MICC
Media Integration and Communication Center

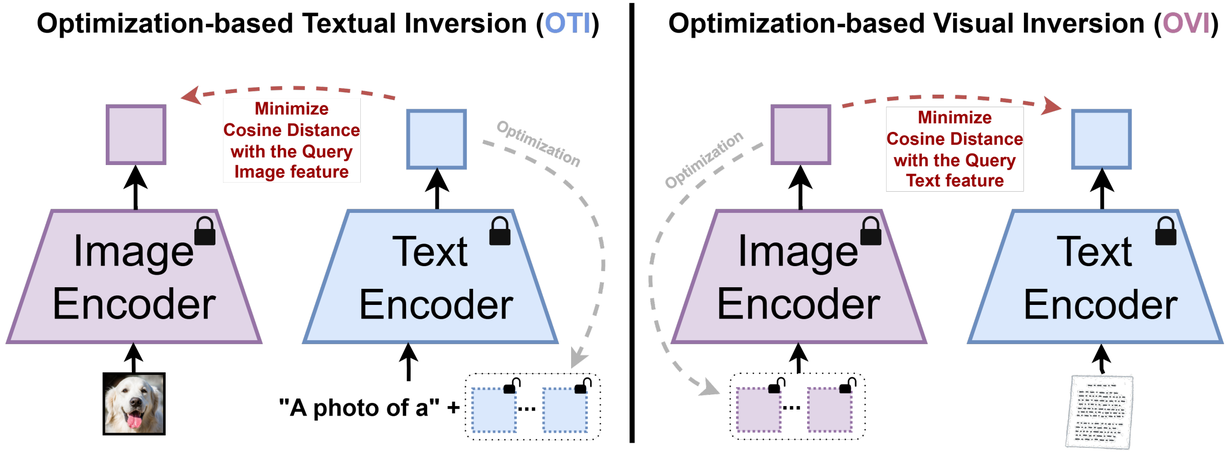
ICLR

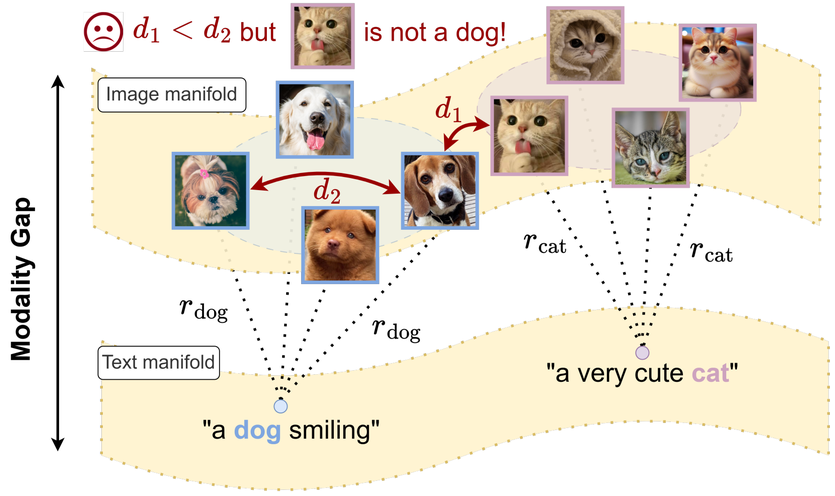




Cross the Gap: Exposing the Intra-modal Misalignment in CLIP via Modality Inversion

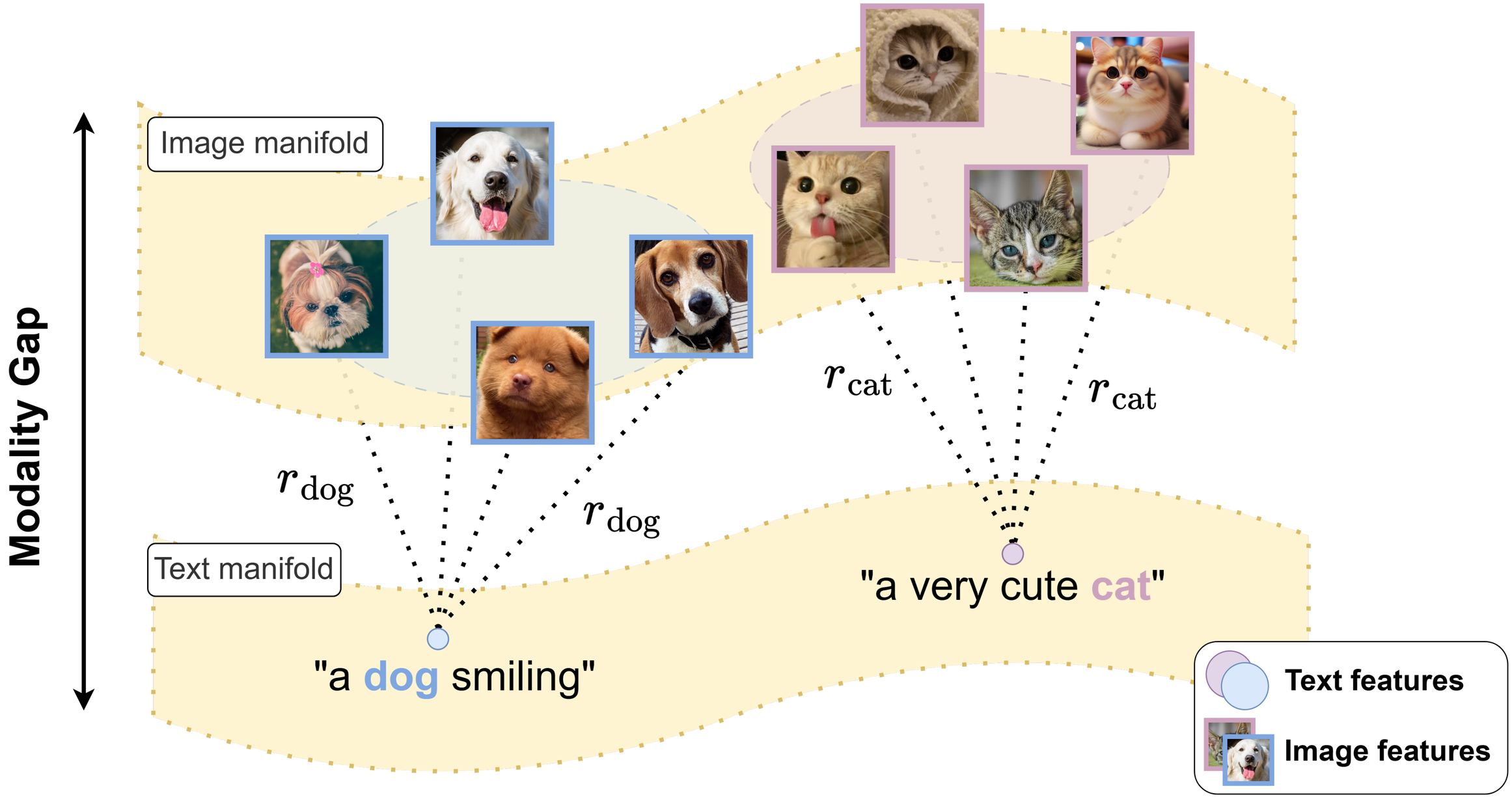
*Marco Mistretta, *Alberto Baldrati, *Lorenzo Agnolucci, Marco Bertini, Andrew D. Bagdanov



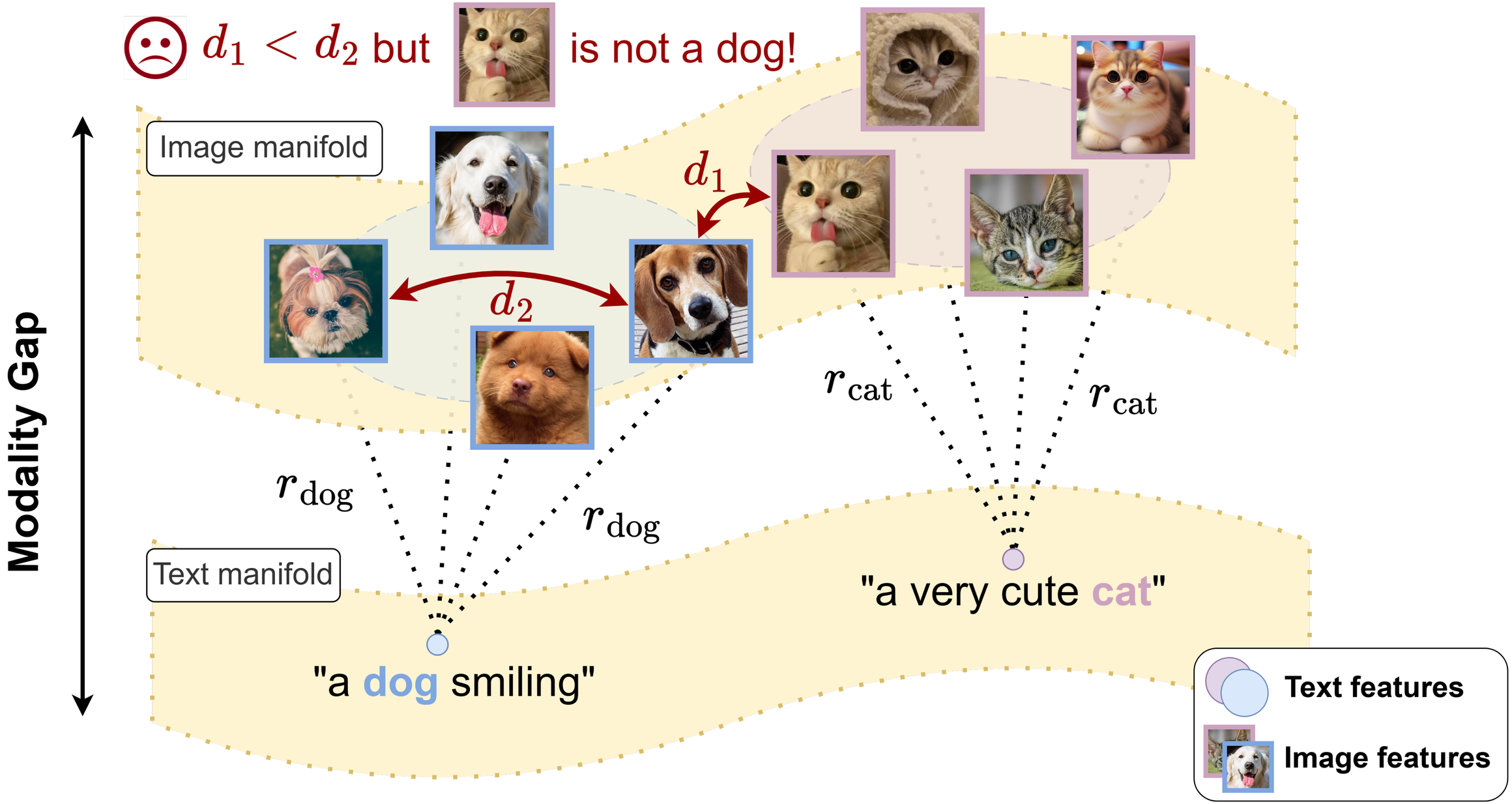


TLDR; *We adapt **OTI**, and we introduce **OVI**...* *...in order to expose **CLIP Intra-modal Misalignment***

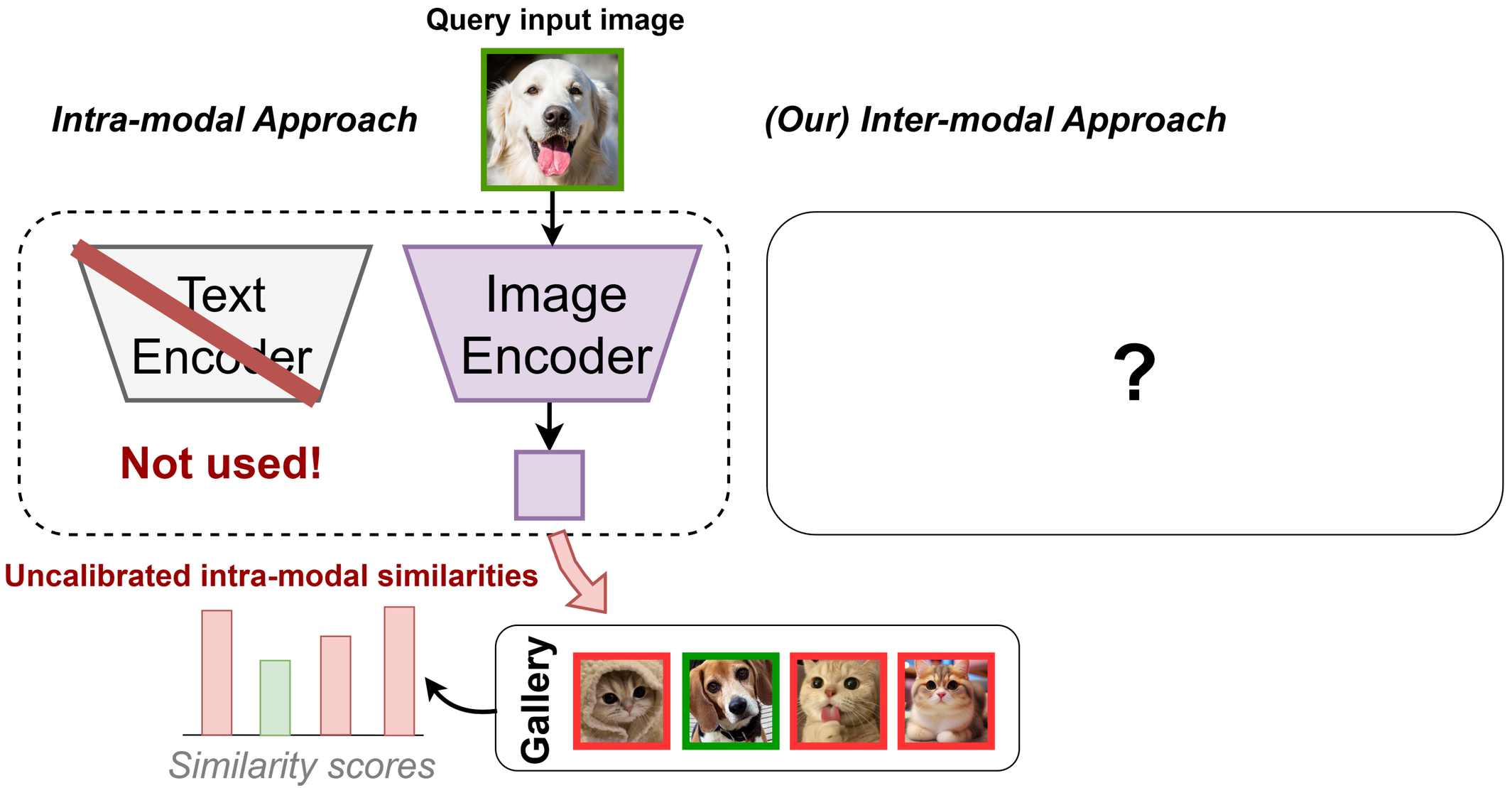
Exposing the Intra-modal misalignment



Exposing the Intra-modal misalignment

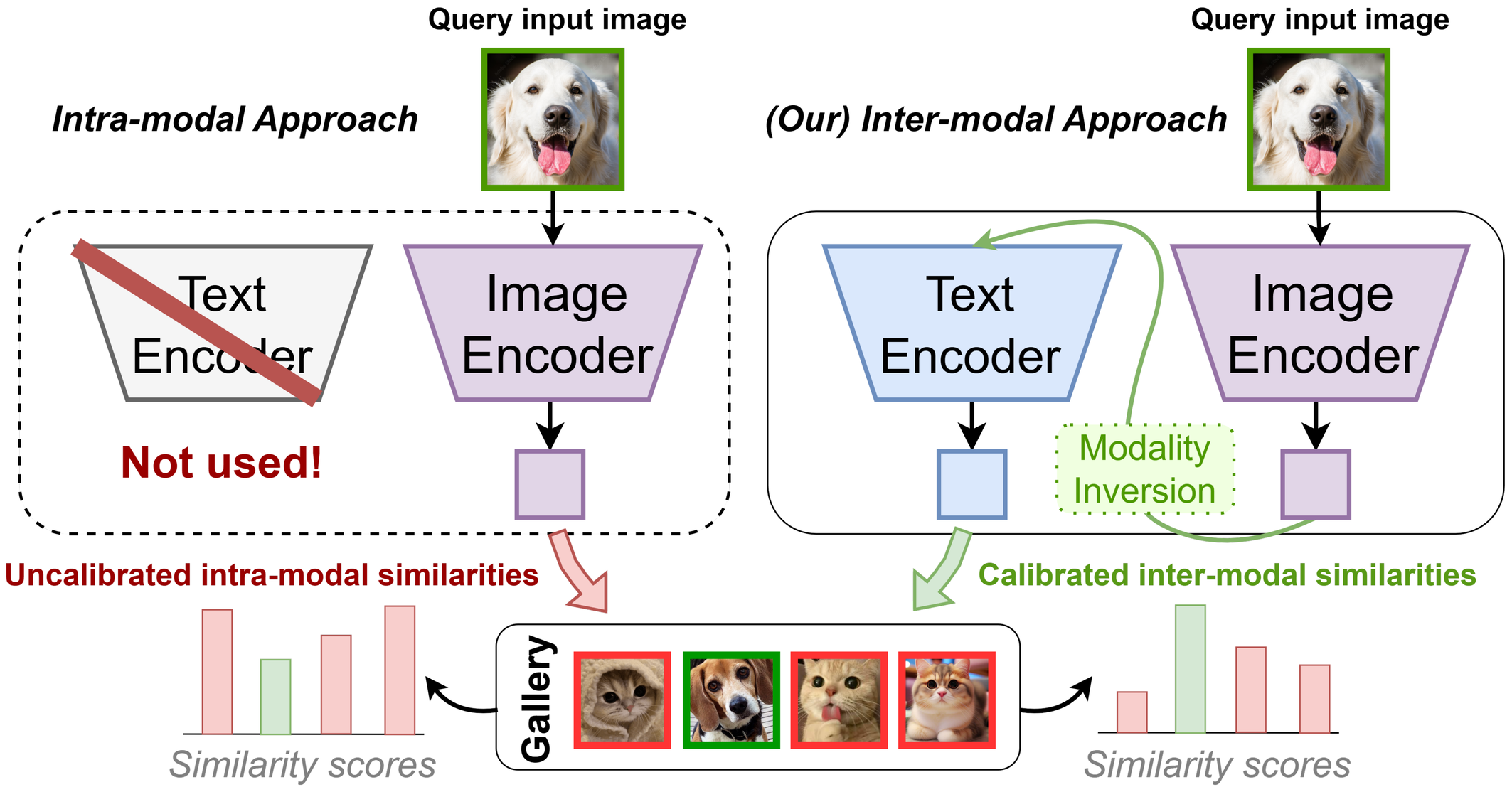


From Intra-modal to Inter-modal via Modality Inversion

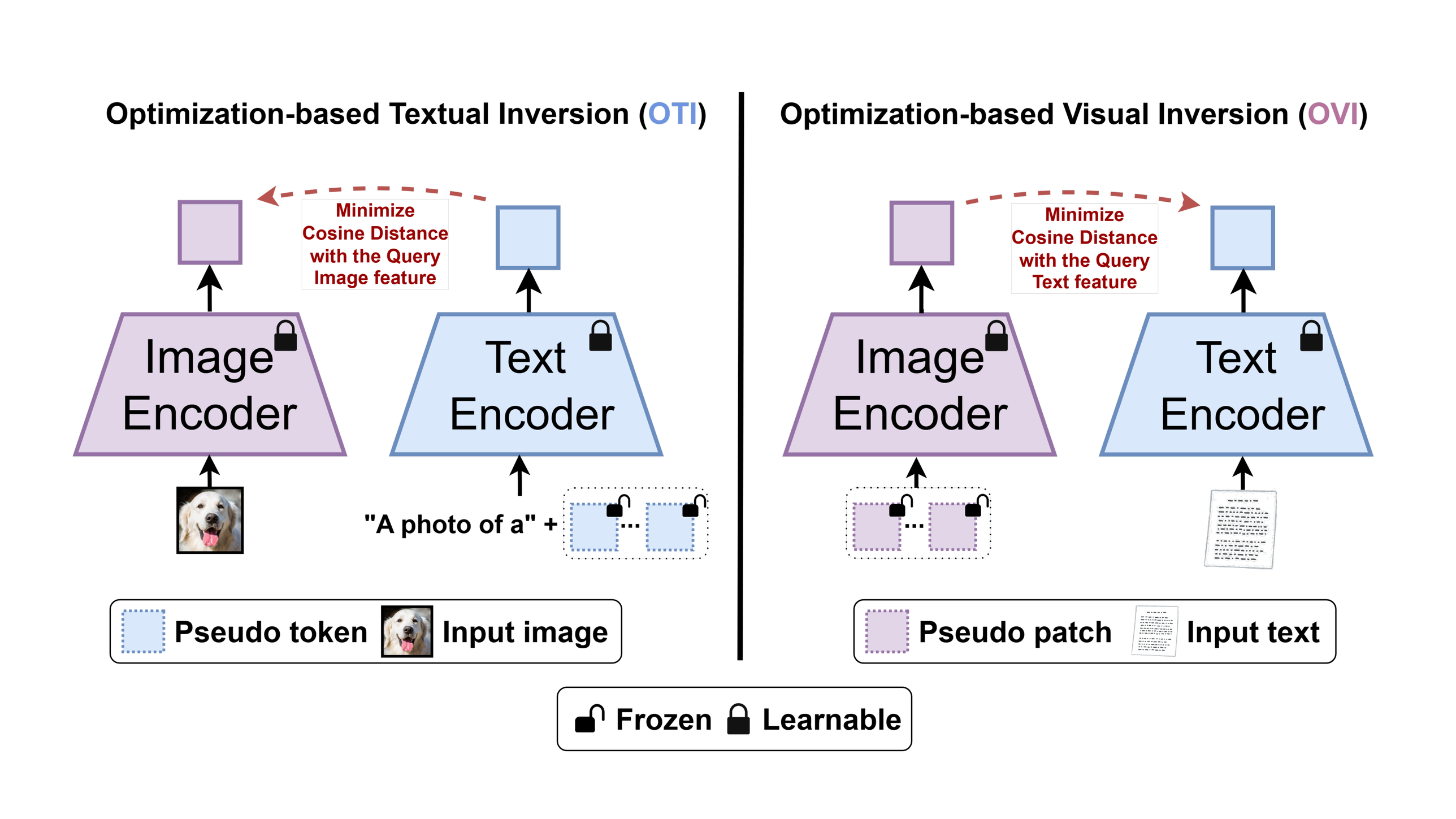


↑*Example for the intra-modal task of image-to-image retrieval. Note that text-to-text retrieval can be approached analogously*

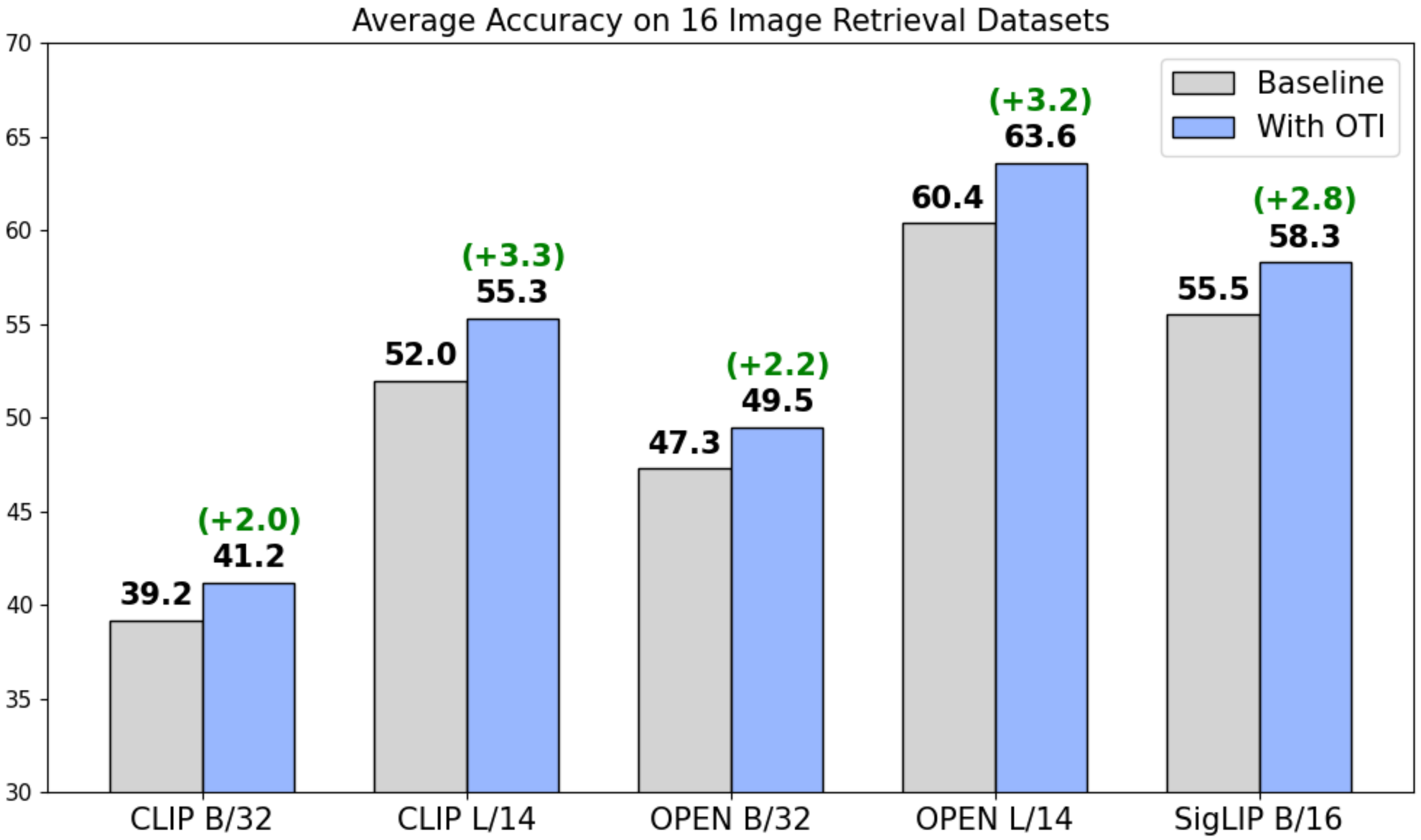
From Intra-modal to Inter-modal via Modality Inversion



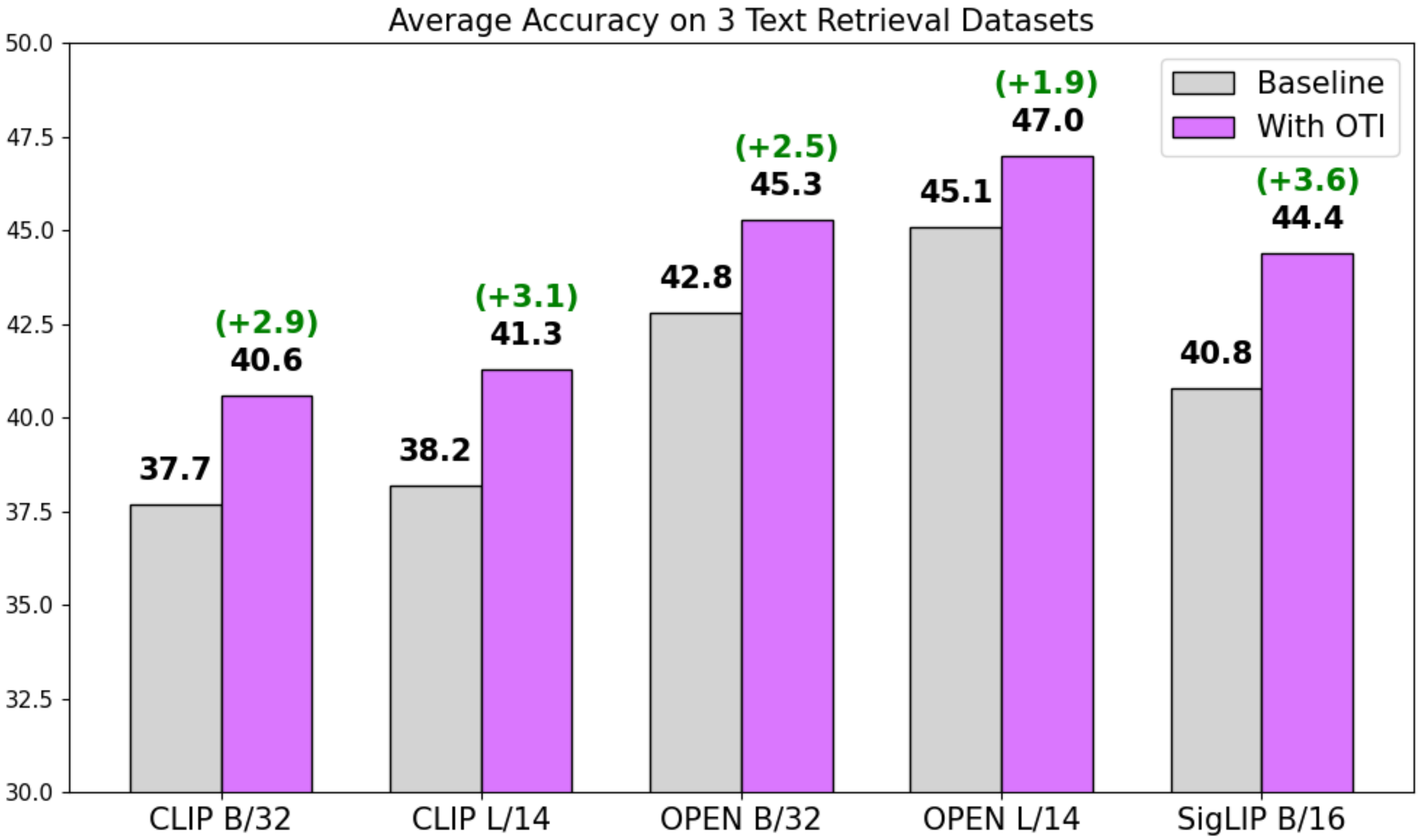
↑*Example for the intra-modal task of image-to-image retrieval. Note that text-to-text retrieval can be approached analogously*


Optimization-based Textual Inversion (OTI)
Minimize Cosine Distance with the Query Image feature
Image Encoder
Text Encoder
"A photo of a" +
Optimization-based Visual Inversion (OVI)
Minimize Cosine Distance with the Query Text feature
Image Encoder
Text Encoder
Pseudo token
Input image
Pseudo patch
Input text
Frozen
Learnable

Approaching Image-to-image retrieval inter-modally with OTI



Approaching Text-to-text retrieval inter-modally with OVI



Cross the Gap:
Exposing the Intra-modal Misalignment in CLIP via Modality Inversion

*Marco Mistretta, *Alberto Baldrati, *Lorenzo Agnolucci, Marco Bertini, Andrew D. Bagdanov

Thank you for your attention!

arXiv

GitHub

Presented by Marco Mistretta

marco.mistretta@unifi.it

MICC
Media Integration and Communication Center

ICLR