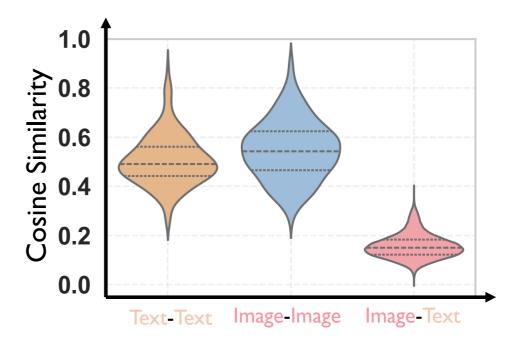
## (a) Mixed Modality Search Text Document d<sub>1</sub> Reviewer #2:This is a very interesting paper! I really like it! Image Document d<sub>2</sub> Query q Mountain Fuji Text and Image Document d<sub>3</sub> Mount Rainier is a large, active stratovolcano.. (b) Embedding Method **Fusion** Weight Mount Rainier is **CLIP** Text $d_3^T$ $\overrightarrow{e_3}^{T}$ a large, active Encoder stratovolcano.. $\overrightarrow{e_3}$ $d_3^I \\$ **CLIP** Image $1 - \alpha$ Encoder (c) Modality Gap Modality Gap Text Embeddings UMAP I Fused **Embeddings** Image **Embeddings** $\overrightarrow{e_2}$ UMAP 2

## (d) Cosine Similarity across Modalities

		Cosine Similarity	Rank	Ground Truth
	$\int d_1$	0.53	I	3
q {	$d_2$	0.26	3	1
	d <sub>2</sub>	0.35	2	2



## (e) Performance on MixBench

