Rebuttal for #2335 - WikiDO: A New Benchmark Evaluating Cross-Modal Retrieval for Vision-Language Models

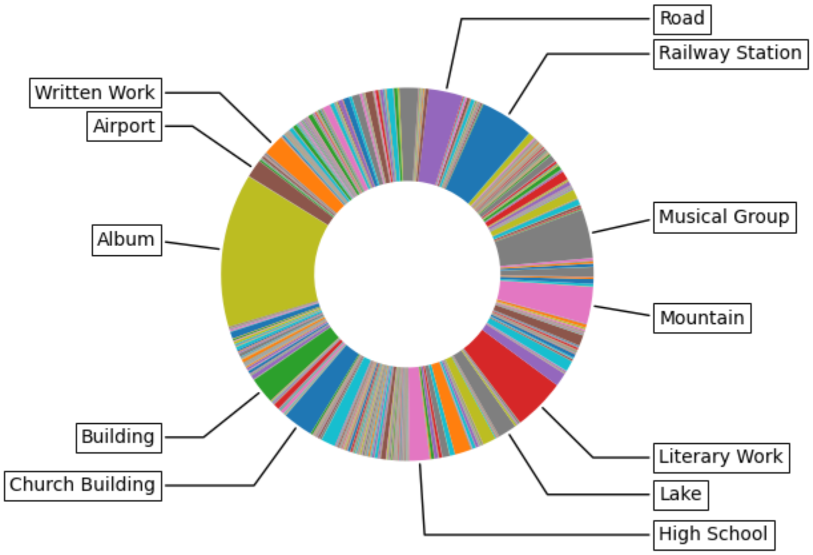

Figure 1: Pie-plot of subtopics in the dataset. Only a few subtopics are labelled to avoid clutter. Each color in the plot denotes a different subtopic.

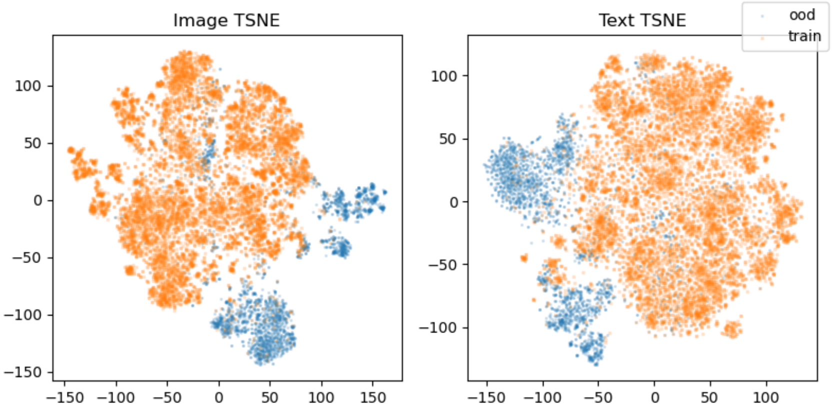

Figure 2: T-SNE plot for embeddings of 1000 random images and texts per topic with perplexity 32 (Figure 3 in the submission).

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.