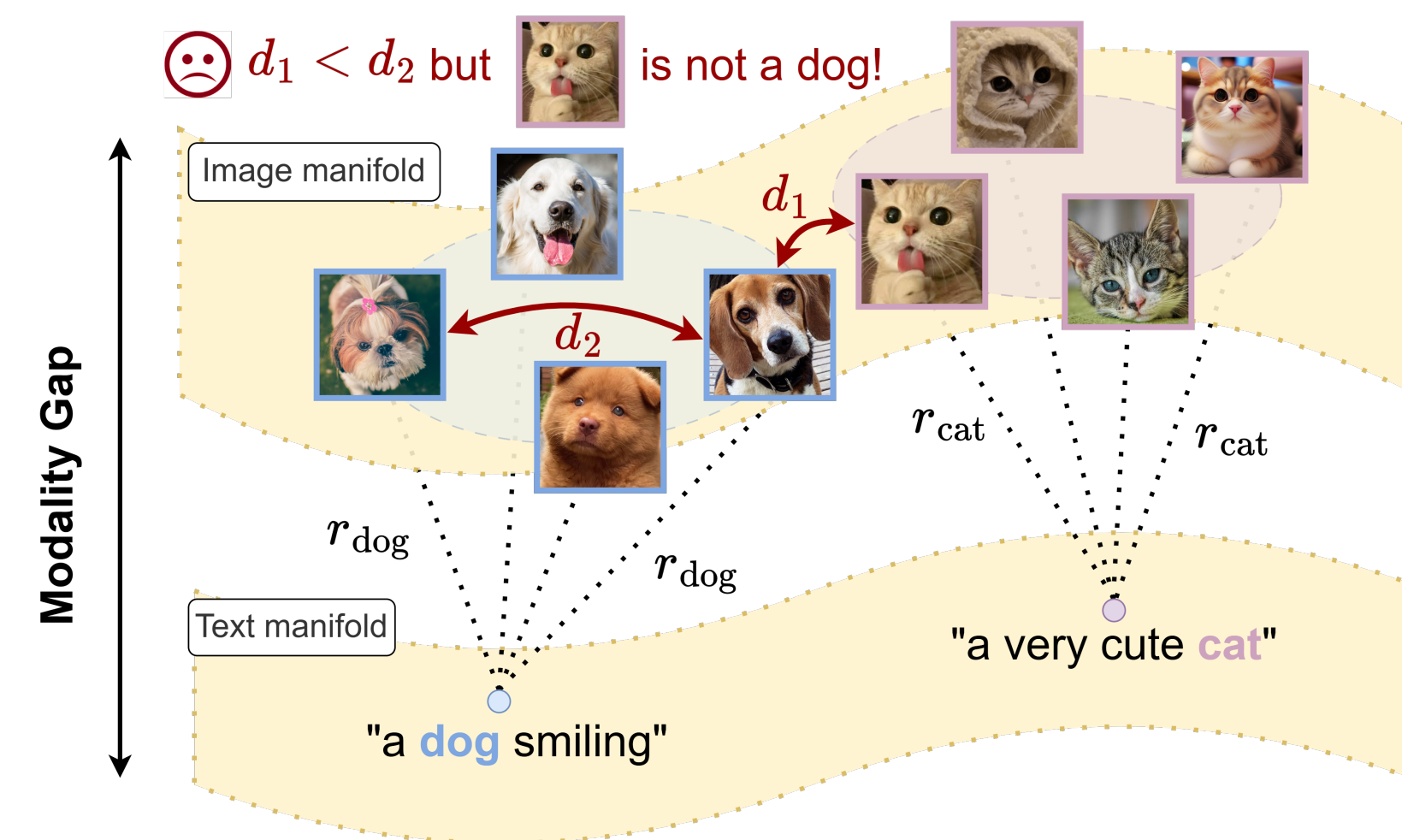
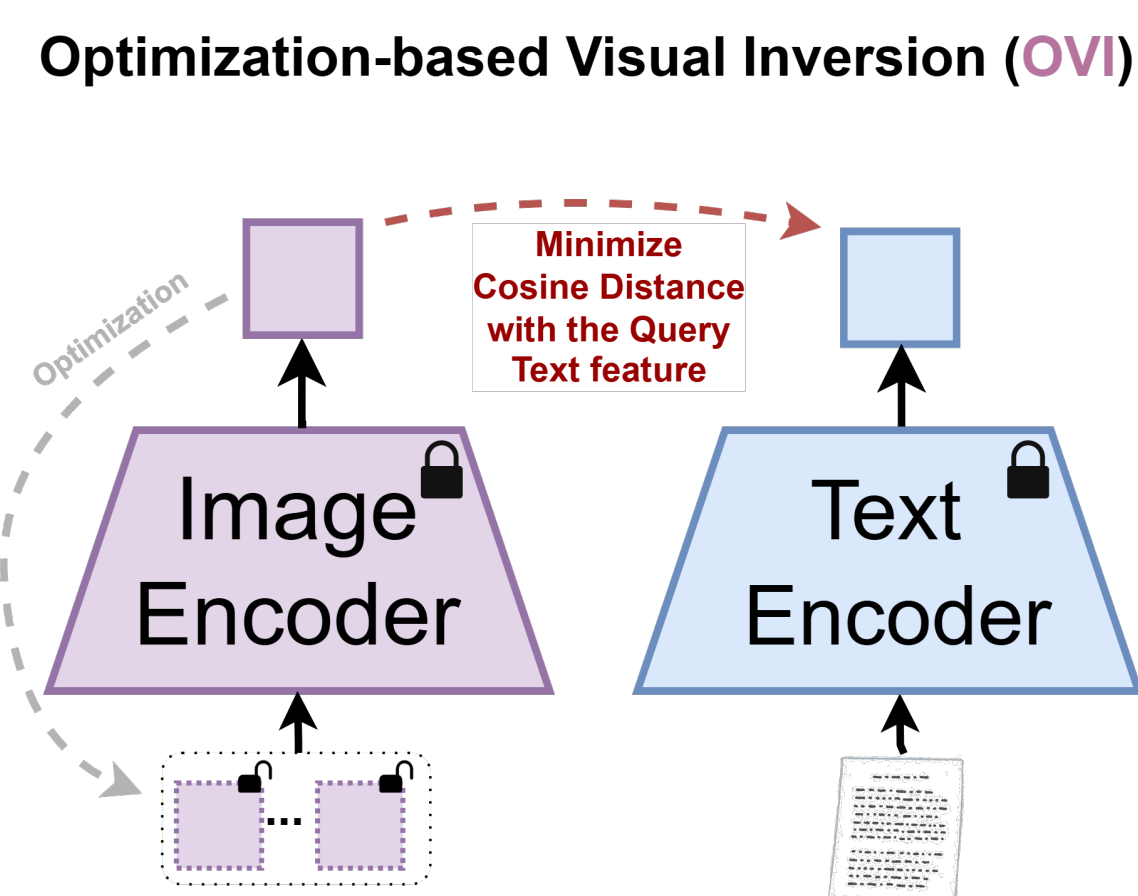
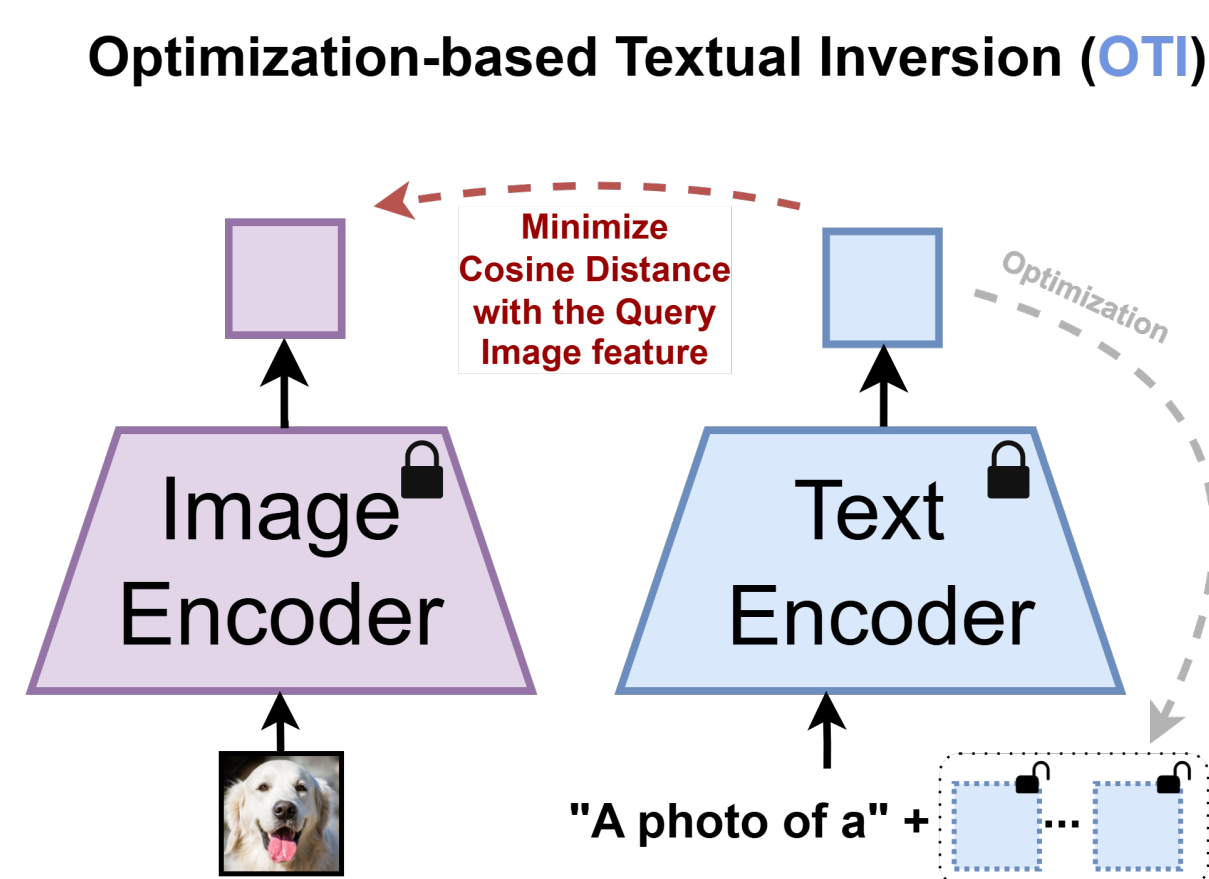




Cross the Gap: Exposing the Intra-modal Misalignment in CLIP via Modality Inversion

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



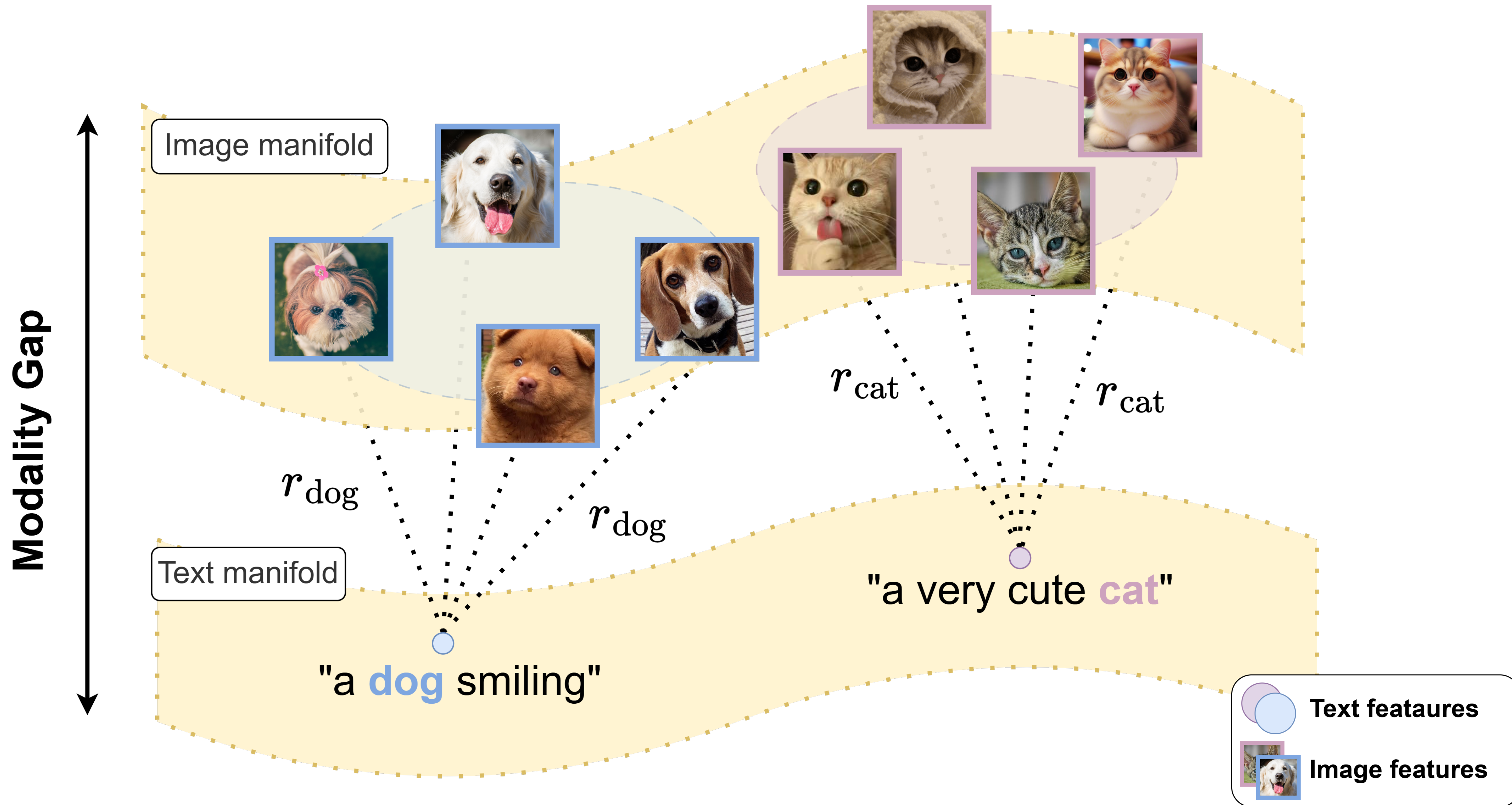
TLDR;

We adapt **OTI**, and we introduce **OVI**...

...in order to expose **CLIP Intra-modal Misalignment**

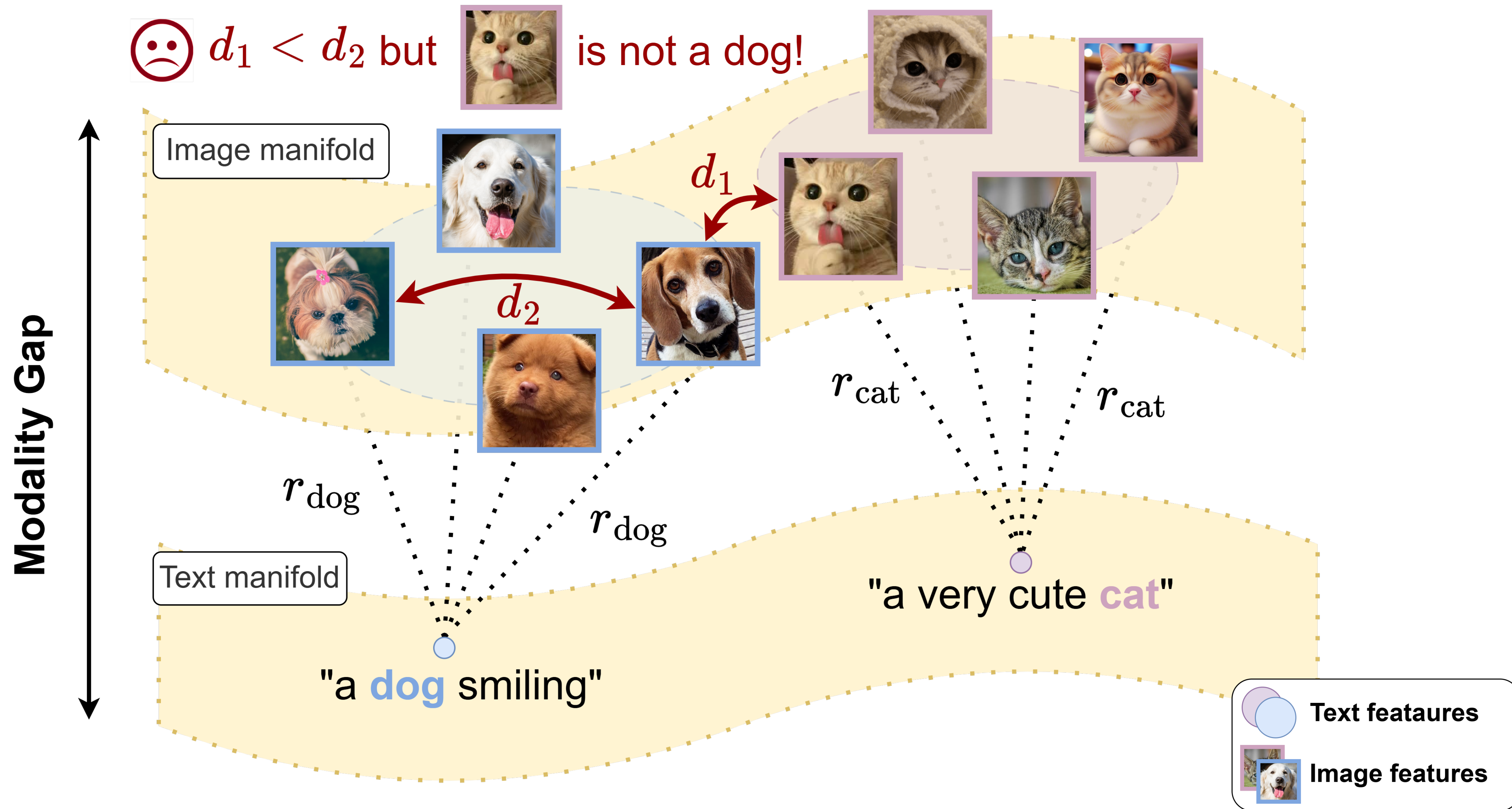
Exposing the Intra-modal misalignment

1



Exposing the Intra-modal misalignment

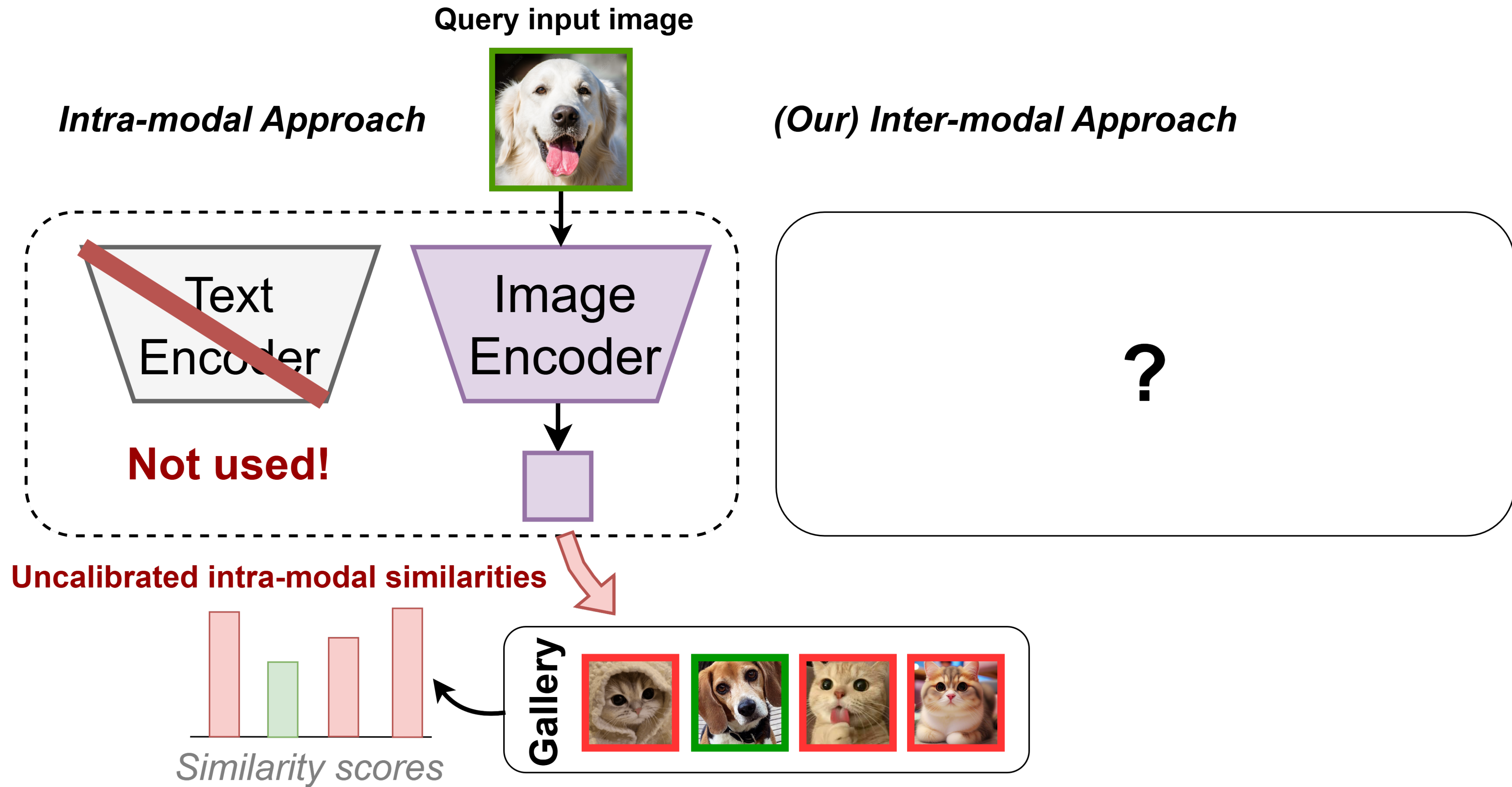
1



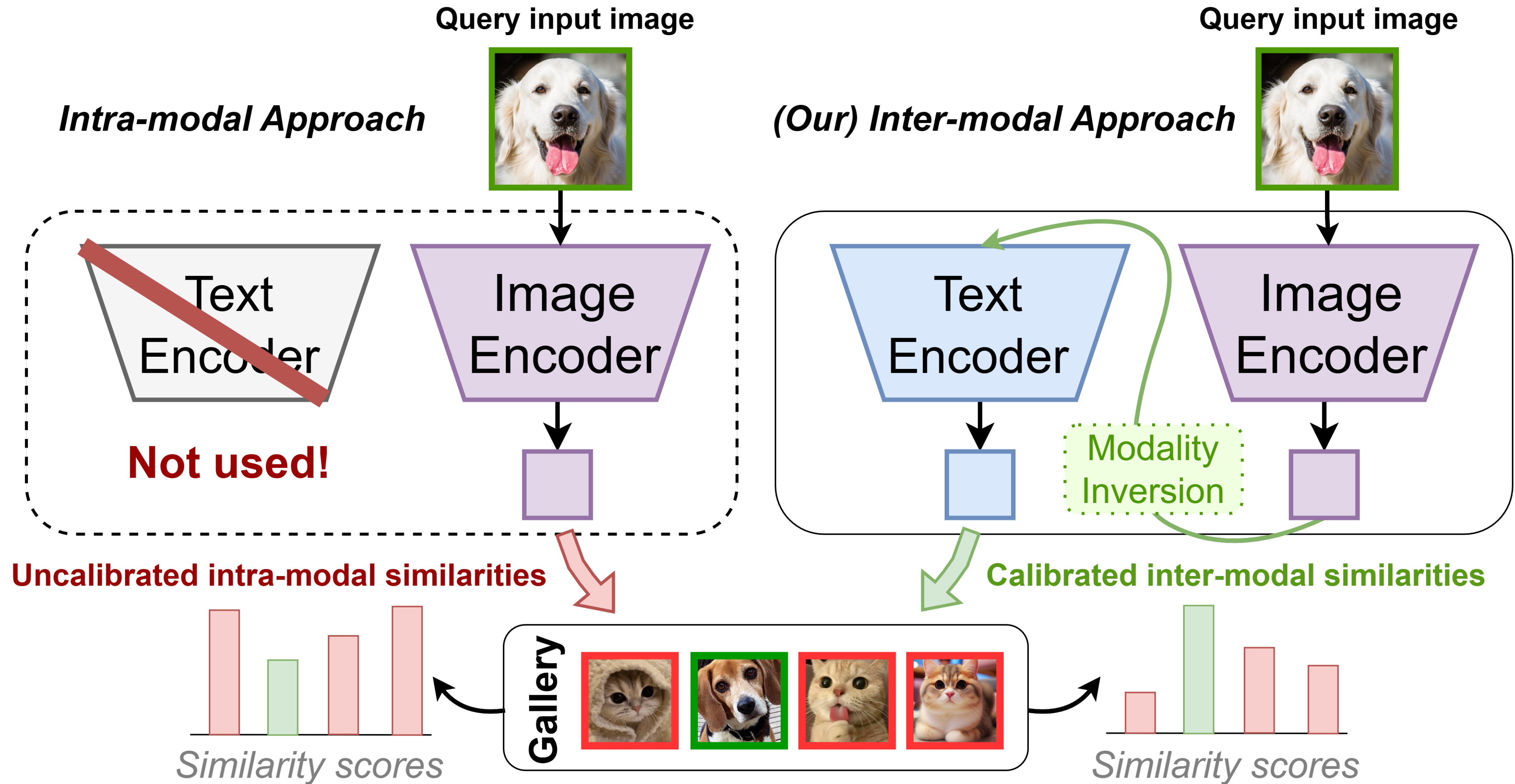
Exposing the Intra-modal misalignment



From Intra-modal to Inter-modal via Modality Inversion



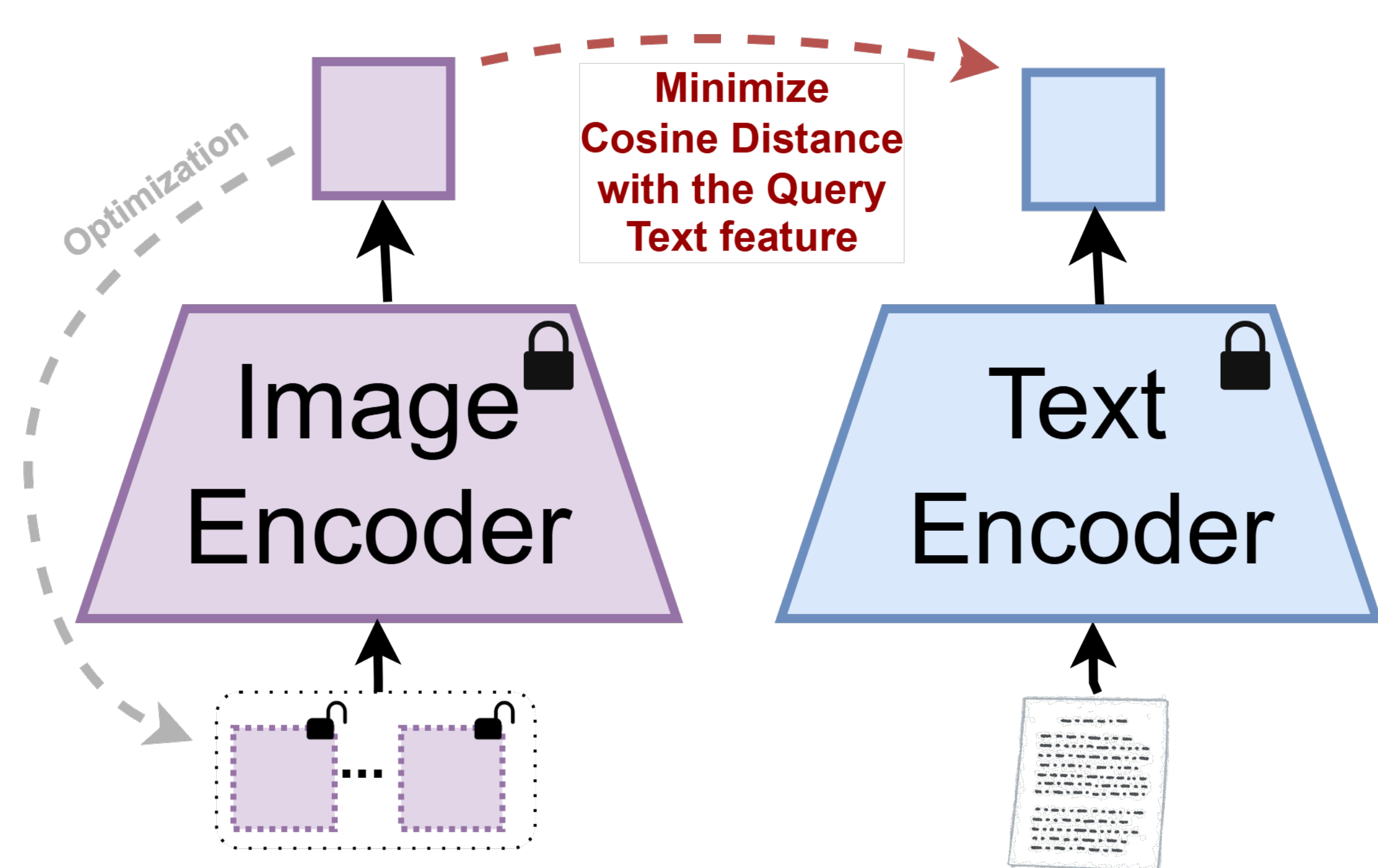
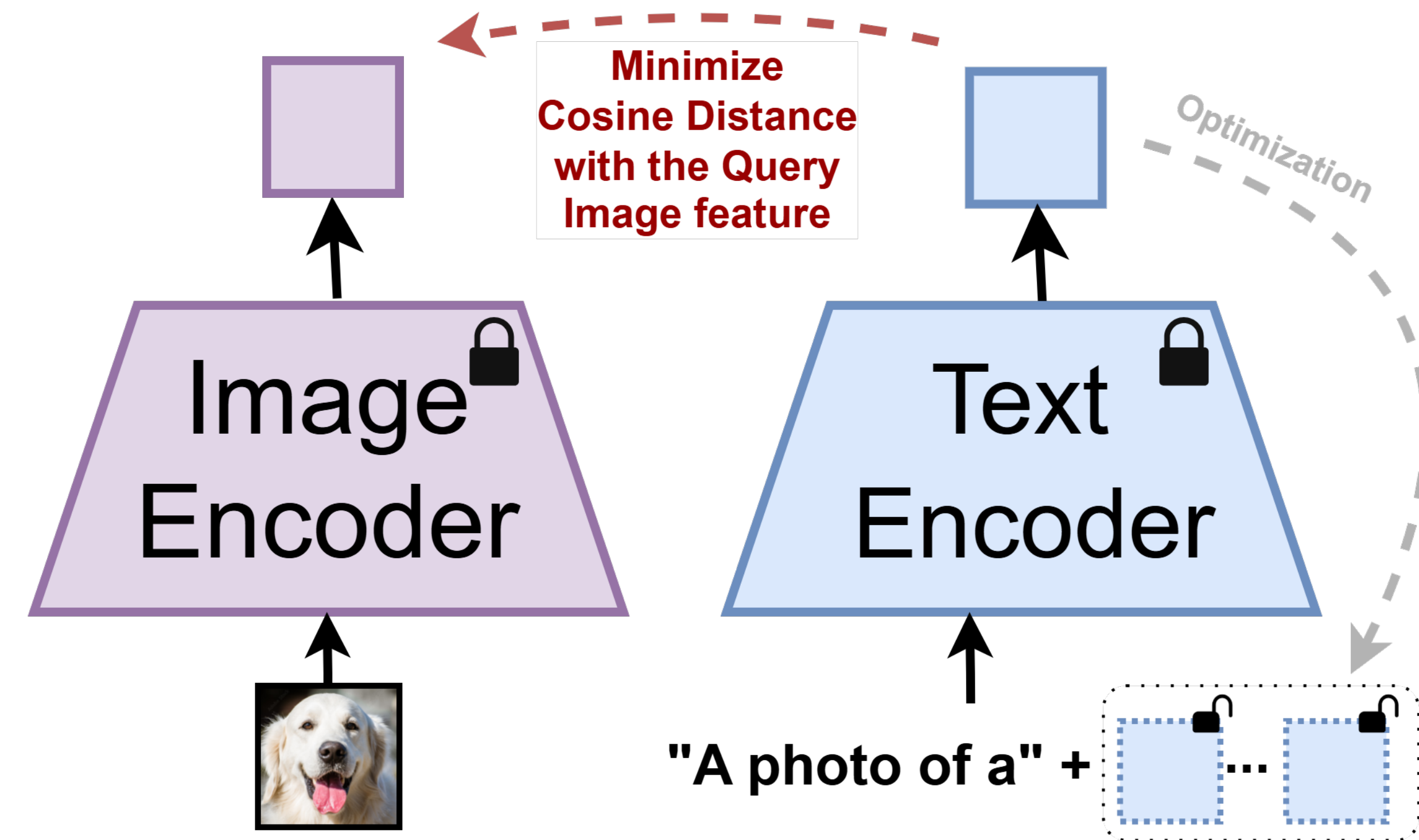
From Intra-modal to Inter-modal via Modality Inversion



↑ Example for the intra-modal task of image retrieval, text retrieval can be approached analogously

Optimization-based Textual Inversion (OTI)

Optimization-based Visual Inversion (OVI)

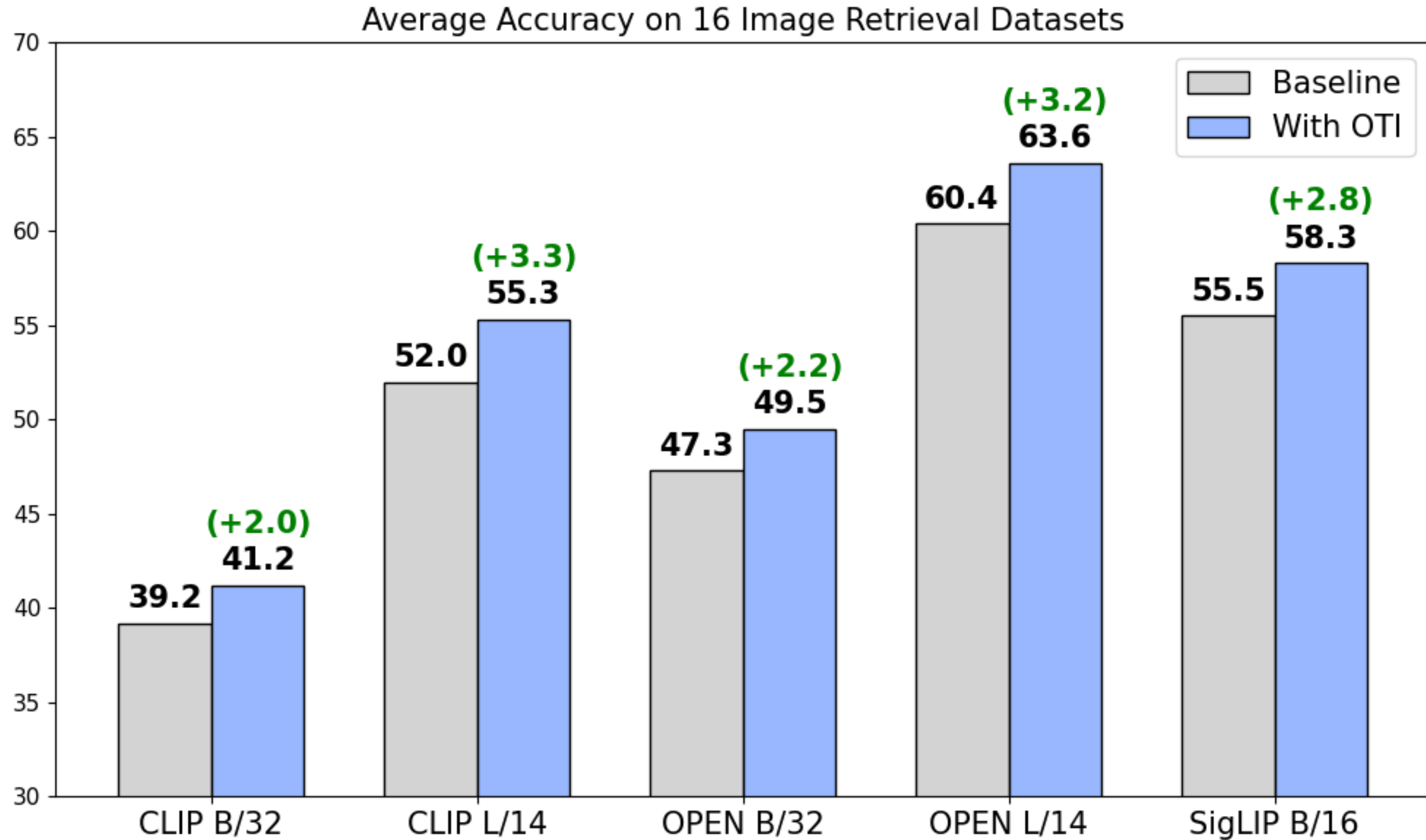


 **Pseudo token**  **Query input image**

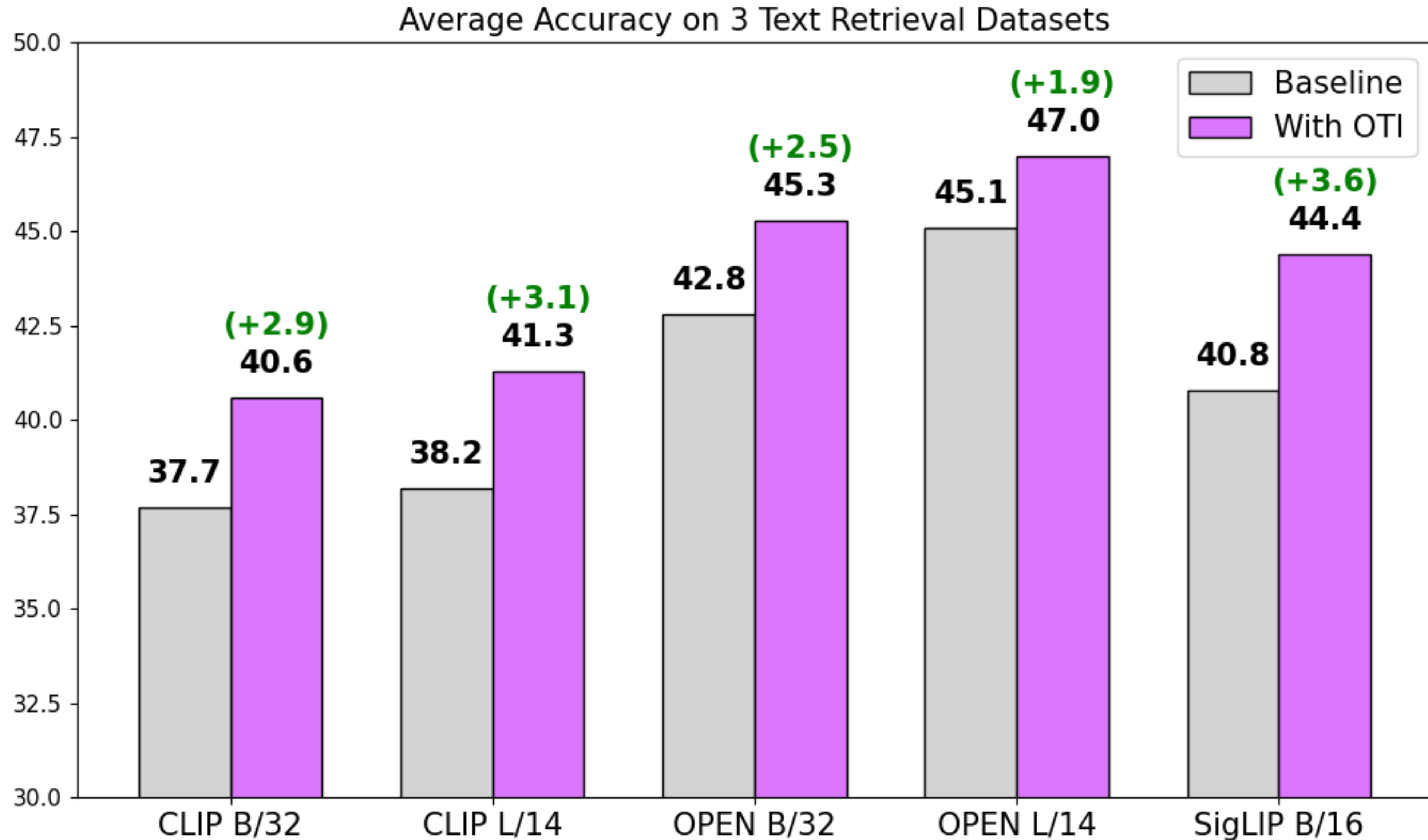
 **Pseudo patch**  **Query input text**

 **Frozen**  **Learnable**

Approaching Image-to-image retrieval intermodally with *OTI*



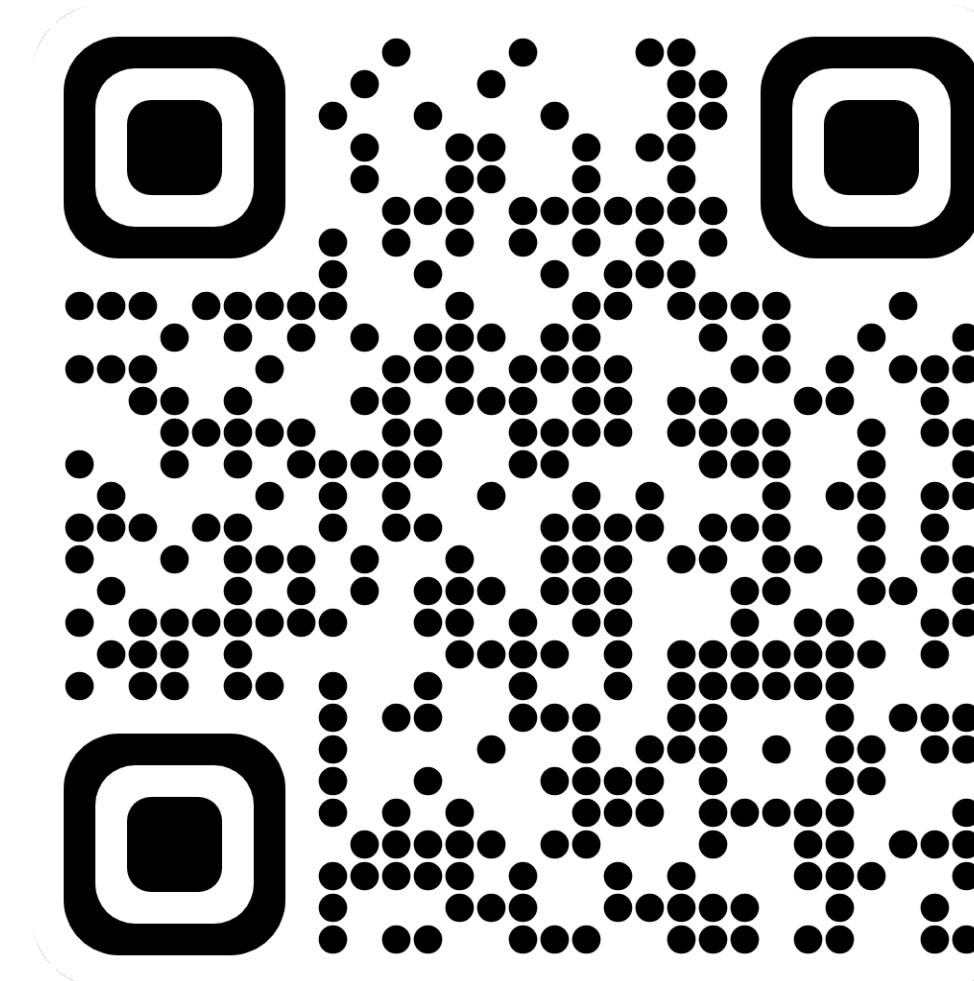
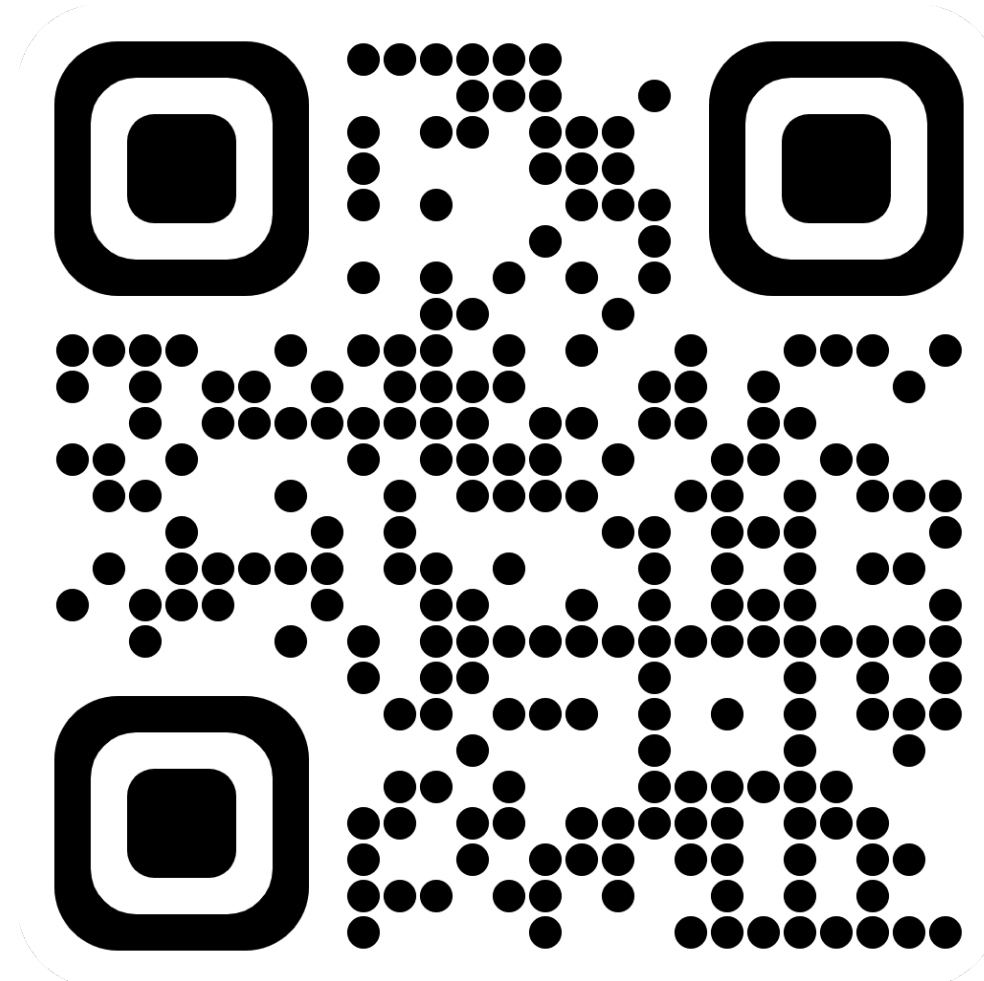
Approaching Text-to-text retrieval intermodally with *OVI*



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!



Presented by Marco Mistretta
marco.mistretta@unifi.it



ICLR