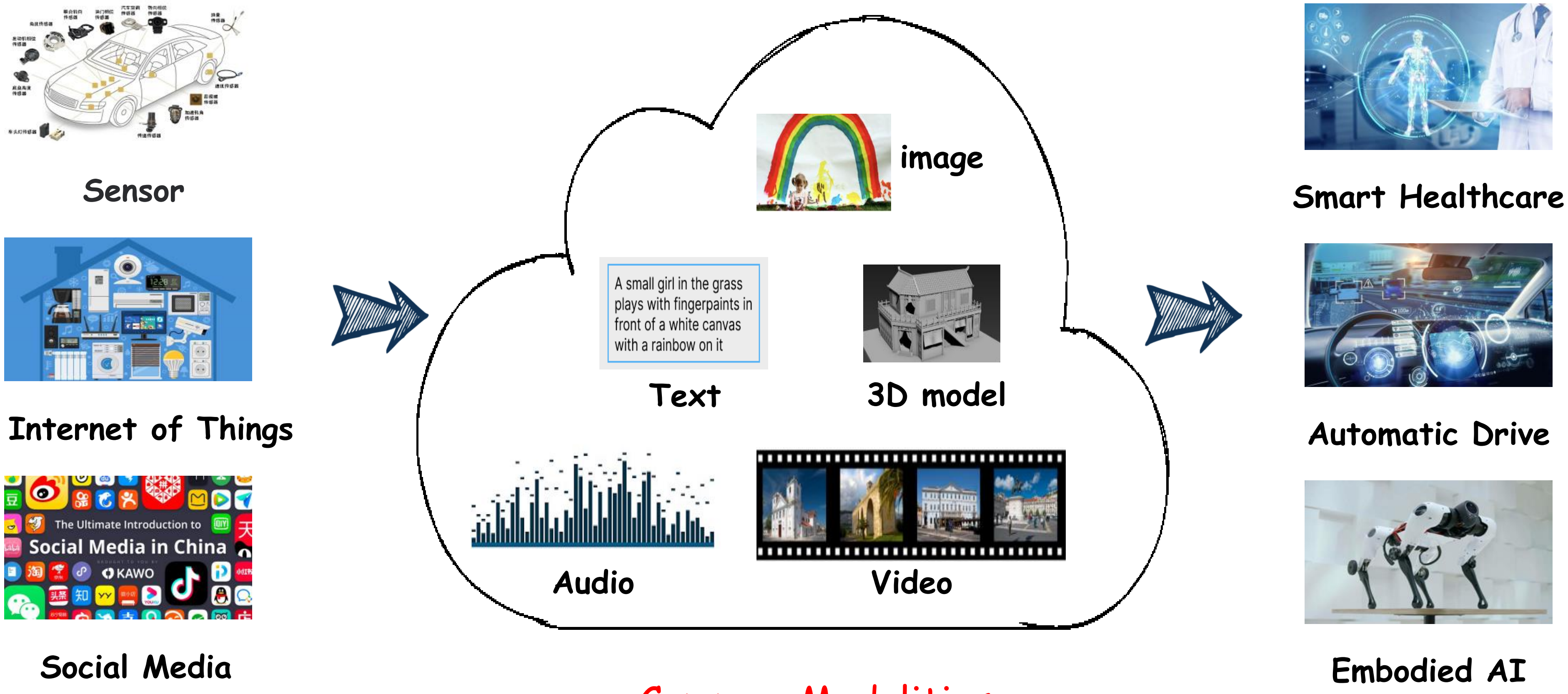
Test-time Adaptation for Cross-modal Retrieval with Query Shift

Haobin Li, Peng Hu, Qianjun Zhang, Xi Peng,
XitingLiu, Mouxing Yang

ICLR 2025 Spotlight

Background

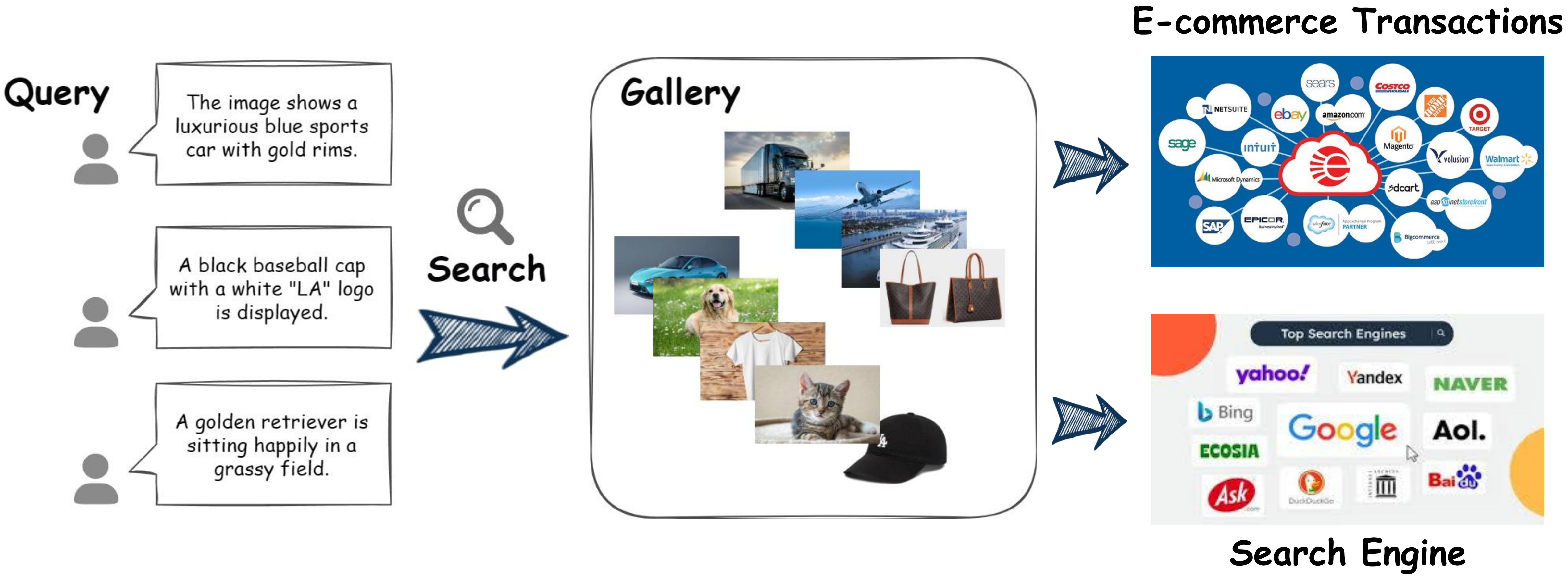
With the evolution of sensors, the popularization of smart devices, and the rise of the internet and social media, multi-modal data is showing a rapidly growing trend.



Common Modalities

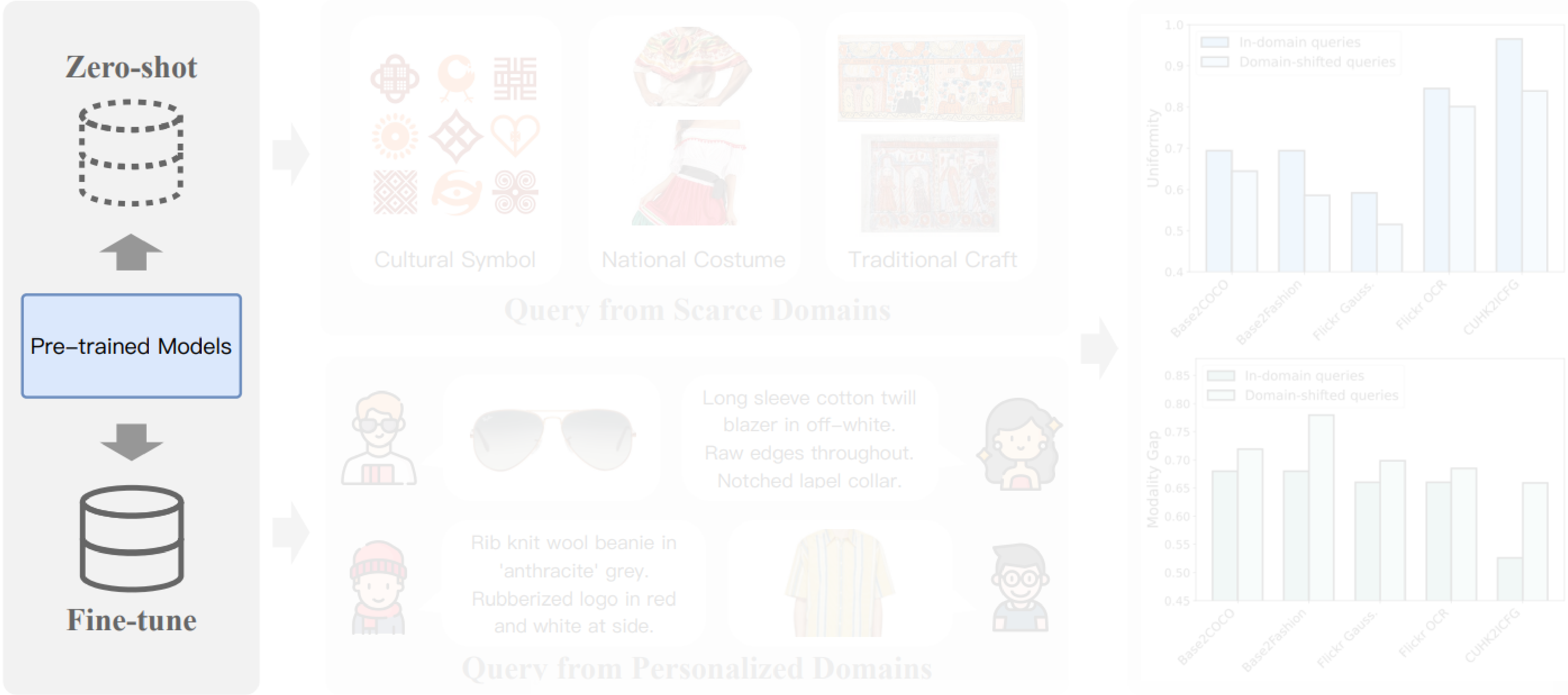
Background

Given queries of interest, cross-modal retrieval try to associate some relevant samples from the gallery set across various modalities, supporting numerous applications such as e-commerce transactions and search engine.



Background

Recently, the pre-trained models have emerged as the dominant paradigm for cross-modal retrieval.



(a) Dominant Paradigm (b) Query Shift (c) Observations

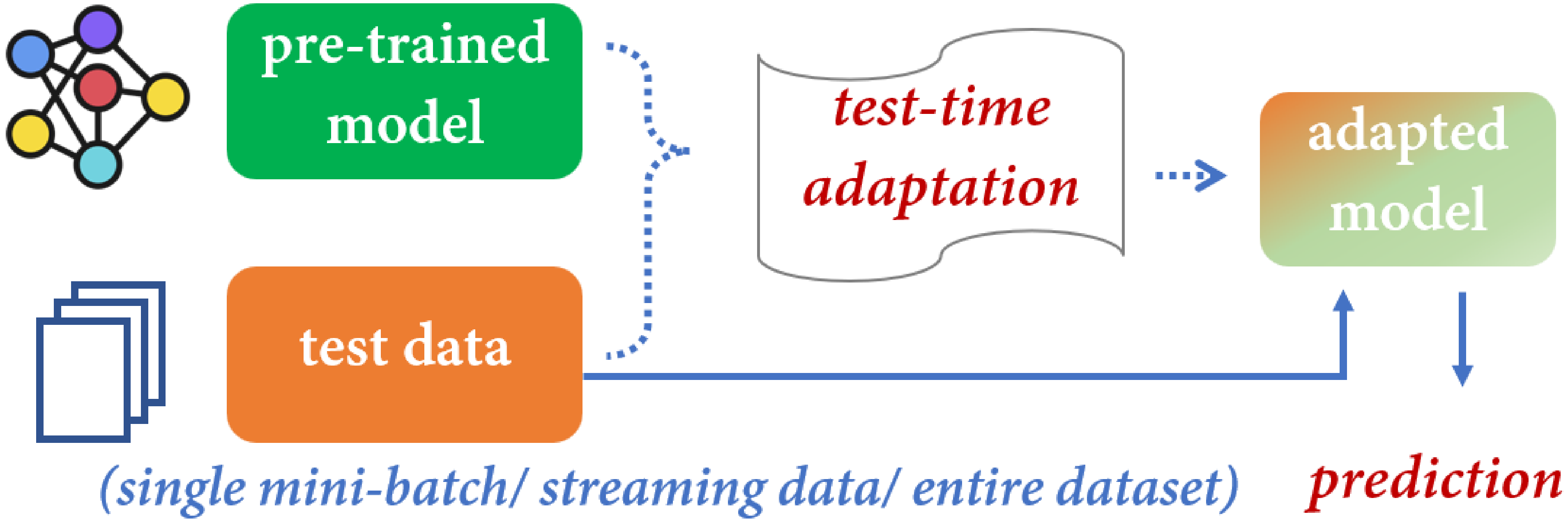
Motivation

Query shift refers to the online query stream originating from the domain that follows a different distribution with the source one.

(a) **Dominant Paradigm** (b) **Query Shift** (c) **Observations**

Previous Works

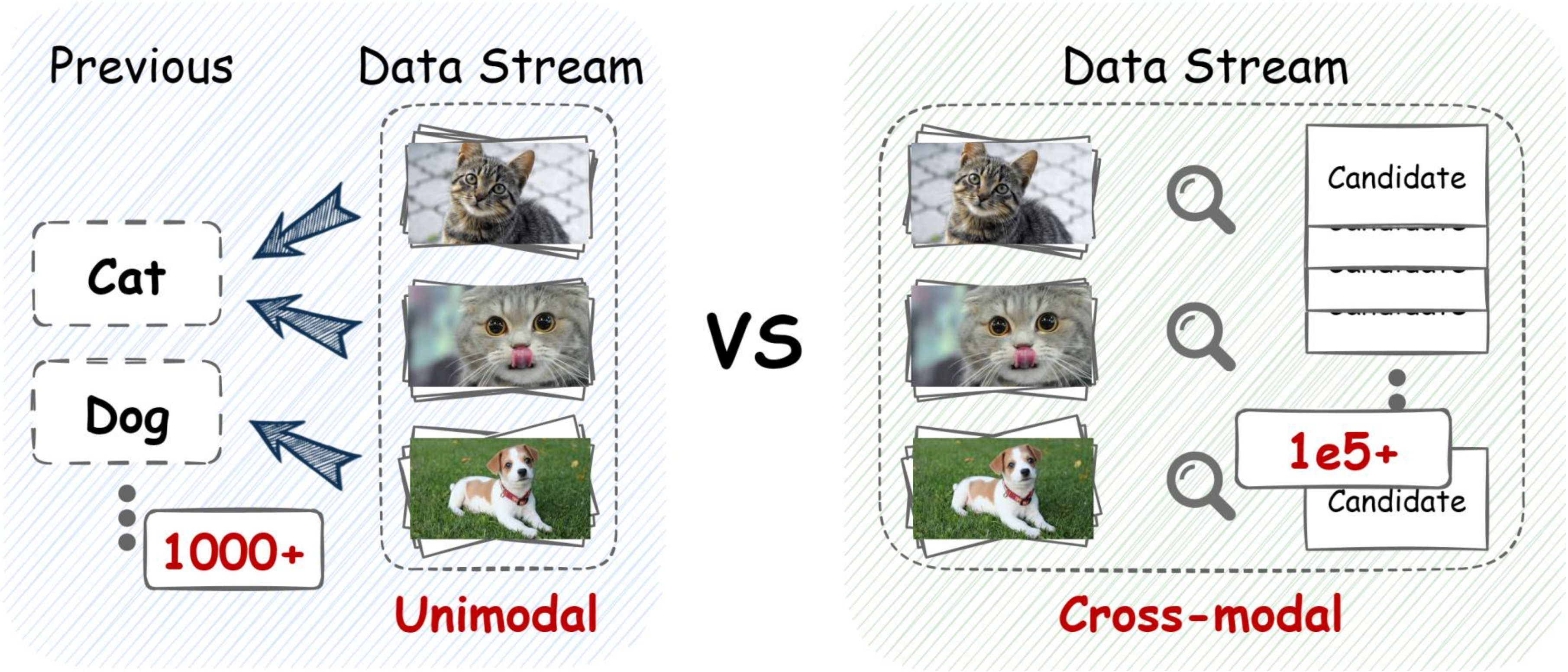
As one of the most effective paradigms in reconciling distribution shifts, **Test-Time Adaptation** (**TTA**) methods work by continually updating the given source model using the online target data stream.



Previous Works

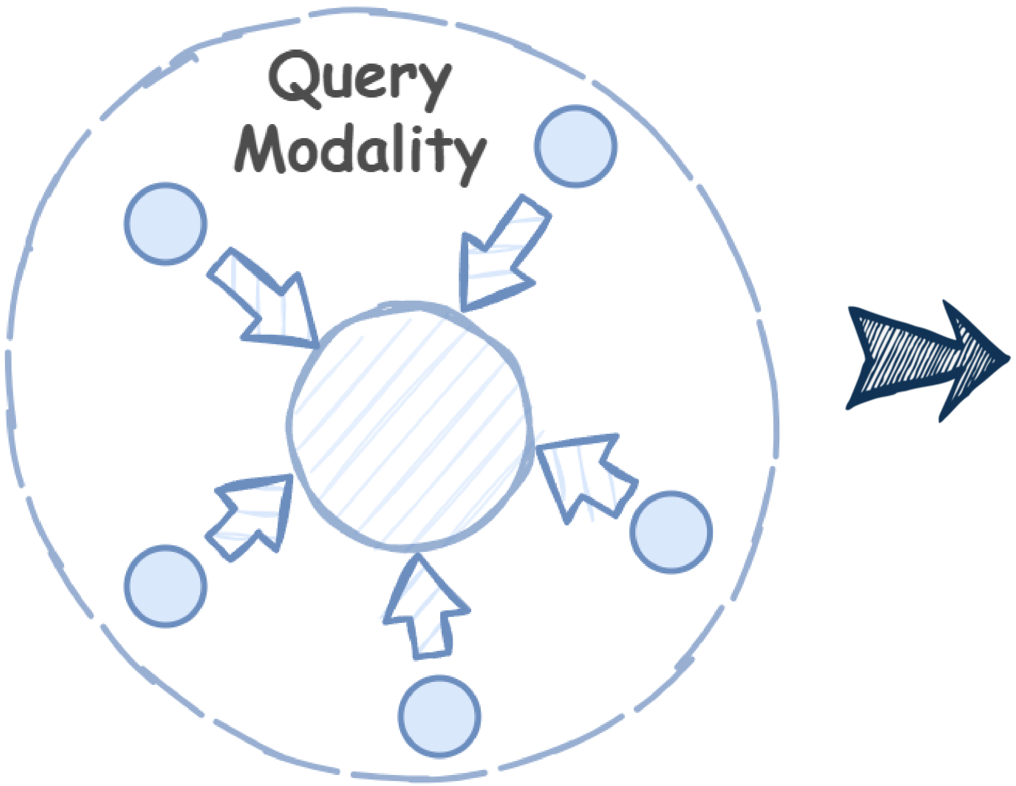
However, most existing TTA methods:

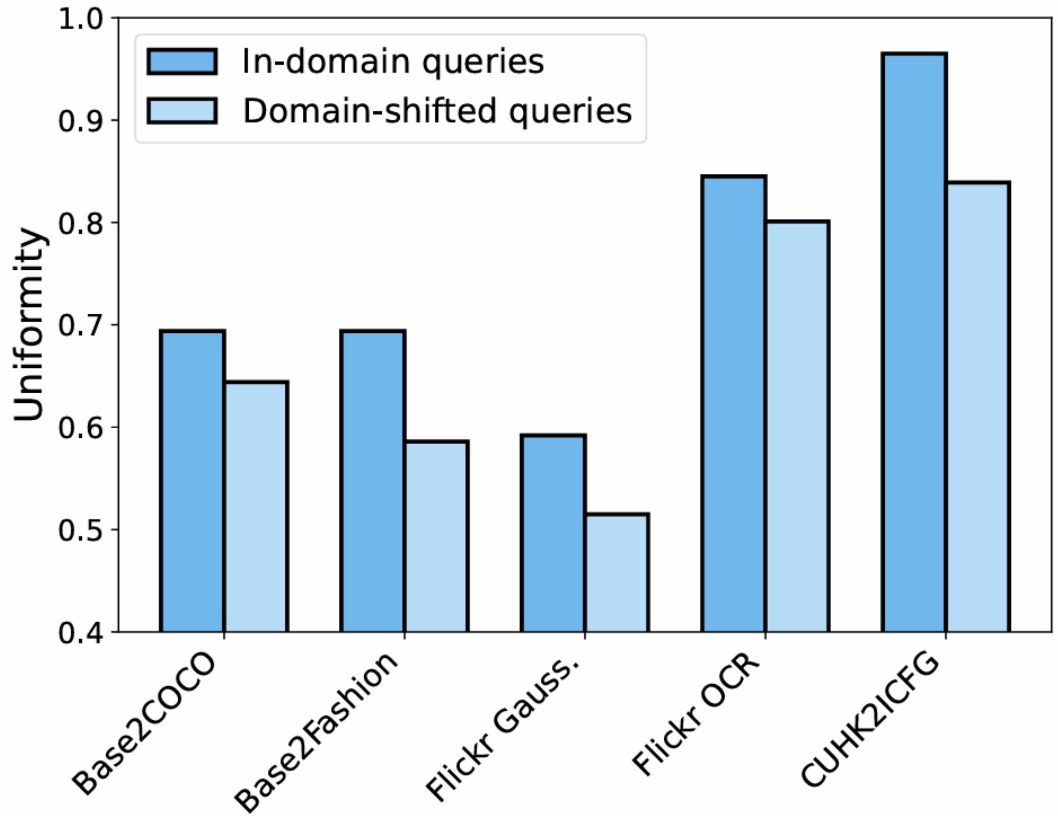
- Focus on the unimodal setting while **overlooking the complexity of the query shift in the cross-modal setting**, which would affect cross-modal relationship.
- Are specifically designed for the recognition task, which would **struggle with the heavy noise** from the query predictions if simply applied to the retrieval task.



Observation

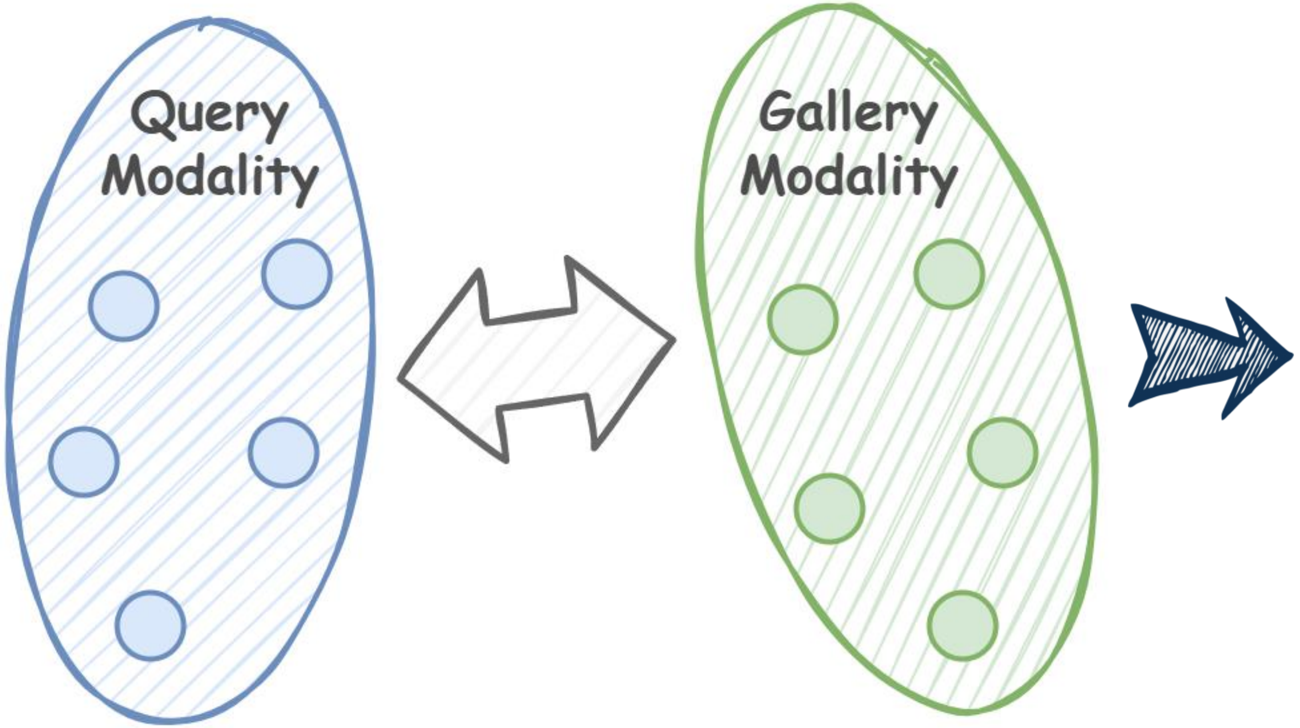
1. Query shift would **diminish the uniformity of the query modality**, prohibiting discrimination between diverse queries in the common space.

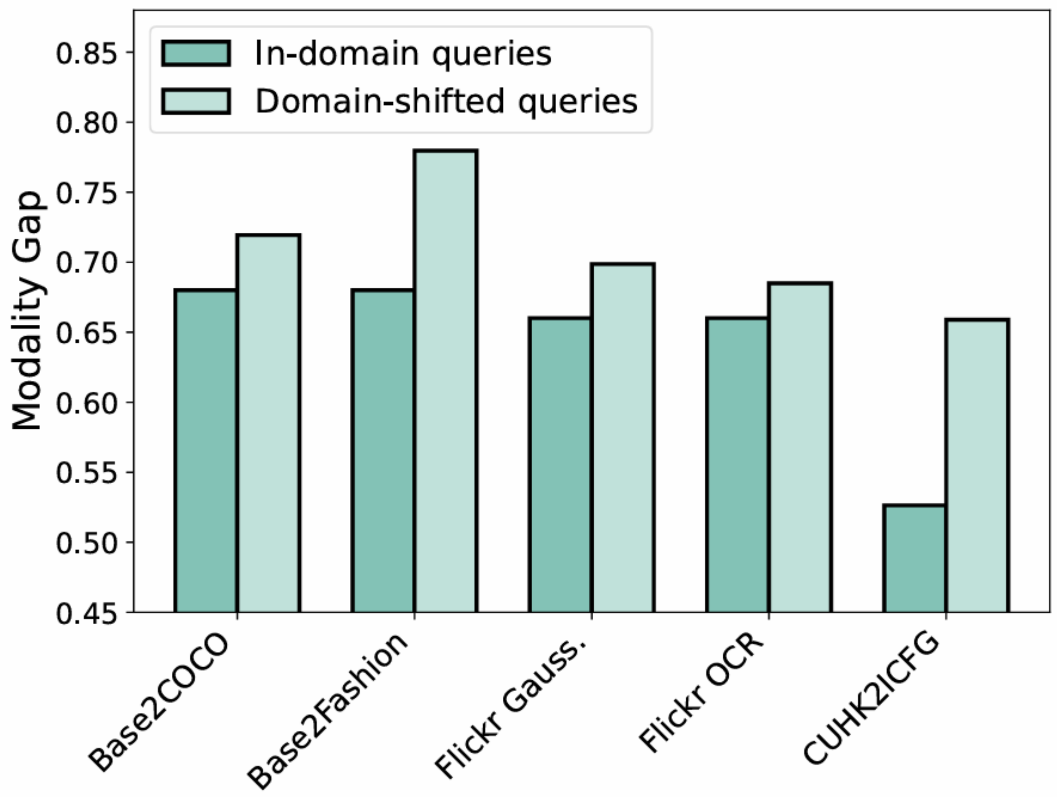




Observation

2. Query shift would **amplify the modality gap between query and gallery modalities**, undermining the well-constructed common space established by the pre-trained models.





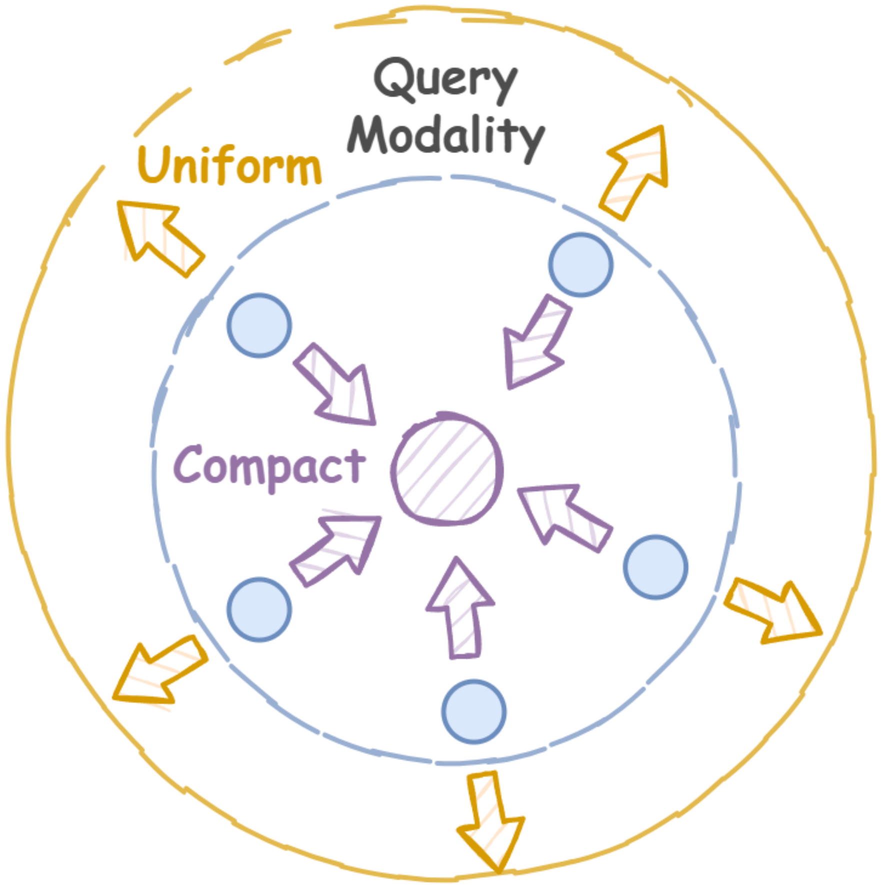
Key Idea

Challenge 1: negative impacts of query shift within and between modalities

- Investigate how the intra-modality uniformity affect the retrieval performance.

Intra-modality Uniformity

$$\left(\mathbf{z}_i^Q\right)^{\text{scale}} = \overline{\mathbf{Z}}^Q + \lambda^{\text{scale}}\left(\mathbf{z}_i^Q - \overline{\mathbf{Z}}^Q\right)$$



Key Idea

Challenge 1: negative impacts of query shift within and between modalities

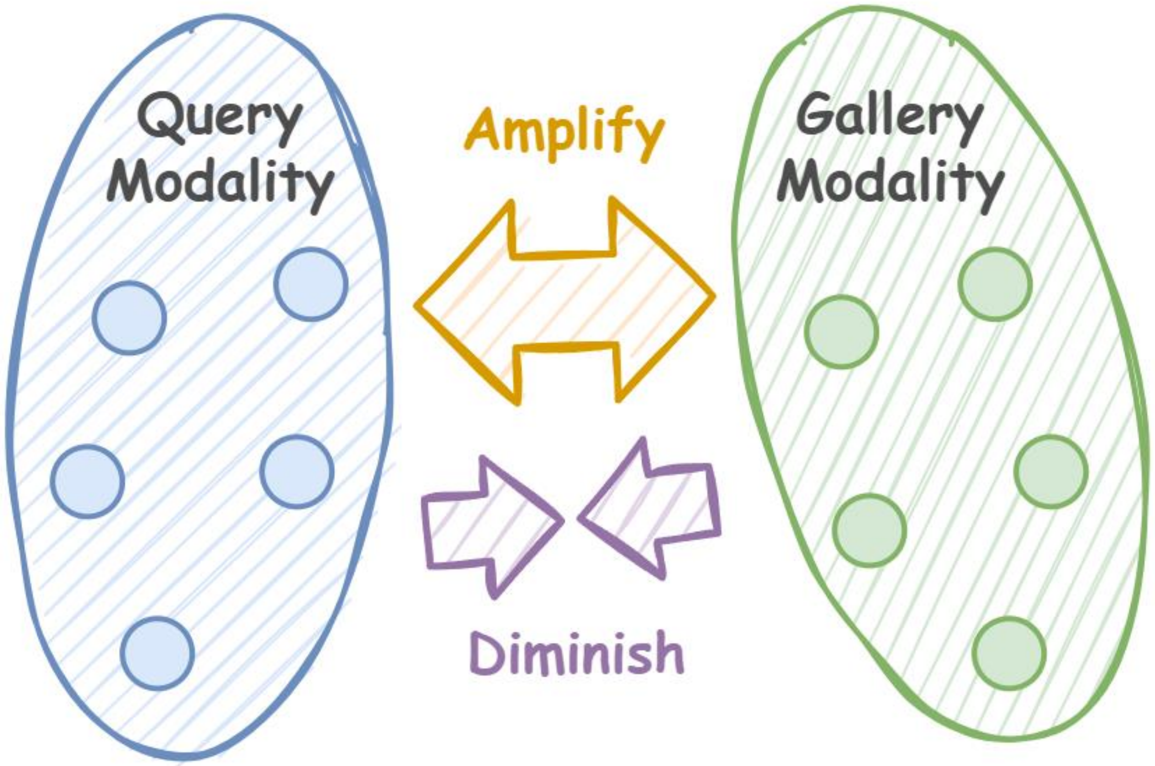
- Investigate how the inter-modality gap affect the retrieval performance

Intra-modality Uniformity

$$\left(\mathbf{z}_i^Q\right)^{\text{scale}} = \overline{\mathbf{Z}}^Q + \lambda^{\text{scale}}\left(\mathbf{z}_i^Q - \overline{\mathbf{Z}}^Q\right)$$

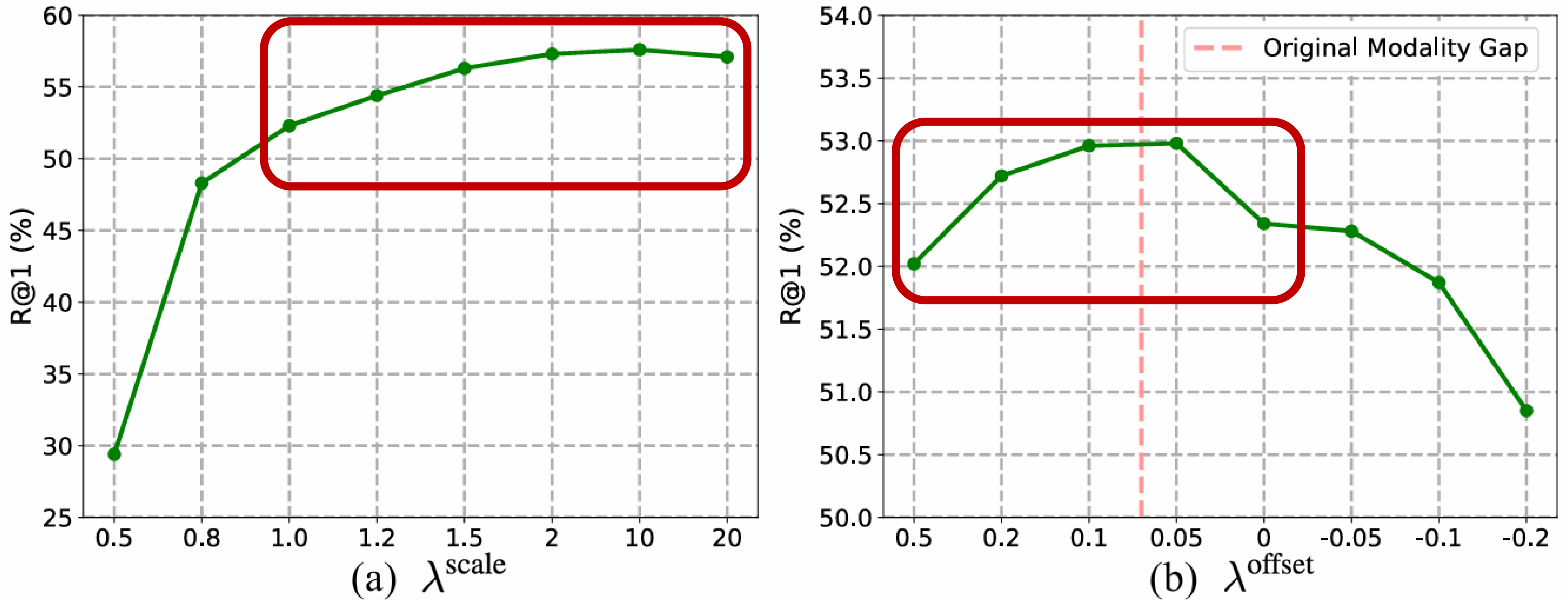
Inter-modality Gap

$$\left(\mathbf{z}_i^Q\right)^{\text{offset}} = \mathbf{z}_i^Q - \lambda^{\text{offset}}\left(\overline{\mathbf{Z}}^Q - \overline{\mathbf{Z}}^G\right)$$



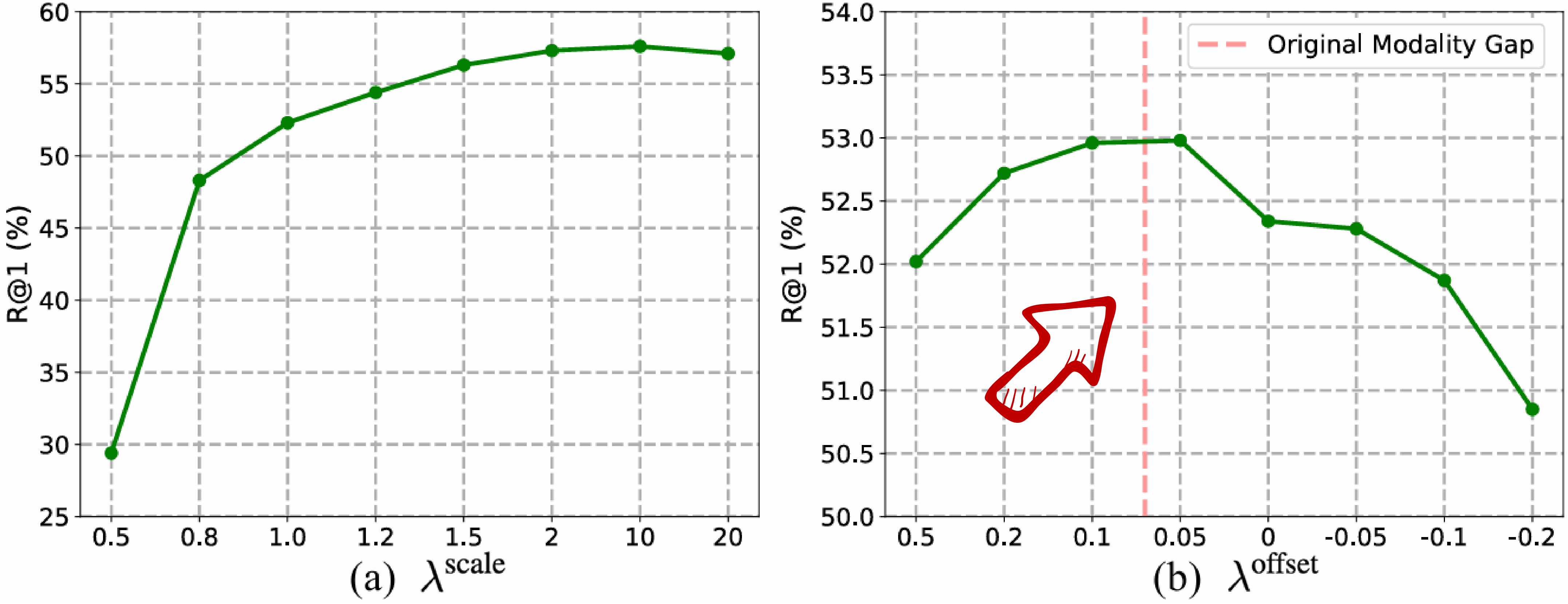
Key Idea

- Enlarging intra-modal uniformity $\lambda^{\mathbf{scale}} > \mathbf{1}$ and reducing inter-modal discrepancy $\lambda^{\mathbf{offset}} > \mathbf{0}$ would enhance retrieval performance, the reverse does not.



Key Idea

- *As discussed in [1], excessively eliminating inter-modal gap does not improve and may even degrade model performance.*
- **Modality gap of the source model** might be a good choice.



[1] Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning, NeurIPS, 2022.

Method

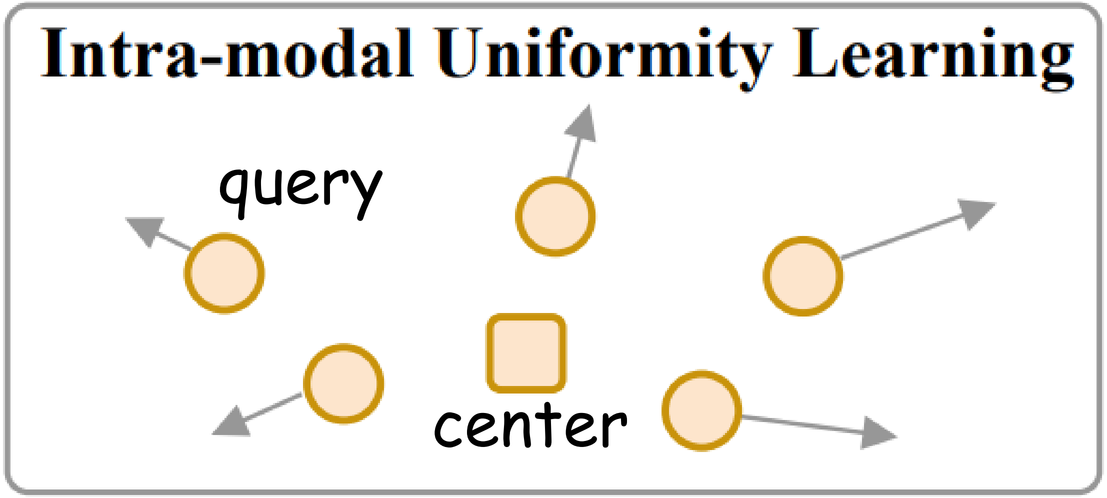
- Intra-modality Uniformity Learning

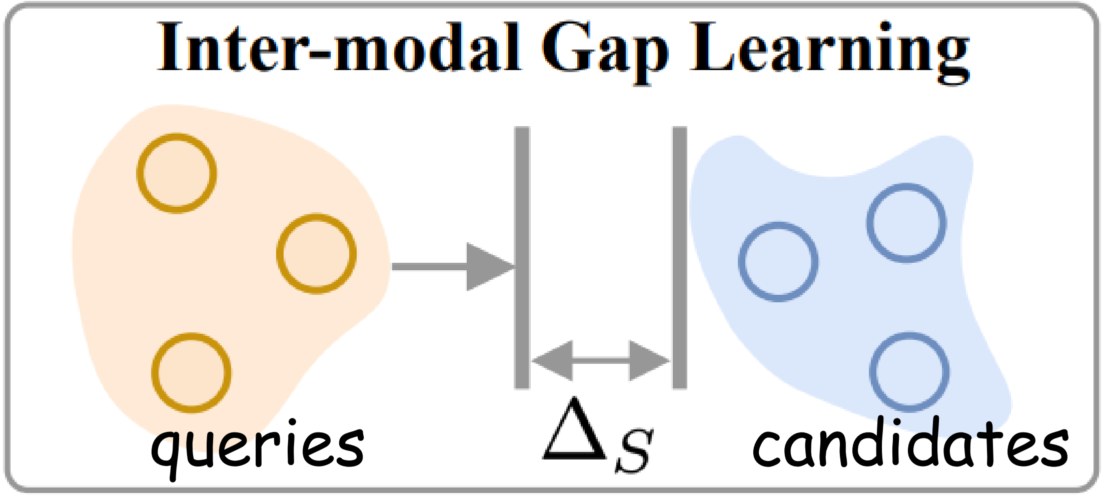
$$\mathcal{L}_{MU} = \frac{1}{B}\sum_{i}^{B} exp\left(-\|\mathbf{z}_i^Q - \overline{\mathbf{z}}^Q\|/t\right)$$

- Inter-modality Gap Learning

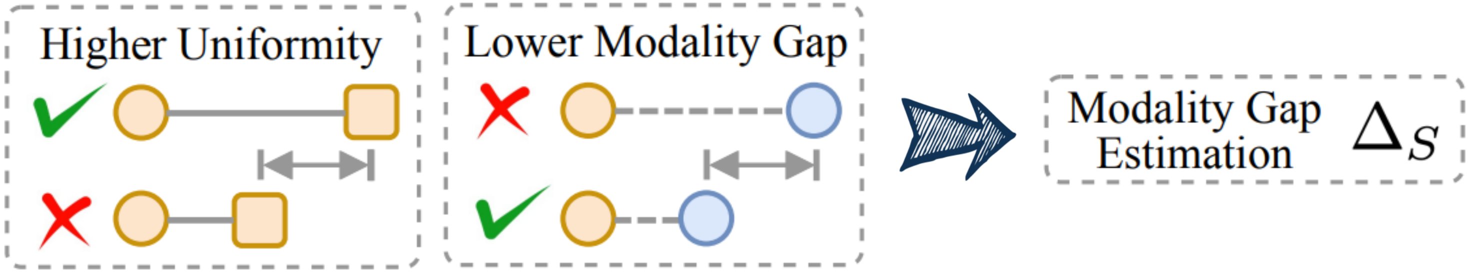
$$\mathcal{L}_{MG} = (\Delta_T - \Delta_S)^2$$

$$\Delta_T = \left\|\overline{\mathbf{z}}^Q - \overline{\mathbf{z}}^{G'}\right\|$$





Source-domain-like data



$$\Delta_S = \left\|\frac{1}{M}\sum_{i}^{M}\mathbf{z}_i^{Q_m} - \frac{1}{M}\sum_{j}^{M}\mathbf{z}_j^{G'_m}\right\|$$

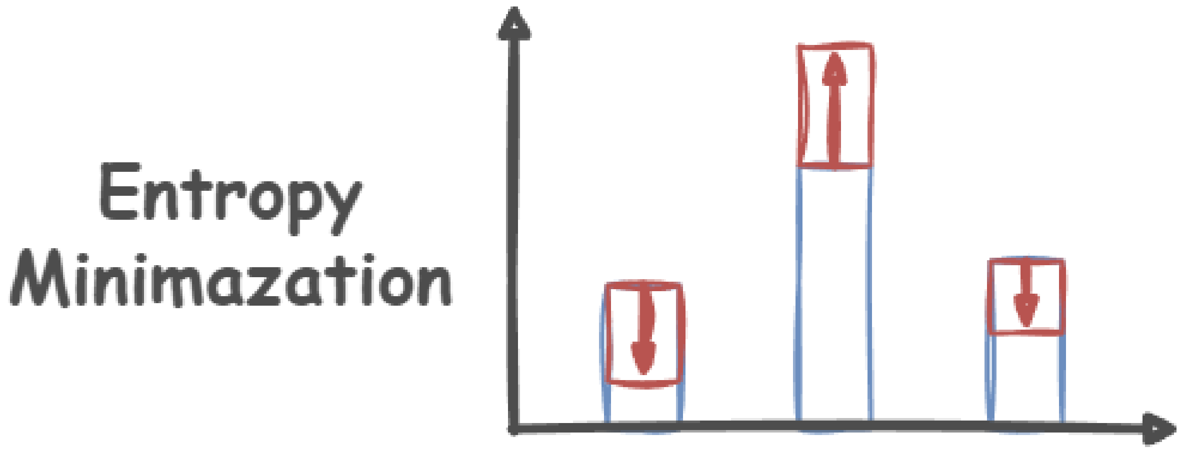
Method

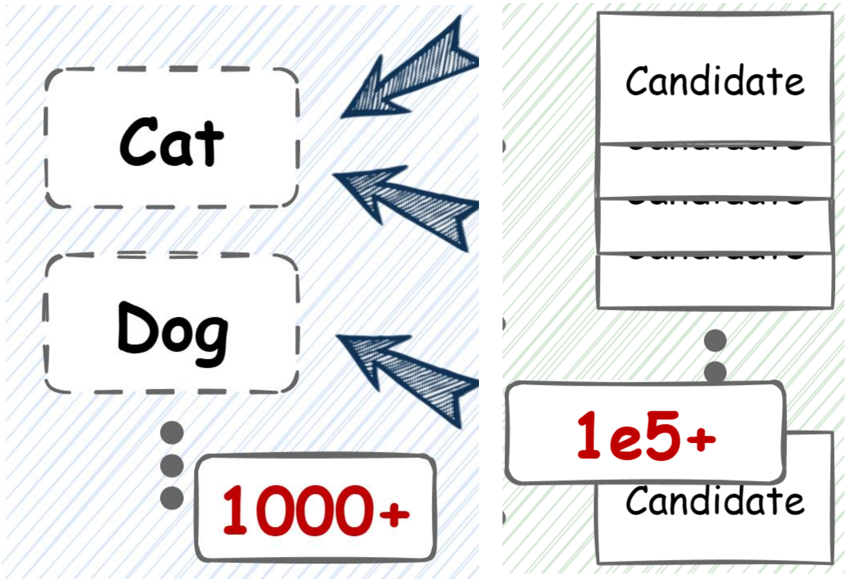
Challenge 2: query shift would result in **heavy noise** in cross-modal retrieval.

- Previous TTA methods rely on **entropy minimization** paradigm and are mainly designed for unimodal classification task.

$$\mathbf{p} = \mathrm{Softmax}\left(\mathbf{z}^Q \left(\mathbf{Z}^G\right)^T / \tau\right)$$

- Retrieval (N: 1e5+) vs Classification (K: 1000+)





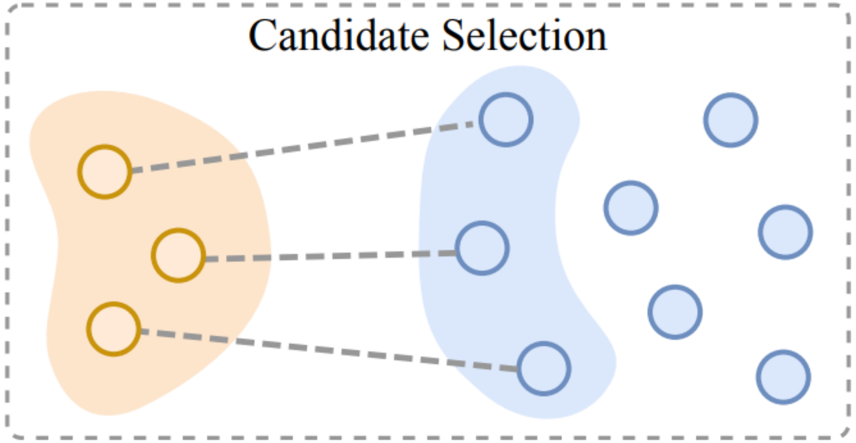
Method

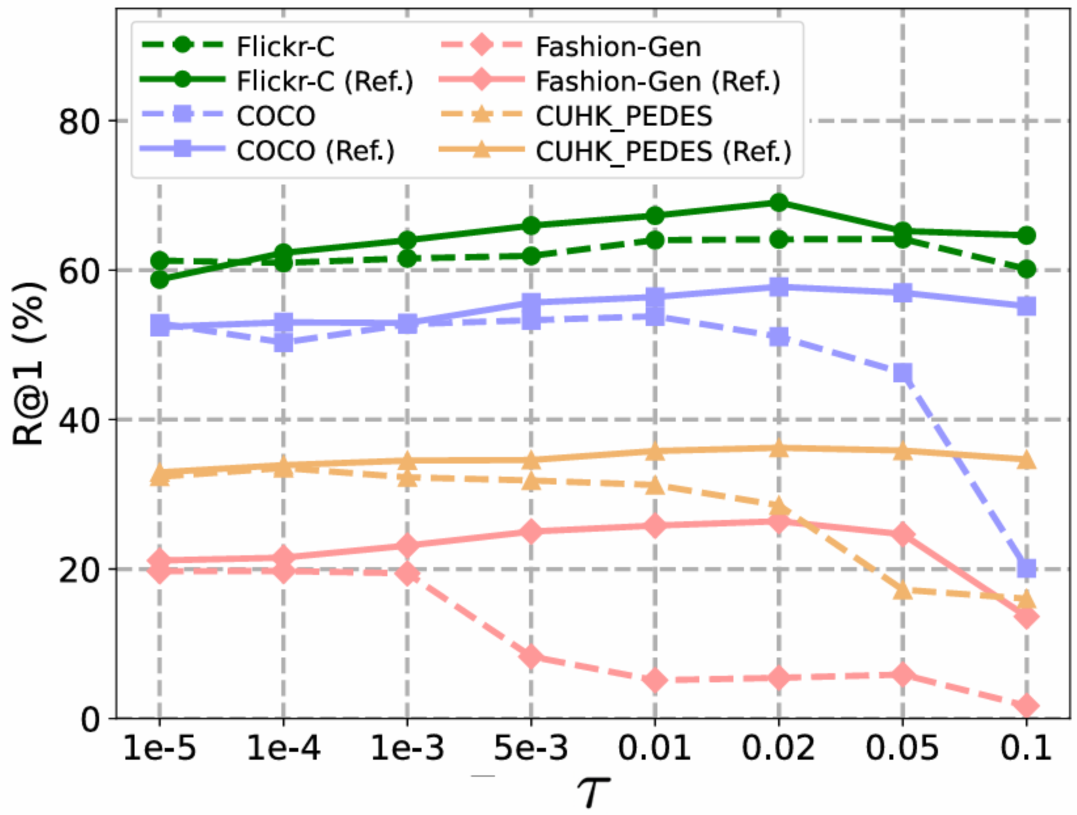
Query Prediction Refinement

$$\mathbf{x}_i^{G'} = \mathcal{N}(\mathbf{x}_i^{Q})$$

$$\hat{\mathbf{p}} = \mathrm{Softmax}\left(\mathbf{z}^{Q}\left(\mathbf{Z}^{G'}\right)^{T}/\tau\right)$$

- Exclude some irrelevant samples in the gallery, thus preventing the model from overfitting.
- The excluded irrelevant samples would avoid looking for a needle in a bottle of hay for queries, thus alleviating the model underfitting issue.



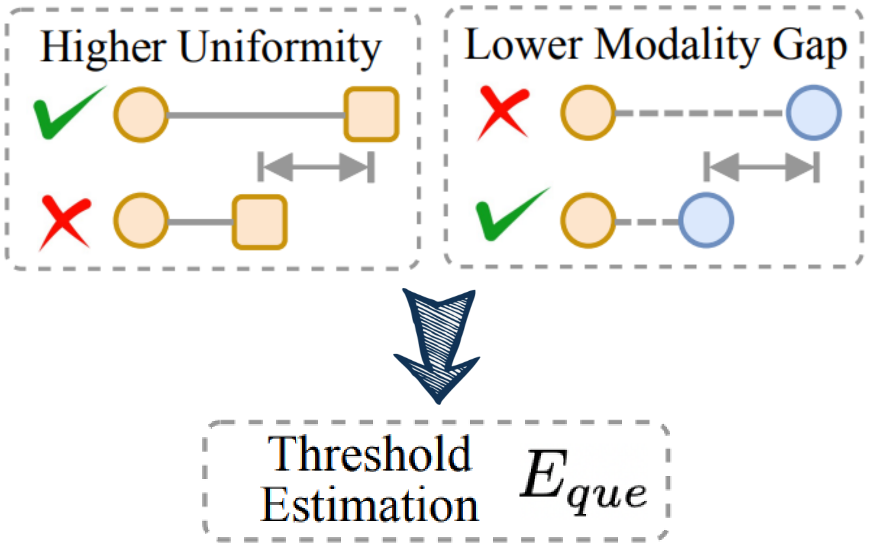


Method

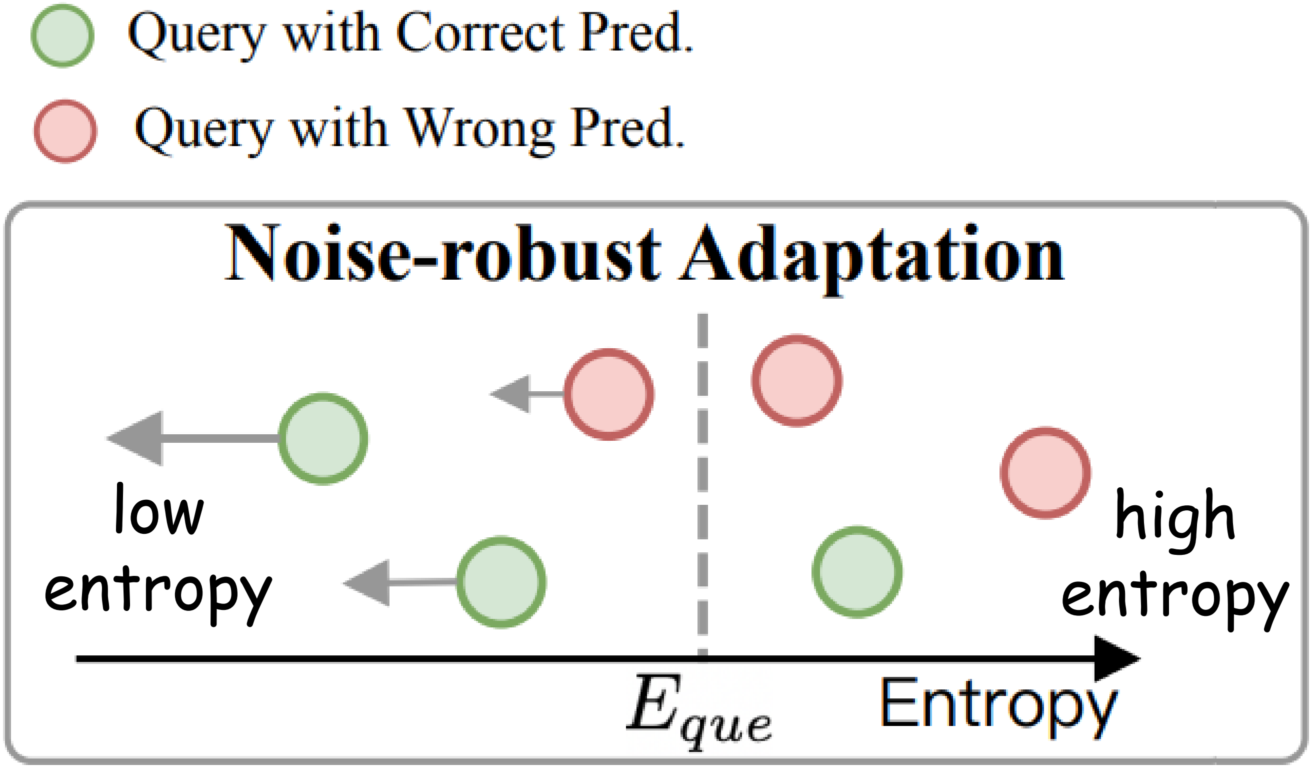
Noisy Robust Adaptation

$$\mathcal{L}_{NA} = \frac{1}{\sum_i \mathbb{I}_{\{S(\mathbf{x}_i^Q)\neq 0\}}} \sum_{i=1}^{N^Q} S(\mathbf{x}_i^Q) E(\mathbf{x}_i^Q), \text{ where } S(\mathbf{x}_i^Q) = \max\left(1 - \frac{E(\mathbf{x}_i^Q)}{E_m}, 0\right)$$

Source-domain-like data



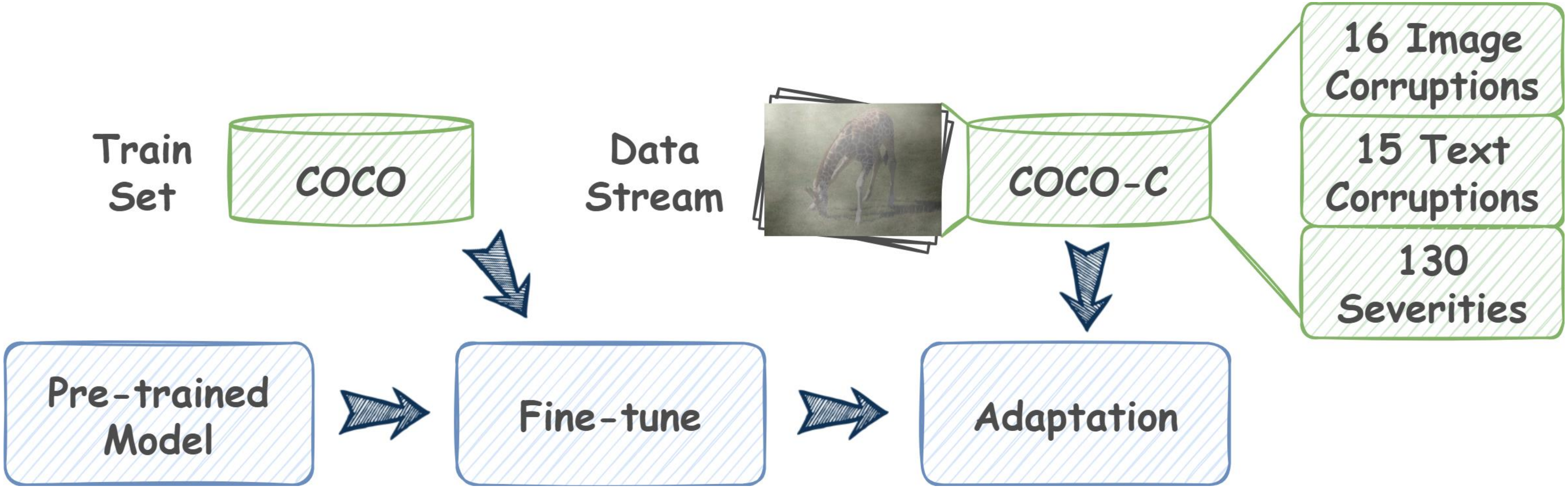
$$E_m = \max_{i=1,\ldots,M} E\left(\mathbf{x}_i^{Q_m}\right)$$



Experiments

Query Shift (QS): **only the queries** come from different distributions with the source-domain data.

- COCO-C benchmark
- Flickr-C benchmark



Experiments

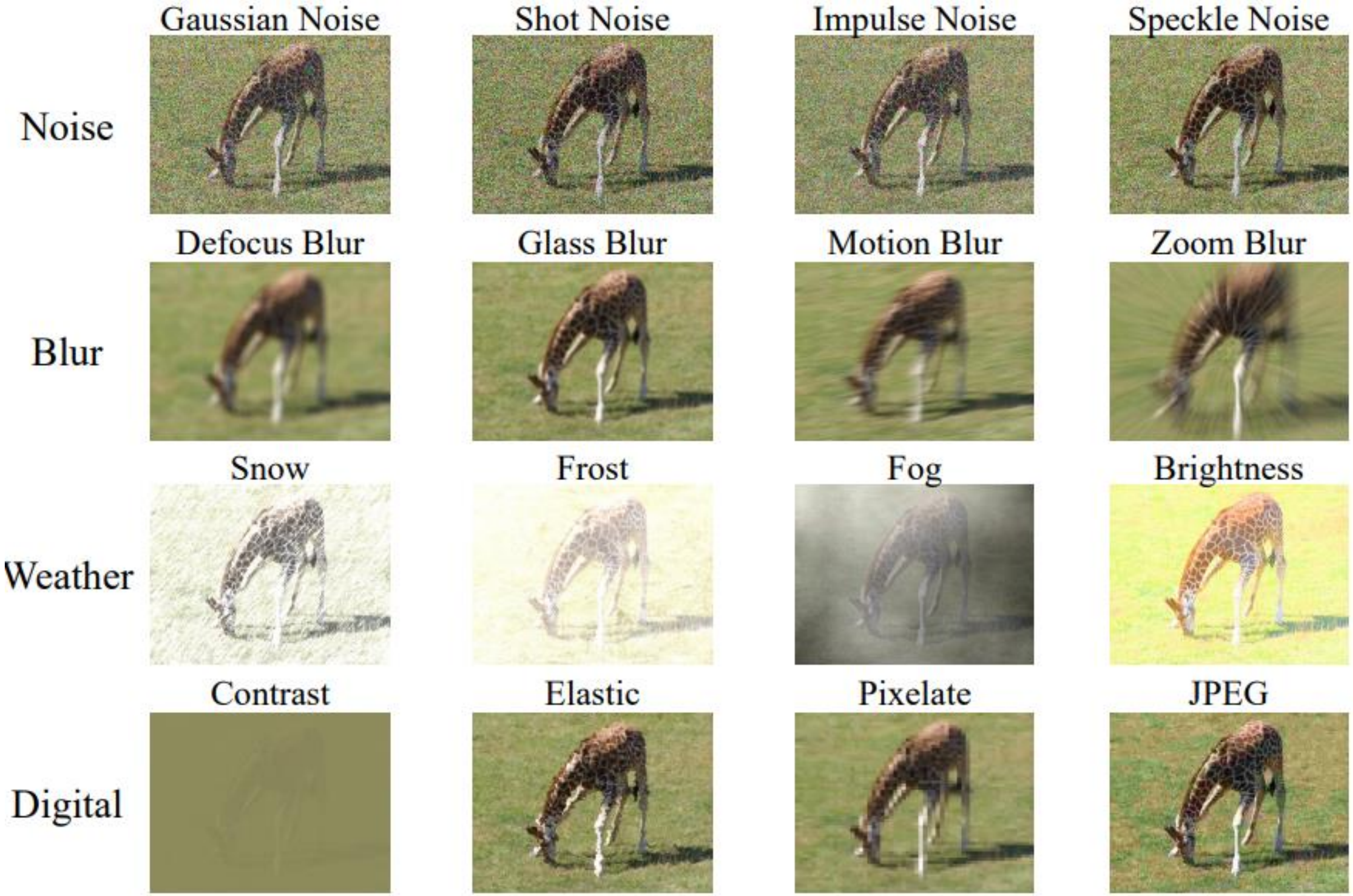


Image Corruptions

Experiments

Category	Perturbation	Example
Original	Clean	A train traveling down tracks next to a brick building.
Character	OCR	A train travelin9 down track8 next to a brick building.
	CI	A train traveling down traGcks next to a brick bui1lding.
	CR	A train traveling doPn tracks next to a brick buildirg.
	CS	A train rtaveilng down tracks next to a brick building.
	CD	A train tr[X]veling down tr[X]cks next to a brick building.
Word	SR	A train jaunt down running adjacent to a brick building.
	RI	A train pass traveling down tracks next to go a brick building
	RS	A building traveling down tracks next to a brick train.
	RD	A train [X] down tracks [X] to a brick building.
	IP	A : train traveling down tracks next to , a brick building.
Sentence	Formal	A train moving down tracks next to a brick building.
	Casual	A train that goes down tracks next to a brick building.
	Passive	Tracks next to a brick building are being traveled down by a train.
	Active	There is a train traveling down tracks next to a brick building.
	Backtrans	A train runs down the tracks next to a brick building.

Text Corruptions

Experiments

Query Shift (QS): Image2Text

robustness against severe query shift

	Noise				Blur				Weather				Digital				
Query Shift	Gauss.	Shot	Impul.	Speckle	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Avg.
BLIP ViT-B/16	43.4	46.3	43.2	57.3	43.3	68.0	39.7	8.4	32.3	52.2	57.0	66.8	36.0	41.3	20.6	63.7	45.0
• Tent	41.6	40.5	37.9	54.0	44.7	65.1	39.6	8.3	31.9	48.7	56.3	66.5	31.8	40.3	19.2	62.3	43.0
• EATA	41.4	50.3	35.7	63.1	49.8	72.2	46.2	6.9	45.6	56.7	62.5	71.4	43.6	51.3	25.6	67.0	49.3
• SAR	42.3	51.5	37.5	61.8	40.3	71.5	32.8	6.2	38.0	56.2	59.1	70.6	31.1	53.5	17.5	66.4	46.0
• READ	45.8	48.4	37.2	59.9	44.5	71.8	46.6	11.5	39.9	49.9	58.4	70.3	35.8	45.0	18.8	66.2	46.9
• DeYO	47.9	53.5	46.8	63.4	42.9	72.1	36.7	3.2	37.5	59.7	66.4	71.2	40.3	49.0	13.1	67.6	48.2
• Ours	**53.2**	**56.2**	**54.8**	**64.6**	**58.0**	**73.7**	**56.4**	**32.2**	**56.5**	**64.1**	**71.0**	**73.4**	**57.9**	**63.7**	**41.8**	**68.4**	**59.1**
BLIP ViT-L/16	50.3	51.8	51.1	61.6	53.7	72.1	49.4	14.5	44.0	57.5	61.8	70.5	37.3	50.6	32.0	70.5	51.8
• Tent	46.3	49.3	46.7	58.4	52.2	71.8	47.5	12.3	41.9	56.2	60.9	69.7	35.7	48.3	29.4	69.6	49.8
• EATA	46.2	53.5	49.5	63.8	56.5	73.8	52.6	18.4	50.6	59.1	64.5	72.1	40.7	55.4	43.5	70.7	54.4
• SAR	45.9	50.2	47.3	63.1	51.1	73.8	47.2	11.6	40.8	58.9	60.7	71.6	33.6	54.0	34.4	70.5	50.9
• READ	38.1	48.0	43.3	63.5	43.6	73.4	43.6	22.0	44.5	56.5	62.2	71.9	32.9	49.6	27.5	70.6	49.5
• DeYO	39.9	50.2	43.5	63.8	50.4	74.0	52.4	5.4	49.5	59.3	62.8	71.8	34.0	54.7	34.4	69.7	51.0
• Ours	**58.2**	**60.7**	**59.8**	**66.6**	**61.5**	**74.9**	**60.3**	**36.8**	**59.0**	**65.2**	**72.1**	**73.5**	**56.3**	**65.7**	**50.2**	**71.6**	**62.0**

marginal improvements

stable improvements

Experiments

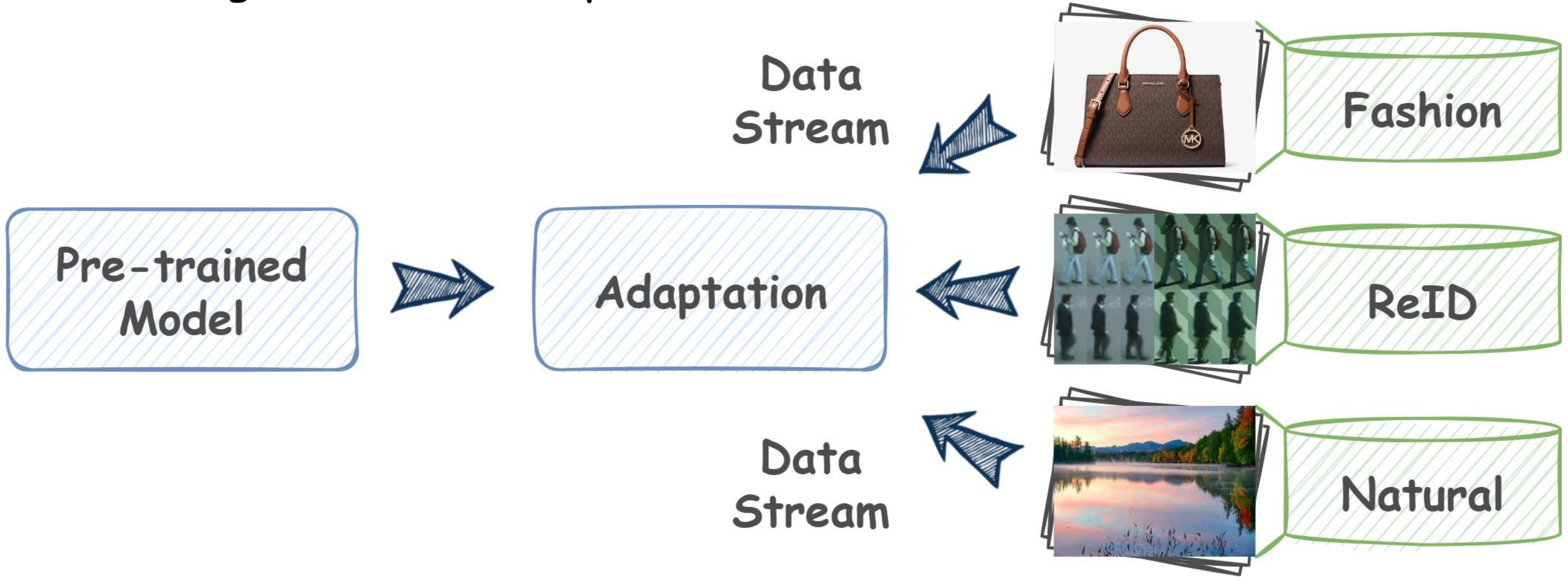
Query Shift (QS): Text2Image

	Character-level					Word-level					Sentence-level					
Query Shift	OCR	CI	CR	CS	CD	SR	RI	RS	RD	IP	Formal	Casual	Passive	Active	Backtrans	Avg.
BLIP ViT-B/16	31.4	11.3	9.4	18.9	11.4	43.6	51.5	50.3	50.6	56.8	56.6	56.2	54.9	56.8	54.2	40.9
• Tent	31.4	11.0	9.5	17.7	11.3	43.2	51.3	50.3	50.6	56.6	56.2	56.0	54.9	56.9	53.9	40.7
• EATA	33.1	11.9	10.5	18.4	12.0	44.9	53.0	51.6	50.3	56.2	56.8	**56.8**	**56.0**	56.8	54.3	41.5
• SAR	31.8	11.6	9.9	18.5	11.7	43.6	51.5	50.3	50.6	56.8	56.5	56.2	54.9	56.8	54.2	41.0
• READ	32.3	11.4	9.6	18.2	11.2	44.3	52.9	51.7	51.1	57.6	57.1	56.7	55.9	57.1	**54.7**	41.4
• DeYO	31.4	11.3	9.4	17.9	11.4	43.6	51.5	50.3	50.6	56.8	56.5	56.2	54.9	56.7	54.2	40.9
• Ours	**34.1**	**13.7**	**11.8**	**19.5**	**13.2**	**45.3**	**53.8**	**51.8**	**51.5**	**57.3**	**57.1**	**56.8**	**56.0**	**57.3**	**54.7**	**42.3**
BLIP ViT-L/16	34.5	12.3	11.1	19.7	12.9	46.0	54.4	54.0	53.5	59.4	59.1	58.8	57.8	59.4	56.7	43.3
• Tent	34.0	12.3	11.0	19.6	12.9	46.5	54.2	53.8	53.4	59.4	59.1	58.8	57.6	58.9	56.5	43.2
• EATA	35.6	13.3	11.3	20.3	13.2	47.2	55.4	54.2	53.8	59.2	59.1	59.4	57.9	59.4	56.8	43.7
• SAR	34.5	13.1	11.2	20.3	13.1	46.7	54.4	54.0	53.5	59.5	59.1	58.8	57.8	59.4	56.7	43.5
• READ	35.3	12.2	10.9	19.1	12.7	47.3	55.1	55.0	53.3	59.7	59.3	**59.1**	58.1	**59.6**	56.7	43.6
• DeYO	34.5	12.3	11.1	19.7	12.9	46.7	54.4	54.0	53.5	59.5	59.1	58.8	57.8	59.4	56.7	43.4
• Ours	**36.8**	**14.7**	**13.4**	**21.3**	**14.3**	**47.9**	**56.3**	**54.8**	**53.9**	**59.5**	**59.4**	59.0	**58.2**	**59.6**	**56.9**	**44.4**

Experiments

Query-Gallery Shift (QGS): both the query and gallery samples are drawn from distributions different from the source-domain data.

- E-commerce domain: Fashion-Gen
- ReID domain: CUHK-PEDES, ICFG-PEDS
- Natural image domain: Nocaps, COCO, Flickr



Experiments

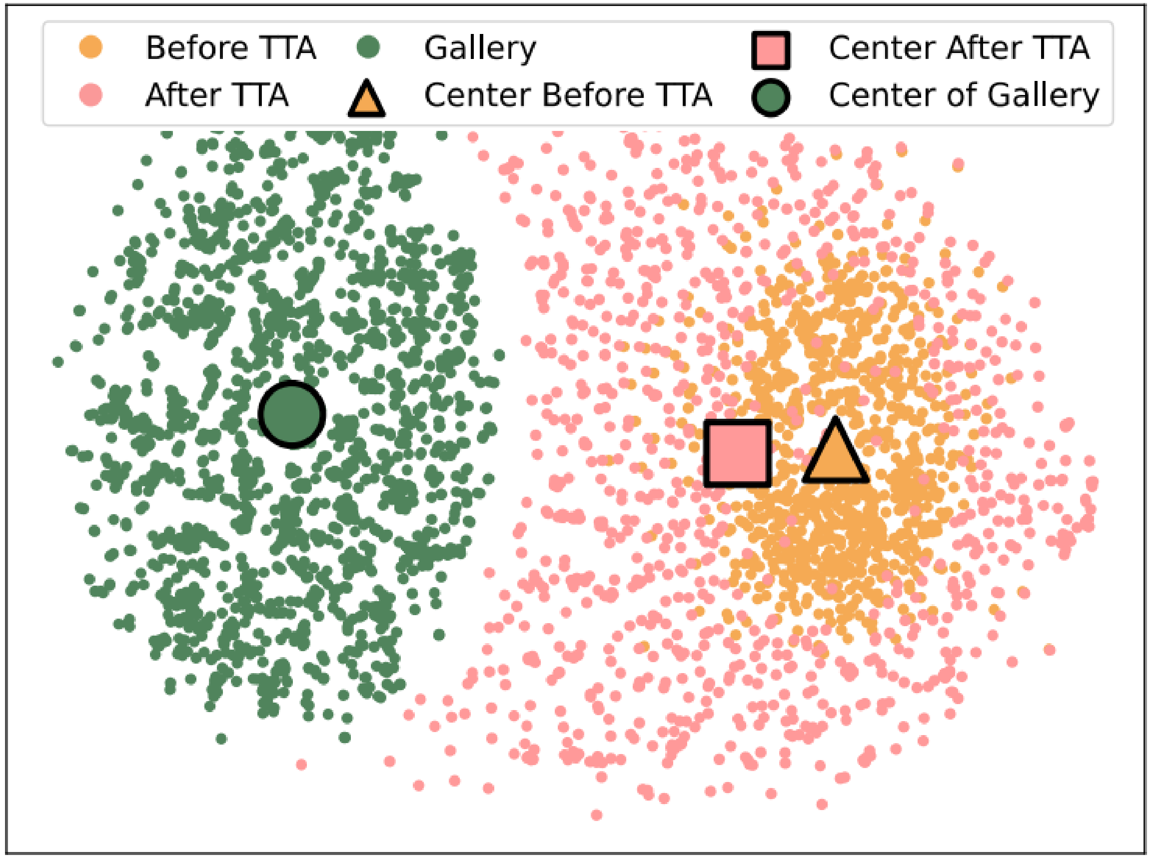
Query-Gallery Shift (QGS): **both the query and gallery** samples are drawn from distributions different from the source-domain data.

	Gallery size increases						In domain -> Out-domain						
	Base2Flickr		Base2COCO		Base2Fashion		Base2Nocaps(ID)		Base2Nocaps(ND)		Base2Nocaps(OD)		
Query Shift	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1	IR@1	Avg.
CLIP ViT-B/16	80.2	61.5	52.5	33.0	8.5	13.2	84.9	61.4	75.4	49.2	73.8	55.8	54.1
• Tent	81.4	64.0	48.8	27.6	5.6	10.7	**85.1**	61.7	74.6	48.6	71.8	56.1	53.0
• EATA	80.4	63.4	52.1	34.8	8.1	12.0	84.7	62.0	75.1	52.3	74.1	56.9	54.7
• SAR	80.3	62.2	51.8	33.9	8.0	13.3	84.7	61.3	75.4	51.3	73.7	56.1	54.3
• READ	80.6	64.4	46.0	35.7	5.8	11.2	**85.1**	63.0	75.0	52.1	73.5	57.0	54.1
• DeYO	80.1	64.0	51.5	33.4	6.9	10.9	84.4	62.2	75.1	52.0	73.2	57.3	54.3
• Ours	**82.4**	**64.8**	**52.9**	**36.5**	**8.9**	**14.0**	**85.1**	**63.5**	**75.7**	**54.0**	**74.4**	**58.0**	**55.9**
BLIP ViT-B/16	70.0	68.3	59.3	45.4	19.9	26.1	88.2	74.9	79.3	63.6	81.9	67.8	62.1
• Tent	81.9	68.5	61.7	41.7	14.1	26.1	88.5	75.4	82.6	64.1	82.7	68.9	63.0
• EATA	82.3	69.4	64.2	47.9	12.8	25.2	87.8	75.1	82.8	63.9	81.5	67.9	63.4
• SAR	81.7	68.3	63.5	46.6	17.9	26.1	88.2	75.6	81.0	65.4	81.2	69.3	63.7
• READ	80.0	69.9	62.1	46.4	5.6	24.1	87.3	75.1	80.6	63.9	80.7	67.9	62.0
• DeYO	83.5	69.9	65.0	47.3	12.2	24.1	89.2	75.6	83.7	65.7	84.3	69.4	64.2
• Ours	**86.8**	**70.3**	**68.9**	**48.9**	**23.6**	**30.3**	**89.7**	**76.0**	**86.3**	**66.1**	**87.2**	**69.5**	**67.0**

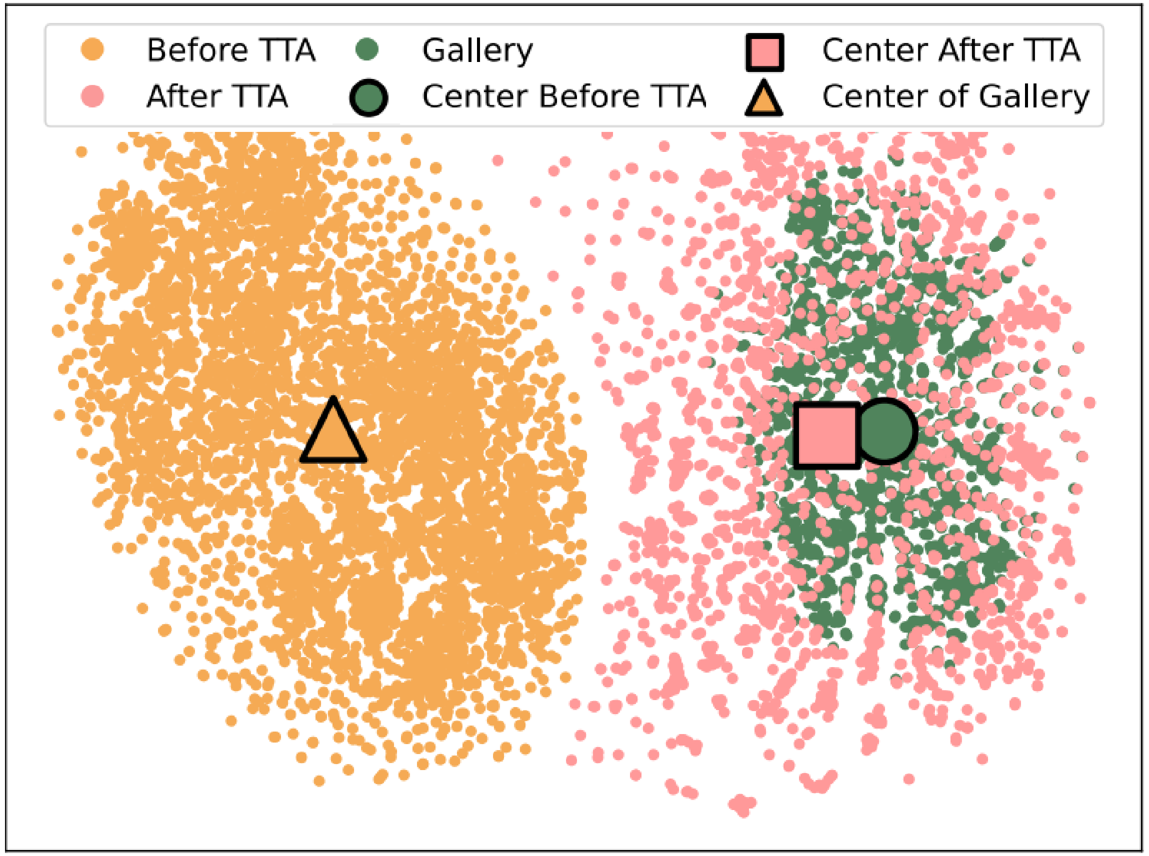
Experiments

Visualization Result

Image2Text



Text2Image



Experiments

An alternative implementation of TCR without training

Input: Test samples $\mathcal{D}_T = \left\{ \{\mathbf{x}_i^Q\}_{i=1}^{N^Q}, \{\mathbf{x}_j^G\}_{j=1}^{N^G} \right\}$, the source model f_{Θ_s}, batch size B, scaling factor λ^{scale}.

Output: Predictions $\{\mathbf{p}_i\}_{i=1}^{N^Q}$.

Initialize $\tilde{\Theta}_0 = \Theta_s$;

for *given queries* $\mathbf{x}^Q \in \mathcal{D}_T$ **do**

- Select a subset of candidates $\mathbf{x}^{G'}$ from the gallery using Eq. 4 ; `// Candidate Selection`
- `// Update the queue`
- Compute the criterion SI in Eq. 6;
- Select the 30% query-candidate pairs with the smallest SI;
- Maintain a queue of size B to save the pairs;
- Scale $\mathbf{x}^Q$ using Eq. 12 with λ^{scale} ; `// Scaling up Intra-modality Uniformity`
- Estimate the modality gap Δ_S using Eq. 7 ; `// Constraint Estimation`
- Rectify the modality gap to Δ_S using Eq. 13 ; `// Rectifying between-modality Gap`
- Perform $\ell 2$-normalization on the embeddings in the query modality;
- Obtain the query predictions $\mathbf{p}$ in Eq. 1;

end

		Noise				Blur				Weather				Digital				
Dataset	Query Shift	Gauss.	Shot	Impul.	Speckle	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Avg.
Flickr-C	EATA	55.5	60.5	55.8	75.8	64.6	86.2	52.2	8.5	72.0	83.7	82.5	87.9	68.4	60.1	45.9	81.6	65.1
	Ours (untrain)	58.7	63.2	58.1	78.8	65.9	87.8	61.2	34.6	79.2	84.8	84.4	89.1	68.2	67.4	46.0	83.0	69.4
	Ours	**62.0**	**66.6**	**61.4**	**80.0**	**68.1**	**87.9**	**65.2**	**39.9**	**78.2**	**85.2**	**85.7**	**89.5**	**75.1**	**73.1**	**56.8**	**83.3**	**72.4**
COCO-C	EATA	41.4	50.3	35.7	63.1	49.8	72.2	46.2	6.9	45.6	56.7	62.5	71.4	43.6	51.3	25.6	67.0	49.3
	Ours (untrain)	48.8	51.7	49.8	61.5	53.9	72.6	49.4	18.7	49.7	60.5	67.1	71.4	43.9	49.9	26.7	67.4	52.7
	Ours	**53.2**	**56.2**	**54.8**	**64.6**	**58.0**	**73.7**	**56.4**	**32.2**	**56.5**	**64.1**	**71.0**	**73.4**	**57.9**	**63.7**	**41.8**	**68.4**	**59.1**

Experiments

Examples in real life: personalized queries in e-commerce domain

	Query Shift	TOPS	SWEATERS	JACKETS	PANTS	JEANS	SHIRTS	DRESSES	SHORTS	SNEAKERS	SKIRTS	Avg.
TR	CLIP ViT-B/32	18.0	19.3	19.9	12.0	5.5	18.3	38.1	**17.9**	37.3	29.6	21.6
	• Ours	**22.9**	**25.2**	**21.6**	**14.3**	**6.0**	**22.8**	**44.3**	8.5	**41.7**	**37.4**	**24.5**
IR	CLIP ViT-B/32	24.9	27.9	29.2	16.9	6.7	25.4	51.8	25.7	47.1	47.8	30.3
	• Ours	**28.2**	**31.7**	**32.8**	**19.5**	**9.6**	**28.5**	**57.1**	**29.1**	**53.6**	**50.7**	**34.1**

Conclusions

- From the perspectives of intra-modal distribution and inter-modal alignment relationship, we reveal the **underlying impacts of query shift** on cross-modal retrieval.

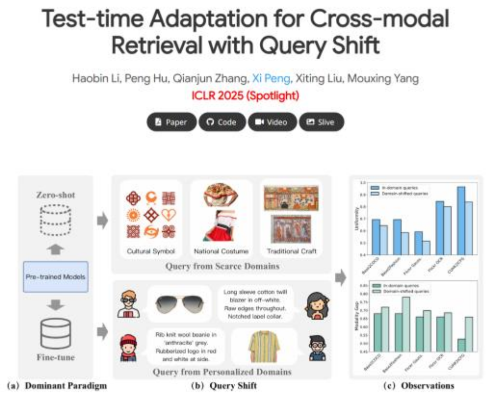
- **Extend TTA to cross-modal retrieval**. TCR not only manipulates both the modality uniformity and modality gap but also prevents the model from overfitting noisy query predictions, thus achieving robust adaptation.

- **Benchmark the existing TTA methods** on cross-modal retrieval with query shift across **six datasets and 130 diverse corruptions** of varying severity. The proposed TCR **supports mainstream pre-trained models**, including BLIP and CLIP.

Thanks for your attention!

Project Page

https://hbinli.github.io/TCR/



Code

https://github.com/XLearning-SCU/2025-ICLR-TCR