ICLR The Thirteenth International Conference on Learning Representations

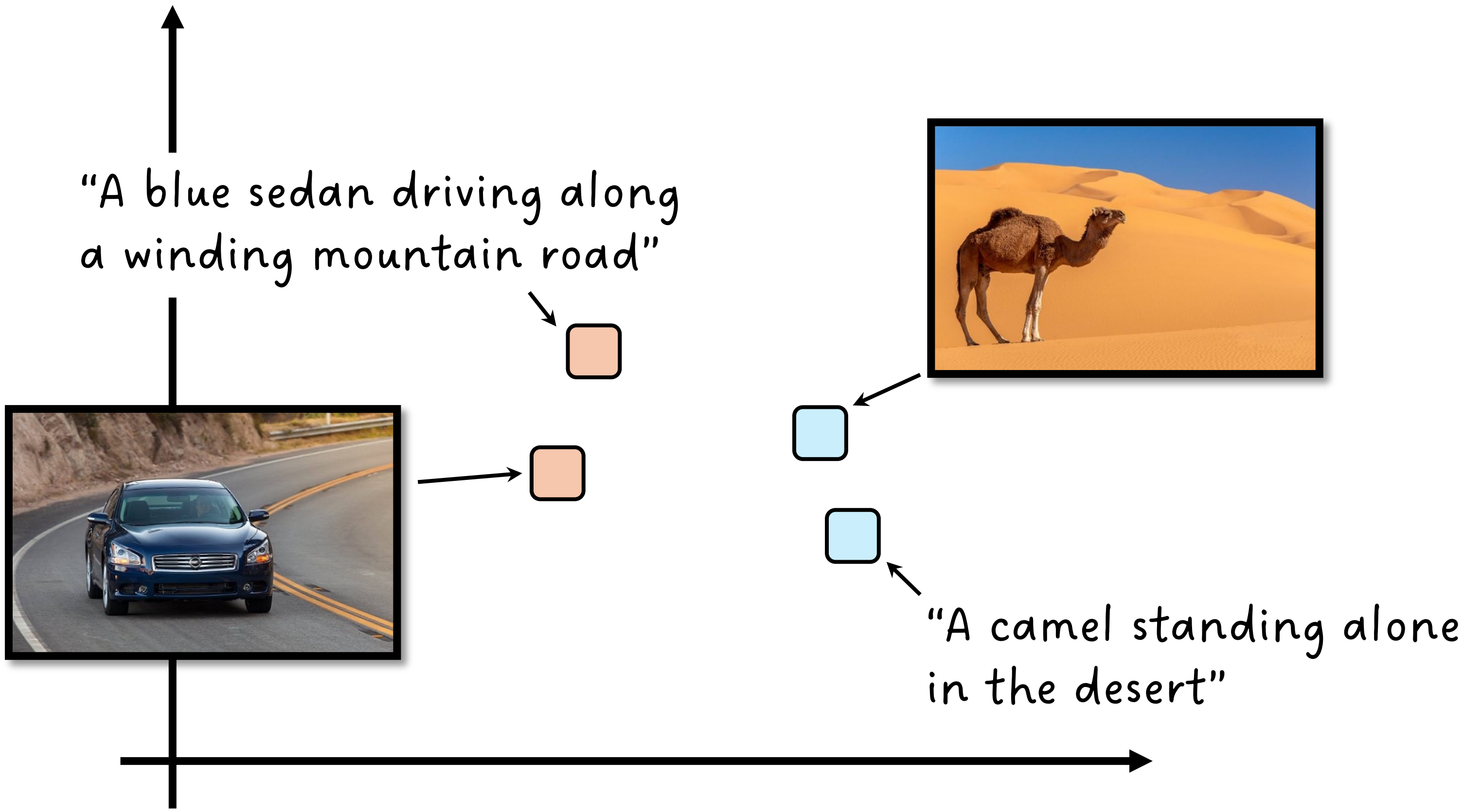
RA-TTA: Retrieval-Augmented Test-Time Adaptation for Vision-Language Models

Youngjun Lee[1] , Doyoung Kim[1] , Junhyeok Kang[2] , Jihwan Bang[1] , Hwanjun Song[1] , Jae-Gil Lee[1,*]

[1] KAIST [2] LG AI Research

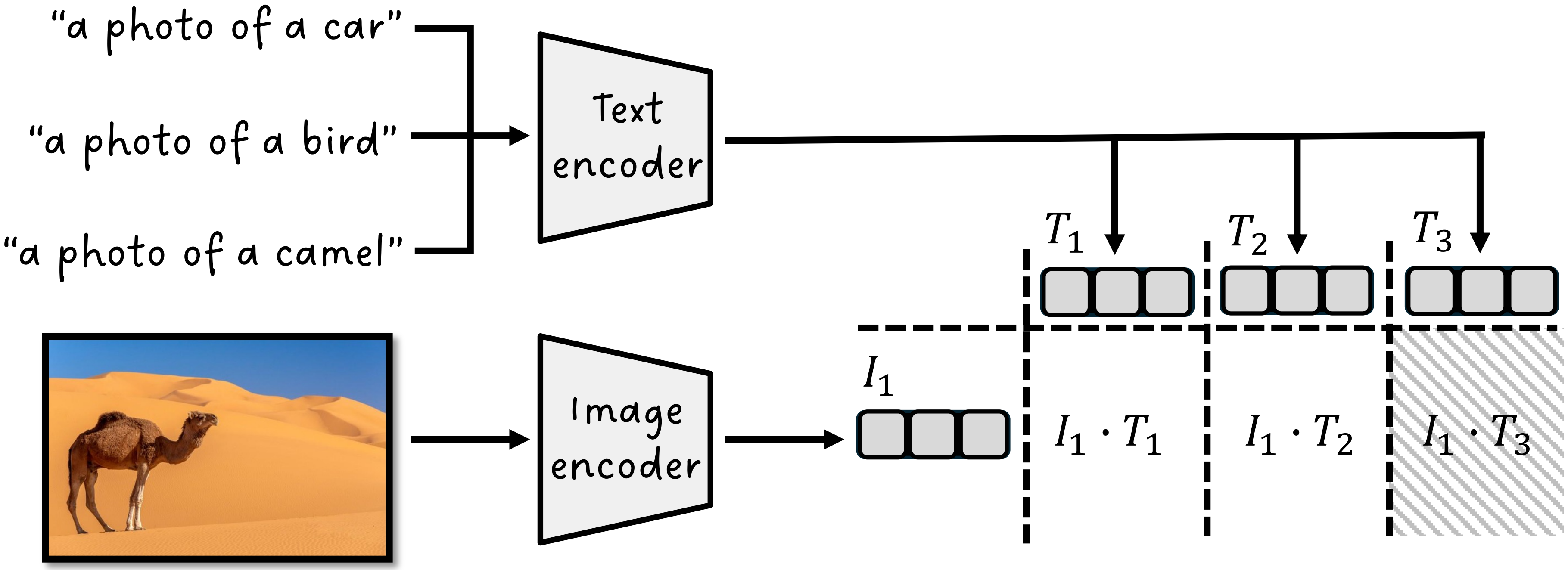
Vision-Language Models

- **Vision-language models (VLMs)** are multi-modal models that can understand **both visual and textual information**.



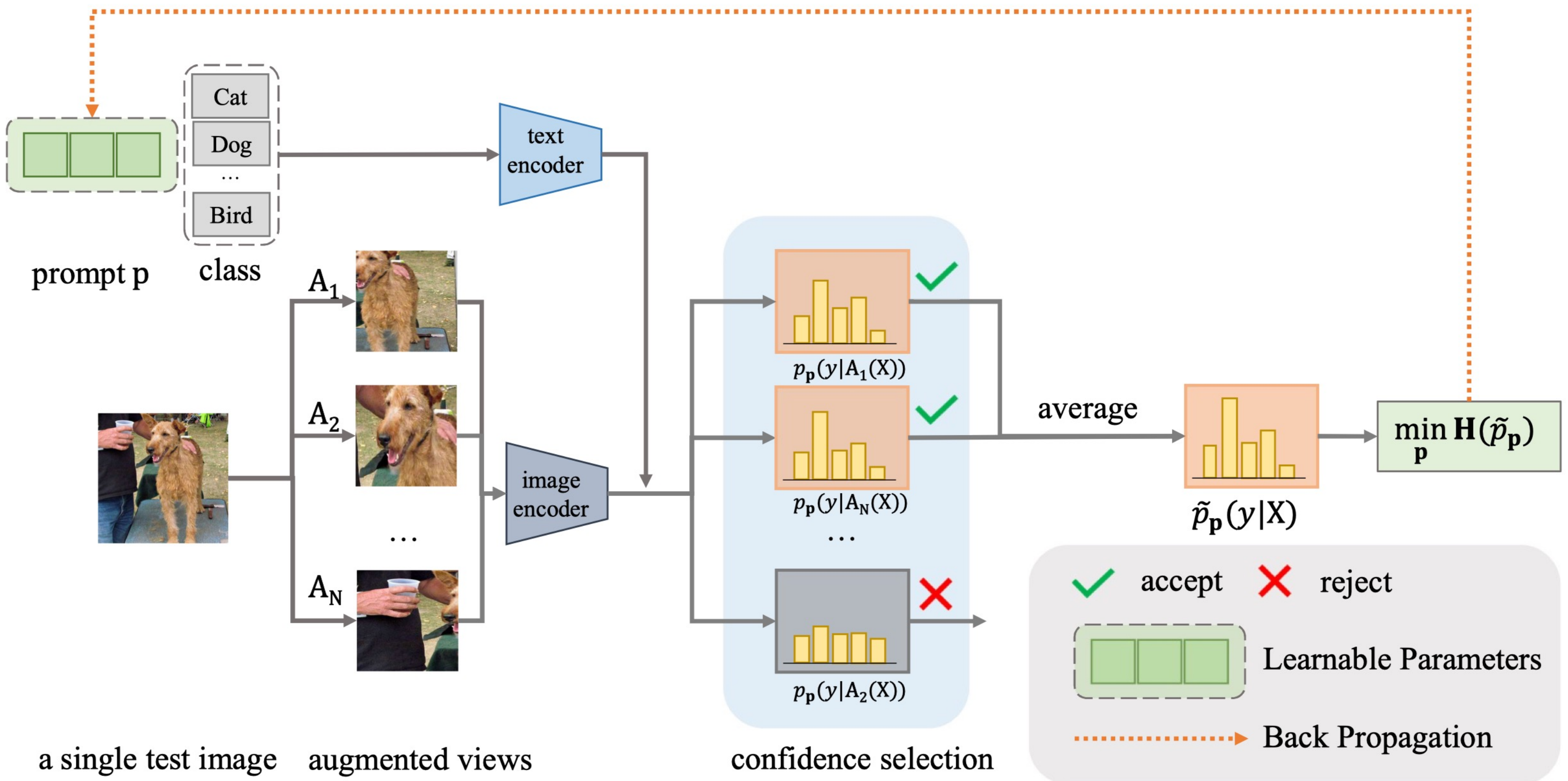
Zero-Shot Transfer of VLMs

- VLMs have demonstrated excellent **zero-shot transferability** for image classification tasks, where classes are represented by text prompts (e.g., a photo of a/an).



Test-Time Adaptation (TTA)

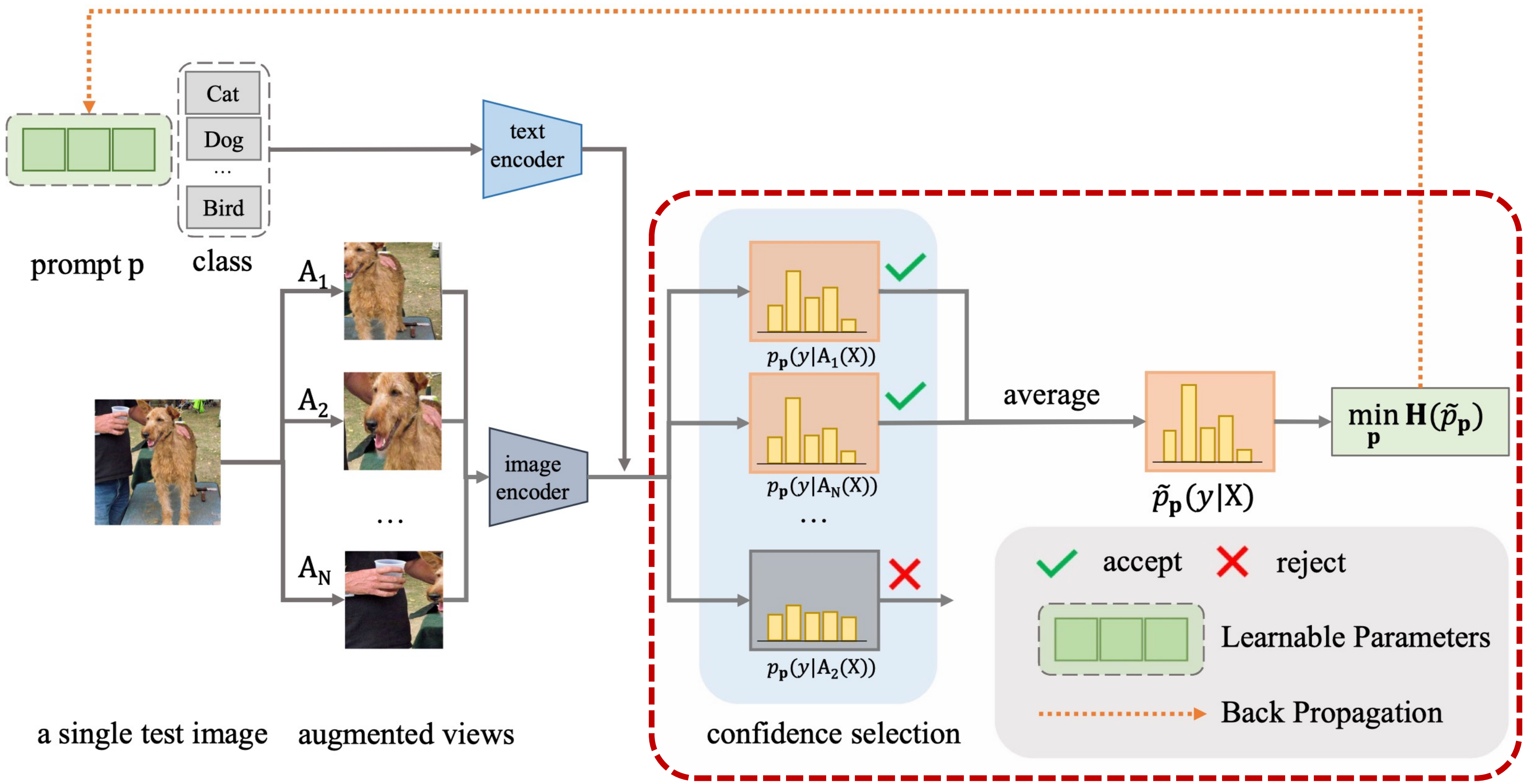
- When transferring the zero-shot capability of VLMs, **test-time adaptation (TTA)** methods for VLMs have been proposed to mitigate the detrimental impact of the distribution shifts between pre-training and test data.



TPT [1] (*NeurIPS*'22)

A Limitation of Previous TTA methods

- However, the previous TTA methods solely rely on the **internal knowledge** encoded in the VLM parameters, which are constrained to the pre-training data.



TPT [1] (*NeurIPS*'22)

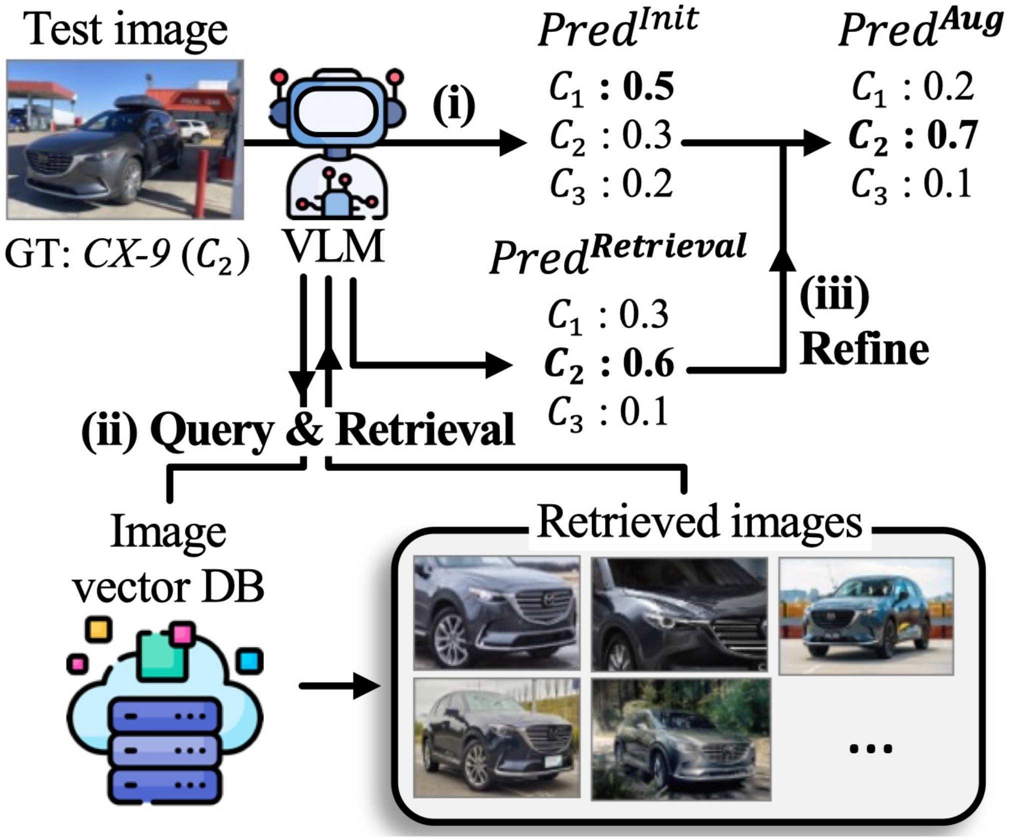
The **internal** knowledge determine the adaptation

⇩

Can we leverage **external** knowledge?

Retrieval-Augmented TTA (RA-TTA)

- Thus, we propose a **retrieval-augmented approach for TTA with VLMs**, which can incorporate **external knowledge** from a web-scale image database.



Overview of the proposed RA-TTA

Proper External Knowledge for RA-TTA

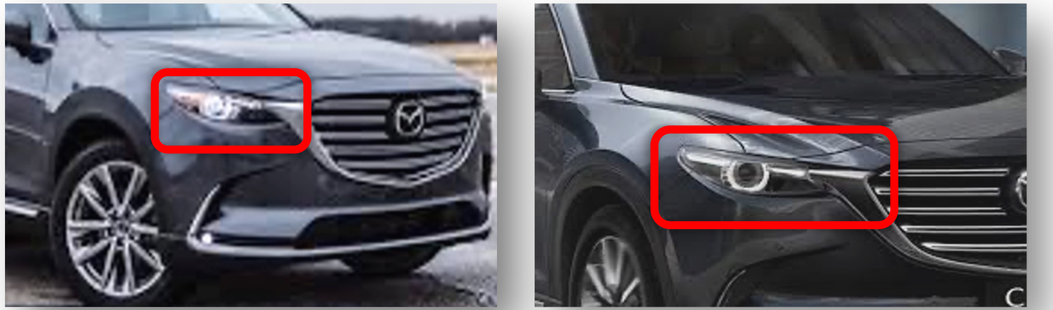
- We assume that the proper external knowledge for a given test image should have **pivotal** features (rather than irrelevant features) that is informative for recognizing the test image.

Test image

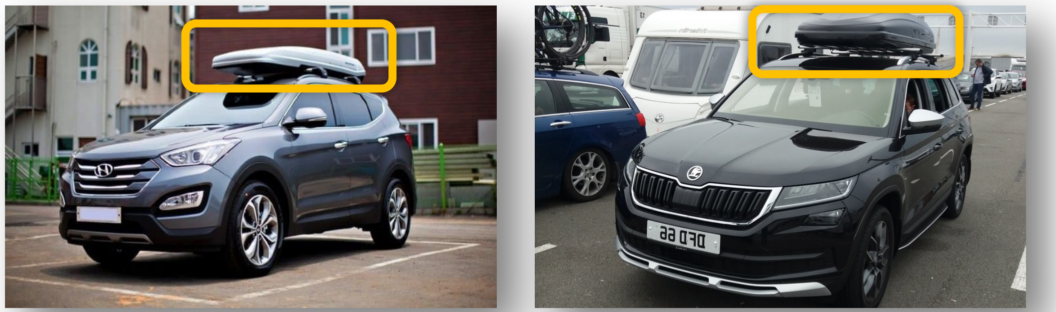
: headlights (pivotal features)

: a ski-box (irrelevant features)

Images with **pivotal** features

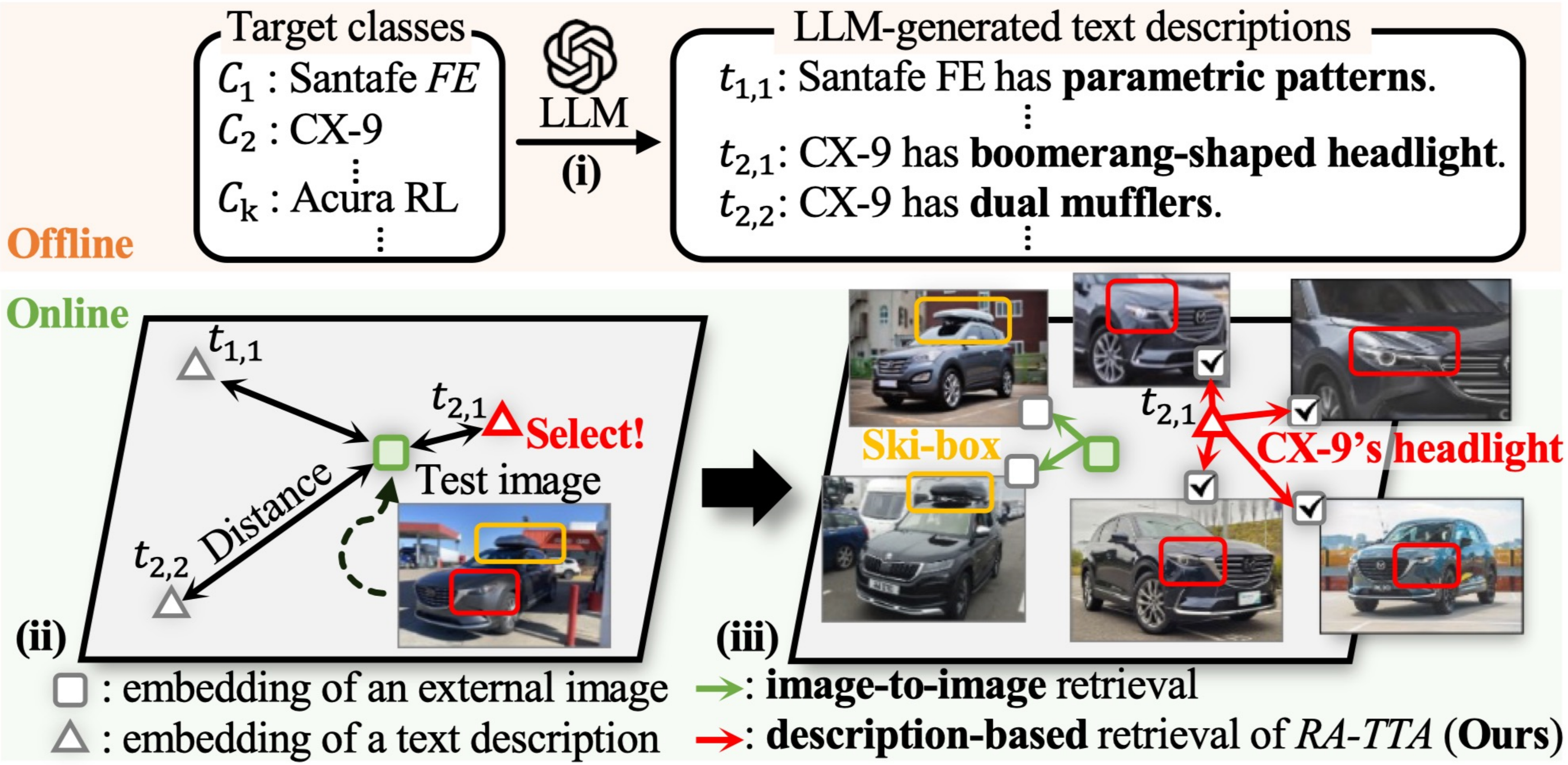


Images with **irrelevant** features



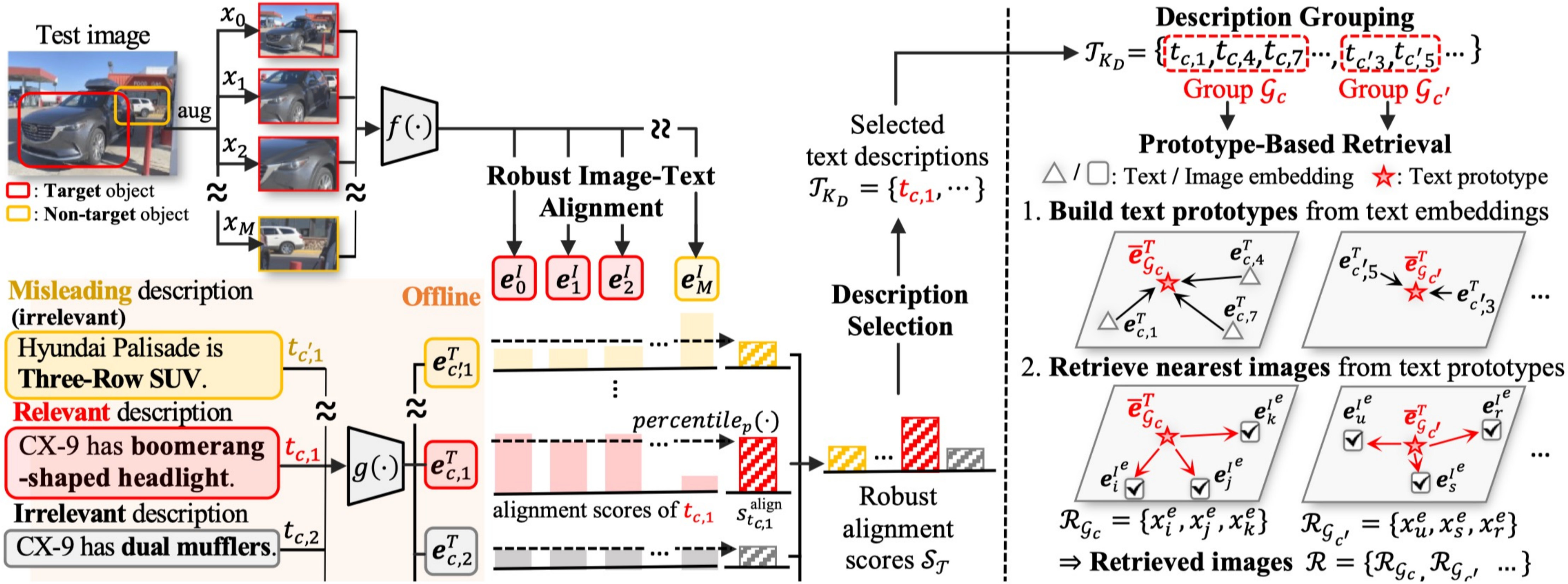
Description-Based Retrieval

- To retrieve proper external images, we propose a **description-based retrieval** approach that fully leverages the **bi-modality of VLMs** through text-to-image search and then through text-to-image retrieval.



Description-Based Retrieval (Cont'd)

- Specifically, the description-based retrieval selects relevant descriptions from multiple perspectives and retrieves external images using the prototypes of the selected descriptions.



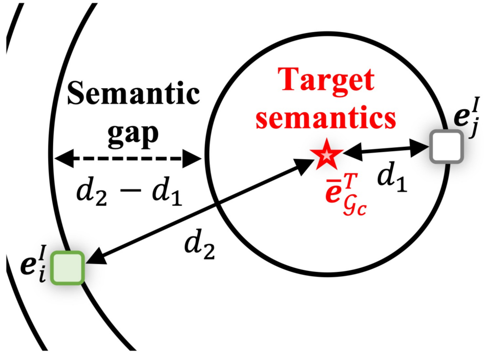
Description-Based Adaptation

- Based on the semantic relevance between the test image and the retrieved images, we calculate a retrieval-based prediction, which is then fused with an initial prediction for an augmented prediction.

Semantic gap　　　　　　　　　　　　　　Retrieved image

1. $\text{gap}(x_i, x_j, \mathcal{G}_c) = |(1 - \cos(\mathbf{e}_i^I, \bar{\mathbf{e}}_{\mathcal{G}_c}^T)) - (1 - \cos(\mathbf{e}_j^I, \bar{\mathbf{e}}_{\mathcal{G}_c}^T))|$

Test image　　　　　　　　　　　　　　Target semantics



2. $\mathbf{C}_{\mathcal{G}_c} = \left[\text{gap}(x_i, x_j^{\text{e}}, \mathcal{G}_c)\ |_{x_i \in \mathcal{A}, x_j^{\text{e}} \in \mathcal{R}_{\mathcal{G}_c}}\right] \in \mathbb{R}^{(M+1)\times K_S}$ Pair-wise semantic gaps

3. $s_{\mathcal{G}_c}^{\text{rel}} = \dfrac{1}{\text{OT}_{\text{dist}}(\mathbf{C}_{\mathcal{G}_c}, \mathcal{U}, \mathcal{V}) + 1}$ Semantic relevance considering the significance of each image in each set

4. $\hat{p}(c|x^{\text{test}}) = \begin{cases} \dfrac{\exp(s_{\mathcal{G}_c}^{\text{rel}})/\tau)}{\sum_{c\in\mathcal{C}} \exp(s_{\mathcal{G}_c}^{\text{rel}})/\tau)} & \text{if } c \in \mathcal{C} \\ 0 & \text{otherwise} \end{cases}$

5. $p^{\text{aug}}(c|x^{\text{test}}) = \alpha \times p(c|x^{\text{test}}) + (1-\alpha) \times \hat{p}(c|x^{\text{test}})$

Experiments

	IN-1k	Flowers102	DTD	Oxford pets	Stanford cars	UCF101	Caltech101	Food101	SUN397	FGVC aircraft	RESISC45	Caltech256	CUB200	Avg. (13)
CLIP	66.76	67.19	44.50	88.14	65.27	64.92	92.78	85.40	62.55	24.60	55.70	82.80	58.08	66.05
Ensemble	68.37	65.85	45.21	88.20	66.34	67.41	93.77	85.41	65.79	24.39	58.35	85.81	58.61	67.19
TPT	69.08	69.18	47.04	87.44	66.55	68.04	93.79	86.34	65.32	23.31	56.84	85.37	60.11	67.57
C-TPT	68.32	69.43	45.27	88.25	65.48	65.50	93.39	84.95	64.55	24.39	56.02	85.25	58.84	66.90
RLCF	68.61	67.72	46.40	86.73	66.51	66.98	93.83	86.09	64.92	23.43	56.89	85.18	57.91	67.02
VisDesc	69.09	71.86	50.41	88.55	65.48	69.52	94.81	86.43	68.25	25.59	57.81	88.17	60.13	68.93
WaffleCLIP	69.05	72.59	48.33	89.79	64.60	69.13	94.61	86.85	67.17	25.25	63.31	88.10	59.83	69.12
CuPL	69.78	75.92	58.22	91.47	66.92	67.80	94.24	86.39	67.38	28.98	65.17	88.07	60.18	70.81
SuS-X-LC	69.45	76.23	59.23	91.83	67.55	67.12	93.78	86.13	67.78	29.41	65.22	88.75	59.12	70.89
Neural Priming	69.38	73.22	55.98	89.76	66.13	68.02	94.71	87.01	67.86	27.32	63.11	88.50	57.14	69.86
RA-TTA (Ours)	**70.58**	**78.65**	**60.98**	**92.78**	**70.11**	**73.28**	**94.84**	**87.10**	**70.38**	**32.34**	**66.95**	**89.50**	**62.73**	**73.09**

	IN-A	IN-V2	IN-R	IN-K	Avg. (4)
CLIP	47.51	60.80	73.98	46.19	57.12
Ensemble	50.04	61.89	77.58	48.29	59.45
TPT	54.39	63.48	77.27	47.95	60.77
C-TPT	50.28	62.47	75.68	47.42	58.96
RLCF	56.52	63.37	77.04	48.09	61.26
VisDesc	50.17	62.76	75.25	48.25	59.11
WaffleCLIP	50.51	62.68	75.81	48.73	59.43
CuPL	50.23	63.00	78.16	49.60	60.25
SuS-X-LC	49.91	63.22	77.82	49.18	60.03
Neural Priming	49.68	62.79	76.70	49.03	59.55
RA-TTA (Ours)	**59.21**	**64.16**	**79.68**	**50.83**	**63.47**

- RA-TTA outperforms all existing methods.
- RA-TTA is particularly effective for a specialized domain or fine-grained datasets because RA-TTA retrieves a **customized** set of external images for each test image by identifying its **pivotal** features through fine-grained descriptions.

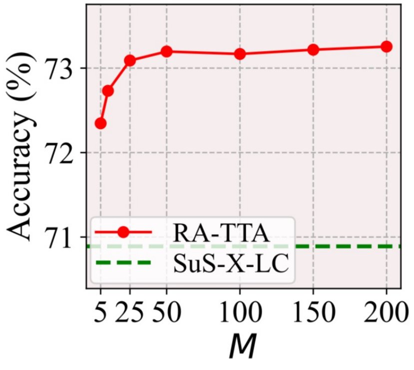
Empirical Analyses

Table 3: **Ablation studies.** We report the top-1 accuracy (%) on the FGVC aircraft dataset, where the benefit of RA-TTA is significant. Description-based retrieval, description-based adaptation, and image weighting are disabled separately.

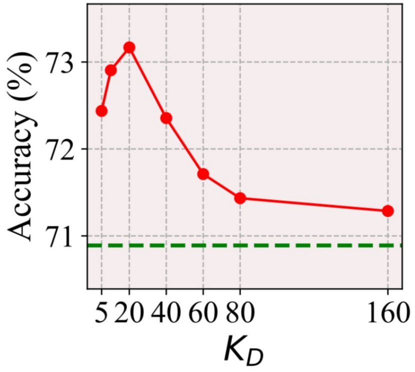
	Retrieval	Adaptation	Weighting	Accuracy
Var. 1	✗	✗	✗	29.39
Var. 2	✓	✗	✗	30.91
Var. 3	✓	✓	✗	31.96
RA-TTA	✓	✓	✓	**32.34**

Table 4: GPU inference time per sample (s/sample)

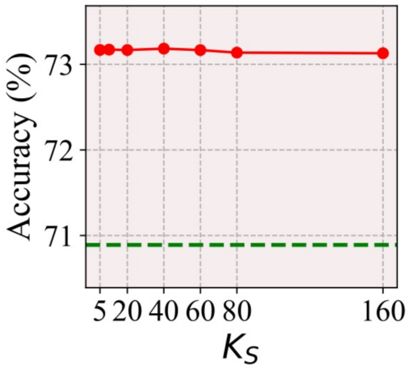
	FGVC aircraft	Stanford cars	RESISC45	Avg. (3)
TPT	**0.103**	0.155	**0.95**	0.118
RA-TTA (Ours)	0.113	**0.117**	0.121	**0.117**



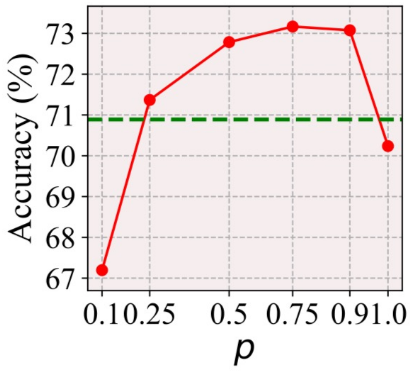
(a) Augmentation size.



(b) # of selected descriptions.



(c) # of retrieved images.



(d) Score percentile.

E.O.F