Distilling from Vision-Language Models for Improved OOD Generalization in Vision Tasks

Sravanti Addepalli*, Ashish Asokan*, Lakshay Sharma, R. Venkatesh Babu
Vision and AI Lab, Indian Institute of Science, Bangalore, India

ODRUM @ CVPR '23
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

1. Contributions

- We investigate the robustness of image and text embeddings of a Vision Language Model (VLM), and highlight the importance of text embeddings for better OOD generalization.
- We propose the following white-box and black-box distillation methods to improve OOD generalization of vision classification models using VLMs:
  - **VL2V-SD** (**Vision-Language to Vision - Self Distillation**), a white-box Self-Distillation approach that imparts the generalization of text embeddings to the image encoder.
  - **VL2V-ADiP** (**Vision-Language to Vision - Align, Distill, Predict**), a black-box distillation approach that combines the features of a pre-trained vision model with the text and vision encoders of the VLM teacher for better OOD generalization.
- We demonstrate SOTA results on DomainBed in both black-box and white-box settings of the VLM.

2. Motivation for a black-box setting

- General purpose open-sourced models cannot be used in specialized applications like healthcare, which need application-specific data with expert annotation.
- Security-critical applications demand the use of clean training data that is free from data-poisoning attacks and biases, while maintaining data privacy.
- This motivates the need for training highly specialized VLMs, which can be expensive in terms of data collection, annotation and curation, in addition to the training costs. Thus, the trained VLM is valuable IP.
- This motivates a vendor-client setting, where vendor trains a model and grants only black-box access to the client on a pay-per-query basis. The client minimizes inference costs by using a distilled model.

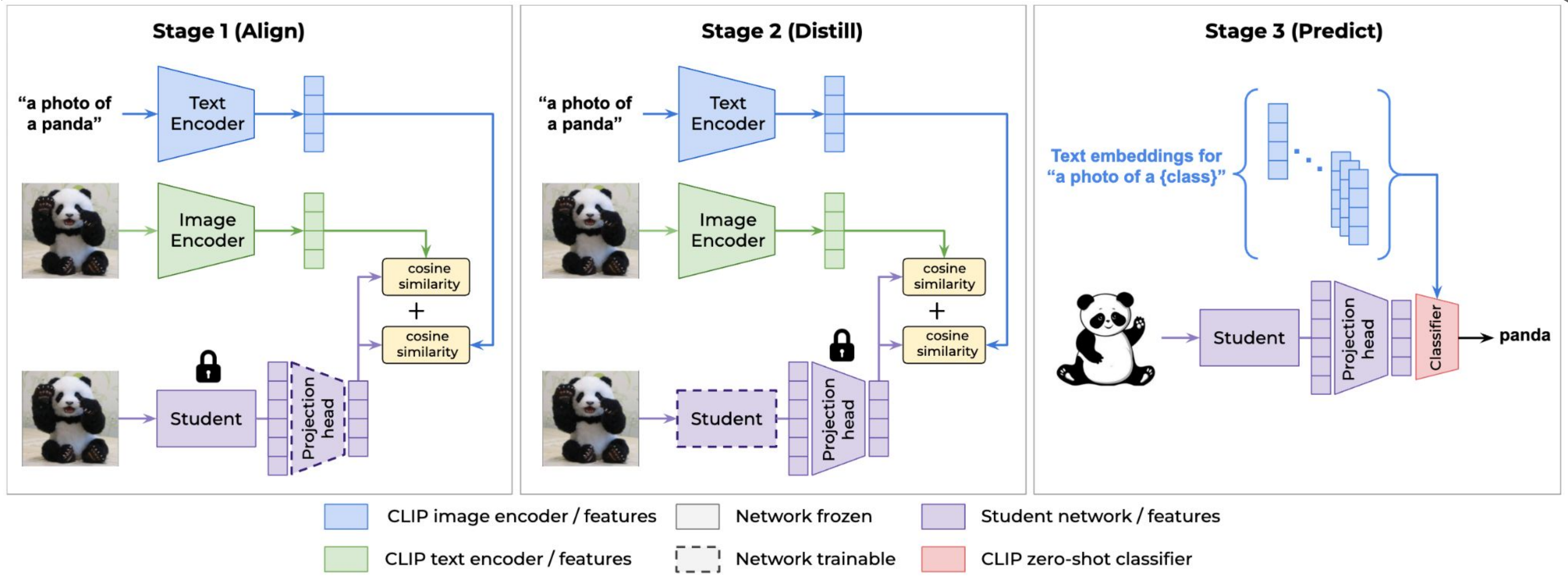
3. Nearest Neighbor evaluation using Image and Text embeddings

- [E1, E2]: Text embeddings can be used effectively for zero-shot classification since the text encoder learns a **generalized representation** for a **concept,** that is consistent across various distributions.
- [E3 - E6]: The image encoder of CLIP learns **unique representations** for variations in pose, color and background, and thus does not yield generalized representations that are common across both source and target domains.

Embedding used for computing similarity	OH	TI	VLCS	PACS	Average
E1: T.E. for "A photo of a {class}"	82.36	34.19	82.08	96.10	73.68
E2: Avg. T.E. for "A {domain} of a {class}" across all train domains	83.70	35.55	82.28	96.21	74.44
E3: Avg. I.E. of all images in each class (Source domain)	71.37	33.99	48.21	79.03	58.15
E4: Avg. I.E. of all images in each class (Target domain)	78.21	38.69	69.31	93.08	69.82
E5: Avg. I.E. of 10 images per class closest to test image (Source domain)	76.42	39.33	76.42	92.15	71.08
E6: Avg. I.E. of 10 images per class closest to test image (Target domain)	84.86	85.38	87.88	98.32	89.11

T.E. : Text Embedding, I.E.: Image Embedding

4. Proposed Approach : VL2V-ADiP (Align, Distill, Predict)



Algorithm Steps:

- **Align -** Train the projection head using $\mathcal{L}_{ADiP}$
- **Distill -** Train the student backbone using $\mathcal{L}_{ADiP}$
- **Predict -** Inference using VLM's zero-shot classifier

Training loss: $\mathcal{L}_{\text{ADiP}} = -\frac{1}{2n}\sum_{i=1}^{n}\left\{\cos(\mathbf{PF}^s_{x_i}, \mathbf{T}_{y_i}) + \cos(\mathbf{PF}^s_{x_i}, \mathbf{I}^t_{x_i})\right\}$

- $\mathbf{PF}^s_{x_i}$ - Projection head features of student s for sample x_i
- $\mathbf{T}_{y_i}$ - Text embedding of VLM for ground truth label y_i
- $\mathbf{I}^t_{x_i}$ - Image embedding for sample x_i of VLM teacher t

5. Results on DomainBed

(a) VL2V-SD - The white-box setting, where the ViT-B/16 student is initialized with the image-encoder of CLIP, and distilled from the ViT-B/16 CLIP model. (S): SWAD [2]

Method	OH	TI	VLCS	PACS	DN	Avg.
CLIP Zero-Shot	82.40	34.10	82.30	96.50	57.70	70.60
ERM (S) Finetuning	81.01	42.92	79.13	91.35	57.92	70.47
MIRO [1]	82.50	54.30	82.20	95.60	54.00	73.72
MIRO (S) [1, 2]	84.80	**59.30**	82.30	96.44	60.47	76.66
VL2V - SD (Ours)	**87.38**	58.54	**83.25**	**96.68**	**62.79**	**77.73**

(b) VL2V-ADiP - The black-box setting, where the client has only input-output access to CLIP. The ViT-B/16 student is initialized with ImageNet pre-trained model, and distilled from CLIP ViT-B/16. (S) : SWAD [2]

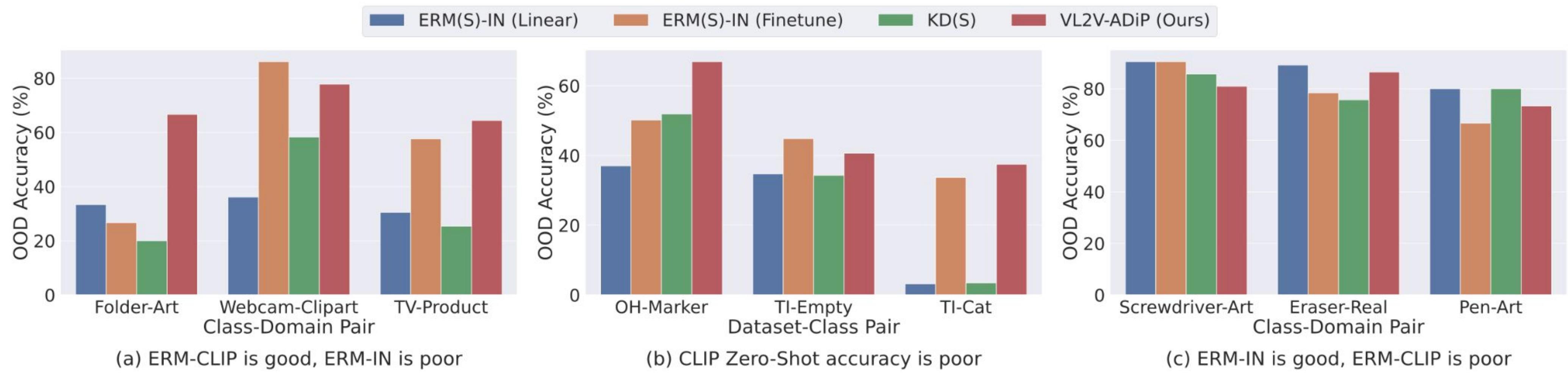
Method	OH	TI	VLCS	PACS	DN	Avg-ID	Avg-OOD
ERM (linear)	71.8	24.4	78.6	66.4	36.7	74.3	58.7
ERM (finetune)	78.0	42.5	78.1	85.3	50.8	86.9	70.3
MIRO [1]	74.9	44.5	80.4	81.6	49.9	86.6	69.6
KD [3]	77.6	38.7	79.7	84.9	50.7	87.0	69.8
ERM (S) (linear)	71.5	31.4	77.5	67.0	36.6	74.0	56.8
ERM (S) (finetune)	83.2	50.0	80.3	90.3	56.1	**89.3**	72.0
MIRO (S) [1, 2]	80.1	50.3	81.1	89.5	55.7	88.7	71.3
KD (S) [3, 2]	82.7	48.4	80.5	91.5	56.1	89.2	71.8
Ours	**85.7**	**55.4**	**81.9**	**94.9**	**59.4**	89.0	**75.5**

(c) VL2V-ADiP (lower capacity student models) - Teacher is ViT-B/16 CLIP model, and student has ImageNet initialization.

Student	Method	OH	TI	VLCS	PACS	DN	Avg-OOD
ViT-B/16 (86M)	KD (S)	82.7	48.4	80.5	91.5	56.2	71.8
	Ours	**85.7**	**55.4**	**81.9**	**94.9**	**59.4**	**75.5**
ViT-S/16 (22M)	KD (S)	78.1	50.1	79.1	86.0	52.0	69.1
	Ours	**81.2**	**52.5**	**81.4**	**89.3**	**54.2**	**71.7**
DeiT-S/16 (22M)	KD (S)	74.7	48.1	78.9	88.1	49.1	67.8
	Ours	**77.6**	**48.7**	**81.9**	**89.0**	**50.4**	**69.5**
ResNet-50 (26M)	KD (S)	70.7	51.2	78.6	**87.2**	46.3	66.8
	Ours	**74.4**	**53.5**	**79.2**	86.7	**47.7**	**68.3**

6. Comparison of OOD accuracy in extreme cases (VL2V-ADiP)

Comparison of OOD accuracy in three extreme cases: (a) CLIP image encoder is much better than the ImageNet pretrained model, (b) CLIP Zero-shot accuracy is poor, (c) ImageNet pretrained model is much better than CLIP image encoder. VL2V-ADiP (Ours) is better than baselines in (a) and (b), and is comparable to baselines in (c) where the base model is better than the VLM teacher.



7. Ablation study on VL2V-ADiP

Method - Changes done w.r.t. VL2V-ADiP (Ours)	OH	TI	VLCS	PACS	Avg-ID	Avg-OOD
VL2V-ADiP (Ours)	85.7	55.4	81.9	94.9	92.7	79.5
A1: Combining "Align" and "Distill" stages	74.5	53.0	80.1	85.3	90.5	73.2
A2: Without freezing projection head in Stage-2	86.1	56.7	81.8	93.7	93.0	79.6
A3: Distilling only from Text encoder in Stages-1 and 2	83.2	47.0	79.8	90.9	91.9	75.2
A4: Distilling only from Image encoder in Stages-1 and 2	79.0	29.0	82.2	90.0	74.1	70.0
A5: Finetuning CLIP classifier-head in Stage-3 (CE loss) - CLIP init classifier	84.5	49.3	81.3	93.5	93.1	77.1
A6: Finetuning full network in Stage-3 (CE loss) - CLIP init classifier	83.9	49.8	80.1	92.3	93.4	76.5
A7: Finetuning classifier-head in Stage-3 (CE loss) - random init classifier	84.6	49.6	81.3	93.6	93.0	77.3
A8: Finetuning full network in Stage-3 (CE loss) - random init classifier	83.2	50.0	79.8	92.2	93.1	76.3

References: [1] Cha, Junbum, et al. "Domain generalization by mutual-information regularization with pre-trained models." ECCV '22. [2] Cha, Junbum, et al. "Swad: Domain generalization by seeking flat minima." NeurIPS '21. [3] Hinton et al. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).

Acknowledgements: This work was supported by the research grant CRG/2021/005925 from SERB, DST, Govt. of India. Sravanti Addepalli is supported by a Google PhD Fellowship in Machine Learning.