Leveraging Vision-Language Models for Improving Domain Generalization in Image Classification

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

This supplementary material presents further details on the proposed approach, datasets, and results. To ensure the reproducibility of our results, we share the code on our project page - https://val.cds.iisc.ac.in/VL2V-ADiP/. The supplementary material is structured as follows:

- **Section 1:** Training Algorithm
- **Section 2:** Details on Datasets
- **Section 3:** Additional Results
  - **Section 3.1:** Variance re-runs
  - **Section 3.2:** Comparison with Additional Baselines
  - **Section 3.3:** Distillation using diverse VLMs
  - **Section 3.4:** Domain-wise Results
- **Section 4:** Analysis on Loss Weighting

1. Training Algorithm

The detailed training algorithm of the proposed approach VL2V-ADiP is presented in Algorithm-1. We additionally incorporate SWAD [1] during training, which detects the onset of the optimal basin and performs weight-averaging across several model snapshots in the basin. To enable a fair comparison, we present results across all baselines as well using SWAD, denoted using "(S)" in Tables- 2, 3, and 4 of the main paper.

2. Details on Datasets

We evaluate the proposed approaches VL2V-SD and VL2V-ADiP on five Domain Generalization datasets that are widely used in literature and recommended on the DomainBed benchmark [6]. The details of these five datasets are presented in Table-1. This includes diverse datasets with several unique aspects such as - less training data [4, 11, 19] and a larger amount of training data [15], small domain-shifts [4] and larger domain shifts [15], lesser number of classes [4, 11] and a higher number of classes [15, 19]. We

---

Algorithm 1 VL2V - Align, Distill, Predict (ADiP)

Input: Let $\mathcal{D}_s = \{D_i,\ \forall i = 1, 2, \dots d-1\}$ be the data from $d-1$ source domains, $(x_i, y_i) \sim D_s$ be an image-label pair from source domains, x_i^{target} be a test image from the target domain, f_T^{text} and f_T^{img} be the text and image encoders of the VLM Teacher respectively, f_S^{fe} and f_S^{proj} be the feature extractor and linear projection layer of the student vision model respectively, h_{VLM} be the zero-shot classifier of the VLM teacher, and C be the set of all class names in dataset $\mathcal{D}_s$. For the data sample (x_i, y_i), let $\mathbf{I}^t_{x_i}$ and $\mathbf{T}_{y_i}$ be the image and text embeddings from the VLM teacher respectively, and $\mathbf{PF}^s_{x_i}$ be the projected features from the student.

$P_c =$ "A photo of a c" $\ \forall\, c \in C$

$\mathbf{T}_c = f_T^{text}(P_c)\ \ \forall\, c \in C$

Stage 1 - Align ▷ Projection layer trained

for *iter* $<$ *MaxIters* **do**:

  Sample batch (x_i, y_i) from $\mathcal{D}_s, \forall\ 0 \le i < n$

  $\mathbf{I}^t_{x_i} \leftarrow f_T^{img}(x_i), \forall\ 0 \le i < n$

  $\mathbf{PF}^s_{x_i} \leftarrow f_S^{proj}(f_S^{fe}(x_i)), \forall\ 0 \le i < n$

  $\mathcal{L} = -\frac{1}{2n} \sum_i \left\{ \cos(\mathbf{PF}^s_{x_i}, \mathbf{T}_{y_i}) + \cos(\mathbf{PF}^s_{x_i}, \mathbf{I}^t_{x_i}) \right\}$

  $\theta_{proj} \leftarrow \theta_{proj} - \nabla_{\theta_{proj}} \mathcal{L}$

end for

Stage 2 - Distill ▷ Feature extractor trained

for *iter* $<$ *MaxIters* **do**:

  Sample batch (x_i, y_i) from $\mathcal{D}_s, \forall\ 0 \le i < n$

  $\mathbf{I}^t_{x_i} \leftarrow f_T^{img}(x_i), \forall\ 0 \le i < n$

  $\mathbf{PF}^S_{x_i} \leftarrow f_S^{proj}(f_S^{fe}(x_i)), \forall\ 0 \le i < n$

  $\mathcal{L} = -\frac{1}{2n} \sum_i \left\{ \cos(\mathbf{PF}^s_{x_i}, \mathbf{T}_{y_i}) + \cos(\mathbf{PF}^s_{x_i}, \mathbf{I}^t_{x_i}) \right\}$

  $\theta_{fe} \leftarrow \theta_{fe} - \nabla_{\theta_{fe}} \mathcal{L}$

end for

Stage 3 - Predict

$h_{\text{VLM}}(\mathbf{x}) := [\ \cos(\mathbf{x}, \mathbf{T}_c),\ \forall\, c \in C\ ]$

$\mathbf{PF}^s_{x_i^{\text{target}}} \leftarrow f_S^{proj}(f_S^{fe}(x_i^{\text{target}}))$

$\hat{y}_i = \text{argmax}_c\ h_{\text{VLM}}(\mathbf{PF}^s_{x_i^{\text{target}}})$

---

[*] Equal Contribution.

Correspondence to Sravanti Addepalli <sravantia@iisc.ac.in>, Ashish Ramayee Asokan <ashish.ramayee@gmail.com>

Table 1. **Domain Generalization Datasets:** Details of the five DG datasets recommended by the DomainBed benchmark [6]

Dataset	No. of classes	No. of domains	No. of images	Domains	Domain shift
Office-Home (OH)	65	4	15,588	Art, Clipart, Product, Real	Style
Terra-Incognita (TI)	10	4	24,788	L100, L38, L43, L46	Camera location
VLCS	5	4	10,729	Caltech101, LabelMe, SUN09, VOC2007	Photography
PACS	7	4	9,991	Art, Cartoons, Photos, Sketches	Style
DomainNet (DN)	345	6	586,575	Clipart, Infograph, Painting, Quickdraw, Real, Sketch	Style

Table 2. **Variance across re-runs:** Mean and standard deviation of the OOD accuracy (%) of our proposed approach VL2V-ADiP when compared to the ERM and KD [7] baselines across the five DG datasets. All results are presented with SWAD (S) [1].

Method	Office-Home	Terra-Incognita	VLCS	PACS	DomainNet	Avg-OOD
ERM-FFT (S)	82.33 $\pm$ 0.87	48.87 $\pm$ 0.74	80.12 $\pm$ 0.30	90.15 $\pm$ 0.51	56.09 $\pm$ 0.09	71.51 $\pm$ 0.50
KD (S)	81.90 $\pm$ 0.78	48.90 $\pm$ 1.32	79.95 $\pm$ 0.49	90.70 $\pm$ 0.67	56.01 $\pm$ 0.10	71.49 $\pm$ 0.67
VL2V-ADiP (Ours)	**85.82** $\pm$ 0.27	**55.32** $\pm$ 0.74	**82.31** $\pm$ 0.37	**94.32** $\pm$ 0.56	**59.29** $\pm$ 0.11	**75.41** $\pm$ 0.41

compare against several baselines on each of these individual datasets and also report the average performance across all datasets as is the standard practice [6].

3. Additional Results

3.1. Variance re-runs

The results in Tables - 2, 3, 4, and 5 of the main paper are reported with a fixed seed of 0, in order to ensure reproducibility of results. In Table-2, we report the mean and standard deviation of the proposed method VL2V-ADiP across 3 re-runs with different random seeds. We additionally present standard deviation for the two standard baselines - ERM Fine-tuned (S) and KD (S) [7], for reference. We note that the standard deviation of the proposed method is comparable to the baselines on the respective datasets.

3.2. Comparison with Additional Baselines

We present additional baseline results corresponding to Tables - 2, 3, and 4 of the main paper in Tables-3, 4 and 5 respectively, for the sake of completeness. In Tables-3 and 4, we additionally present the respective baseline results without including SWAD [1] during training. In Table-5, we compare the performance of the proposed approach VL2V-ADiP on the OfficeHome dataset, with all the baselines considered in Table-2 of the main paper, on student models with different architectures. The proposed approaches show gains across baselines in all the tables.

3.3. Distillation using diverse VLMs

We demonstrate the compatibility of the proposed method VL2V-ADiP with diverse VLM teacher models in Table-6. Specifically, we show results by distilling from FLAVA [18], BLIP [12], and the data-efficient versions [13] of CLIP and FILIP [22]. We observe that our method achieves the highest gains over the KD baseline [7] with CLIP, where the teacher VLM has been trained with a large pre-training dataset. However, our method achieves significant gains even with VLMs pre-trained on smaller datasets.

Table 3. **SOTA comparison with CLIP initialization (extended comparisons to show results without integrating the baselines with SWAD):** Performance (%) of the proposed self-distillation approach VLV2-SD, compared to the SOTA DG methods. ViT-B/16 architecture is used with CLIP initialization. (S) denotes SWAD [1]

Method	OH	TI	VLCS	PACS	DN	Avg-ID	Avg-OOD
Zero-shot [16]	82.40	34.10	82.30	96.50	57.70	-	70.60
SWAD [1]	81.01	42.92	79.13	91.35	57.92	89.05	70.47
MIRO [2]	83.36	54.30	81.32	95.60	54.00	89.32	76.32
DART [9]	77.35	46.41	77.04	91.45	56.53	88.65	69.76
SAGM [20]	81.11	54.29	81.11	90.61	53.59	89.56	72.41
LP-FT [10]	69.72	36.04	77.10	86.28	49.00	84.72	67.14
FLYP [5]	75.25	40.22	75.89	92.97	48.90	84.66	69.65
CLIPood [17]	67.51	35.68	78.32	79.61	47.72	82.52	65.23
RISE [8]	70.28	40.15	81.18	91.65	50.81	85.21	66.81
VL2V-SD (Ours)	85.44	41.18	82.67	95.67	58.71	89.50	72.73
Combined with SWAD (S) [1]							
MIRO (S) [2]	84.80	**59.30**	82.30	96.44	60.47	**91.00**	76.66
DART (S) [9]	80.93	51.24	80.38	93.43	59.32	89.25	73.06
SAGM (S) [20]	83.40	58.64	82.05	94.31	59.05	89.74	75.49
LP-FT (S) [10]	81.17	47.26	80.88	92.92	57.04	88.97	71.85
FLYP (S) [5]	82.76	33.25	66.64	78.53	57.41	78.94	63.72
CLIPood (S) [17]	83.31	46.28	77.19	93.16	57.78	69.90	71.55
RISE (S) [8]	78.39	49.61	80.62	93.25	55.37	87.91	71.45
VL2V-SD (Ours)	**87.38**	58.54	**83.25**	**96.68**	**62.79**	89.99	**77.73**

3.4. Domain-wise Results

We present the results of the proposed approaches VL2V-SD and VL2V-ADiP on each of the individual domains in Table-7a and Table-7b. The domain in the column heading indicates the unseen test domain, where the training was done on the remaining $d-1$ domains mentioned in Table-1. We note that the proposed methods VL2V-SD and VL2V-ADiP outperform existing methods across several datasets

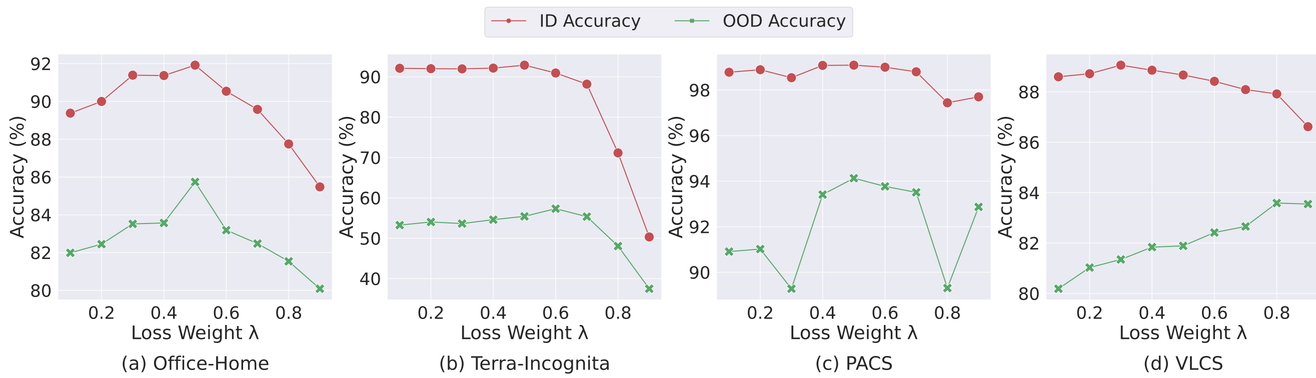

Figure 1. OOD and ID accuracy (%) of the proposed approach VL2V-ADiP across variation in loss weight λ for 4 Domain Generalization datasets. Cosine similarity of the student's projected features w.r.t. the text embeddings of the VLM teacher is given a weight of $(1-\lambda)$, while that w.r.t. the image embeddings of the VLM is given a weight of λ.

Table 4. **SOTA comparison with ImageNet-1K initialization (extended comparisons to show results without integrating the baselines with SWAD):** Performance (%) of the proposed approach VLV2-ADiP, compared to the SOTA DG methods. ViT-B/16 architecture is used with ImageNet-1K initialization. (S) denotes SWAD [1]

Method	OH	TI	VLCS	PACS	DN	Avg-ID	Avg-OOD
ERM-LP	71.45	31.80	77.77	67.52	36.66	74.25	57.04
ERM FFT	78.03	42.53	78.13	85.32	50.84	86.90	66.97
LP-FT [10]	75.23	44.05	76.51	85.08	51.10	87.59	66.40
SimKD [3]	76.89	26.32	80.16	85.66	48.45	68.31	64.30
KD [7]	77.62	38.66	79.73	84.87	50.73	87.04	66.32
MIRO [2]	74.88	44.52	80.39	81.53	49.95	86.56	66.25
DART [9]	82.56	50.70	79.70	89.76	56.13	89.99	71.77
SAGM [20]	80.87	52.38	79.53	87.29	54.04	88.88	73.83
Text2Concept [14]	70.57	26.86	79.03	66.10	23.29	53.22	53.17
RISE [8]	80.34	44.64	84.15	90.99	53.29	87.42	73.47
VL2V-ADiP (Ours)	84.56	49.99	81.53	93.41	56.82	88.74	73.26
Combined with SWAD (S) [1]							
ERM-LP (S)	71.48	31.35	77.52	67.02	36.65	73.99	56.81
ERM FFT (S)	83.22	50.05	80.33	90.28	56.10	89.31	72.00
LP-FT (S) [10]	81.55	51.61	80.17	91.20	56.03	**90.03**	72.11
SimKD (S) [3]	66.76	81.01	**83.92**	28.24	49.42	68.24	61.87
KD (S) [7]	82.73	48.40	80.48	91.46	56.11	89.20	71.84
MIRO (S) [2]	80.09	50.29	81.10	89.50	55.75	88.71	71.35
DART (S) [9]	83.75	49.68	77.29	90.55	58.05	88.54	71.86
SAGM (S) [20]	82.22	53.24	79.60	90.02	55.66	89.22	72.15
Text2Concept (S) [14]	70.24	26.46	64.77	79.03	23.26	53.15	52.82
RISE (S) [8]	83.48	52.55	83.70	93.54	56.58	88.91	73.97
VL2V-ADiP (Ours)	**85.74**	**55.43**	81.90	**94.94**	**59.38**	89.02	**75.48**

and domains.

VL2V-ADiP achieves the highest gains in cases where domain shift is large, highlighting the benefit of using the supervision from CLIP in improving OOD generalization on downstream tasks. The domains with the highest gains include ClipArt (OH), Location-38 (TI), Location-46 (TI), Cartoon (PACS), Infograph (DN), and Painting (OH). The domains with the least gains include Product (OH), Real-World (OH), Art (PACS), Photo (PACS), Quickdraw (DN), and all domains in VLCS. It is intuitive to see that most of the domains with the least gains are the cases where the target distribution is similar to at least one of the source distributions, making them less challenging to evaluate OOD robustness. For example, there is no real domain shift in

Table 5. **Distillation to lower capacity student models:** Performance (%) of the proposed approach VL2V-ADiP when compared to existing SOTA DG methods (rows), with different architectures of the student model (columns) on OfficeHome dataset. The teacher architecture is ViT-B/16. (S) denotes SWAD.

Method	ViT-B/16	ViT-S/16	DeiT-S/16	ResNet-50	Avg.
ERM-LP (S)	71.48	68.47	74.12	68.46	70.63
ERM-FFT (S)	83.22	78.58	74.95	70.85	76.90
LP-FT (S) [10]	81.55	78.77	74.41	70.39	76.28
SimKD (S) [3]	66.76	54.18	58.75	60.88	60.14
KD (S) [7]	82.73	78.14	74.65	70.67	76.55
MIRO (S) [2]	80.09	69.45	73.18	72.40	73.78
DART (S) [9]	83.75	79.67	75.85	71.90	77.79
SAGM (S) [20]	82.22	77.00	73.94	70.10	75.81
Text2Concept (S) [14]	70.24	63.30	66.27	61.89	65.42
RISE (S) [8]	83.48	80.47	76.09	72.40	78.11
VL2V-ADiP (Ours)	**85.74**	**81.22**	**77.63**	**74.42**	**79.75**

Table 6. **Distillation using various VLMs:** Performance (%) of the proposed approach VL2V-ADiP (denoted as Ours) on 4 DG datasets, when distilling from FLAVA [18], BLIP [12], CLIP [16] and the data-efficient versions [13] of CLIP and FILIP [22]. The student architecture is ViT-B/16 in all cases.

Teacher	Dataset	Method	OH	VLCS	PACS	TI	Avg-OOD
FLAVA ViT-B/16	PMD 70M	Zero-shot	69.99	79.21	91.34	28.85	67.35
		KD (S)	82.50	80.41	90.71	50.86	76.12
		Ours	**84.16**	**82.94**	**93.22**	**54.56**	**78.72**
BLIP ViT-B/16	CapFilt 129M	Zero-shot	84.83	71.60	92.23	29.75	69.60
		KD (S)	82.45	80.31	87.73	48.03	74.63
		Ours	**85.86**	**81.60**	**94.10**	**52.07**	**78.41**
CLIP ViT-B/16	CLIP 400M	Zero-shot	81.57	82.55	95.99	31.15	72.81
		KD (S)	82.73	80.48	91.49	48.33	75.76
		Ours	**85.74**	**81.89**	**94.13**	**55.43**	**79.30**
DeCLIP ViT-B/32	YFCC 15M	Zero-shot	43.46	77.79	83.69	27.70	58.16
		KD (S)	81.84	79.95	89.96	49.49	75.31
		Ours	**82.85**	**81.40**	**92.16**	**50.50**	**76.73**
DeFILIP ViT-B/32	YFCC 15M	Zero-shot	46.97	74.08	82.02	16.34	54.85
		KD (S)	82.14	79.53	90.68	50.96	75.83
		Ours	**83.11**	**81.43**	**92.03**	**51.69**	**77.06**

VLCS, apart from the fact that each split is obtained from a different dataset, with a possible domain shift due to

Table 7. **Domain-wise performance** (%) of the proposed approaches VL2V-SD and VL2V-ADiP when compared to the respective baselines, on individual domains of all Domain Generalization datasets on the DomainBed benchmark [6].

(a) Domain-wise OOD accuracy for the approach VL2V-SD compared to the baselines combined with SWAD (S) [1] for the white box setting.

Method / Dataset	Domains					
OfficeHome	*Art*	*Clipart*	*Product*	*Real*	*Avg.*	
ERM Full Fine-Tuning (S)	80.12	70.25	86.18	87.49	81.01	
MIRO (S) [2]	83.57	75.72	89.70	90.22	84.80	
DART (S) [9]	78.79	72.71	86.04	86.17	80.93	
SAGM (S) [20]	82.60	72.94	88.94	89.13	83.40	
LP-FT (S) [10]	80.18	71.94	86.35	86.23	81.17	
CLIPood (S) [17]	84.86	70.93	88.09	89.39	83.31	
RISE (S) [8]	75.08	69.16	84.35	84.97	78.39	
WiSE-FT [21]	85.15	76.17	**92.90**	91.04	86.32	
VL2V-SD (Ours)	**87.33**	**78.55**	91.98	**91.65**	**87.38**	
TerraIncognita	*L100*	*L38*	*L43*	*L46*	*Avg.*	
ERM Full Fine-Tuning (S)	38.94	38.03	54.03	40.68	42.92	
MIRO (S) [2]	67.15	50.75	**66.63**	52.71	**59.30**	
DART (S) [9]	61.09	39.48	58.97	45.42	51.24	
SAGM (S) [20]	**72.21**	50.10	62.50	49.73	58.64	
LP-FT (S) [10]	54.23	38.09	56.52	40.20	47.26	
CLIPood (S) [17]	47.32	38.12	55.73	43.94	46.28	
RISE (S) [8]	60.30	37.43	56.77	43.94	49.61	
WiSE-FT [21]	56.75	**51.93**	61.71	47.62	54.50	
VL2V-SD (Ours)	69.10	48.40	63.10	**53.56**	58.54	
VLCS	*Caltech*	*LabelMe*	*Pascal*	*Sun*	*Avg.*	
ERM Full Fine-Tuning (S)	99.12	63.31	79.01	75.08	79.13	
MIRO (S) [2]	97.53	66.59	81.57	**83.53**	82.30	
DART (S) [9]	99.12	65.73	81.07	75.63	80.38	
SAGM (S) [20]	97.73	65.86	83.77	80.85	82.05	
LP-FT (S) [10]	98.24	65.66	81.23	78.40	80.88	
CLIPood (S) [17]	82.22	66.47	84.19	75.90	77.19	
RISE (S) [8]	**99.50**	67.27	81.44	74.27	80.62	
WiSE-FT [21]	98.99	66.04	83.47	83.01	82.88	
VL2V-SD (Ours)	99.24	**67.81**	**86.89**	79.05	**83.25**	
PACS	*Art*	*Cartoon*	*Photo*	*Sketch*	*Avg.*	
ERM Full Fine-Tuning (S)	91.34	89.07	97.53	87.47	91.35	
MIRO (S) [2]	98.05	97.50	99.78	90.46	96.44	
DART (S) [9]	94.45	92.27	98.80	88.20	93.43	
SAGM (S) [20]	95.18	93.60	99.03	89.41	94.31	
LP-FT (S) [10]	91.46	92.59	99.10	88.52	92.92	
CLIPood (S) [17]	92.68	91.05	98.95	89.98	93.16	
RISE (S) [8]	92.25	93.82	98.65	88.26	93.25	
WiSE-FT [21]	**98.29**	**98.50**	**100.00**	**92.36**	**97.29**	
VL2V-SD (Ours)	98.05	98.19	99.93	90.55	96.68	
DomainNet	*clp*	*inf*	*pnt*	*qkdr*	*real*	*skt*	*Avg.*
ERM Full Fine-Tuning (S)	77.10	38.32	66.13	25.02	75.19	65.76	57.92
MIRO (S) [2]	79.70	43.50	67.36	24.62	79.22	68.42	60.47
DART (S) [9]	78.51	39.99	66.89	**25.85**	76.37	68.29	59.32
SAGM (S) [20]	78.78	40.21	67.31	24.18	76.29	67.54	59.05
LP-FT (S) [10]	77.37	33.88	65.27	24.82	74.60	66.32	57.04
CLIPood (S) [17]	76.28	38.46	66.98	21.76	75.79	67.43	57.78
RISE (S) [8]	77.80	31.32	57.64	24.60	73.96	66.90	55.37
WiSE-FT [21]	72.74	46.36	64.05	16.79	**82.82**	65.29	58.01
VL2V-SD (Ours)	**79.96**	**49.00**	**71.05**	23.34	82.05	**71.36**	**62.79**

(b) Domain-wise OOD accuracy for the approach VL2V-ADiP compared to the baselines combined with SWAD (S) [1] for the black box setting.

Method / Dataset	Domains					
OfficeHome	*Art*	*Clipart*	*Product*	*Real*	*Avg.*	
ERM Full Fine-Tuning (S)	82.24	72.19	88.43	90.02	83.22	
LP-FT (S) [10]	81.67	65.64	89.22	89.67	81.55	
KD (S) [7]	80.79	70.76	89.08	90.30	82.73	
MIRO (S) [2]	78.89	64.15	87.70	89.62	80.09	
DART (S) [9]	81.72	73.17	89.64	90.48	83.75	
SAGM (S) [20]	80.23	70.16	88.46	90.02	82.22	
Text2Concept (S) [14]	71.06	48.97	77.48	83.45	70.24	
RISE (S) [8]	81.87	72.42	89.27	90.33	83.48	
VL2V-ADiP (Ours)	**84.81**	**75.92**	**90.65**	**91.60**	**85.74**	
TerraIncognita	*L100*	*L38*	*L43*	*L46*	*Avg.*	
ERM Full Fine-Tuning (S)	58.98	37.76	58.31	45.15	50.05	
LP-FT (S) [10]	58.29	41.08	**63.22**	43.83	51.61	
KD (S) [7]	61.09	33.21	57.84	41.17	48.33	
MIRO (S) [2]	61.27	38.14	57.68	44.08	50.29	
DART (S) [9]	56.82	37.34	62.31	42.26	49.68	
SAGM (S) [20]	**64.17**	44.42	59.64	44.74	53.24	
Text2Concept (S) [14]	43.55	2.04	31.83	28.43	26.46	
RISE (S) [8]	59.77	43.79	59.45	47.19	52.55	
VL2V-ADiP (Ours)	62.93	**44.83**	60.71	**53.26**	**55.43**	
VLCS	*Caltech*	*LabelMe*	*Pascal*	*Sun*	*Avg.*	
ERM Full Fine-Tuning (S)	98.49	64.05	82.60	76.17	80.33	
LP-FT (S) [10]	96.97	63.58	82.02	78.13	80.17	
KD (S) [7]	98.87	65.32	81.33	76.39	80.48	
MIRO (S) [2]	99.75	64.79	82.66	77.20	81.10	
DART (S) [9]	94.08	63.11	76.12	75.86	77.29	
SAGM (S) [20]	98.49	64.92	79.32	75.68	79.60	
Text2Concept (S) [20]	98.36	68.08	77.21	72.47	79.03	
RISE (S) [8]	**100.00**	**69.15**	**84.03**	**81.61**	**83.70**	
VL2V-ADiP (Ours)	99.62	66.60	82.87	78.46	81.89	
PACS	*Art*	*Cartoon*	*Photo*	*Sketch*	*Avg.*	
ERM Full Fine-Tuning (S)	93.78	86.25	99.18	81.93	90.28	
LP-FT (S) [10]	94.27	86.83	99.48	84.22	91.20	
KD (S) [7]	94.20	86.35	99.25	86.04	91.46	
MIRO (S) [2]	94.69	85.98	99.63	77.70	89.50	
DART (S) [9]	94.45	86.67	99.55	81.52	90.55	
SAGM (S) [20]	93.72	86.57	99.18	80.63	90.02	
Text2Concept (S) [14]	80.17	66.47	96.63	15.81	64.77	
RISE (S) [8]	93.72	**93.23**	99.55	87.66	93.54	
VL2V-ADiP (Ours)	**95.61**	92.38	**99.85**	**88.68**	**94.13**	
DomainNet	*clp*	*inf*	*pnt*	*qkdr*	*real*	*skt*	*Avg.*
ERM Full Fine-Tuning (S)	76.34	30.92	64.76	21.30	77.70	65.60	56.10
LP-FT (S) [10]	76.49	30.75	65.30	20.82	77.83	64.98	56.03
KD (S) [7]	76.56	31.29	64.55	**21.44**	77.62	65.23	56.11
MIRO (S) [2]	76.32	30.96	64.52	20.18	77.88	64.63	55.75
DART (S) [9]	77.62	34.14	67.64	21.05	80.73	67.11	58.05
SAGM (S) [20]	76.67	29.85	64.42	20.68	77.58	64.58	55.63
Text2Concept (S) [14]	22.41	10.14	35.26	0.47	56.55	14.61	23.26
RISE (S) [8]	77.87	32.71	61.03	21.20	79.90	66.77	56.58
VL2V-ADiP (Ours)	**78.80**	**36.86**	**69.21**	21.32	**81.33**	**68.79**	**59.38**

photography differences, which can be considered minor. Hence, taking the supervision of a CLIP model is the least beneficial here.

4. Analysis on Loss Weighting

The training loss of the proposed approach VL2V-ADiP presented in Eq. 6 of the main paper, and in L8 and L15 of Algorithm-1, considers equal weights on both loss terms - cosine similarity of the image embeddings $\mathbf{PF}^s_{x_i}$ w.r.t. text and image embeddings of the VLM teacher respectively. In this section, we explore the impact of varying these weights as a convex interpolation between the cosine similarity w.r.t. text embeddings (weighted by $1-\lambda$) and image embeddings (weighted by λ) respectively as shown below:

$$\mathcal{L} = -\frac{1}{2n}\sum_{i=1}^{n}\left\{(1-\lambda)\cdot\cos(\mathbf{PF}^s_{x_i}, \mathbf{T}_{y_i}) + \lambda\cdot\cos(\mathbf{PF}^s_{x_i}, \mathbf{I}^t_{x_i})\right\} \tag{1}$$

We note from the plots in Fig.1 that while the best OOD accuracy could be achieved at a different λ value, a setting of 0.5 works reasonably well, since the proposed approach is not too sensitive to variations in λ in most cases. Moreover, a value of 0.5 assigns equal weightage to losses w.r.t. both image and text embeddings (since they are of the same scale), which is the best setting to consider in the absence of hyperparameter tuning.

References

[1] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *NeurIPS*, 2021. 1, 2, 3, 4

[2] Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. Domain generalization by mutual-information regularization with pre-trained models. *ECCV*, 2022. 2, 3, 4

[3] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. Knowledge distillation with the reused teacher classifier. In *CVPR*, 2022. 3

[4] Chen Fang, Ye Xu, and Daniel N. Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *ICCV*, 2013. 1

[5] Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *CVPR*, 2023. 2

[6] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021. 1, 2, 4

[7] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 3, 4

[8] Zeyi Huang, Andy Zhou, Zijian Ling, Mu Cai, Haohan Wang, and Yong Jae Lee. A sentence speaks a thousand images: Domain generalization through distilling clip with language guidance. In *ICCV*, 2023. 2, 3, 4

[9] Samyak Jain, Sravanti Addepalli, Pawan Kumar Sahu, Priyam Dey, and R Venkatesh Babu. Dart: Diversify-aggregate-repeat training improves generalization of neural networks. In *CVPR*, 2023. 2, 3, 4

[10] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *ICLR*, 2022. 2, 3, 4

[11] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017. 1

[12] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*. PMLR, 2022. 2, 3

[13] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*, 2022. 2, 3

[14] Mazda Moayeri, Keivan Rezaei, Maziar Sanjabi, and Soheil Feizi. Text-to-concept (and back) via cross-model alignment. *PMLR*, 2023. 3, 4

[15] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019. 1

[16] 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*. PMLR, 2021. 2, 3

[17] Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. Clipood: Generalizing clip to out-of-distributions. *PMLR*, 2023. 2, 4

[18] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, 2022. 2, 3

[19] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, 2017. 1

[20] Pengfei Wang, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. Sharpness-aware gradient matching for domain generalization. In *CVPR*, 2023. 2, 3, 4

[21] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *CVPR*, 2022. 4

[22] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. In *ICLR*, 2022. 2, 3