Spherical Linear Interpolation and Text-Anchoring for Zero-shot Composed Image Retrieval - Supplementary Material

Young Kyun Jang[1], Dat Huynh[1], Ashish Shah[1],
Wen-Kai Chen[2], and Ser-Nam Lim[2]

[1] Meta AI
[2] University of Central Florida

1 More Qualitative Results

In this supplementary material, we provide additional qualitative results for our Slerp + TAT with BLIP-ViT-L/16 [2] backbone on CIR benchmarks. Refer to Figure 1 for CIRR [3], Figure 2 for CIRCO [1], and Figure 3 for FashionIQ [4].

References

1. Baldrati, A., Agnolucci, L., Bertini, M., Del Bimbo, A.: Zero-shot composed image retrieval with textual inversion. In: ICCV (2023)
2. Li, J., et al.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)
3. Liu, Z., Rodriguez-Opazo, C., Teney, D., Gould, S.: Image retrieval on real-life images with pre-trained vision-and-language models. In: CVPR (2021)
4. Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K., Feris, R.: Fashion iq: A new dataset towards retrieving images by natural language feedback. In: CVPR (2021)

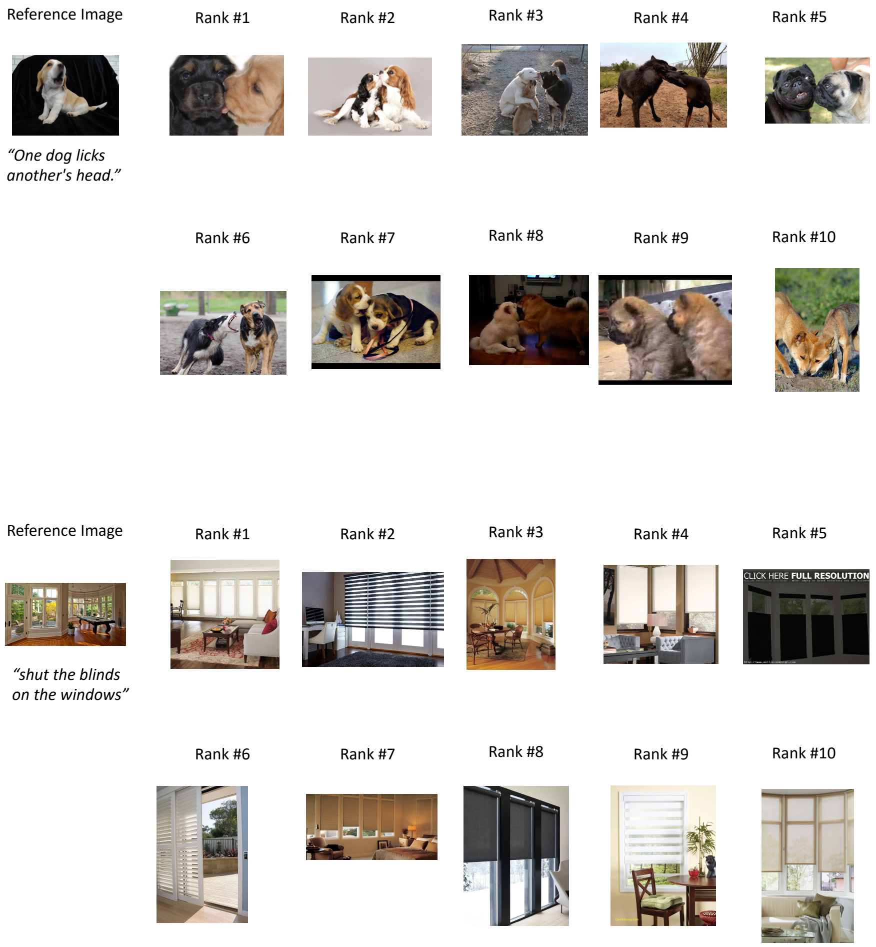

Fig. 1: Retrieval results on *CIRR test set*.

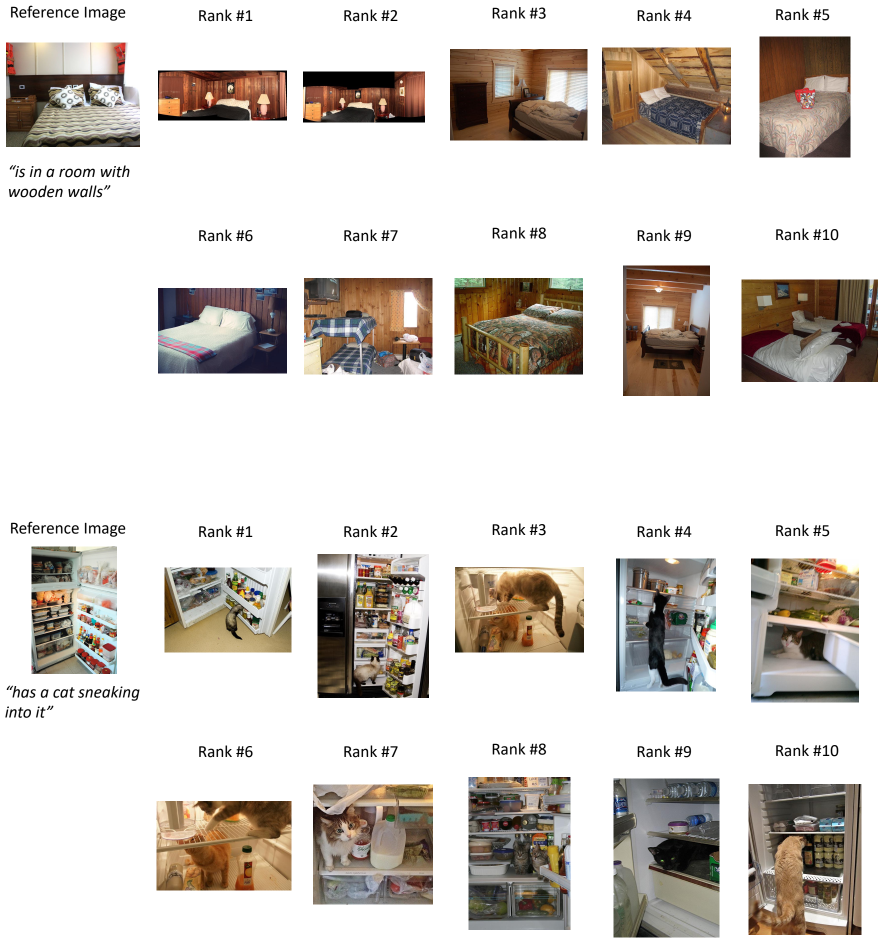

Fig. 2: Retrieval results on *CIRCO test set*.

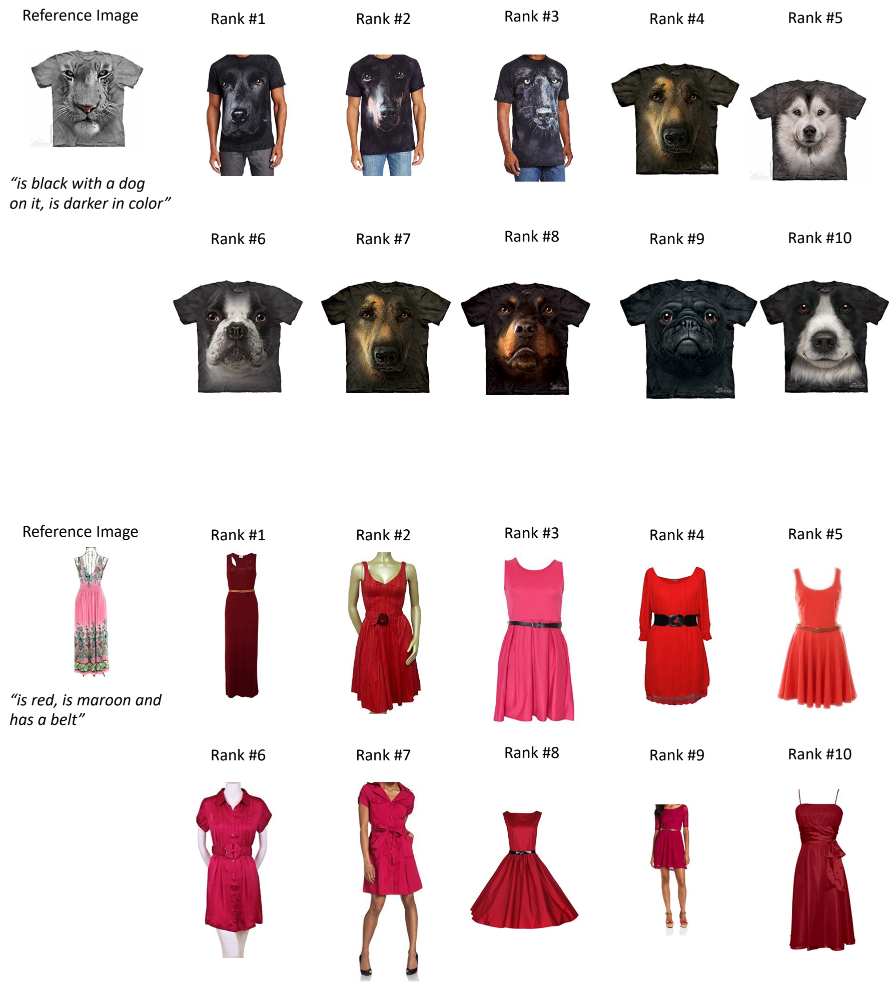

Fig. 3: Retrieval results on *FashionIQ validation set*.