IMA & TMA: Efficient Test-Time Adaptation for VLMs via Linear Transformation in Embedding Space

Rishik Vamshi Rohith Vempati* Eswar Venkata Sai Kadava* Konda Reddy Mopuri
Department of Artificial Intelligence, Indian Institute of Technology Hyderabad
{ai24mtech12003@iith.ac.in, ai24mtech11007@iith.ac.in, krmopuri@ai.iith.ac.in}

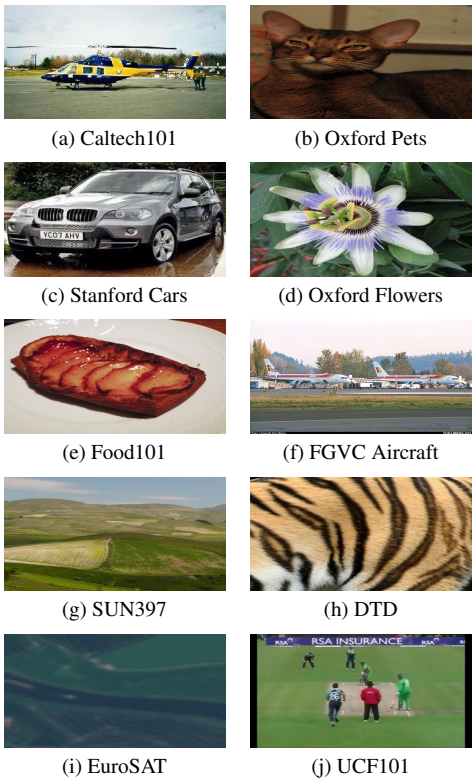
Supplementary Overview

Contents

*Equal contribution

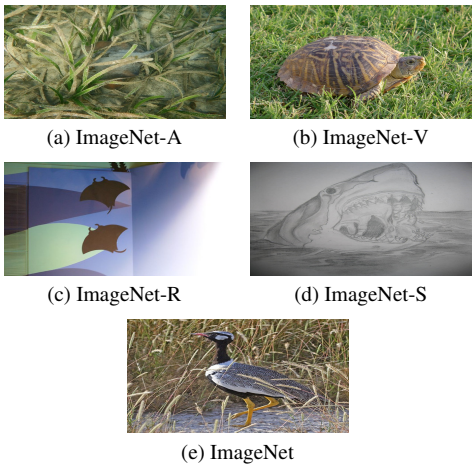
Supplementary Material

A. Overview of Benchmark Details



(a) Caltech101 (b) Oxford Pets (c) Stanford Cars (d) Oxford Flowers (e) Food101 (f) FGVC Aircraft (g) SUN397 (h) DTD (i) EuroSAT (j) UCF101

Figure 1. Visualization on Fine-grained classification datasets.



(a) ImageNet-A (b) ImageNet-V (c) ImageNet-R (d) ImageNet-S (e) ImageNet

Figure 2. Visualization on ImageNet and OOD variants.

A.1. Fine-grained Classification Datasets

To further evaluate cross-dataset generalization, we conduct experiments on ten diverse and publicly available image classification benchmarks spanning a wide range of visual domains. These include fine-grained recognition tasks such as Flowers102 [13] (102 classes, 2,463 test images) and OxfordPets [14] (37 classes, 3,669 test images), transportation categories including StanfordCars [11] (196 classes, 8,041 test images) and FGVC-Aircraft [12] (100 classes, 3,333 test images), and scene understanding with SUN397 [21] (397 classes, 19,850 test images). We also evaluate texture recognition on DTD [2] (47 classes, 1,692 test images), food classification on Food101 [1] (101 classes, 30,300 test images), human action recognition on UCF101 [18] (101 classes, 3,783 test images), satellite imagery classification on EuroSAT [7] (10 classes, 8,100 test images), and general object categorization using Caltech101 [6] (100 classes, 2,465 test images). Together, these datasets vary substantially in granularity, visual complexity, and semantic structure, providing a comprehensive testbed for evaluating the ability of our approach to adapt across heterogeneous domains.

A.2. ImageNet and its OOD variants

We assess robustness to natural distribution shifts using four ImageNet variants that are commonly treated as out-of-distribution (OOD) benchmarks with respect to the original ImageNet [3] dataset. The standard ImageNet validation set contains 1,000 classes and 50,000 test images. These OOD benchmarks provide a standardized and realistic setting for measuring performance degradation under domain changes.

ImageNet-V2 [15] contains 1,000 classes and 10,000 test images and is an independently curated dataset collected from sources distinct from the original ImageNet distribution to capture natural distribution shifts. ImageNet-A [9] consists of 200 classes and 7,500 naturally occurring adversarial images, designed to expose model failures under challenging yet realistic visual conditions. ImageNet-R [8] includes 200 classes and approximately 30,000 artistic and non-photographic renditions of ImageNet categories, such as paintings and sketches, evaluating robustness to significant appearance variations. Finally, ImageNet-Sketch [20] contains 1,000 classes and 50,889 black-and-white sketch images, testing on shape-driven representations.

B. Detailed Algorithmic Descriptions

In this section, we provide detailed algorithmic descriptions of the proposed Image Matrix Adapter (IMA) and Text Matrix Adapter (TMA) methods. The procedures outline the key computational steps involved in performing embedding-space transformations during test-time adaptation.

Algorithm 1: Image Matrix Adapter (IMA)

Input: Input sample x_0
Pre-trained frozen image encoder f_v
Pre-computed textual prototypes $\{t_{c_i} \in R^d\}_{i=1}^K$
Set of augmentations $\mathcal{A}$
Number of additional augmentations $(N-1)$
Aggregation Strategy mode (Filtered/All)
AdamW optimizer Opt

Output: Predicted class label c_i from the K classes

function ADAPT (x_0, f_v, $\{t_{c_i}\}_{i=1}^K$, $\mathcal{A}$, N, Opt, mode)
  Sample $x_1, x_2, ..., x_{N-1} \in \mathcal{U}(\mathcal{A})$
  $v_j = f_v(x_j) \in R^d$, for j = 0 to N - 1
  Compute $p(c_i \mid (x_j, t_{c_i})) = \dfrac{\exp((v_j^T t_{c_i})/\tau)}{\sum_{k=1}^K \exp((v_j^T t_{c_k})/\tau)}$ for $j \in \{0, 1, \ldots, N-1\}$
  $H(x_j) = -\sum_{i=1}^K p(c_i \mid (x_j, t_{c_i})) \log(p(c_i \mid (x_j, t_{c_i})))$ for $j \in \{0, 1, \ldots, N-1\}$
  Initialize $W_v \leftarrow \mathbf{I_{d \times d}}$
  if mode = *Filtered* **then**
    $S = \{x_j : \mathbb{1}\{H(x_j) \le \tau\}\}$ where,
    $\tau = \text{Percentile}_\rho(\{H(x_j)\}_{j=0}^{N-1})$
    The set $\{x'_j\}_{j=0}^{s-1} \equiv S$ denote Filtered Augmentations
    $v'_j = f_v(x'_j)$ for $j \in \{0, 1, \ldots, s-1\}$
    Compute
    $\tilde{p}(c_i \mid (x_0, W_v), t_{c_i}) = \dfrac{1}{s}\sum_{j=0}^{s-1} p(c_i \mid (x'_j, W_v), t_{c_i})$ where
    $p(c_i \mid (x'_j, W_v), t_{c_i}) = \dfrac{\exp(sim((W_v v'_j), t_{c_i})/\tau)}{\sum_{k=1}^K \exp(sim((W_v v'_j), t_{c_k})/\tau)}$
    Compute $\mathcal{L}$ by eq:3 (in main) using $\tilde{p}$
  end
  else if mode = *All* **then**
    Compute α_j by eq:13 (in main), for $j \in \{0, 1, \ldots, N-1\}$
    Compute $\mathcal{L}$ by eq:11 (in main), where $W = W_v$
  end
  Compute $\partial\mathcal{L}$
  Update $W_v := W_v \ - \text{Opt}(\partial\mathcal{L})$
  Return W_v
end

function INFERENCE (x_0, f_v, $\{t_{c_i}\}_{i=1}^K$, $\mathcal{A}$, $\mathcal{N}$, Opt, mode)
  $v_0 = f_v(x_0)$
  $W_v^* = \text{ADAPT}(x_0, f_v, \{t_{c_i}\}_{i=1}^K, \mathcal{A}, \mathcal{N}, \text{Opt}, \text{mode})$
  $v_0^{\text{new}} = \frac{W_v^* v_0}{\|W_v^* v_0\|_2}$
  Compute $p(c_i \mid (x_0, W_v^*), t_{c_i}) = \dfrac{\exp((v_0^{\text{new}})^T t_{c_i})}{\sum_{k=1}^K \exp((v_0^{\text{new}})^T t_{c_k})}$ for $i \in \{1, 2, \ldots, K\}$
  Return $\arg\max_{c_i} p(c_i \mid (x_0, W_v^*), t_{c_i})$
end

Algorithm 2: Text Matrix Adapter (TMA)

Input: Input sample x_0
Pre-trained frozen image encoder f_v
Pre-computed textual prototypes $\{t_{c_i} \in R^d\}_{i=1}^K$
Set of augmentations $\mathcal{A}$
Number of additional augmentations $(N-1)$
Aggregation Strategy mode (Filtered/All)
AdamW optimizer Opt

Output: Predicted class label c_i from the K classes

function ADAPT (x_0, f_v, $\{t_{c_i}\}_{i=1}^K$, $\mathcal{A}$, N, Opt, mode)
  Sample $x_1, x_2, ..., x_{N-1} \in \mathcal{U}(\mathcal{A})$
  $v_j = f_v(x_j) \in R^d$, for j = 0 to N - 1
  Compute $p(c_i \mid (x_j, t_{c_i})) = \dfrac{\exp((v_j^T t_{c_i})/\tau)}{\sum_{k=1}^K \exp((v_j^T t_{c_k})/\tau)}$ for $j \in \{0, 1, \ldots, N-1\}$
  $H(x_j) = -\sum_{i=1}^K p(c_i \mid (x_j, t_{c_i})) \log(p(c_i \mid (x_j, t_{c_i})))$ for $j \in \{0, 1, \ldots, N-1\}$
  Initialize $W_t \leftarrow \mathbf{I_{d \times d}}$
  if mode = *Filtered* **then**
    $S = \{x_j : \mathbb{1}\{H(x_j) \le \tau\}\}$ where,
    $\tau = \text{Percentile}_\rho(\{H(x_j)\}_{j=0}^{N-1})$
    The set $\{x'_j\}_{j=0}^{s-1} \equiv S$ denote Filtered Augmentations
    $v'_j = f_v(x'_j)$ for $j \in \{0, 1, \ldots, s-1\}$
    Compute
    $\tilde{p}(c_i \mid x_0, (t_{c_i}, W_t)) = \dfrac{1}{s}\sum_{j=0}^{s-1} p(c_i \mid x'_j, (t_{c_i}, W_t))$ where
    $p(c_i \mid x'_j, (t_{c_i}, W_t)) = \dfrac{\exp(sim(v'_j, (W_t t_{c_i}))/\tau)}{\sum_{k=1}^K \exp(sim(v'_j, (W_t t_{c_k}))/\tau))}$
    $\mathcal{L} = -\sum_{i=1}^K \tilde{p}(c_i \mid x_0, (t_{c_i}, W_t)) \log(\tilde{p}(c_i \mid x_0, (t_{c_i}, W_t)))$
  end
  else if mode = *All* **then**
    Compute α_j by eq:13 (in main), for $j \in \{0, 1, \ldots, N-1\}$
    Compute $\mathcal{L}$ by eq:11 (in main), where $W = W_t$
  end
  Compute $\partial\mathcal{L}$
  Update $W_t := W_t \ - \text{Opt}(\partial\mathcal{L})$
  Return W_t
end

function INFERENCE (x_0, f_v, $\{t_{c_i}\}_{i=1}^K$, $\mathcal{A}$, $\mathcal{N}$, Opt, mode)
  $v_0 = f_v(x_0)$
  $W_t^* = \text{ADAPT}(x_0, f_v, \{t_{c_i}\}_{i=1}^K, \mathcal{A}, \mathcal{N}, \text{Opt}, \text{mode})$
  $t_{c_i}^{\text{new}} = \frac{W_t^* t_{c_i}}{\|W_t^* t_{c_i}\|_2}$, for $i \in \{1, 2, \ldots, K\}$
  Compute $p(c_i \mid x_0, (t_{c_i}, W_t^*)) = \dfrac{\exp((v_0^T t_{c_i}^{new})}{\sum_{k=1}^K \exp((v_0^T t_{c_k}^{new})}$ for $i \in \{1, 2, \ldots, K\}$
  Return $\arg\max_{c_i} p(c_i \mid x_0, (t_{c_i}, W_t^*))$
end

C. Additional Experimental Results

This section contains results other episodic approaches such as ZERO, MTA and RLCF using our default setting (ViT-B/32 backbone). Also, results are reported using baselines in Section 4 of main paper with CLIP ViT-B/16 and RN50.

APPROACH	TPT	TTL	TPS	ZERO [5]	MTA [23]	RLCF [24]	TMA (FA)	IMA (FA)	TMA (AA)	IMA (AA)
ImageNet [17]	63.44	64.27	64.25	**65.31**	64.79	63.64	64.26	64.13	64.50	64.27
ImageNet OOD [22]	49.88	51.91	51.39	50.88	51.60	51.14	**52.16**	52.12	51.78	51.75

Table 1. Results on cross-dataset benchmarks with ViT-B/32 backbone. FA and AA denote Filtered and All Augmentation Strategies. We report Acc@1 (in %). The best result across all methods is shown in bold.

APPROACH	Caltech101	Pets	Cars	Flowers	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Average
CLIP-**ViT-B/16**	93.31	88.25	65.33	67.40	83.64	23.91	63.05	44.39	42.22	65.24	63.68
Existing Back-propagation based TTA approaches											
TPT [17]	**94.04**	87.71	66.48	**69.47**	**84.46**	24.00	**65.19**	**46.04**	42.37	67.27	**64.70**
C-TPT [22]	93.63	**88.83**	65.96	69.18	83.92	24.03	64.42	45.45	39.65	66.03	64.11
R-TPT [16]	89.70	86.81	62.87	67.19	80.46	24.39	63.76	42.85	32.37	61.93	61.23
TTL [10]	93.83	87.74	**66.78**	67.24	84.08	**25.11**	65.14	45.15	**42.73**	**67.51**	64.53
TPS [19]	94.00	87.14	67.03	67.64	84.25	24.27	64.64	45.69	43.54	66.69	64.49
Ours											
TMA (FA)	93.47	86.35	66.26	66.75	83.20	24.60	64.77	44.62	40.56	66.27	<u>63.69</u>
IMA (FA)	93.35	86.29	65.69	66.50	82.84	24.54	64.53	44.62	39.58	66.27	63.42
TMA (AA)	93.43	87.08	66.47	66.54	83.55	**25.11**	65.17	44.27	38.01	66.40	63.60
IMA (AA)	93.23	86.86	65.86	66.30	83.26	24.75	65.17	44.03	36.05	66.48	63.20

Table 2. Results on cross-dataset benchmarks with ViT-B/16 backbone. FA and AA denote Filtered and All Augmentation Strategies. We report Acc@1 (in %). The best result across all methods is shown in bold, while the best result of our method is underlined.

APPROACH	Caltech101	Pets	Cars	Flowers	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Average
CLIP-**RN50**	85.68	83.62	55.29	61.67	73.96	15.69	59.25	40.43	23.69	58.90	55.81
Existing Back-propagation based TTA approaches											
TPT [17]	86.82	**84.71**	57.11	**62.48**	**74.93**	16.20	**60.87**	**41.55**	24.00	**60.64**	**56.93**
C-TPT [22]	86.61	83.43	55.47	62.44	74.29	16.68	60.56	41.25	23.19	59.34	56.33
R-TPT [16]	71.08	82.99	50.54	59.40	66.95	15.96	55.67	37.71	19.69	50.30	51.03
TPS [19]	**87.22**	84.19	57.16	61.75	74.54	17.22	60.42	40.84	**26.00**	59.82	56.92
Ours											
TMA (FA)	86.41	83.40	57.28	59.44	71.89	16.62	59.92	40.19	23.17	58.97	<u>55.73</u>
IMA (FA)	86.21	83.43	56.87	59.07	71.60	15.87	59.50	39.30	22.60	58.79	55.32
TMA (AA)	85.35	83.97	**57.82**	59.72	72.32	**17.79**	60.49	40.96	16.07	59.37	55.39
IMA (AA)	85.07	83.81	57.18	59.32	71.80	17.10	60.16	40.48	15.74	59.00	54.97

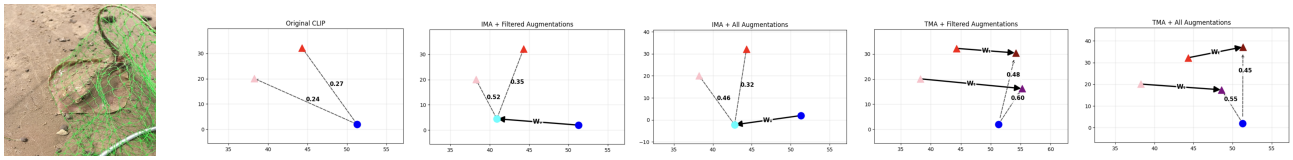
Table 3. Results on cross-dataset benchmarks with ResNet-50 backbone. FA and AA denote Filtered and All Augmentation Strategies. We report Acc@1 (in %). The best result across all methods is shown in bold, while the best result of our method is underlined.

Method	ImageNet-A	ImageNet-V	ImageNet-R	ImageNet-S	OOD Average
CLIP-**ViT-B/16**	47.80	60.84	73.99	46.15	57.20
Existing Back-propagation based TTA approaches					
TPT [17]	52.89	62.57	76.92	47.36	59.94
C-TPT [22]	50.55	62.43	75.67	47.19	58.96
R-TPT [16]	47.71	60.59	74.14	42.75	56.30
TTL [10]	55.21	62.97	**77.24**	47.56	**60.75**
TPS [19]	55.23	62.88	76.61	**47.66**	60.60
Ours					
TMA (FA)	56.47	62.77	76.53	47.24	**60.75**
IMA (FA)	**56.55**	62.62	76.46	47.19	60.71
TMA (AA)	54.57	**63.10**	76.22	47.51	60.35
IMA (AA)	54.41	63.04	76.22	47.32	60.25

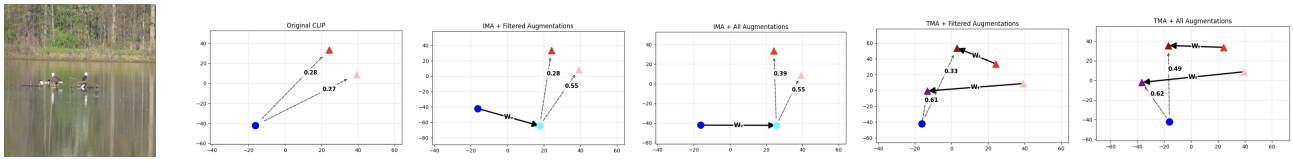
Table 4. Results on ImageNet-OOD datasets using the ViT-B/16 backbone. We report Acc@1 (in %). FA and AA denote the Filtered and All augmentation strategies, respectively. The best result across all methods is shown in bold.

Method	ImageNet-A	ImageNet-V	ImageNet-R	ImageNet-S	OOD Average
CLIP-**RN50**	21.84	51.52	56.09	33.34	40.70
Existing Back-propagation based TTA approaches					
TPT [17]	25.54	52.61	**58.93**	**35.17**	43.06
C-TPT [22]	23.96	**54.14**	56.67	34.68	42.36
R-TPT [16]	24.35	54.12	57.73	33.87	42.52
TPS [19]	**27.18**	52.84	57.34	34.92	**43.07**
Ours					
TMA (FA)	26.40	53.14	57.25	33.56	42.64
IMA (FA)	26.29	52.68	56.78	32.91	42.17
TMA (AA)	25.39	53.36	57.54	34.28	<u>42.65</u>
IMA (AA)	25.65	52.97	56.93	34.00	42.39

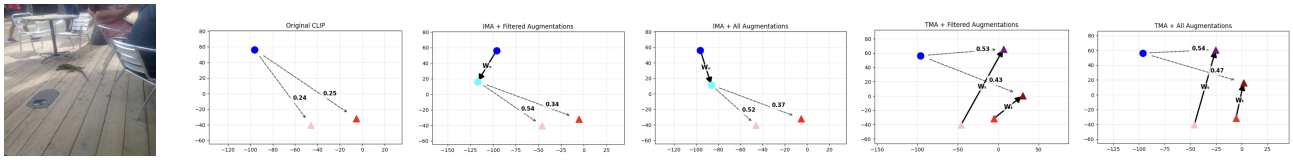
Table 5. Results on ImageNet-OOD datasets using the RN50 backbone. We report Acc@1 (in %). FA and AA denote the Filtered and All augmentation strategies, respectively. The best result across all methods is shown in bold, while the best result of our method is underlined.



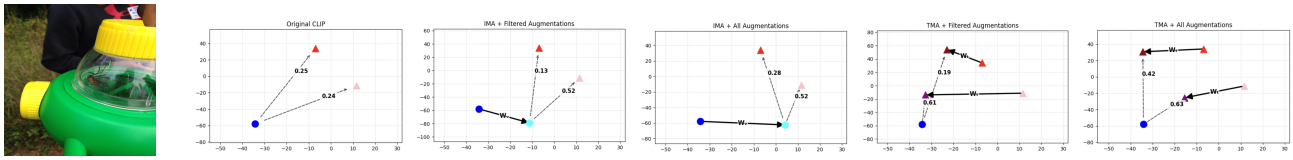
Ground Truth : **stingray**; CLIP Zero-Shot Prediction : hermit crab



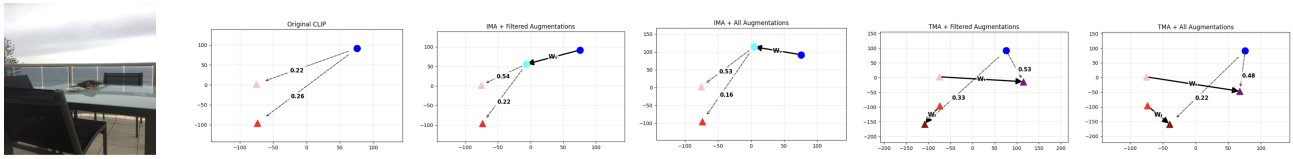
Ground Truth : **bald eagle**; CLIP Zero-Shot Prediction : goose



Ground Truth : **green iguana**; CLIP Zero-Shot Prediction : american bullfrog



Ground Truth : **harvestman**; CLIP Zero-Shot Prediction : rhinoceros beetle



Ground Truth : **lorikeet**; CLIP Zero-Shot Prediction : wine bottle

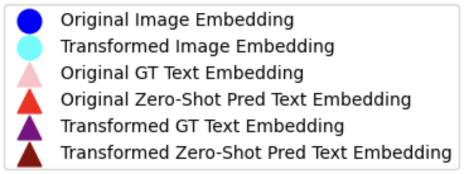


Figure 3. Visualization of embedding transformations under different adaptation strategies. In each row, the leftmost image corresponds to the raw test sample from the ImageNet-A [9] dataset. All subsequent plots visualize the corresponding image embeddings and textual prototypes projected onto a 2D space using t-SNE for interpretability. The second plot shows the original zero-shot embedding configuration of CLIP ViT-B/32 [4], where it predicts incorrectly i.e., it is giving lesser similarity score to ground truth class label. The remaining four plots illustrate the effect of the proposed linear adaptation strategies: IMA with Filtered Augmentations (FA), IMA with All Augmentations (AA), TMA with Filtered Augmentations (FA), and TMA with All Augmentations (AA). Through entropy minimization, these adaptations adjust the image or text embeddings to increase the confidence of the ground-truth label, effectively making it **argmax** in the final inference aligning the model prediction with the correct class. The markers of legend denote image and text embeddings. Dashed arrows denote cosine similarity between embeddings, while bold arrows signify the direction of embedding transformations in the projected space.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 1

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 5

[5] Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3

[6] 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 *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 1

[7] 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*, 12(7): 2217–2226, 2019. 1

[8] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 1

[9] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021. 1, 5

[10] Raza Imam, Hanan Gani, Muhammad Huzaifa, and Karthik Nandakumar. Test-time low rank adaptation via confidence maximization for zero-shot generalization of vision-language models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5449–5459. IEEE, 2025. 3, 4

[11] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 1

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

[13] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 1

[14] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 1

[15] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 1

[16] Lijun Sheng, Jian Liang, Zilei Wang, and Ran He. R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29958–29967, 2025. 3, 4

[17] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022. 3, 4

[18] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 1

[19] Elaine Sui, Xiaohan Wang, and Serena Yeung-Levy. Just shift it: Test-time prototype shifting for zero-shot generalization with vision-language models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 825–835. IEEE, 2025. 3, 4

[20] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019. 1

[21] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 119(1):3–22, 2016. 1

[22] Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark A. Hasegawa-Johnson, Yingzhen Li, and Chang D. Yoo. C-TPT: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *The Twelfth International Conference on Learning Representations*, 2024. 3, 4

[23] Maxime Zanella and Ismail Ben Ayed. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23783–23793, 2024. 3

[24] Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. Test-time adaptation with CLIP reward for zero-shot generalization in vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024. 3