ALIP: Adaptive Language-Image Pre-training with Synthetic Caption

Kaicheng Yang[1], Jiankang Deng[* 2], Xiang An[1], Jiawei Li[1], Ziyong Feng[1],
Jia Guo[3], Jing Yang[3], Tongliang Liu[4]
[1]DeepGlint [2]Huawei UKRD [3]InsightFace [4]University of Sydney
{kaichengyang,xiangan,jiaweili,ziyongfeng}@deepglint.com, tongliang.liu@sydney.edu.au
{jiankangdeng,guojia,y.jing2016}@gmail.com

A. More Analysis

A.1. Ablation on Different Caption Model

In Tab. 1, we present the performance using different caption models OFA vs. BLIP. The BLIP model we use is BLIP$_{CapFilt-L}$. Experimental results show that ALIP$_{BLIP}$ is on par with ALIP$_{OFA}$.

Table 1. Comparison with different caption models.

METHODS	MSCOCO I2T	MSCOCO T2I	LINEAR PROBE AVERAGE	ZERO-SHOT AVERAGE
ALIP$_{OFA}$	46.8	29.3	72.2	41.7
ALIP$_{BLIP}$	45.6	27.2	72.4	40.6

A.2. Analysis on Additional Costs

ALIP-ViT-B/32 consumes 40% extra FLOPs and 30% more memory than CLIP-ViT-B/32. As the model size escalates, these extra requirements become less significant. ALIP-ViT-L/14 necessitates only about a 7% increase in FLOPs and a mere 5% additional memory, compared to CLIP-ViT-L/14.

A.3. Analysis on Proportion of Noisy Pairs

To better demonstrate the effectiveness of the ALIP, we conducted a statistical analysis on the number of data pairs where $W_c > W_t$. There are 4,153,279 pairs, which accounts for about 27.6% of the total dataset.

B. Detail Experimental Settings

B.1. Experimental Settings

We show the settings in Tab. 2 for ALIP pre-training.

B.2. Model Architectures

We follow the same architecture as CLIP. Tab. 3 describe the detail of the ALIP-ViT-B/32 and ALIP-ViT-B/16.

B.3. Prompts for Zero-shot Classification

In this work, we evaluate the zero-shot performance of ALIP on 11 downstream datasets. All the prompts for the 11 downstream datasets are presented in Tab. 4.

Table 2. Hyperparameters used for ALIP pre-training.

Hyperparameter	Value
Initial temperature	0.07
Adam β_1	0.9
Adam β_2	0.98
Adam ϵ	10^{-6}
Weight decay	0.2
Batch size	4096
Learning rate	0.001
Learning rate scheduler	OneCycleLR
Pct start	0.1
Training epochs	32
GPU	16×V100

C. Detail Linear Probe on LAION

C.1. Downstream Datasets

We use 26 image classification datasets to prove the effectiveness of our method. These datasets include Food101 [2], CIFAR10 [15], CIFAR100 [15], Birdsnap [1], SUN397 [24], Stanford Cars [14], FGVC Aircraft [17], VOC2007 [8], DTD [5], Pets [19], Caltech101 [9], Flowers102 [18], MNIST [16], SLT10 [6], EuroSAT [11], RESISC45 [4], GTSRB [22], KITTI [10], Country211 [20], PCAM [23], UCF101 [21], Kinetics700 [3], CLEVR [12], Hateful Memes [13], SST2 [20], ImageNet [7].Details on each dataset and the corresponding evaluation metrics are provided in Tab. 5.

C.2. Detail Linear Probe results

We conduct experiments on randomly selected subsets of 10M and 30M from the LAION400M dataset. To provide a comprehensive comparison, we report the performance of the linear probe on 26 downstream datasets, the complete experimental results are shown in Tab. 6. The experimental results indicate that ALIP demonstrates both robustness and extensibility.

Table 3. The architecture parameters for ALIP models.

MODEL	EMBEDDING	INPUT	IMAGE ENCODER				TEXT ENCODER		
	DIMENSION	RESOLUTION	LAYERS	WIDTH	HEADS	PATCHS	LAYERS	WIDTH	HEADS
ALIP-VIT-B/32	512	224×224	12	768	12	32	12	512	8
ALIP-VIT-B/16	512	224×224	12	768	12	16	12	512	8

Table 4. Full list of prompts to evaluate the performance of zero-shot classification on 11 visual recognition datasets.

CIFAR 10 & CIFAR 100			
a photo of a {label}.	a blurry photo of a {label}.	a black and white photo of a {label}.	a low contrast photo of a {label}.
a high contrast photo of a {label}.	a bad photo of a {label}.	a good photo of a {label}.	a photo of a small {label}.
a photo of a big {label}.	a photo of the {label}.	a blurry photo of the {label}.	a black and white photo of the {label}.
a low contrast photo of the {label}.	a high contrast photo of the {label}.	a bad photo of the {label}.	a good photo of the {label}.
a photo of the small {label}.	a photo of the big {label}.		
Food101			
a photo of {label}, a type of food.			
Caltech101			
a photo of a {label}.	a painting of a {label}.	a plastic {label}.	a sculpture of a {label}.
a sketch of a {label}.	a tattoo of a {label}.	a toy {label}.	a rendition of a {label}.
a embroidered {label}.	a cartoon {label}.	a {label} in a video game.	a plushie {label}.
a origami {label}.	art of a {label}.	graffiti of a {label}.	a drawing of a {label}.
a doodle of a {label}.	a photo of the {label}.	a painting of the {label}.	the plastic {label}.
a sculpture of the {label}.	a sketch of the {label}.	a tattoo of the {label}.	the toy {label}.
a rendition of the {label}.	the embroidered {label}.	the cartoon {label}.	the {label} in a video game.
the plushie {label}.	the origami {label}.	art of the {label}.	graffiti of the {label}.
a drawing of the {label}.	a doodle of the {label}.		
Stanford Cars			
a photo of a {label}.	a photo of the {label}.	a photo of my {label}.	i love my {label}!
a photo of my dirty {label}.	a photo of my clean {label}.	a photo of my new {label}.	a photo of my old {label}.
DTD			
a photo of a {label} texture.	a photo of a {label} pattern.	a photo of a {label} thing.	a photo of a {label} object.
a photo of the {label} texture.	a photo of the {label} pattern.	a photo of the {label} thing.	a photo of the {label} object.
FGVC Aircraft			
a photo of a {label}, a type of aircraft.	a photo of the {label}, a type of aircraft.		
Flowers102			
a photo of a {label}, a type of flower.			
Pets			
a photo of a {label}, a type of pet.			
SUN39			
a photo of a {label}.	a photo of the {label}.		
ImageNet			
a bad photo of a {label}.	a photo of many {label}.	a sculpture of a {label}.	a photo of the hard to see {label}.
a low resolution photo of the {label}.	a rendering of a {label}.	graffiti of a {label}.	a bad photo of the {label}.
a cropped photo of the {label}.	a tattoo of a {label}.	the embroidered {label}.	a photo of a hard to see {label}.
a bright photo of a {label}.	a photo of a clean {label}.	a photo of a dirty {label}.	a dark photo of the {label}.
a drawing of a {label}.	a photo of my {label}.	the plastic {label}.	a photo of the cool {label}.
a close-up photo of a {label}.	a black and white photo of the {label}.	a painting of the {label}.	a painting of a {label}.
a pixelated photo of the {label}.	a sculpture of the {label}.	a bright photo of the {label}.	a cropped photo of a {label}.
a plastic {label}.	a photo of the dirty {label}.	a jpeg corrupted photo of a {label}.	a blurry photo of the {label}.
a photo of the {label}.	a good photo of the {label}.	a rendering of the {label}.	a {label} in a video game.
a photo of one {label}.	a doodle of a {label}.	a close-up photo of the {label}.	a photo of a {label}.
the origami {label}.	the {label} in a video game.	a sketch of a {label}.	a doodle of the {label}.
a origami {label}.	a low resolution photo of a {label}.	the toy {label}.	a rendition of the {label}.
a photo of the clean {label}.	a photo of a large {label}.	a rendition of a {label}.	a photo of a nice {label}.
a photo of a weird {label}.	a blurry photo of a {label}.	a cartoon {label}.	art of a {label}.
a sketch of the {label}.	a embroidered {label}.	a pixelated photo of a {label}.	itap of the {label}.
a jpeg corrupted photo of the {label}.	a good photo of a {label}.	a plushie {label}.	a photo of the nice {label}.
a photo of the small {label}.	a photo of the weird {label}.	the cartoon {label}.	art of the {label}.
a drawing of the {label}.	a photo of the large {label}.	a black and white photo of a {label}.	the plushie {label}.
a dark photo of a {label}.	itap of a {label}.	graffiti of the {label}.	a toy {label}.
itap of my {label}.	a photo of a cool {label}.	a photo of a small {label}.	a tattoo of the {label}.

D. More Visualization

D.1. Sample Visualization

In Fig. 1, we present visualizations of samples with raw images, raw texts, synthetic captions generated by OFA_{base}, and synthetic captions generated by OFA_{large}. It can be observed that the synthetic captions contain supplementary information that can potentially enhance representation learning. Moreover, the captions generated by OFA_{base} and OFA_{large} exhibit minimal differences.

D.2. Class Activation Maps

In Fig. 2, we present additional class activation maps of ALIP and CLIP for different classes from ImageNet. The

Table 5. List of linear probe datasets with the data distribution and evaluation metrics.

Dataset	Classes	Train size	Test size	Evaluation metric
Food101	102	75,750	25,250	accuracy
CIFAR10	10	50,000	10,000	accuracy
CIFAR100	100	50,000	10,000	accuracy
Birdsnap	500	42,138	2,149	accuracy
SUN397	397	19,850	19,850	accuracy
Cars	196	8,144	8,041	accuracy
Aircraft	100	6,667	3,333	mean per class
VOC2007	20	5011	4952	11-point mAP
DTD	47	3,760	1,880	accuracy
Pets	37	3,680	3,669	mean per class
Caltech101	101	3,000	5,677	mean-per-class
Flowers	102	2,040	6,149	mean per class
MNIST	10	60,000	10,000	accuracy
STL10	10	5,000	8,000	accuracy
EuroSAT	10	10,000	5,000	accuracy
RESISC45	45	3,150	25,200	accuracy
GTSRB	43	26,640	12,630	accuracy
KITTI	4	6770	711	accuracy
Country211	211	42,200	21,100	accuracy
PCAM	2	294,912	32,768	accuracy
UCF101	101	9,537	1,794	accuracy
Kinetics700	700	530,779	33,944	mean(top1,top5)
CLEVR	8	2,000	500	accuracy
Memes	2	8,500	500	ROC AUC
SST2	2	7,792	1,821	accuracy
ImageNet	1000	1,281,167	50,000	accuracy

Table 6. Top-1 accuracy(%) of linear probe on 26 image classification datasets.

Method	Pre-train data	Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech101	Flowers	MNIST	STL10	EuroSAT	RESISC45	GTSRB	KITTI	Country211	PCAM	UCF101	Kinetics700	CLEVR	Memes	SST2	ImageNet	Average
CLIP-ViT B/32	LAION10M	66.9	91.2	74.8	33.1	63.0	71.1	40.3	80.9	68.5	71.0	84.7	89.5	98.0	93.6	95.7	78.4	78.9	72.0	12.7	83.2	70.1	41.5	49.0	53.8	56.4	54.8	68.2
ALIP-ViT B/32	LAION10M	71.5	92.2	76.1	36.3	67.3	70.1	41.8	85.3	71.3	74.3	86.9	90.7	98.0	94.6	95.4	84.3	84.1	70.0	12.9	83.4	75.9	46.4	51.0	54.8	56.5	59.6	70.4
CLIP-ViT B/16	LAION10M	74.2	91.6	76.2	44.1	65.5	80.5	42.9	83.2	70.0	74.5	85.5	92.8	98.2	94.5	96.2	85.0	79.2	70.5	14.9	85.4	75.5	44.9	49.0	55.0	58.3	60.8	71.1
ALIP-ViT B/16	LAION10M	77.2	93.3	77.0	45.1	69.4	77.3	48.6	87.7	74.5	79.0	88.1	93.0	98.3	96.3	96.3	86.4	83.7	72.2	14.2	85.2	80.1	50.1	55.4	55.7	57.3	64.8	73.3
CLIP-ViT B/32	LAION30M	73.1	94.1	79.6	40.9	66.4	79.4	41.5	83.3	71.6	76.7	87.4	92.4	97.8	95.2	95.3	82.6	82.3	72.2	14.6	82.7	73.0	45.7	44.0	54.3	57.8	59.8	70.9
ALIP-ViT B/32	LAION30M	76.7	94.0	79.3	44.2	70.6	77.7	48.4	87.6	74.4	80.4	90.0	93.8	98.3	96.3	96.0	86.7	84.7	72.3	15.0	85.0	81.0	50.6	55.6	56.1	59.8	65.0	73.8

visualizations demonstrate that ALIP is superior in effectively aligning image patches and textual tokens.

References

[1] 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 *CVPR*, 2014. 1

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 1

[3] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv:1907.06987*, 2019. 1

[4] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 2017. 1

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 1

[6] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on*

artificial intelligence and statistics. JMLR, 2011. 1

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[8] Mark Everingham. The pascal visual object classes challenge,(voc2007) results. *http://pascallin. ecs. soton. ac. uk/challenges/VOC/voc2007/index. html.*, 2007. 1

[9] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR*, 2004. 1

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1

[11] 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*, 2019. 1

[12] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 1

[13] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *NeurIPS*, 2020. 1

[14] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, 2013. 1

[15] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[16] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998. 1

[17] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv:1306.5151*, 2013. 1

[18] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 2008. 1

[19] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *ICCV*, 2012. 1

[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. In *ICML*, 2021. 1

[21] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*, 2012. 1

[22] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 2012. 1

[23] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *MICCAI*, 2018. 1

[24] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *ICCV*, 2010. 1

Raw Image	**Raw Text**	**Synthetic Caption Generated by OFA$_{base}$**	**Synthetic Caption Generated by OFA$_{large}$**
	*"Northwest trip."*	*"A woman sitting in a chair talking on a cell phone."*	*"A woman sitting in a chair talking on her cell phone."*
	*"Beulah and rob on the platform sep visit to newcastle sep."*	*"A man standing next to a statue at a train station."*	*"A man standing at a train station with a stuffed animal ."*
	*"Beach day."*	*"A man standing on the beach looking at the ocean."*	*"A young boy standing on the beach holding up a cell phone."*
	*"Christmas eve sunset on the deck."*	*"A man and a woman sitting on a chair with a glass of wine."*	*"A man and a woman sitting on a bench with a glass of wine."*
	*"First birthday cake."*	*"A baby sitting in a high chair eating food."*	*"A baby sitting in a high chair eating a piece of cake."*
	*"Seven day."*	*"A yellow flower with a train in the background."*	*"A yellow flower in front of a train."*
	*"Tom being tom klo good thing we are at a stop light since he does not have his hands on the wheel…."*	*"A man wearing sunglasses sitting in a car."*	*"A man in a car with a beard and sunglasses."*
	*"Is he really milk baba?"*	*"A person laying on a blanket with a cat on the street."*	*"A man laying on the ground under a tent with a cat."*
	*"Switzerland autumn is coming high aperture and exposure 3 in 1"*	*"A stream in a forest with rocks and trees."*	*"A stream running through a forest with moss covered rocks."*

Figure 1. Examples of the image-text-caption triplet pairs from YFCC15M. We present the synthetic captions generated by the OFA$_{base}$ and OFA$_{large}$.

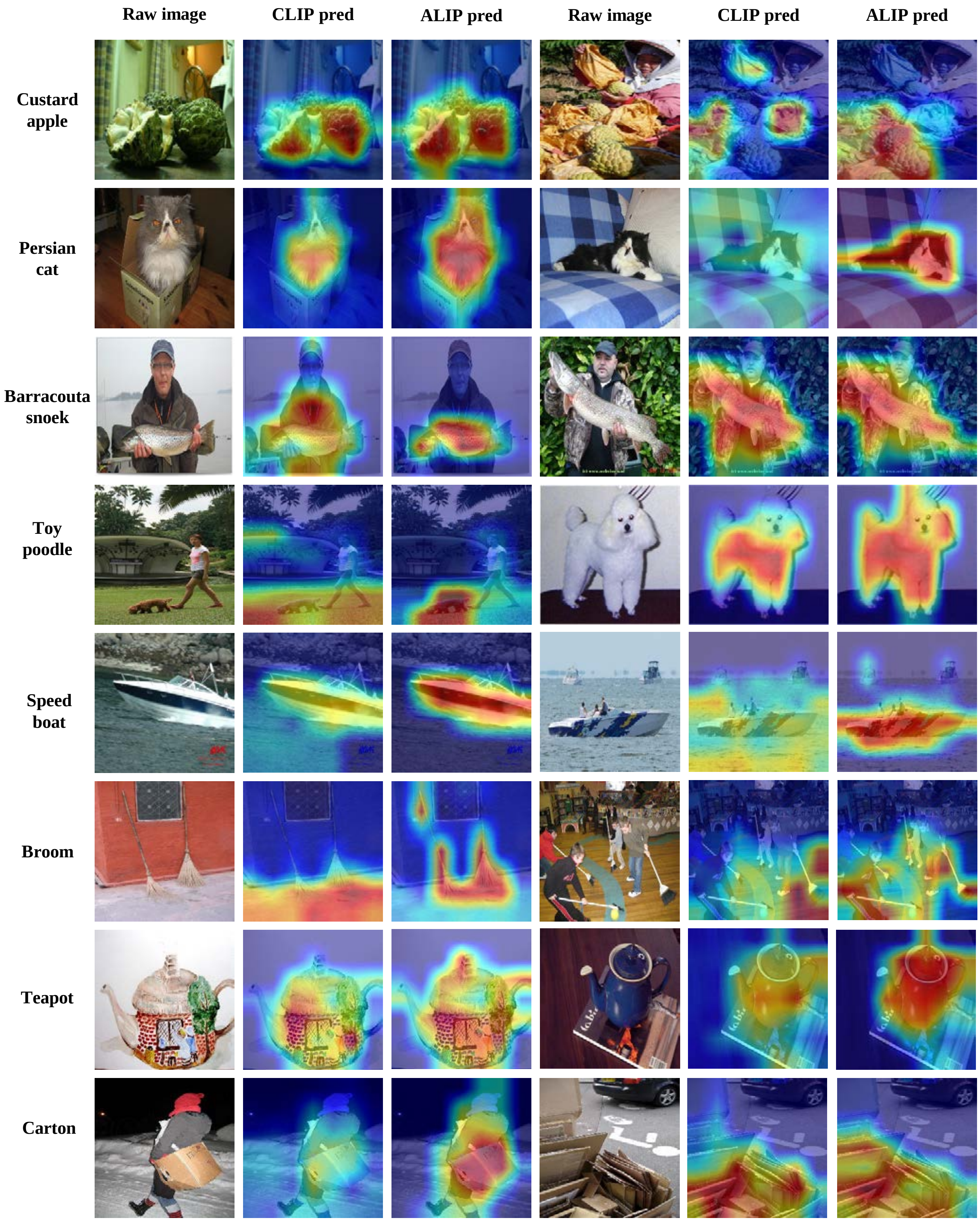

Figure 2. More class activation maps for CLIP and ALIP on different classes from ImageNet.