ICCV HONOLULU HAWAII OCT 19-23, 2025

Test-Time Retrieval-Augmented Adaptation for Vision-Language Models

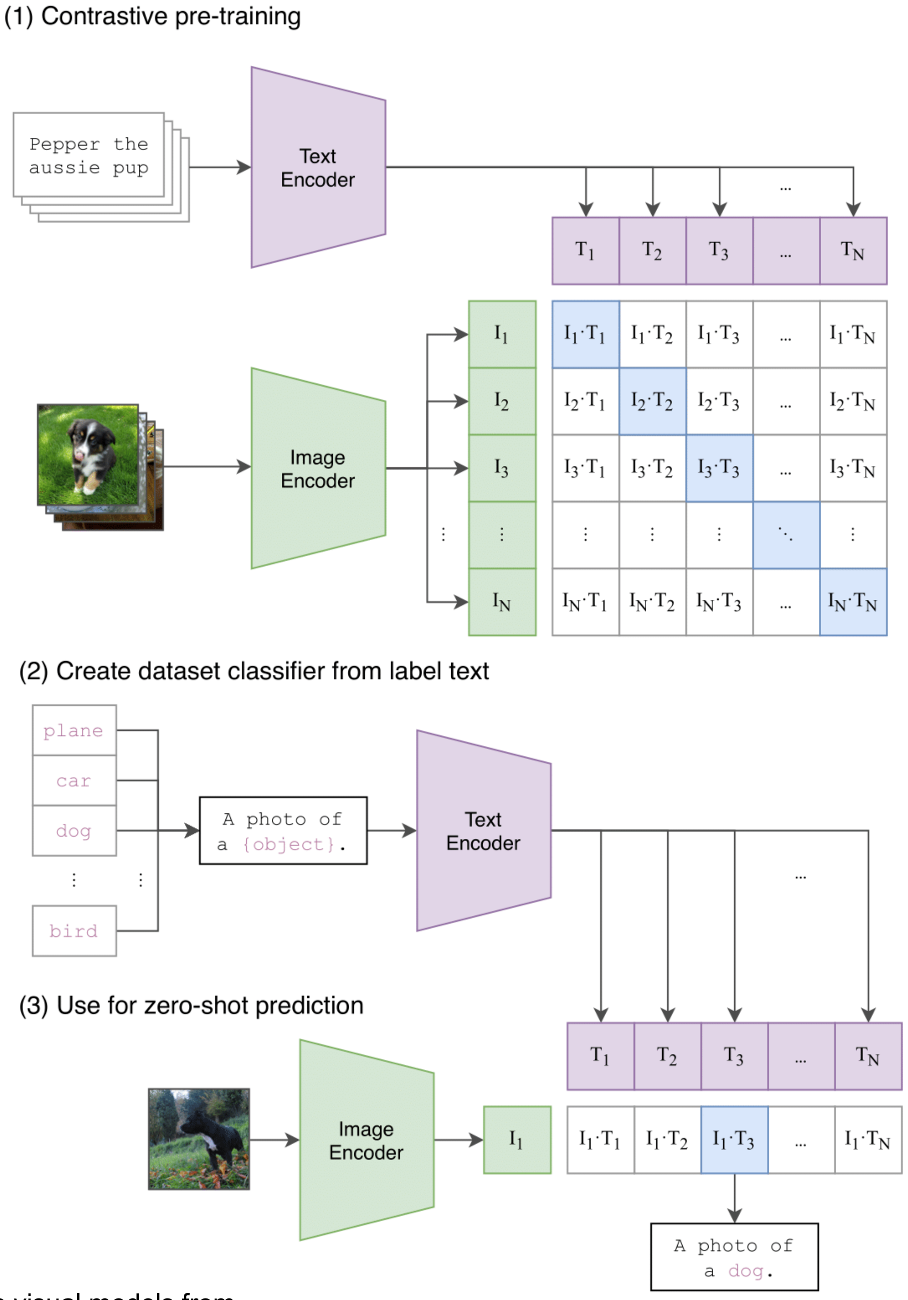
Xinqi Fan[1], Xueli Chen[2], Luoxiao Yang[3], Chuin Hong Yap[1], Rizwan Qureshi[4], Qi Dou[5], Moi Hoon Yap[1], Mubarak Shah[4]

[1]Manchester Metropolitan University, [2]Hong Kong Metropolitan University, [3]Xi'an University of Technology, [4]University of Central Florida, [5]Chinese University of Hong Kong

https://github.com/xinqi-fan/TT-RAA

Introduction

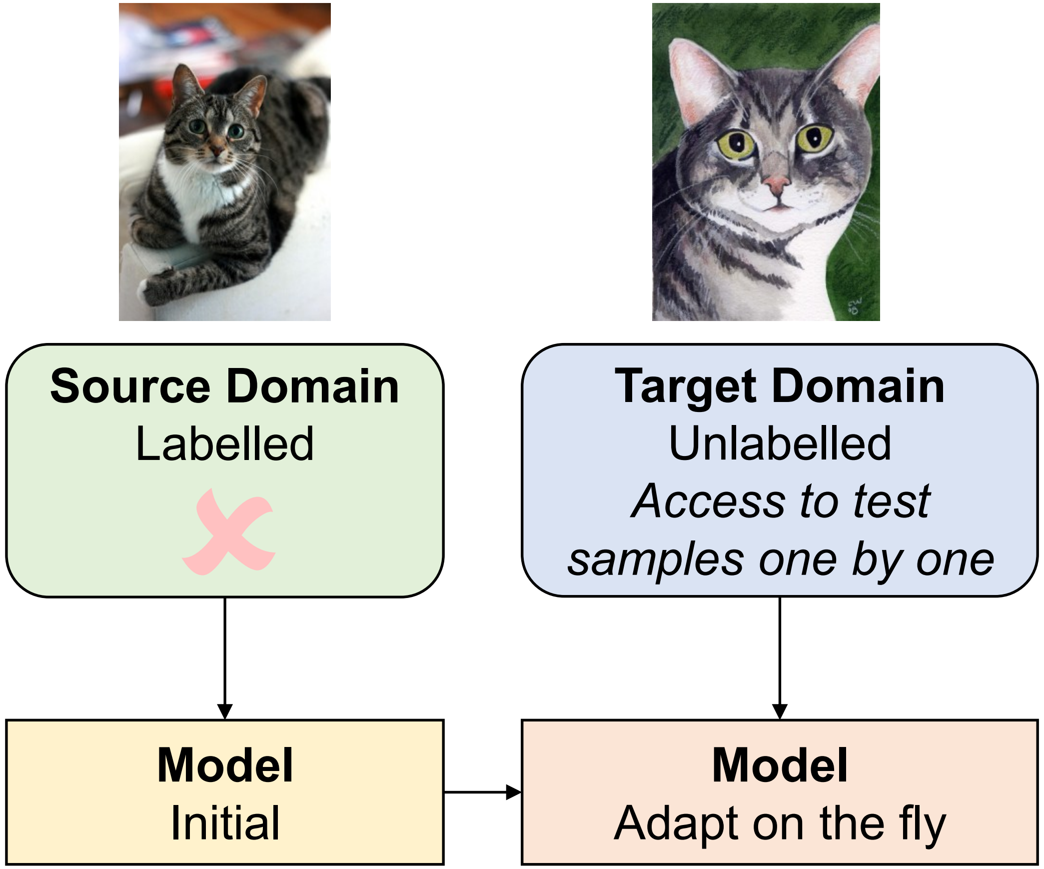
- Contrastive Language-Image Pre-training (CLIP)
- Pre-training
  - Aligned vision and language representations
  - Contrastive learning
  - Paired images and texts
- Zero-shot prediction
  - Use label text
  - Create a classifier
- Various computer vision tasks
  - Integrate human languages



Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML),* 2021.

Motivation: Test-Time Adaptation

- Distribution shift
  - Different source and target domains
  - Degrade performance
- Domain adaptation
  - Access to labelled source data
  - Access to unlabelled target data (before testing)
- Problems
  - Cannot access the source data due to privacy or data retention policies
  - Cannot access the target data (before testing) due to time-consuming collection or the constantly changing environment
- Solution: Test-time adaptation
  - No access to source data
  - Only access to unlabelled test samples one by one



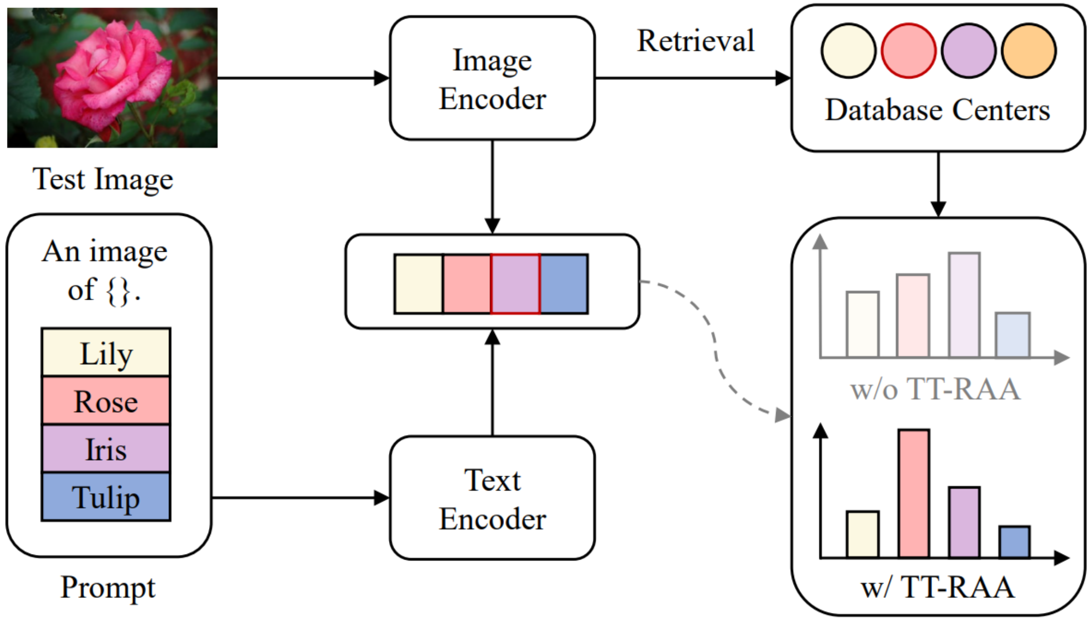
Test-time Adaptation

J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2009
Dan Hendrycks, Steven Basart, Norman Mu, 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 (ICCV),* 2021.

Motivation: Retrieval Augmentation

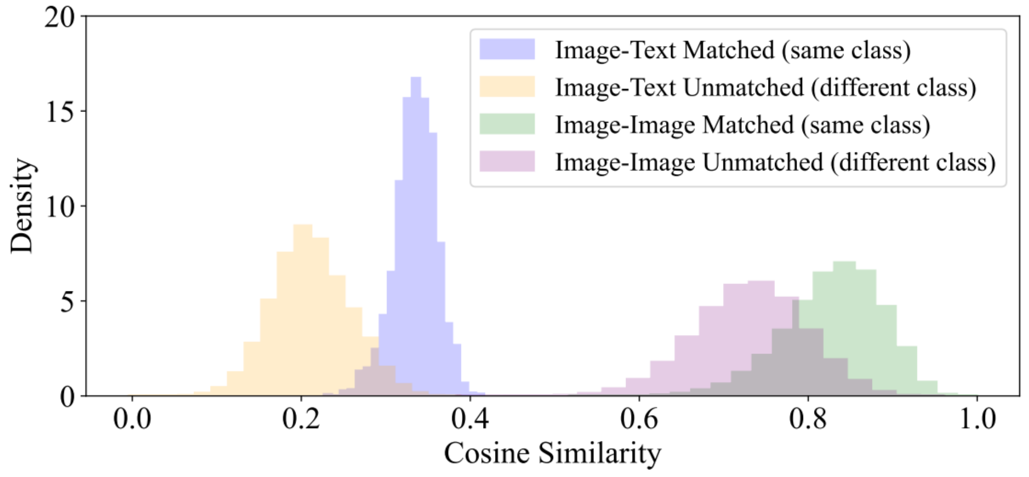
- Training-based adaptation
  - Retrain the model using test samples
  - High computational cost
  - Not affordable in computationally resource-limited real-world applications
- Solution: Retrieval augmentation
  - Construct a test-time database
  - Store important test information
  - Retrieve information in the database
  - Training-free with lower computational cost

Method	Speed (ms/sample)	GPU Usage (MB)
Training-based adaptation (prompt tuning)	103	2213.53
Training-free adaptation (our method)	12.93	535.10



Motivation: Multimodal Retrieval

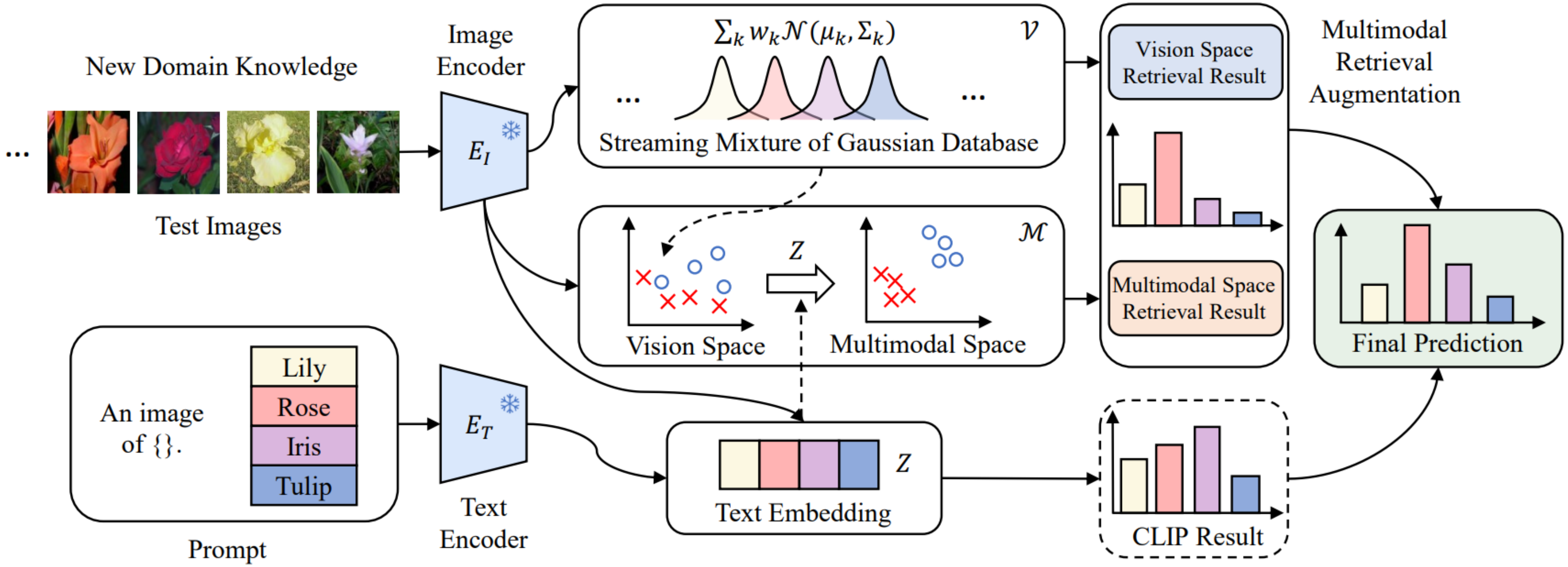
- Limitations of CLIP
  - Optimized to reduce the inter-modal (vision-text) similarities rather than the intra-modal (vision-vision or text-text) similarities
  - Similar images in the vision (image) feature space are not well clustered
  - Cosine similarity distribution indicates that matched and unmatched pairs are more easily distinguishable in the multimodal (CLIP) space
- Solution: Multimodal Retrieval
  - Vision space retrieval
  - Multimodal space retrieval



Cosine similarity distributions of matched and unmatched image-text pairs (inter-modal) exhibit less overlap than those of matched and unmatched image-image pairs (intra-modal)

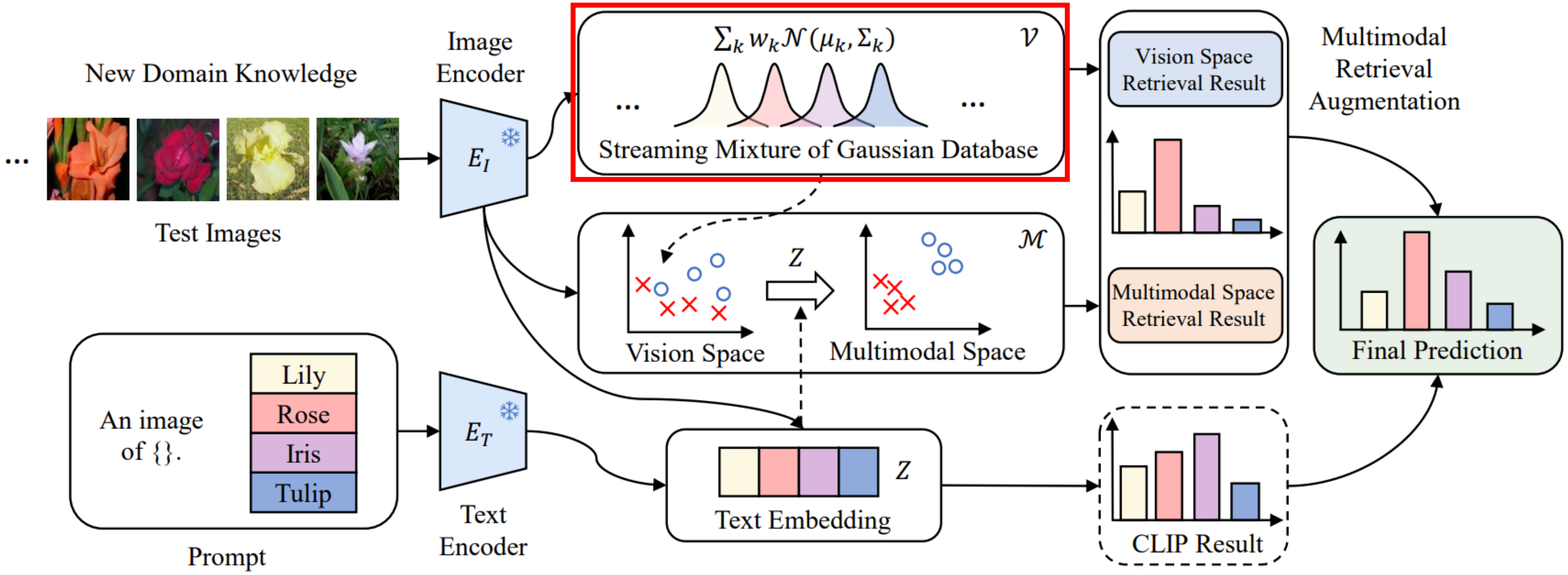
Method: Overview

- Test-time Retrieval-augmented Adaptation
  - Contrastive Language-Image Pre-training (CLIP)
  - Streaming Mixture of Gaussian Database (SMGD)
  - Multimodal Retrieval Augmentation (MRA)



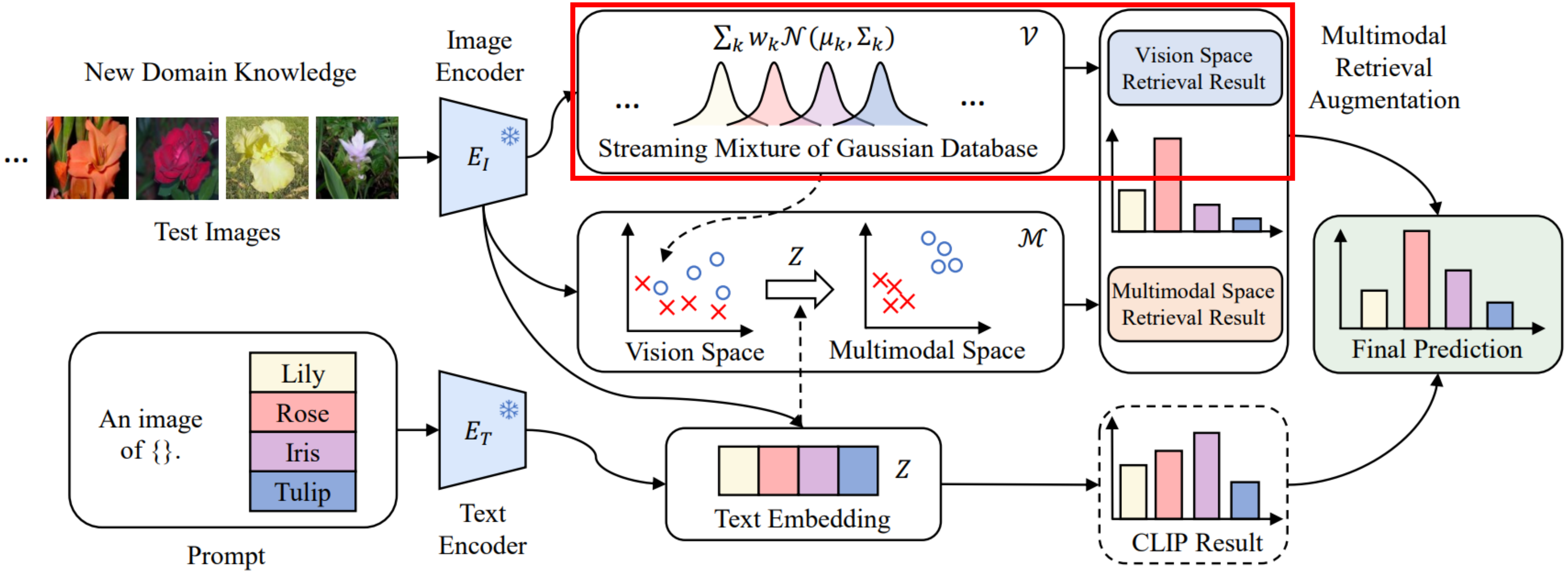
Method: Streaming Mixture of Gaussian Database

- Estimate the test distribution from streaming data
- Test sample embedding f^t draws from a mixture of Gaussian distribution $f^t \sim \sum_k w_k \mathcal{N}(\mu_k, \Sigma_k)$
  - Pseudo labels are obtained from the CLIP prediction to determine the class
- Updates of SMGD
  - Mean update: $\mu_k^t = (1-\eta)\mu_k^{t-1} + \eta f^t$
  - Covariance update: $\Sigma_k^t = (1-\eta)\Sigma_k^{t-1} + \eta(f^t - \mu_k^t)(f^t - \mu_k^t)^T$
  - Entropy update: $h_k^t = (1-\eta)h_k^{t-1} + \eta H(P_{CLIP}^t)$
  - Only update SMGD if new test sample's entropy is lower than the current SMGD entropy



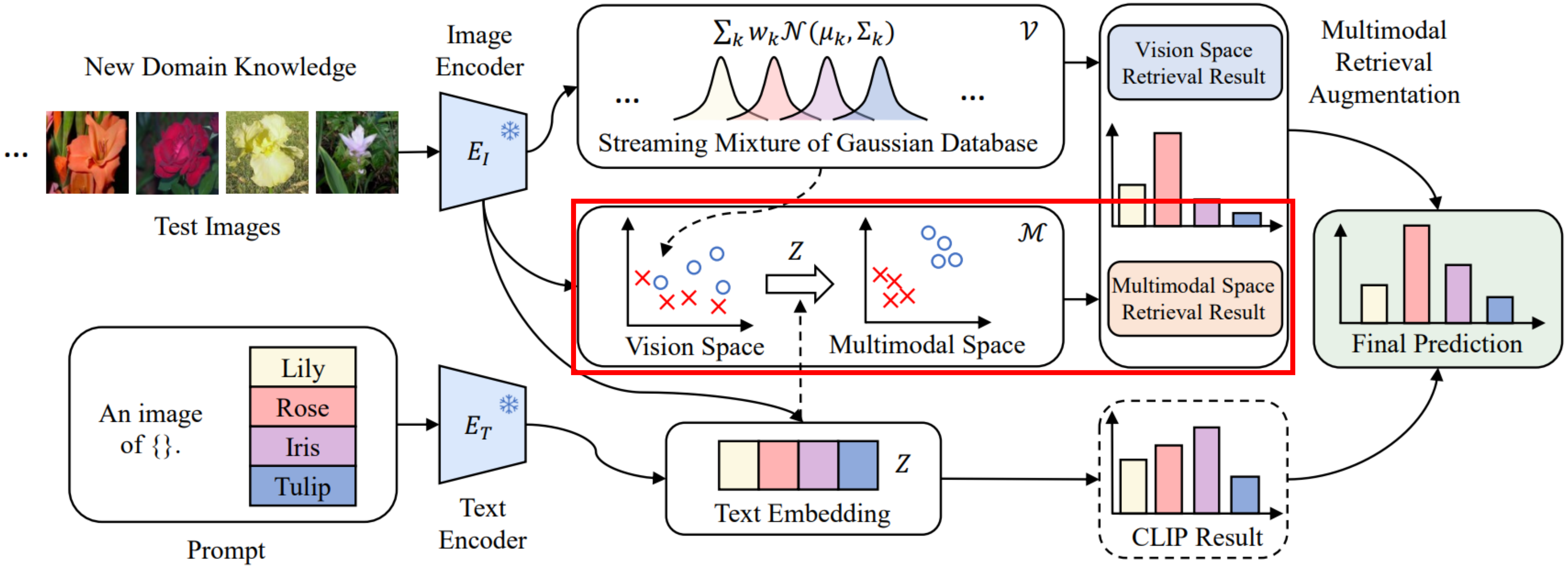
Method: Multimodal Retrieval Augmentation

- Vision-space retrieval
  - Similarity retrieval $P_{sim}(f^t) = LA(G^T f^t)$, where $G = [\mu_1, \mu_2, \dots, \mu_K]$
  - Discriminant analysis
    $\Omega_{disc}(f^t) = G^T {\Sigma^t}^{-1} G - \frac{1}{2} diag\left(G^T {\Sigma^t}^{-1} G\right) + log \frac{1}{K} \mathbf{1}_K$, and $P_{disc}(f^t) = L\Omega_{disc}(f^t)$
  - Vision space prediction $P_R^{\mathcal{V}}(f^t) = P_{sim}(f^t) + P_{disc}(f^t)$



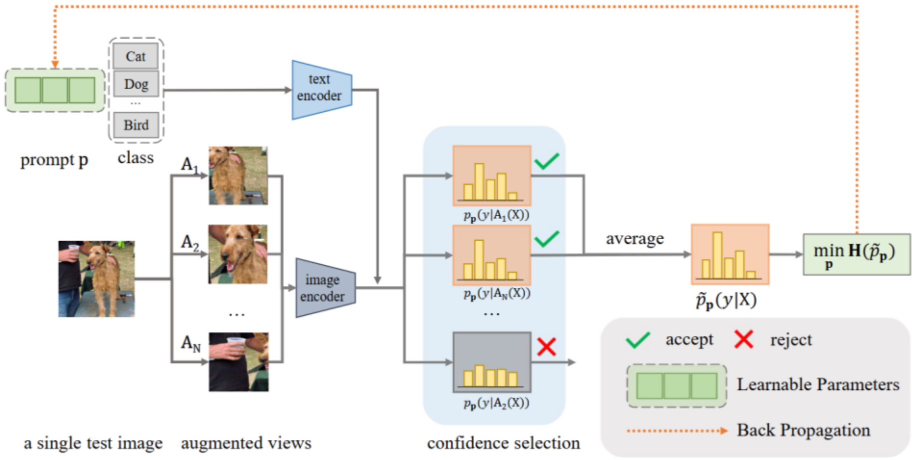
Method: Multimodal Retrieval Augmentation

- Multimodal-space retrieval
  - Transform SMGD centers from vision to multimodal space $\Psi = \sigma(Z^T G)$
  - Transform test sample embedding from vision to multimodal space $\psi = \sigma(Z^T f^t)$
  - Compare the similarity of the test sample and each center $\Phi_k = KL(\psi || \Psi_k)$
  - Obtain the multimodal space prediction $P_R^{\mathcal{M}} = -L\Phi$

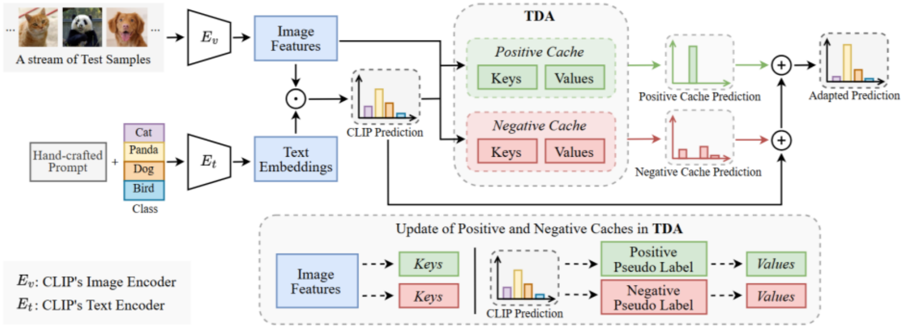


Experiment: Competing Methods

- Training-based adaptation
  - CoOp and CoCoOp tune prompts using training samples
  - TPT and DiffTPT tune prompts using test samples
- Training-free adaptation
  - Distribution-based methods
    - MTA uses MeanShift algorithm
    - DN uses distribution normalization
  - Cache-based methods
    - TDA employs a positive and a negative cache
    - DMN comprises a dynamic and a static cache
  - Entropy-based method
    - ZERO sets the temperature of most confident predictions as zero to approximate marginal entropy minimization



TPT



TDA

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 (NeurIPS),* 2022
Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 2024

Experiment: Comparisons

- Better than both training-based and training-free adaptation approaches on average
- Achieved SOTA performance on the cross-domain (CD) and out-of-distribution (OOD) benchmarks

Method	ImageNet-A	ImageNet-V2	ImageNet-R	ImageNet-S	Average
CLIP-ViT-B/16	49.89	61.88	77.65	48.24	59.42
CoOp	49.71	64.20	75.21	47.99	59.28
CoCoOp	50.63	64.07	76.18	48.75	59.91
TPT	54.77	63.45	77.06	47.94	60.81
DiffTPT	55.68	65.10	75.00	46.80	60.52
MTA	57.41	63.61	76.92	48.58	61.63
DN	58.71	62.89	80.20	48.94	62.69
ZERO	59.61	64.16	77.22	48.40	62.35
DMN	58.28	**65.17**	78.55	**53.20**	63.80
TDA	60.11	64.67	80.24	50.54	63.89
TT-RAA (Ours)	**60.59**	64.69	**80.58**	49.98	**63.96**

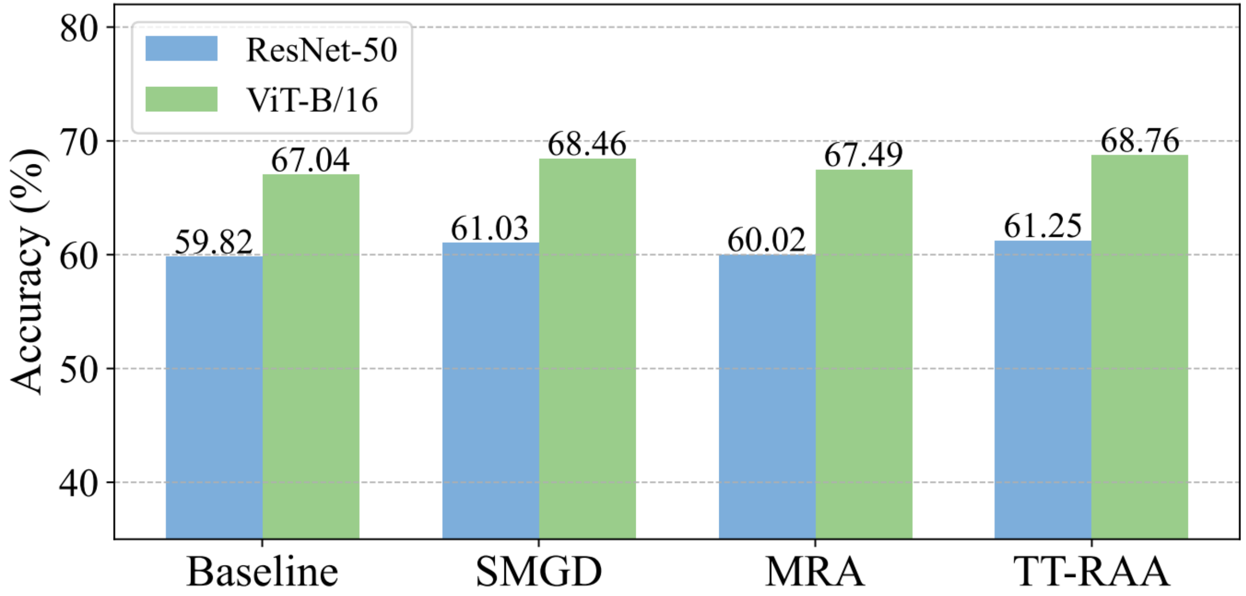
Out-Of-Distribution (OOD) Benchmark

Method	Aircraft	Caltech101	Cars	DTD	EuroSAT	Flower102	Food101	Pets	SUN397	UCF101	Average
CLIP-ViT-B/16	23.22	93.55	66.11	45.04	50.42	66.99	82.86	86.92	65.63	65.16	64.59
CoOp	18.47	93.70	64.51	41.92	46.39	68.71	85.30	89.14	64.15	66.55	63.88
CoCoOp	22.29	93.79	64.90	45.45	39.23	70.85	83.97	**90.46**	66.89	68.44	64.63
TPT	24.78	94.16	66.87	47.75	42.44	68.98	84.67	87.79	65.50	68.04	65.10
DiffTPT	**25.60**	92.49	67.01	47.00	43.13	70.10	**87.23**	88.22	65.74	62.67	65.47
MTA	25.32	94.13	**68.05**	45.59	38.71	68.26	84.95	88.22	64.98	68.11	64.63
DN	24.30	93.60	64.00	45.70	53.30	68.00	86.00	87.70	66.50	68.40	65.75
ZERO	25.21	93.66	68.04	46.12	34.33	67.68	86.53	87.75	65.03	67.77	67.72
DMN	24.84	94.12	65.64	44.39	47.77	71.38	84.48	89.07	66.28	66.75	65.47
TDA	23.91	**94.24**	67.28	47.40	58.00	71.42	86.14	88.63	67.62	70.66	67.53
TT-RAA (ours)	25.38	94.08	66.42	**47.99**	**66.12**	**72.68**	86.09	89.83	**67.69**	**71.29**	**68.76**

Cross-Domain (CD) Benchmark

Experiment: Ablation Studies

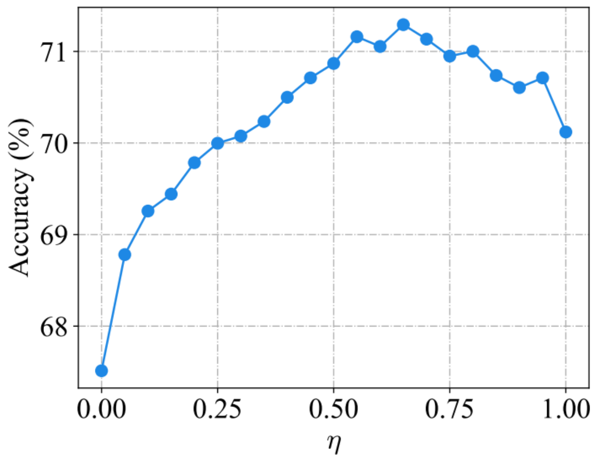
- SMGD contributes more than 1% improvement
- MRA and SMGD show complementary benefits
- Consistent improvements on both ViT-B/16 and ResNet-50
- Detailed analysis of MRA shows the effectiveness of each component



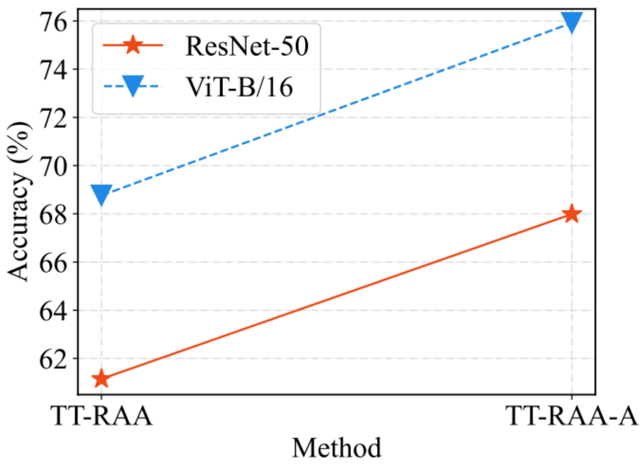
SMGD	VSSR	VSDA	MSRA	CD Average
-	-	-	-	64.59
-	✓	-	-	67.04
-	-	-	✓	65.85
-	✓	-	✓	67.49
✓	✓	-	-	68.46
✓	-	-	✓	65.84
✓	-	✓	-	66.23
✓	✓	✓	-	68.53
✓	✓	✓	✓	68.76

Experiment: Parameter Analysis and Access to More Data

- η balances the historical and new information
- Empirical optimal value $\eta = 0.65$ on UCF101, suggesting 35% historical information and 65% new information

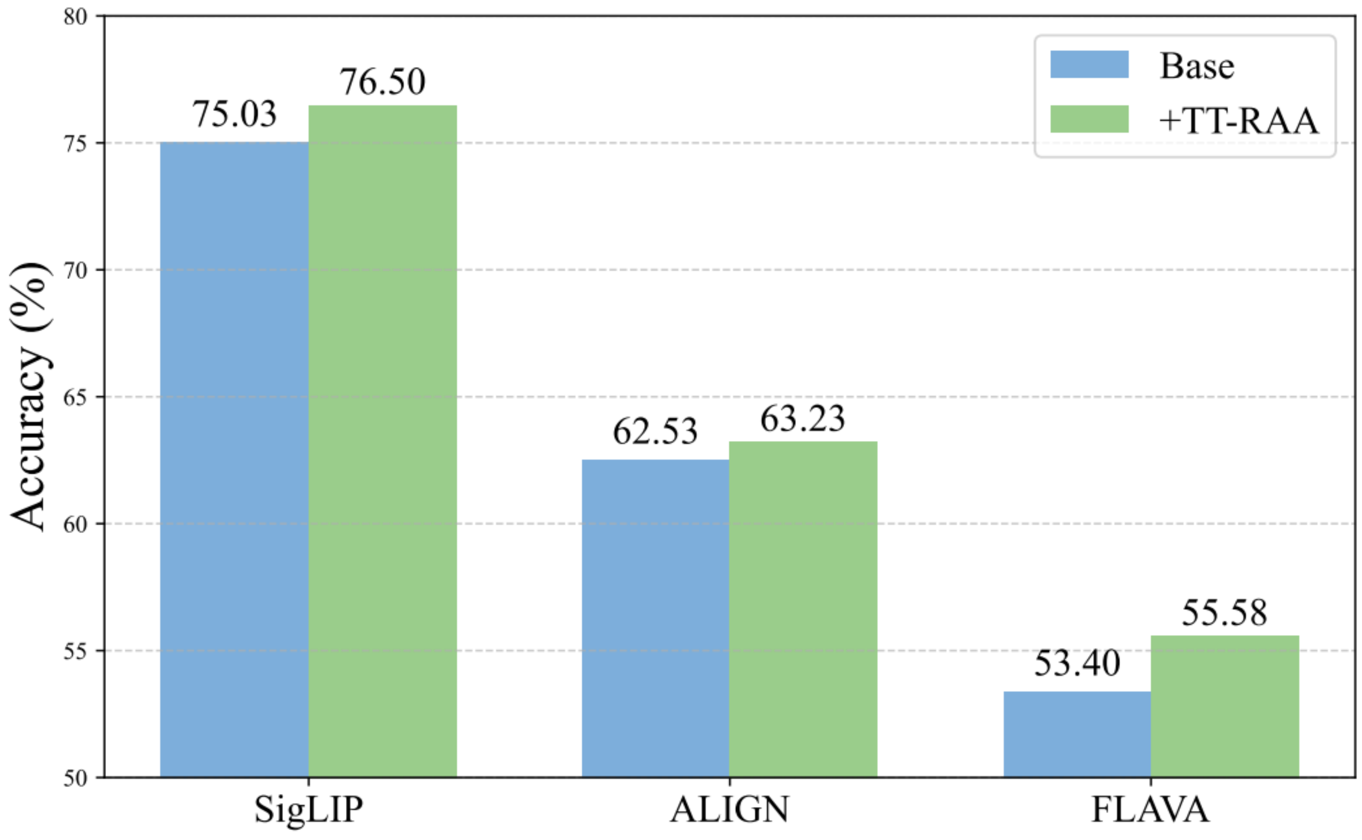


- Accessing to the additional target domain's training data allows us to directly estimate target domain statistics
- Perform the same retrieval augmentation
- Significant performance boost without training



Experiment: Generalizations

- Experiments with other VLMs
  - SigLIP
  - ALIGN
  - FLAVA
- Consistent improvements demonstrate generalization of our method



Thank You!

Please feel free to discuss and ask questions.