Finetune Like You Pretrain

Improving finetuning of zero-shot vision models

CVPR Poster Session : THU-AM-272

Sachin Goyal

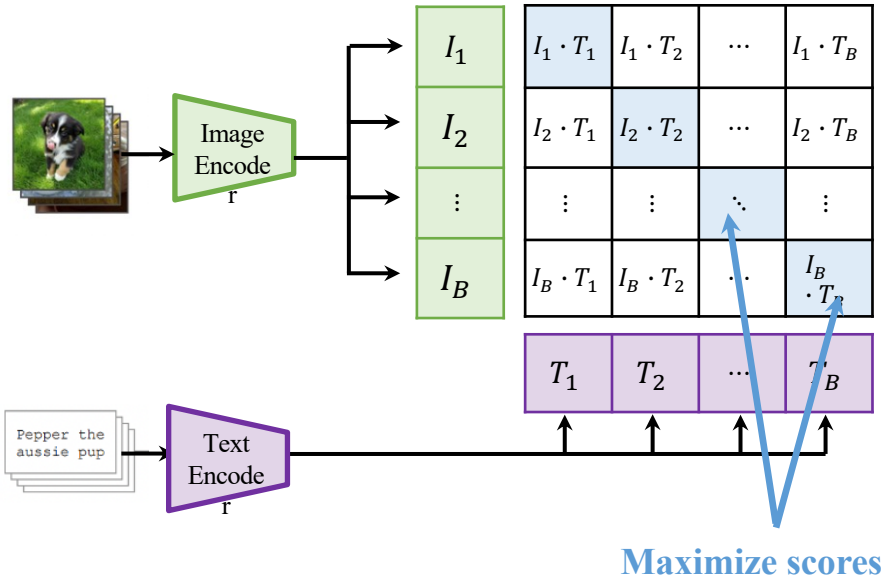
Ananya Kumar

Sankalp Garg

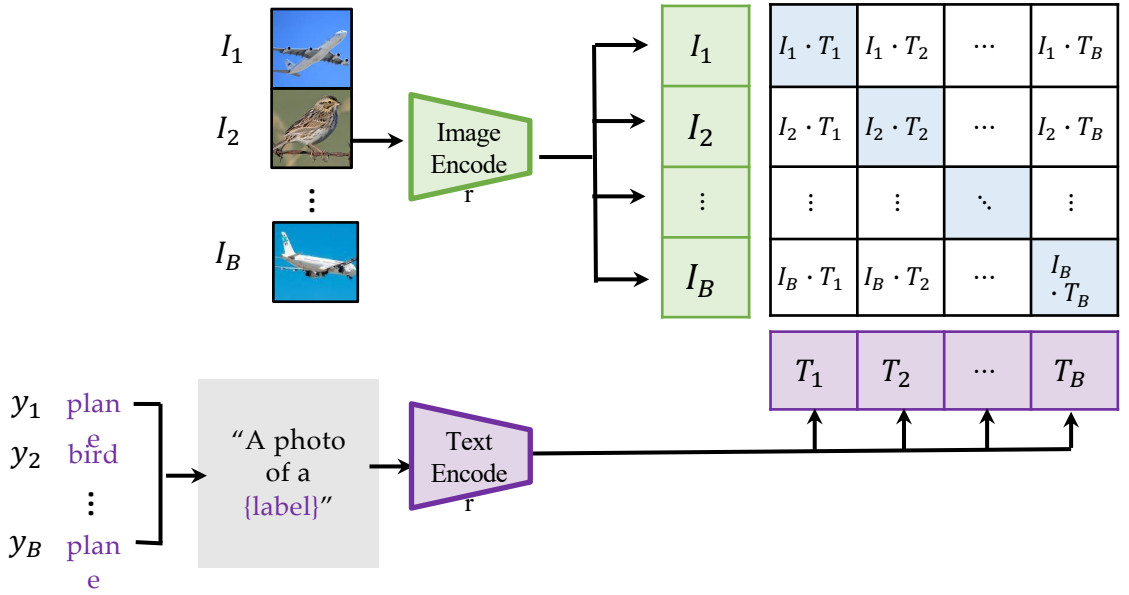
Zico Kolter

Aditi Raghunathan

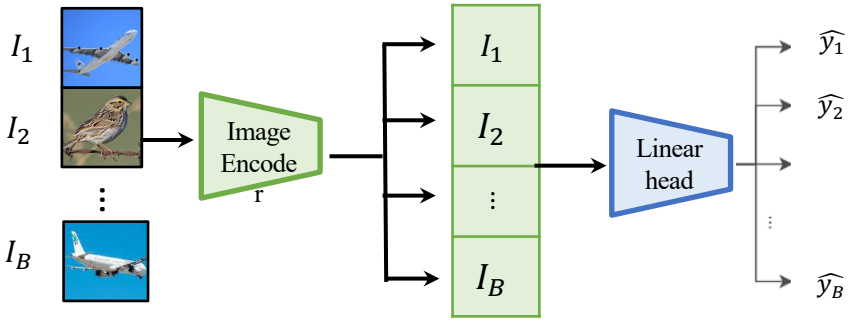
Contrastive pretraining of CLIP



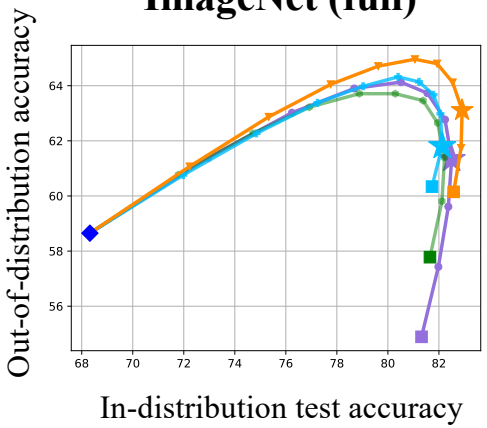
OURS : Finetune like you pretrain (FLYP)



Standard finetuning hurts robustness



ImageNet (full)



ImageNet (4-shot)

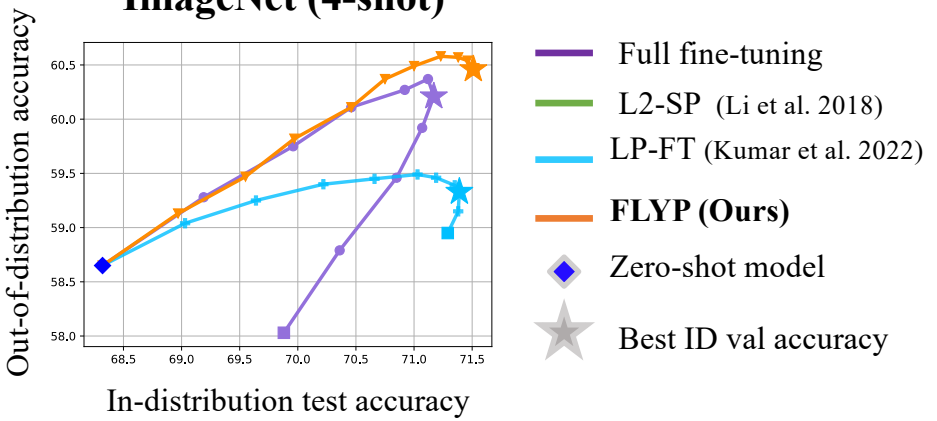


Image-text contrastive pretraining

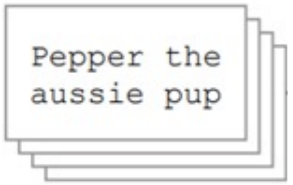


Image-text contrastive pretraining

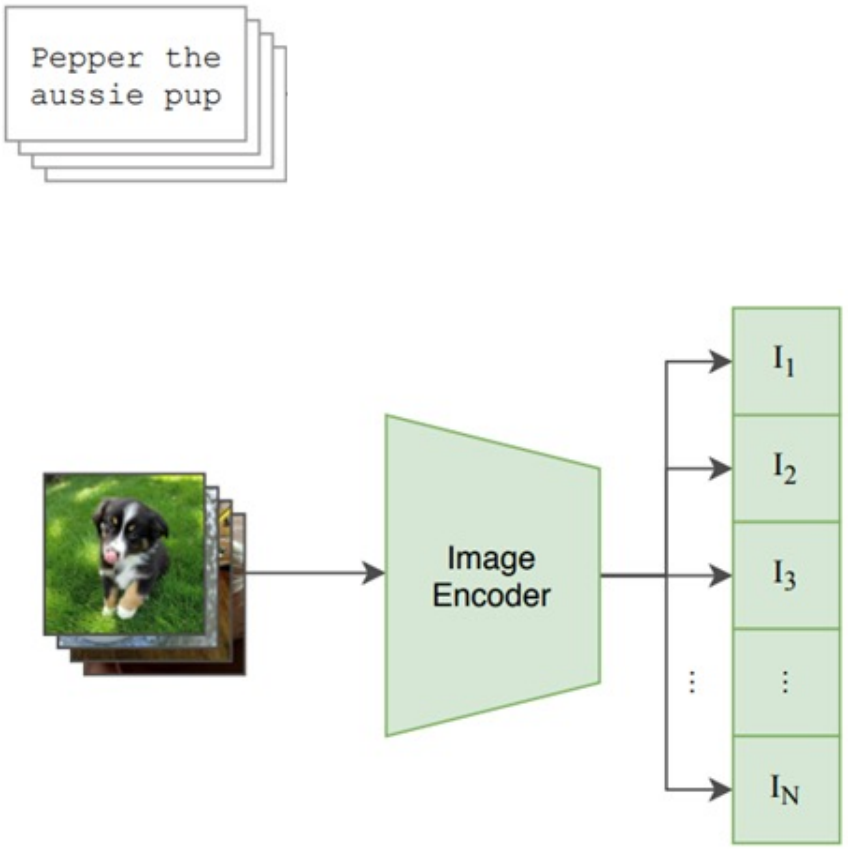


Image-text contrastive pretraining

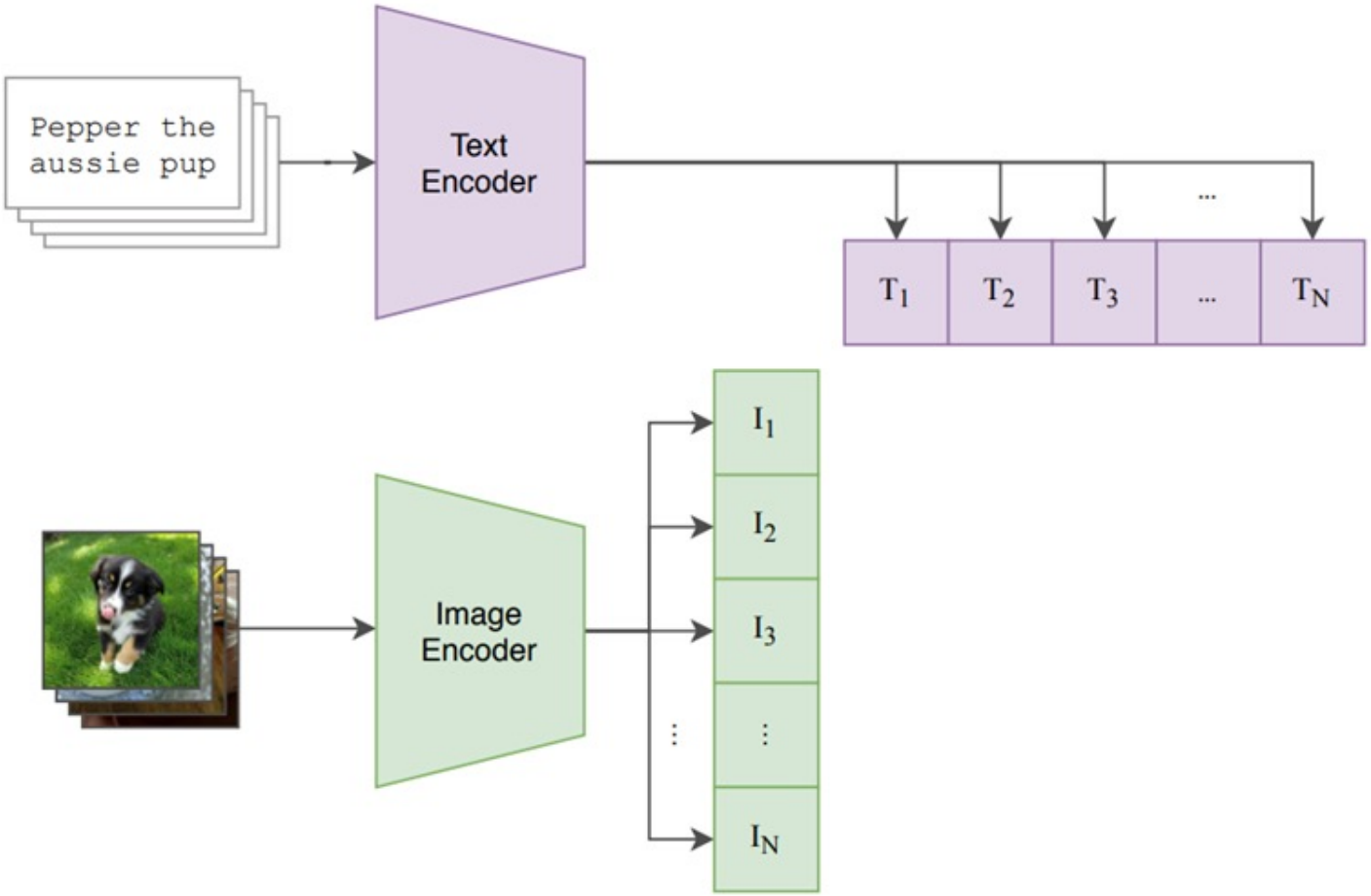
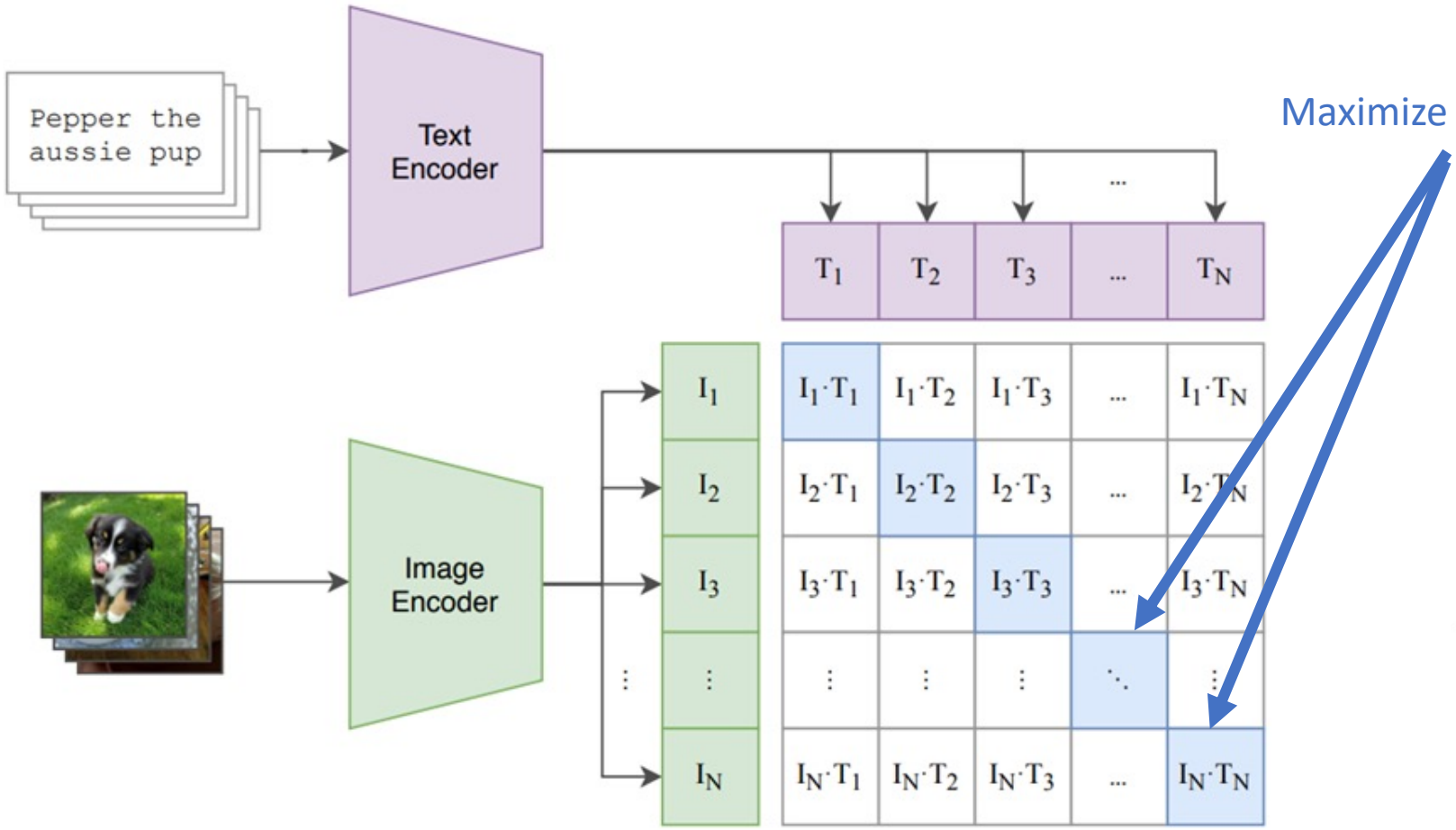


Image-text contrastive pretraining

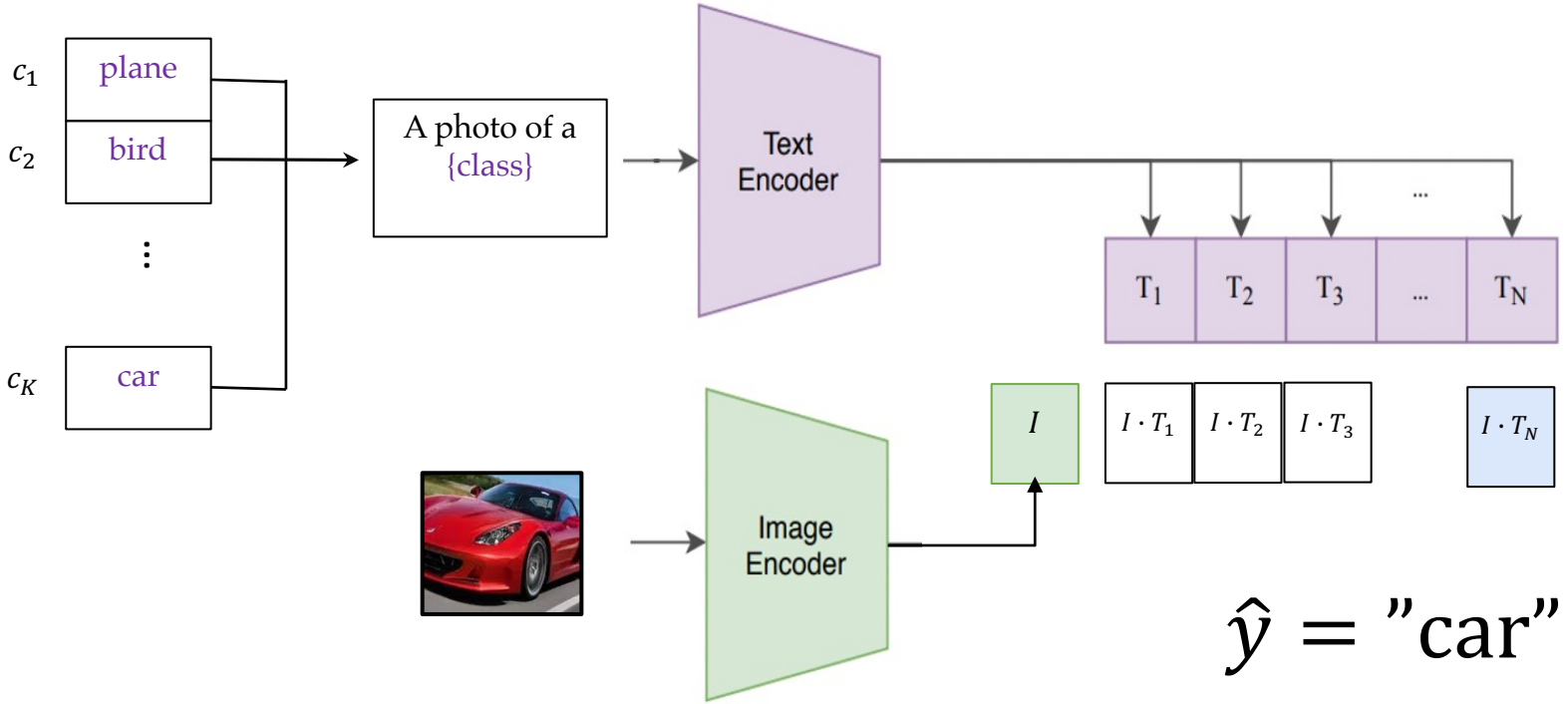


Optimizing the **contrastive loss**

$$\ell_{pre} = -\sum_{i=1}^{B} I_i \cdot T_i + log \sum_{i,j=1}^{B} I_i \cdot T_j$$

Zero-shot prediction

- Surprisingly amazing zero-shot classification performance!



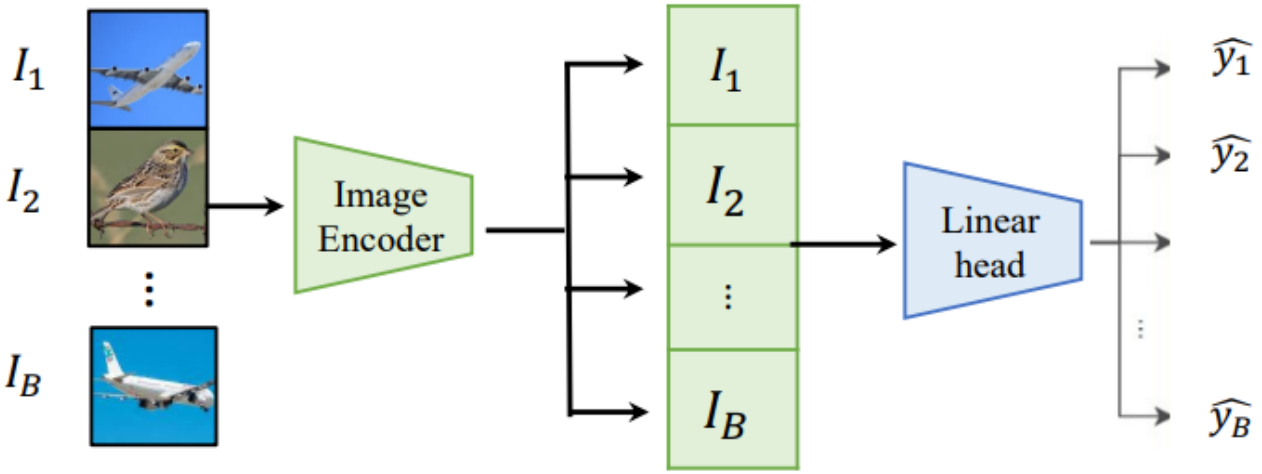
Zeroshot Performance

- Performs quite well.

CLIP ViT-B-16	ImageNet ID Accuracy	ImageNet OOD Accuracy
Zero-Shot	68.3	58.7

- In practice however, one would like to further finetune on task-specific data.

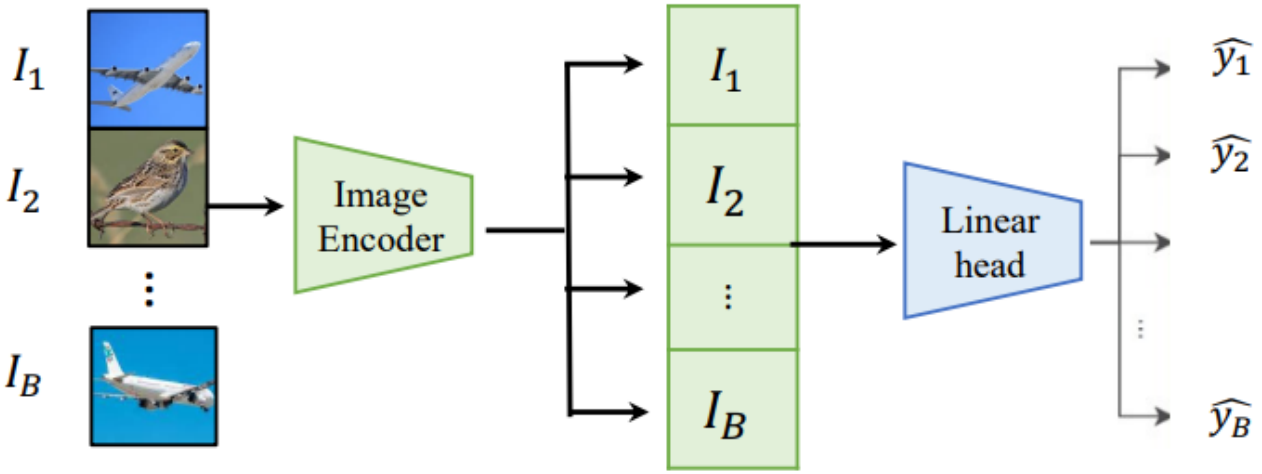
Standard Finetuning of CLIP



Standard approaches involve finetuning the image encoder + linear head **using the cross-entropy loss**

$$\ell_{fine} = -\sum_{i=1}^{B} \ell_{xent}(\hat{y}_i, y_i)$$

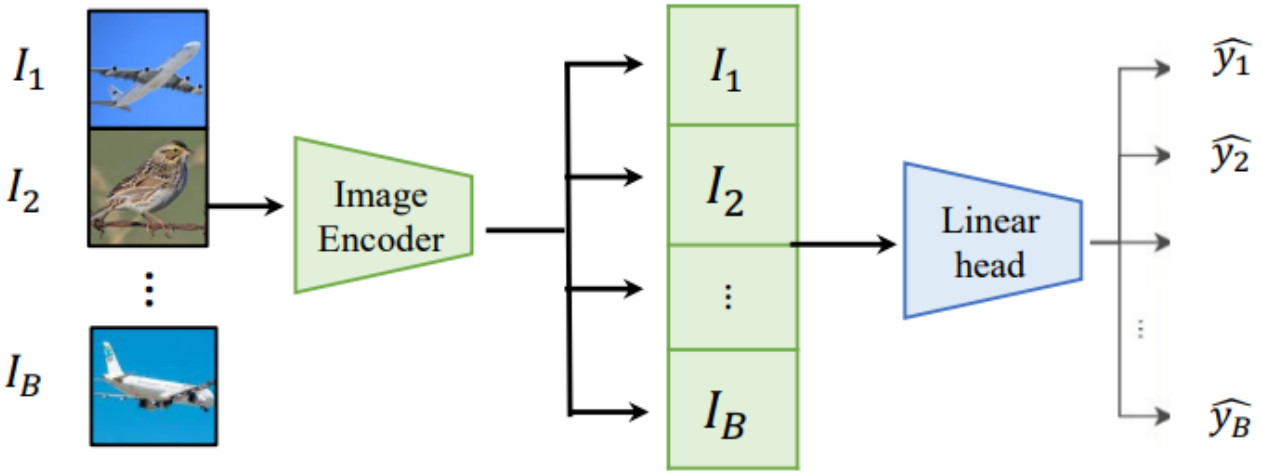
Standard Finetuning of CLIP



Standard approaches involve finetuning the image encoder + linear head **using the cross-entropy loss**

$$\ell_{fine} = -\sum_{i=1}^{B} \ell_{xent}(\hat{y}_i, y_i)$$

Standard Finetuning of CLIP

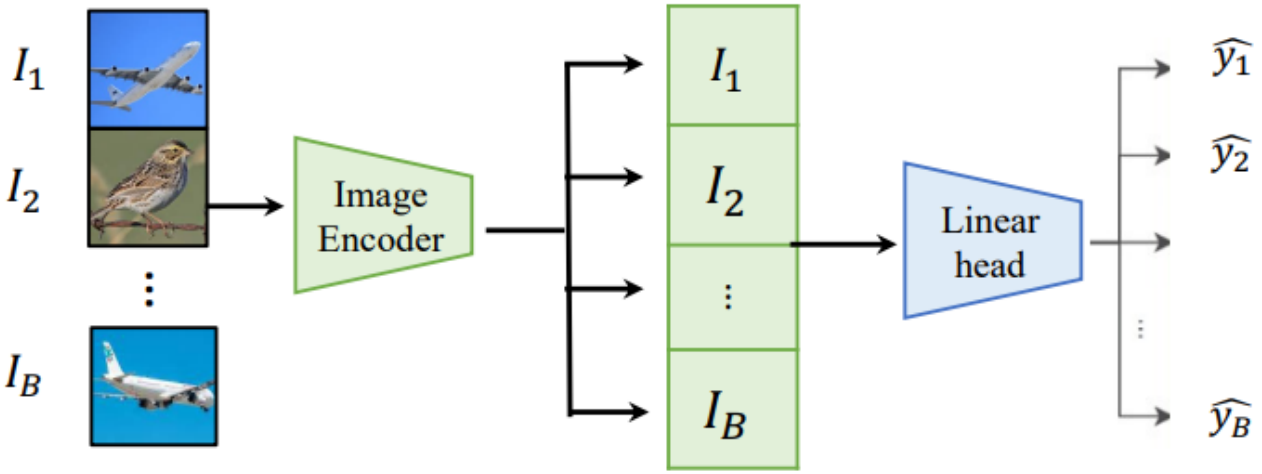


Standard approaches involve finetuning the image encoder + linear head **using the cross-entropy loss**

$$\ell_{fine} = -\sum_{i=1}^{B} \ell_{xent}(\hat{y}_i, y_i)$$

	ImageNet ID Accuracy	ImageNet OOD Accuracy
Zero-Shot	68.3	58.7
Finetuning	81.4	

Standard Finetuning of CLIP



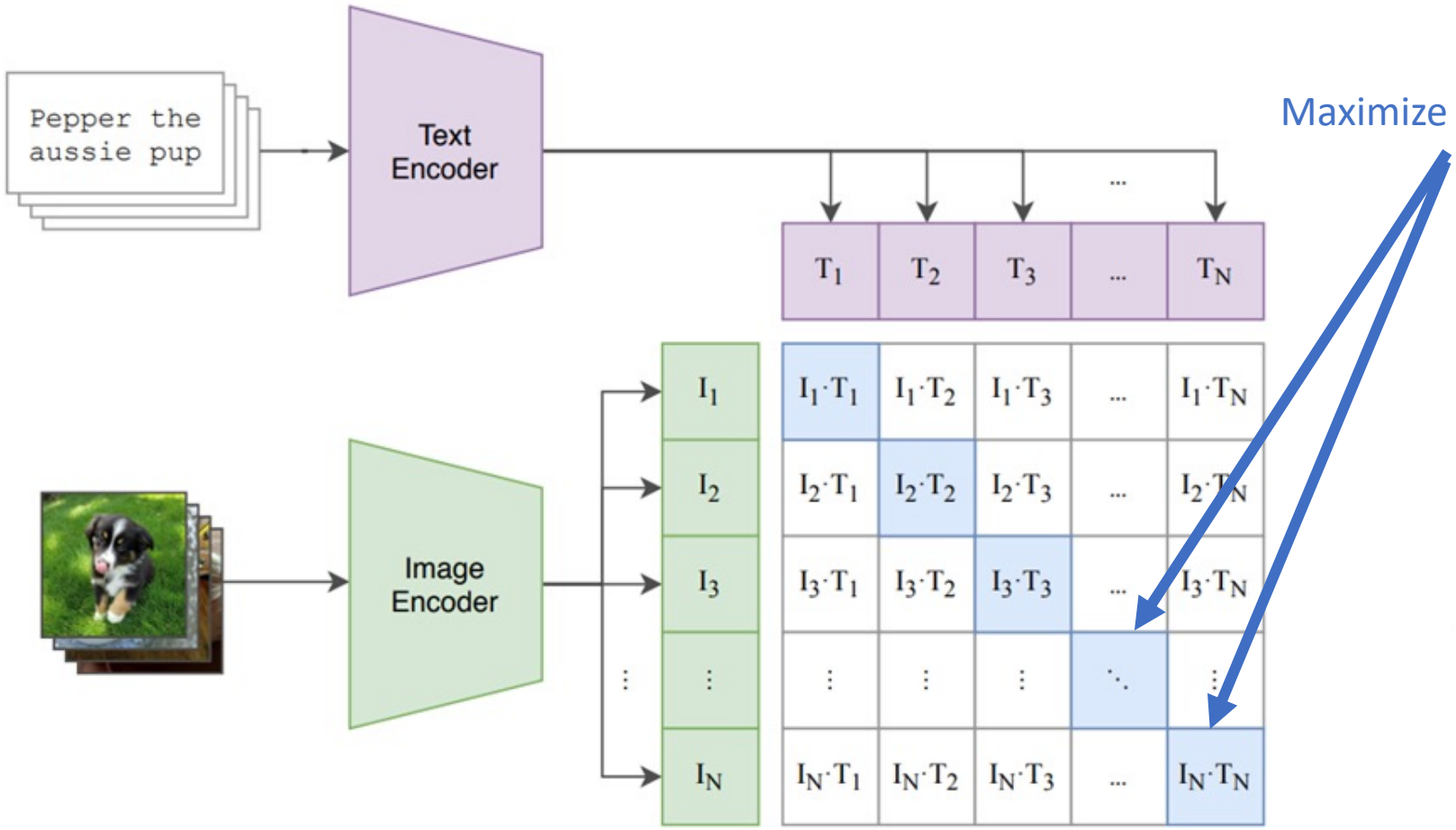
Standard approaches involve finetuning the image encoder + linear head **using the cross-entropy loss**

$$\ell_{fine} = -\sum_{i=1}^{B} \ell_{xent}(\hat{y}_i, y_i)$$

- Mismatch between the pretraining objective and the finetuning objective.

	ImageNet ID Accuracy	ImageNet OOD Accuracy
Zero-Shot	68.3	58.7
Finetuning	81.4	**54.8**

Image-text contrastive pretraining

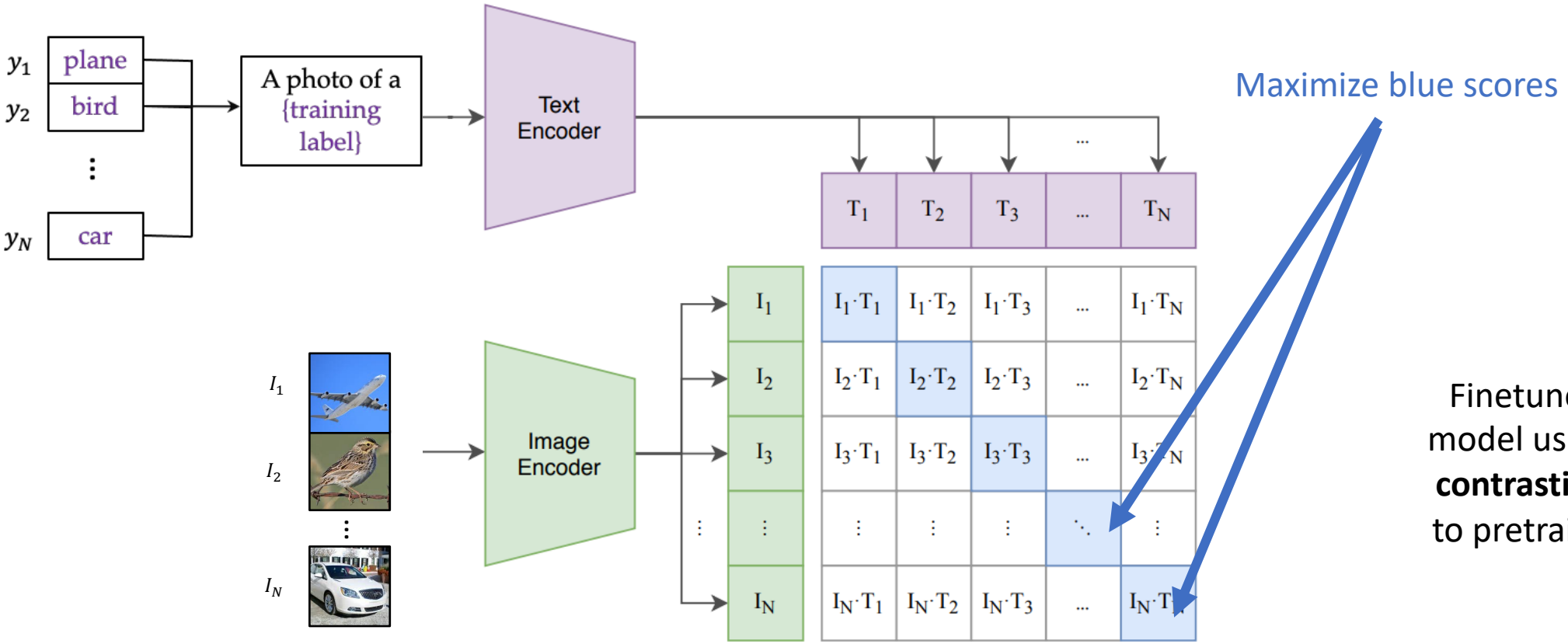


Maximize blue scores

Optimizing the **contrastive loss**

$$\ell_{pre} = -\sum_{i=1}^{B} I_i \cdot T_i + log \sum_{i,j=1}^{B} I_i \cdot T_j$$

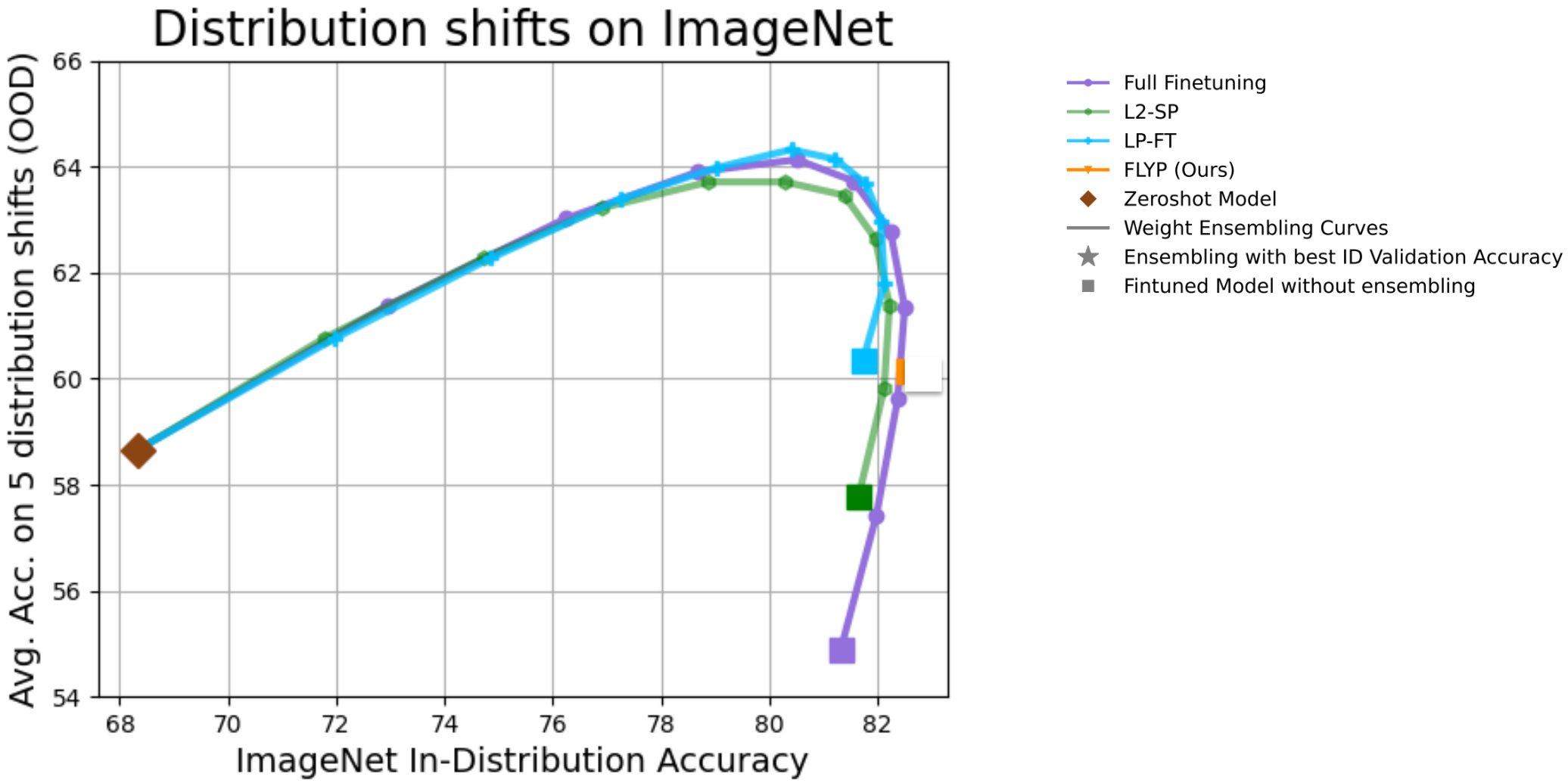
Finetune Like You Pretrain



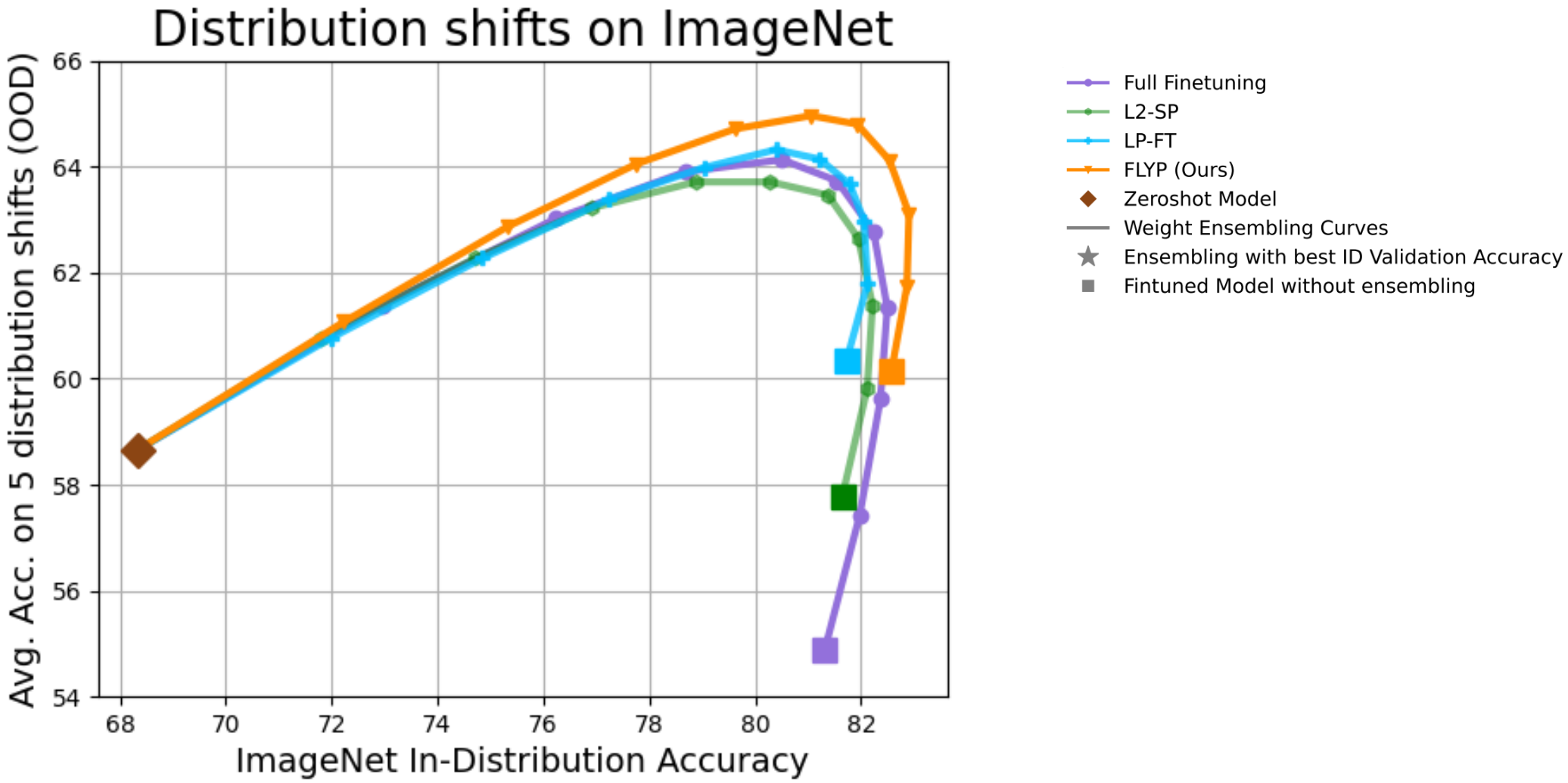
Finetune the entire model using the **same contrastive loss** used to pretrain the model

Construct captions from labeled data, and continue pretraining

FLYP outperforms other methods (I)



FLYP outperforms other methods (I)



State-of-the-Art Accuracies

- FLYP on CLIP-large: SOTA on iWildCam (wildlife conservation dataset)
- Highest ever reported accuracy

	Architecture	ID	OOD
FLYP	ViTL-336px	**59.9 (0.7)**	**46.0 (1.3)**
Model Soups	ViTL	57.6 (1.9)	43.3 (1.)
ERM	ViTL	55.8 (1.9)	41.4 (0.5)
ERM	PNASNet	52.8 (1.4)	38.5 (0.6)
ABSGD	ResNet50	47.5 (1.6)	33.0 (0.6)

FLYP outperforms other methods (II)

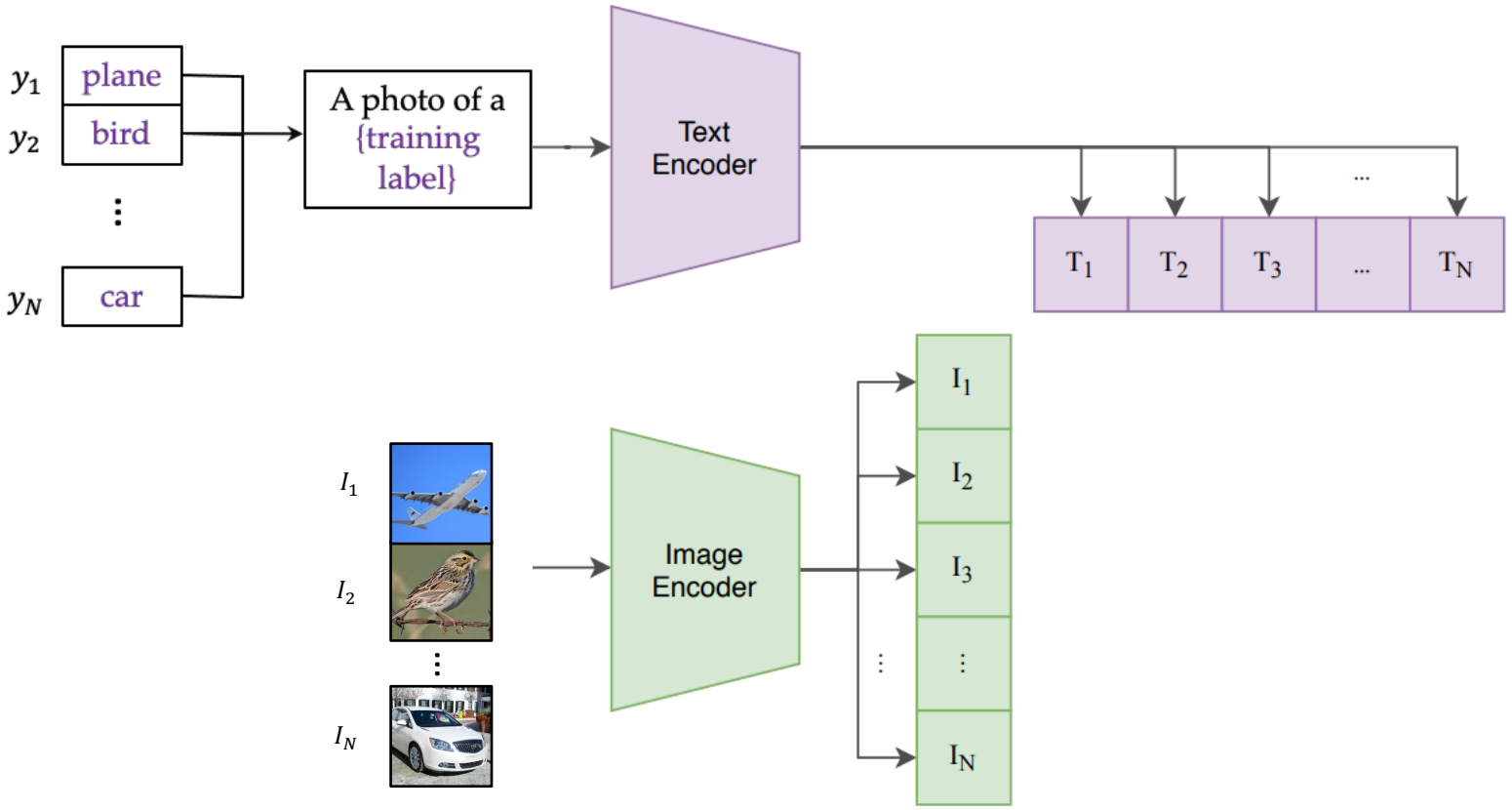
Methods	iWILDCam		FMoW	
	ID	OOD	ID	OOD
Zeroshot	08.7	11.0	20.4	18.6
FT	48.1	35.0	68.5	39.2
LP-FT	50.2	35.7	68.4	40.4
FLYP (OURS)	**52.5**	**37.1**	**68.6**	**41.3**

- More results: Few-shot learning, standard transfer learning datasets

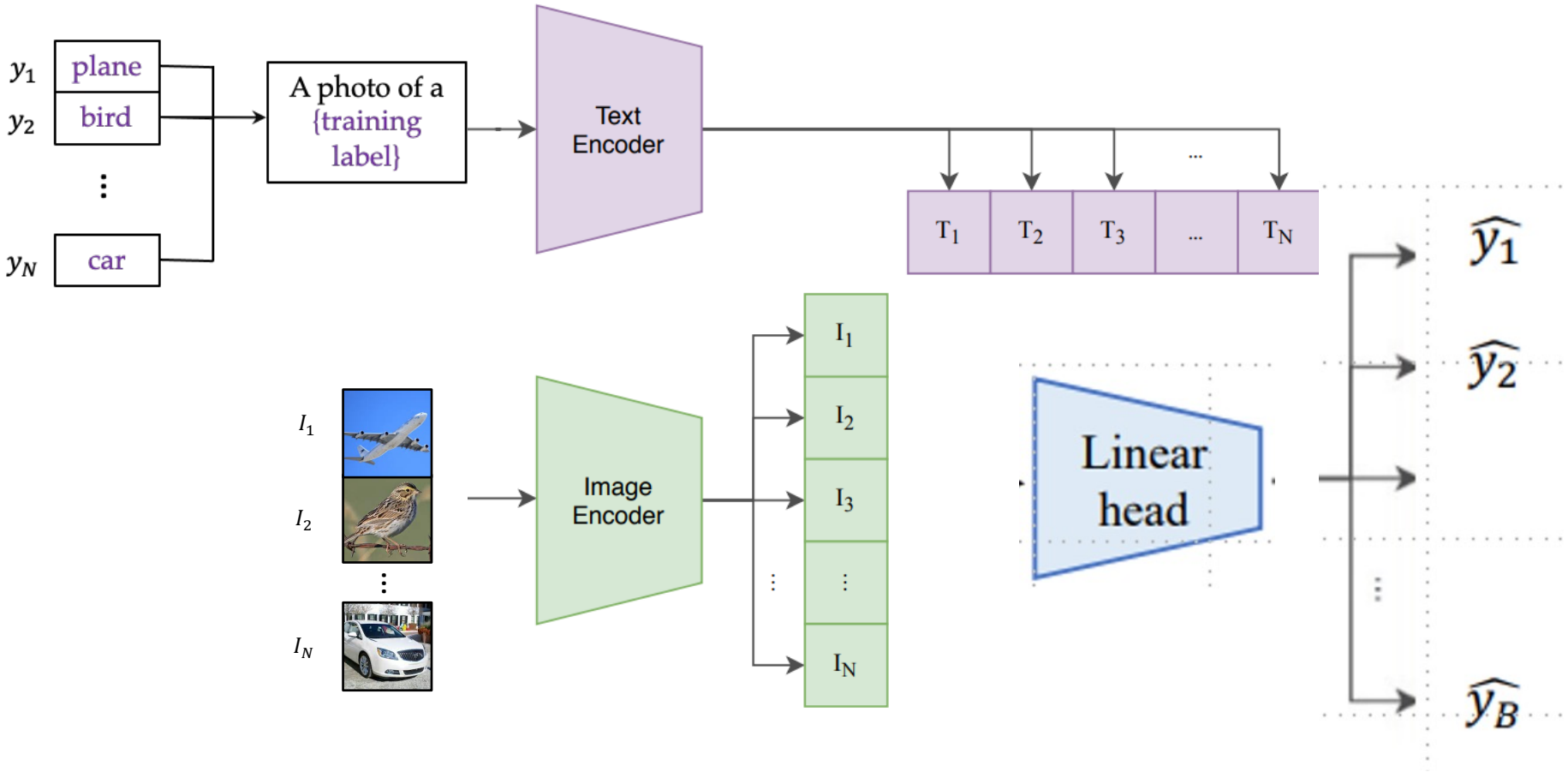
Using fine-tuning = pretraining is key

- Subtle deviations from pretraining worsen accuracy
- FLYP finetunes both the encoders. Are the gains simply due to finetuning both the encoders?
- **Ablation 1 :** Finetuning both the image and language encoders using cross-entropy loss.

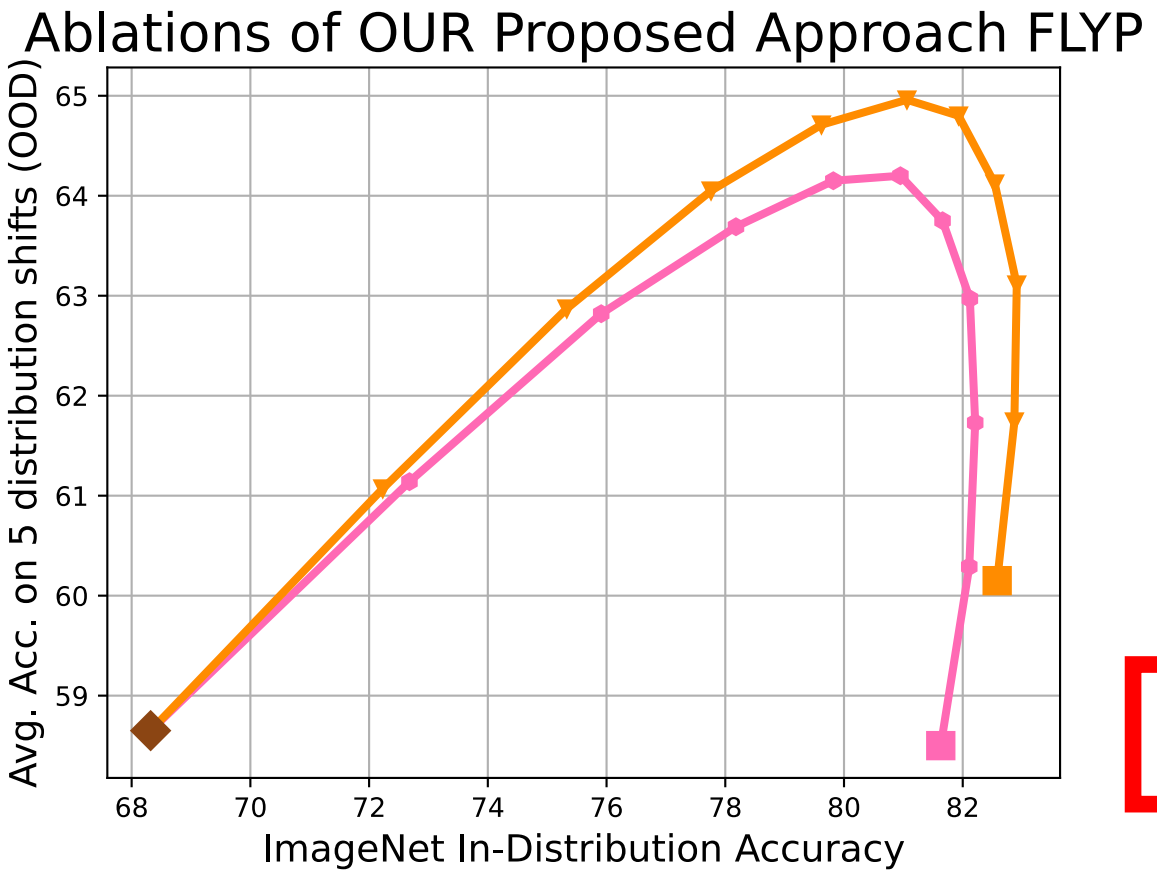
Finetuning both encoders using CE loss

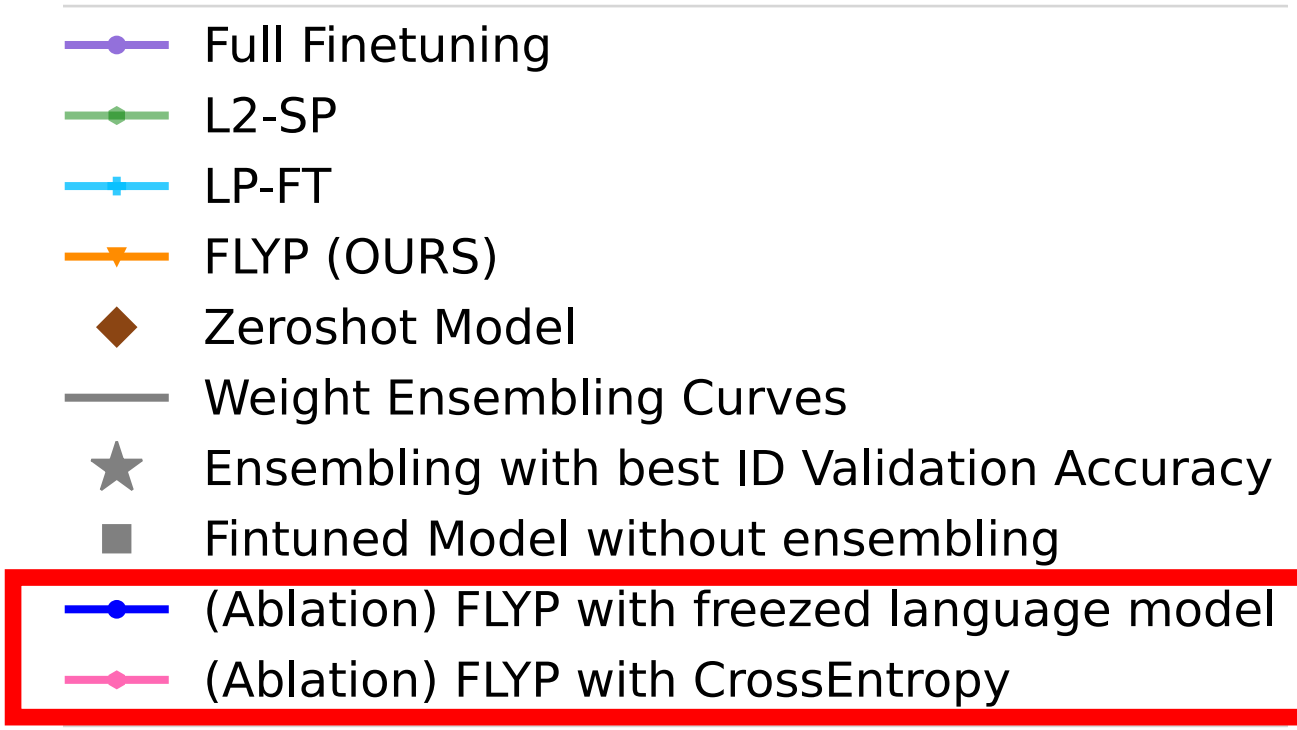


Finetuning both encoders using CE loss

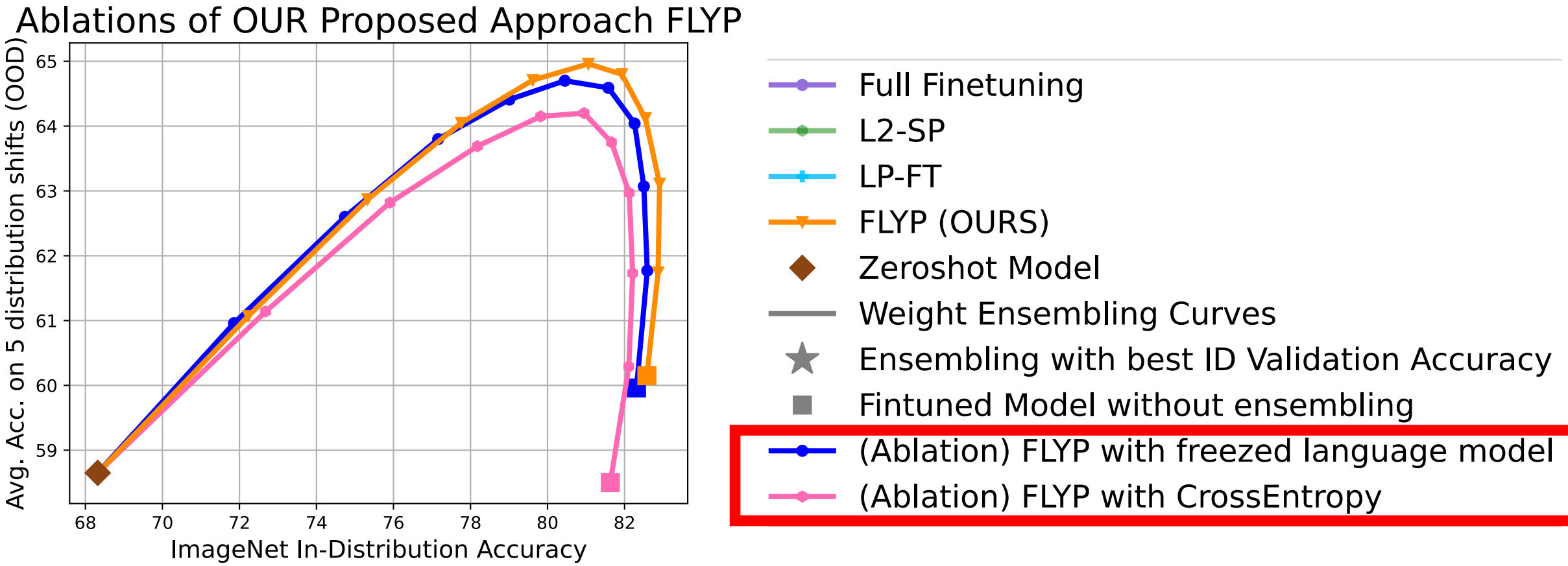


Finetuning both encoders using CE loss





Finetuning only the image encoder?



Using fine-tuning = pretraining is key

- Correcting for class-collisions in contrastive finetuning.
  - There can be multiple samples from the same class in a mini-batch
  - The contrastive loss would also separate embeddings for such samples, which can be sub-optimal.
  - However, removing such collisions only hurts the accuracy.

Summary

- We need to think carefully about finetuning procedures for the pretrained models.

- In this work, we showed a simple change to the finetuning procedure for CLIP can greatly improve the robustness of the finetuned model.