quality control, we consider to use the CLIP score to select the high relevant retrieved image-text pairs. The distribution

Figure 3. Mismatch between the query (Volvo) and the retrieved image (Audi). The retrieved caption helps to correct the mistake.

of the CLIP score is visualized at Fig. 2. We choose CLIP score 0.3 and 0.32 as two thresholds (γ): a threshold of 0.3 filters the low-quality samples while keeping the total number of retrieved samples roughly the same (93.5% retrieved samples are kept), a threshold of 0.32 performs a more aggressive filtering and keeps around 6M samples (sufficient for REACT customization according to Sec. B.1).

As shown in Table 6, REACT is robust towards noise in the pretraining data. When filtering using a CLIP score threshold of 0.3, the model customization performance roughly remains the same. When filtering with a threshold of 0.32, the customization performance drops by around 1%, which suggests that the filtered samples contain useful information for model customization. In conclusion, our REACT model customization is robust against the noises in the retrieval dataset. Therefore, we leave a more sophisticated quality control approach to future work.

A.2. ELEVATER

Breakdown results. We present the full-spectrum breakdown results for zero-shot, few-shot, and full-shot experiments on the ELEVATER benchmark in Table 5. We mark the experiment runs that REACT yields gains compared with baseline CLIP in green and with bold font.

First, across zero-shot, few-shot, and full-shot settings, REACT consistently improves over the baseline CLIP. However, on the ELEVATER benchmark, locked-text and locked-text-gated-image strategy works better in different cases: for zero-shot, locked-text-gated-image works much better than locked-text with 1.4% improvement; while for other cases, they perform similarly well, and locked-text is slightly better in 3/4 cases. This may be partly due to that we are training a unified checkpoint across different domains, and random noises during the final adaptation stage can cause small variations on different datasets.

Second, we find that across different data regimes and different tuning strategies, the gains and losses are mostly consistent on a fixed set of datasets. This provides another clue for that the gains and losses are highly correlated with the retrieval quality, and the gain/loss conclusion can generally transfer to different tuning configurations.

Visualizations. We present more retrieved samples from the

Strategy	Score	Caltech101	CIFAR10	CIFAR100	Country211	DTD	EuroSat	FER2013	FGVCAircraft	Food101	GTSRB	HatefulMemes	KittiDistance	MNIST	Flowers102	OxfordPets	PatchCamelyon	SST2	RESISC45	StanfordCars	VOC2007
Zero-Shot Adaptation																					
	56.8	87.5	89.9	65.1	17.2	44.4	45.6	42.1	19.6	84.0	32.8	56.0	29.0	48.1	66.5	87.1	60.7	58.4	60.0	59.7	82.6
Locked-Text	**59.3**	**90.7**	**94.0**	**73.7**	17.1	**47.6**	**53.4**	**46.8**	**28.1**	83.0	27.2	54.8	25.2	**50.6**	**74.0**	**88.3**	48.1	54.5	59.7	**87.3**	81.0
Locked-Text Gated-Image	**60.7**	**90.6**	**91.7**	**70.7**	**19.1**	**49.2**	**53.8**	**49.0**	**30.0**	**85.1**	30.5	54.1	22.2	**53.9**	**76.1**	**90.1**	53.9	**58.4**	**62.9**	**88.9**	**82.8**
Few-Shot Linear Probe																					
	65.3	89.8	90.0	67.4	17.5	59.6	73.2	47.4	28.4	84.2	52.5	56.0	44.9	71.1	90.5	88.0	63.2	57.5	76.6	65.0	84.0
Locked-Text	**69.4**	**92.2**	**94.1**	**76.3**	**17.6**	**66.6**	**82.5**	**49.9**	**42.1**	**84.2**	**55.0**	54.9	42.2	**78.7**	**96.7**	**89.0**	58.5	54.2	**80.5**	**89.3**	83.3
Locked-Text Gated-Image	**68.9**	**92.5**	**92.3**	**71.6**	**18.9**	**66.2**	**74.5**	**51.8**	**44.1**	**85.6**	51.9	54.1	42.9	68.7	**97.0**	**90.6**	60.5	**60.7**	**78.8**	**90.0**	**84.4**
Few-Shot Full-Model Finetune																					
	63.3	88.8	91.3	73.0	16.6	51.8	79.3	52.3	23.1	84.0	60.4	55.8	44.3	60.5	67.3	86.9	61.8	59.3	70.8	56.3	82.4
Locked-Text	**68.8**	**93.4**	**94.2**	**79.4**	**16.9**	**61.2**	76.0	52.2	**41.1**	83.2	**77.3**	54.9	44.0	**67.5**	**90.0**	**88.9**	57.8	53.3	**78.0**	**89.4**	77.9
Locked-Text Gated-Image	**68.4**	**91.3**	**92.2**	**77.2**	**18.1**	**60.1**	**81.2**	**52.6**	**31.8**	**85.4**	**69.4**	54.1	40.0	**68.3**	**88.8**	**89.7**	61.0	**59.9**	**77.2**	**87.0**	**83.3**
Full-Shot Linear Probe																					
	78.4	86.0	95.1	79.8	25.9	75.3	93.8	67.8	44.7	88.6	86.9	63.1	65.8	98.8	94.5	91.0	83.2	71.6	88.1	82.1	86.0
Locked-Text	**80.1**	**94.5**	**96.6**	**84.1**	24.2	**77.4**	**95.7**	66.0	**57.1**	88.0	86.4	59.7	**68.1**	98.6	**98.1**	**92.5**	**83.4**	63.5	**89.4**	**93.1**	85.0
Locked-Text Gated-Image	**80.4**	**94.5**	**95.6**	**81.6**	**26.3**	**77.8**	**95.3**	67.2	**56.5**	**89.2**	83.3	62.6	65.3	98.4	**98.2**	**93.6**	**83.3**	70.4	**89.7**	**93.5**	**86.0**
Full-Shot Full-Model Finetune																					
	80.3	94.0	97.8	87.0	19.1	70.1	98.1	68.9	50.7	87.7	98.6	61.9	81.0	99.5	88.5	91.6	91.0	70.6	89.4	75.8	85.7
Locked-Text	**82.2**	**95.3**	**98.3**	**89.0**	**20.6**	**75.1**	98.0	**71.6**	**60.2**	**88.0**	**98.7**	58.1	79.2	**99.7**	**95.3**	**93.4**	90.4	65.1	**90.2**	**92.6**	85.1
Locked-Text Gated-Image	**81.8**	**94.6**	**98.3**	**87.8**	**19.5**	**72.5**	97.9	**70.5**	**59.7**	**88.4**	**98.7**	58.5	73.4	**99.6**	**94.5**	**93.0**	89.1	70.5	**89.7**	**93.0**	**86.6**

Table 5. Full-spectrum breakdown results on ELEVATER using CLIP (ViT-B/32) and 10M retrieved image-text pairs from LAION-400M.

Method	Accuracy $\gamma = 0.32$	$\gamma = 0.30$	Unfiltered
CLIP [20]	–	–	63.2
REACT (Locked-Text)	65.2	67.1	66.9
REACT (Locked-Text Gated-Image)	67.3	68.1	68.6

Table 6. Quality control. Comparison between checkpoints trained on CLIP-score-filtered and unfiltered retrieved set. REACT is customized based on CLIP ViT-B/32 checkpoint.

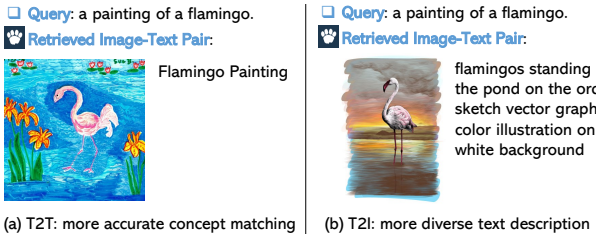

Figure 4. Comparison between T2T and T2I retrieved samples.

ELEVATER datasets to illustrate properties of REACT.

First, we show one example of the benefit of using retrieved image-text pairs for training instead of using pseudo labels. As shown in Fig. 3, the query is `Volvo` sedan, while one of the retrieved sample is an `Audi`. The retrieved text contains the correct brand `Audi`, and the alignment between the retrieved image-text pairs can help correct the retrieval mistake and aid the model training. If the retrieved images are annotated as the same label with query, and used the pseudo-labelled pairs for training, the aforementioned finding suggests that the approach would perform worse than leverage the "true" image-text pair knowledge crawled from the web.

Second, we visualize examples retrieved by text-to-text and text-to-image retrieval, using the same query: "`a painting of a flamingo`". As shown in Fig. 4, text-to-text retrieval can retrieve samples that is more accurately matching the query ("`a painting`"). On the other hand, text-to-image retrieval gives a more diverse text description, while it may not have a perfect match between the query and the retrieved text sample ("`sketch vector`").

B. More Analysis and Discussions

B.1. More ablations on ImageNet

Where to add gated blocks? We conduct an experiment by adding a single GSA block before each Transformer block and continual pretraining the model on the retrieved image-text pairs. We then visualize the learned gated values in Fig. 5 (top): as the network goes deeper, the gate values become larger, and compared with other blocks, the learned alpha gates in the first six layers have a much smaller value. We hypothesize that earlier blocks have a smaller modulation

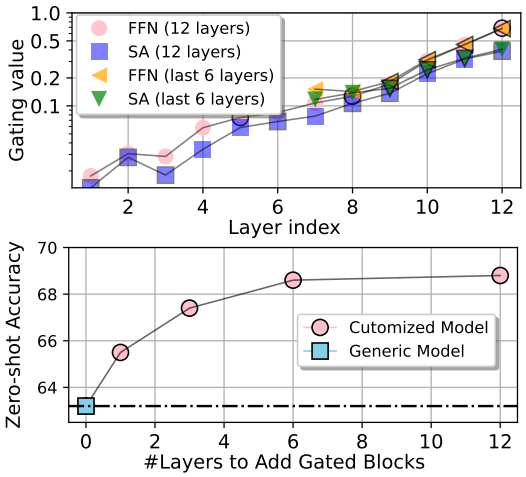

Figure 5. (Top) The learned gated values on the newly added SA and FFN layers for two networks: adding gated blocks in all 12 layers and in the last 6 layers, respectively; (Bottom) the zero-shot accuracy on ImageNet when adding gated blocks into different number of last layers.

to the base network, and removing them has minimal effect on the model's performance. Therefore, we vary the number of *last* layers which we add gated blocks to, and show empirical results in Fig. 5 (bottom). With more layers added to the base network, the model's zero-shot performance gradually increases, and saturates at 6 blocks. Therefore, we add gated blocks to last 6 layers as the default to balance the accuracy and the efficiency.

Search methods. Three search methods are compared: T2I, T2T, and T2I/T2T-combined. We find all modes consistently improve over the baseline CLIP. T2I retrieval alone yields a slightly worse performance, which may be partly due to its retrieved samples being more noisy and less relevant than T2T retrieval. T2T alone or T2I/T2T-combined has a similar performance. We use T2I/T2T-combined as our default strategy.

Search Methods	–	T2I	T2T	T2I/T2T
ImageNet-1K Accuracy	63.2	65.8	**68.6**	**68.6**

Comparison with self-training. We study and compare with the pseudo labeling strategy on retrieval-augmented model customization. After the relevant image-text pairs are retrieved, we use the pretrained CLIP checkpoint to assign pseudo labels to each retrieved image, and finetune the model as a classification task using the retrieved samples and pseudo labels. We optimize the network with UniCL loss [28]. With frozen text, pseudo-labeling can improve the model's performance with T2I/T2T data, while having a decreased performance with T2I data. This can be due to that the relevance of the T2I data is lower than other splits, and pseudo label assigned can be incorrect. Besides, gated self-attention is complimentary to pseudo labeling with ∼2% improvements. In contrast to pseudo labeling, training directly on the retrieved image-text pairs does not use heuristics to create pseudo labels, and the model can receive additional supervision signal, which we empirically find helpful for model adaptation and robustness.

Method	Tuning Strategy	Retrieval Methods T2I	T2T	T2I/T
CLIP	–		63.2	
Self-training	Locked-Text	62.4	63.6	64.6
	Locked-Text Gated-Image	64.3	66.1	66.2
REACT	Locked-Text Gated-Image	**65.8**	**68.6**	**68.6**

B.2. Discussions with Data-Centric Methods

It is recommended in [20] that the task learning capabilities of machine learning (ML) systems can be measured by task-level zero-shot transfer. This recommended evaluation setting is further generalized in [16] by showing that few/full-shot transfer consistently yields higher performance than zero-shot transfer. We argue that the task learning capabilities of ML systems can be improved from both the model and data perspectives. Most existing efforts devote to *model-centric* methods such as efficient network architectures [27], smarter training objectives [28], and scaling up model size [8, 29]. *Data-centric* methods are less explored, where our retrieval-augmented approach attempts to fill this data gap. We discuss the unique properties of REACT and build the connections with existing data-centric paradigms.

Relation to K-LITE. To build transferable visual systems, K-LITE [24] enriches entities in language supervision with structural knowledge in WordNet [19] and Wiktionary [18], in both model training and evaluation stages. It provides the first strong evidence that structural knowledge is effective in task-level transfer for CLIP/UniCL. Our paper is different in two aspects: (i) *Knowledge sources.* K-LITE considers textual common sense knowledge bases, while ours considers the web-scale image-text corpus. (ii) *Motivation.* K-LITE aims to improve the generality of visual models via structural human knowledge, while ours improves the customization of visual models using a plug-and-play task instruction augmentation process.

Relation to Self-Training. As a semi-supervised learning algorithm, self-training [23, 26] provides pseudo labels to the unlabelled images using a pre-trained neural (teacher) model. Though sharing the similarity in expanding the task-relevant data, the two methods are different in the augmented knowledge: (i) For an image, the supervision signal in self-training is based on the teacher model's internal "dark knowledge" [11], which is limited in a fixed prediction space. The supervision signal in our method is the paired text, which is collected from web as the external knowledge, which may contain richer semantics to describe the image. (ii) We build a retrieval process to acquire task-relevant images, which is lacking in self-training. The two methods can mutually benefit: self-training can start from our retrieval-augmented

pool, while we could use pseudo labels from self-training to get additional supervision.

C. Implementation details

C.1. Training Details in Customization

Model architecture. We mainly conduct our experiments on the vision Transformer backbones. For the ViT architecture, we mainly follow the implementation from CLIP [20]. The feature from the `CLS` token from the last visual encoder layer is used as the visual feature. For the gated-image experiments, we only add gated blocks to the last 6 layers. The hidden/embedding dimensions for gated blocks are set the same as the layer that it is added to. Following [1], the gate values are initialized as zero, modulated by `tanh` operator.

Training Protocol. We mainly follow CLIP [20] and UniCL [28] to set up our training hyperparameters. For optimization, we use AdamW [14] with a weight decay of 0.05 for all models. We set the learning rate to 0.0005 for locked-text-gated-image experiments, and 0.00005 for locked-text experiments. We use the same set of data augmentation and regularization as in [28]. For experiments with 10M retrieved samples, the models are trained for 32 epochs with a batch size of 4096. For experiments with fewer retrieved samples, the training epochs are adjusted accordingly so that they have a similar number of optimization steps. For all training, we used a cosine learning rate schedule, with 5000 iterations warmup.

C.2. Our Retrieval System

We implement our retrieval system using FAISS [13]. We use its Hierarchical Navigable Small World (HNSW) approximate k-NN lookup [17] to balance performance and efficiency. Product quantization is used to reduce the index size. We use Autofaiss to select the optimal hyperparameters for the index, and build the index using FAISS index factory. For LAION-400M, the selected configuration is: `OPQ256_768,IVF131072_HNSW32,PQ256x8`. We build two separate indexing systems for T2I and T2T retrieval. For T2I retrieval, CLIP image features are used for building the indexing system. For T2T retrieval, CLIP text features are used.

We benchmark below the latency and the recall of the HNSW k-NN lookup on a server with 64 CPU cores. As shown below, the indexing system is able to retrieve the relevant vectors accurately and efficiently.

Latency	R@1	R@10	R@20
0.57ms	84.8	94.8	97.5

C.3. Cost Estimation

We provide the cost estimation for the REACT pipeline. It includes feature extraction of the retrieval pool, indexing for the retrieval system, querying the indexing system to retrieve relevant image-text pairs, and finally model customization.

Feature extraction. For feature extraction using CLIP ViT-B/32 checkpoint, it takes around 250 T4 GPU hours, which equates to roughly a day on a desktop with 4x RTX 3090 GPUs.

Build index. We build our index system on a cloud VM with 24 CPU cores. Using the selected configuration in Sec. C.2, we can build the index system using FAISS within 20 hours. Note that we use the CPU version of FAISS, and do not leverage GPU acceleration for building the index.

Querying index. As shown in Sec. C.2, generating the retrieval set for model customization using the indexing system is very efficient. Typically, it takes less than 10 minutes to generate 10M indices for our retrieval pool.

Model customization. We train most of our models on a compute node with 16×V100 GPUs. For ViT-L checkpoints, we use 2-node distributed training, each node with 16×V100. It takes around 16/28/42 hours to train the B32/B16/L14 checkpoint, respectively, with either locked-text or locked-text-gated-image tuning strategy.

Remarks. Note that the feature extraction and building index system only needs to be done *once*. They are readily available for *any* queries from any domains. To make this line of research more accessible, our retrieved subsets for both ELEVATER and ImageNet will also be made available, with an easy-touse toolkit to download the subsets without storing the whole dataset locally. Therefore, researchers can directly run experiments on model customization.

In conclusion, REACT framework provides an accessible way to explore the large-scale web-crawled image-text dataset, and effectively and efficiently customize the models to the domain-of-interest.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 6

[2] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*, pages 456–473. Springer, 2022. 2

[3] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael Rabbat. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support sam-

ples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8443–8452, 2021. 2

[4] Zhaowei Cai, Avinash Ravichandran, Paolo Favaro, Manchen Wang, Davide Modolo, Rahul Bhotika, Zhuowen Tu, and Stefano Soatto. Semi-supervised vision transformers at scale. *arXiv preprint arXiv:2208.05688*, 2022. 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *ICCV*, 2021. 1

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 2

[7] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021. 2

[8] Priya Goyal, Dhruv Mahajan, Abhinav Gupta, and Ishan Misra. Scaling and benchmarking self-supervised visual representation learning. In *ICCV*, 2019. 5

[9] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 1

[10] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021. 1

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Dark knowledge. *Presented as the keynote in BayLearn*, 2014. 5

[12] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. July 2021. If you use this software, please cite it as below. 2, 3

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

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[15] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019. 1

[16] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, and Jianfeng Gao. ELEVATER: A benchmark and toolkit for evaluating language-augmented visual models. In *NeurIPS Track on Datasets and Benchmarks*, 2022. 1, 2, 5

[17] Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836, 2018. 6

[18] Christian M Meyer and Iryna Gurevych. *Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography*. na, 2012. 5

[19] George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998. 5

[20] 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. *arXiv preprint arXiv:2103.00020*, 2021. 2, 4, 5, 6

[21] 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. *arXiv preprint arXiv:2210.08402*, 2022. 1

[22] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1, 2

[23] Henry Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 1965. 5

[24] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Anna Rohrbach, Zhe Gan, Lijuan Wang, Lu Yuan, Ce Liu, Kurt Keutzer, Trevor Darrell, and Jianfeng Gao. K-LITE: Learning transferable visual models with external knowledge. In *NeurIPS*, 2022. 5

[25] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 1

[26] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 5

[27] Jianwei Yang, Chunyuan Li, and Jianfeng Gao. Focal modulation networks. *Advances in Neural Information Processing Systems*, 2022. 5

[28] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Lu Yuan, Ce Liu, and Jianfeng Gao. Unified contrastive learning in image-text-label space. *CVPR*, 2022. 5, 6

[29] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 5

[30] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2