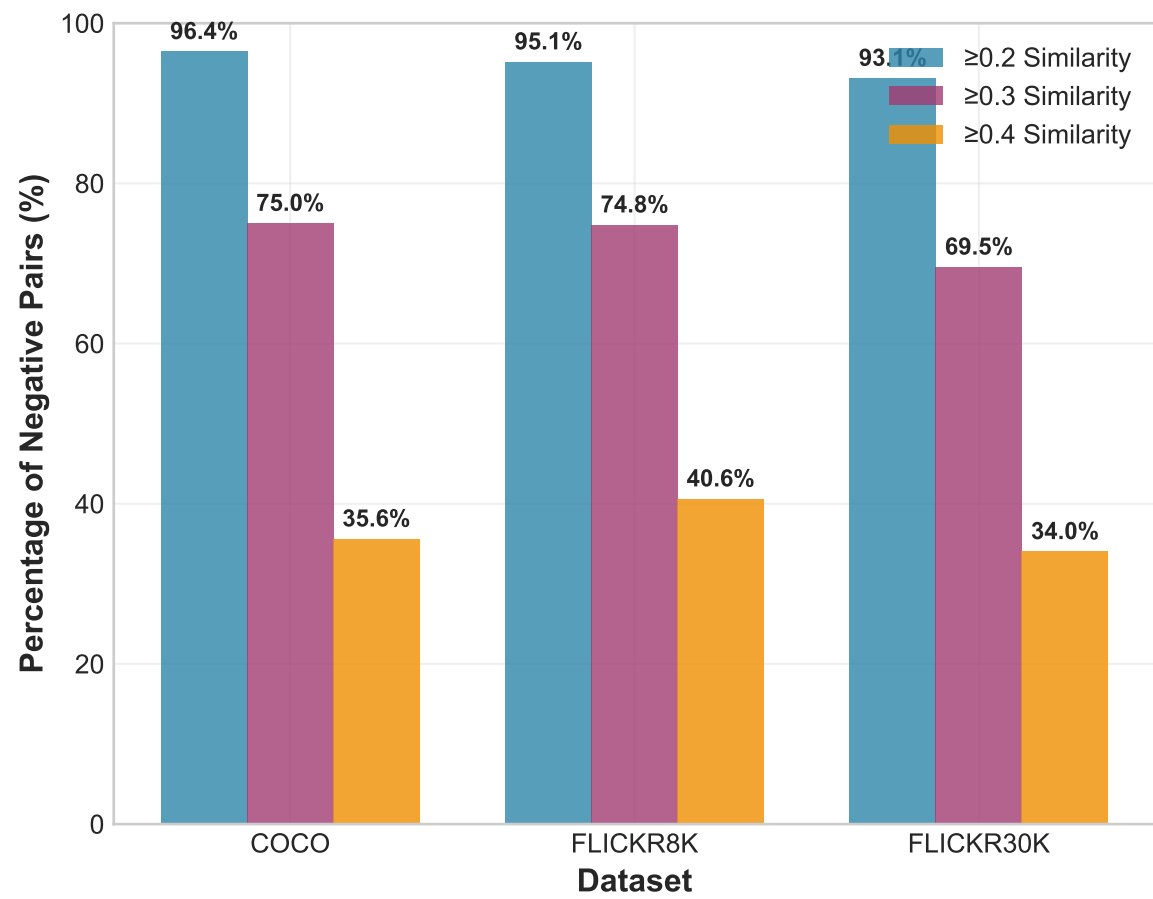
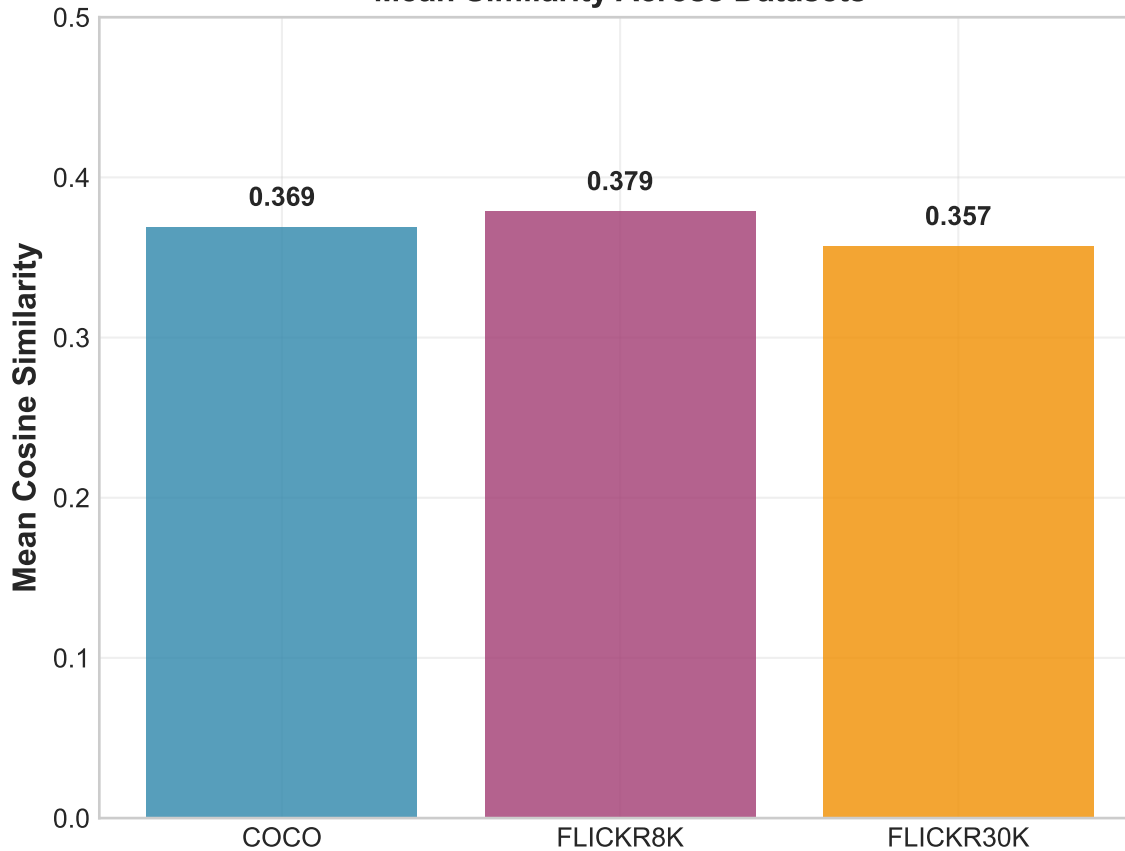


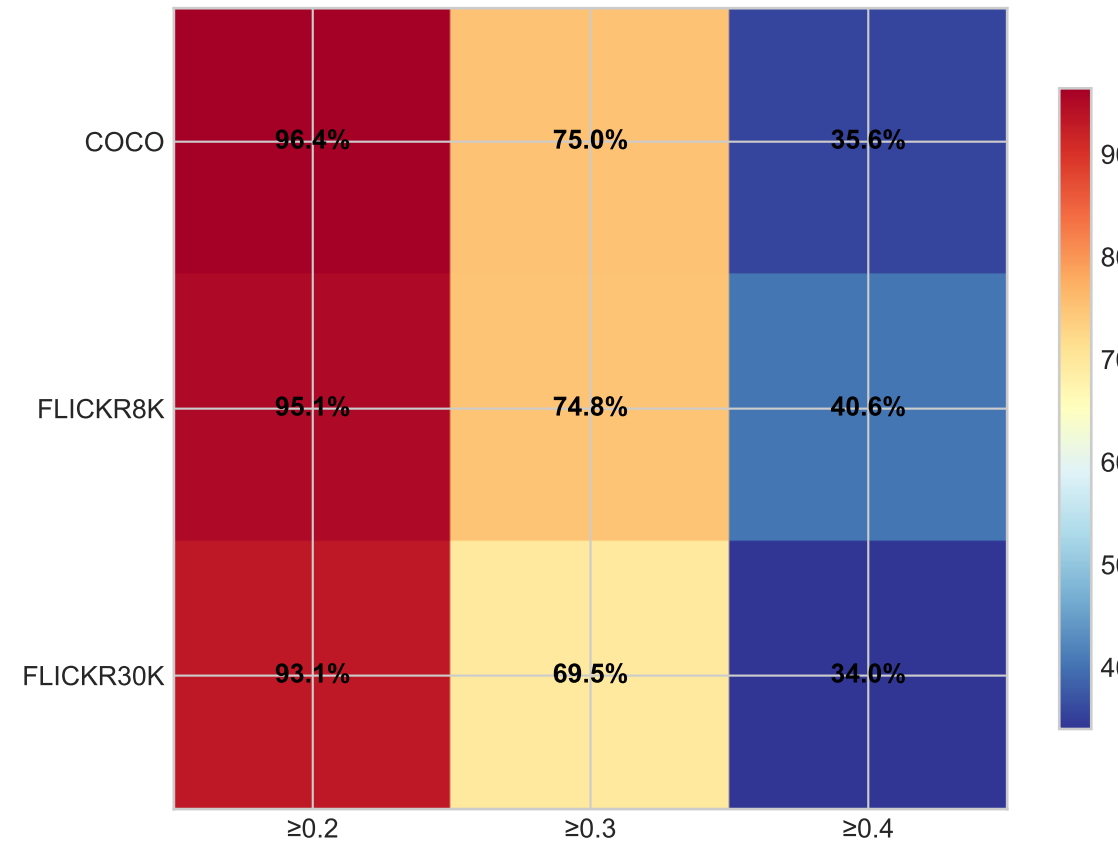
Negative Text Similarity Analysis (SANW Motivation Evidence)



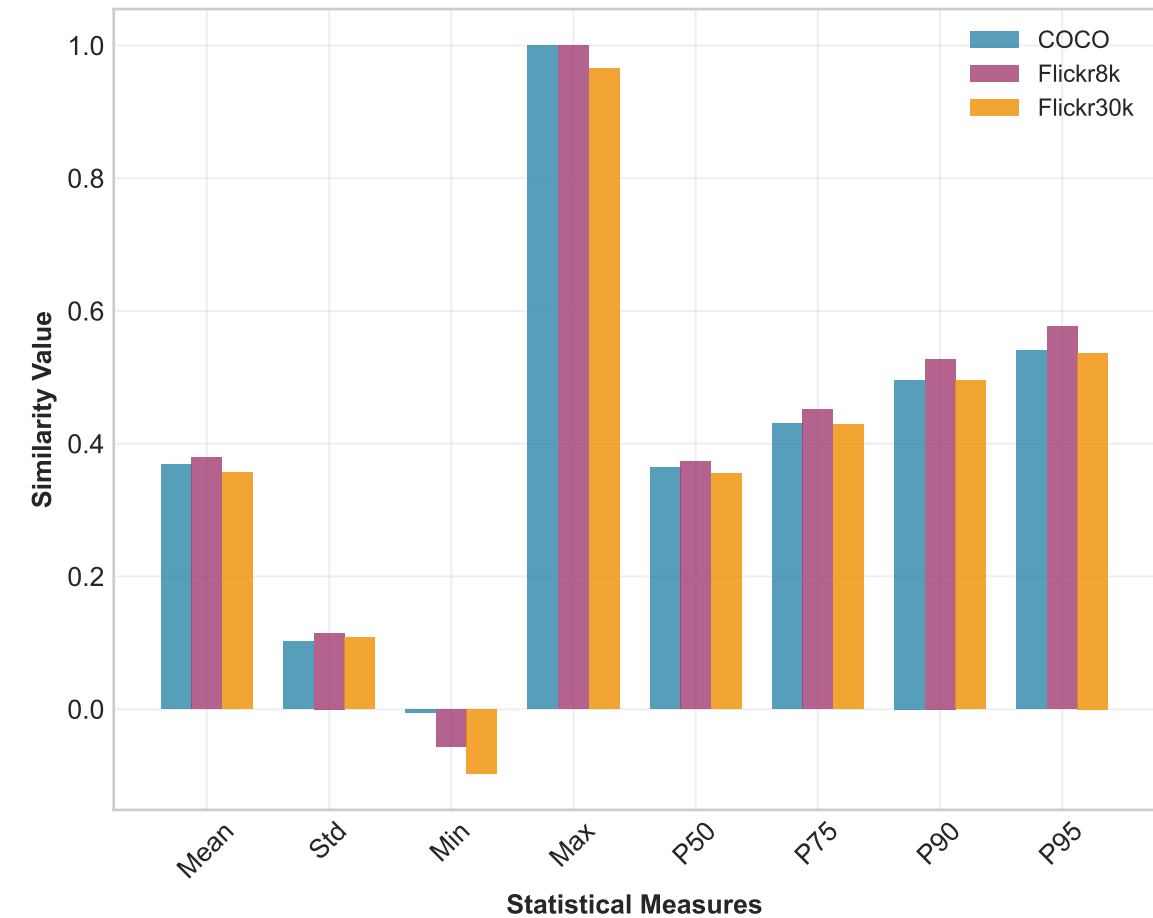
Mean Similarity Across Datasets



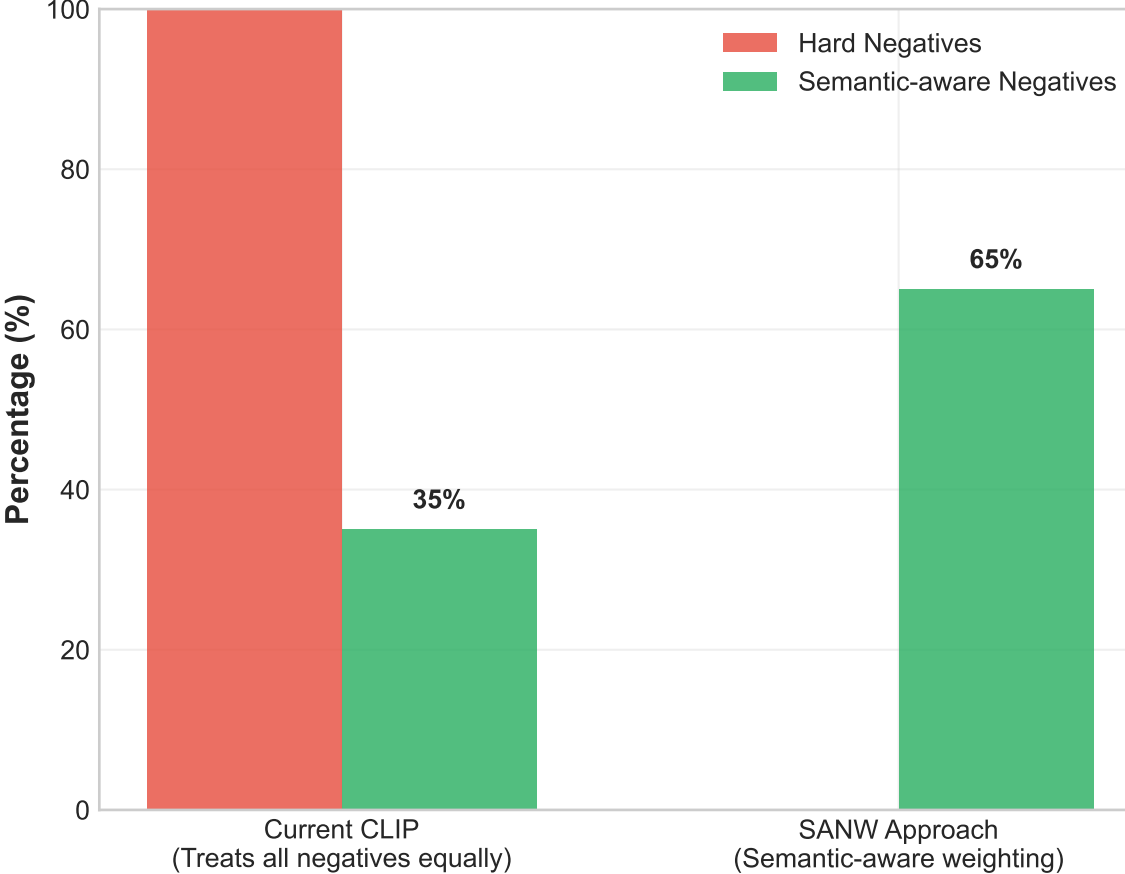
Similarity Threshold Analysis (%)



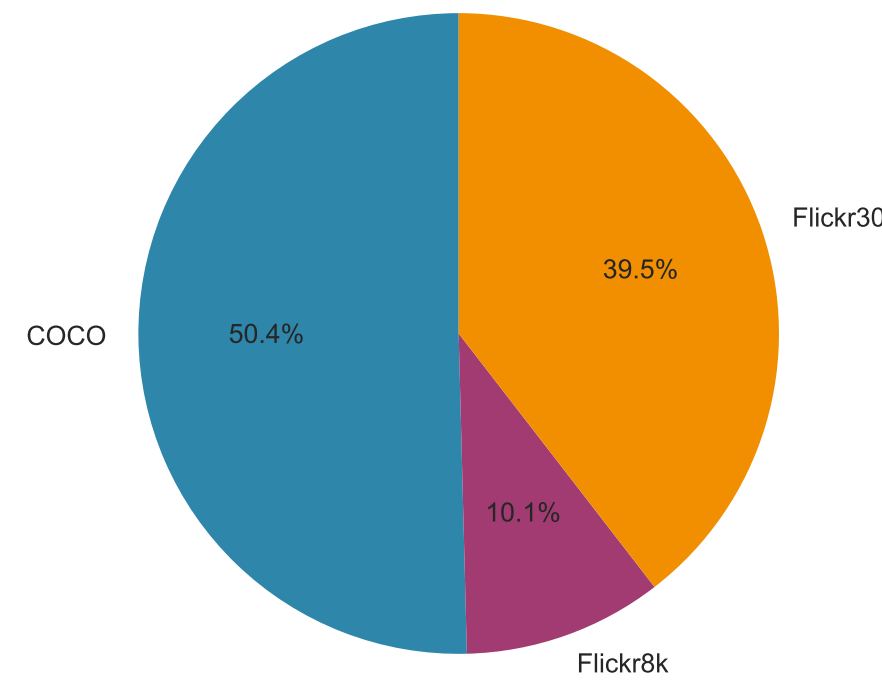
Statistical Distribution Analysis



SANW Motivation: Current vs Proposed Approach



Dataset Size Distribution



KEY FINDINGS:

- 93-96% of "negative" pairs have similarity ≥ 0.2 (very high!)
- 69-75% of "negative" pairs have similarity ≥ 0.3 (extremely high!)
- 34-41% of "negative" pairs have similarity ≥ 0.4 (significant!)

CONCLUSION:

Current CLIP training treats many semantically similar captions as "negatives" when they should be reweighted (SANW motivation).

METHODOLOGY:

- Loaded captions from 3 datasets:
 - COCO (202K captions)
 - Flickr8k (40K captions)
 - Flickr30k (159K captions)
- Used OpenCLIP ViT-B-32 (laion400m_e32) text encoder
- Computed cosine similarities between all caption pairs in random batches (256 captions)
- Analyzed off-diagonal similarities (excluding same-image pairs)
- Calculated statistics across 50 batches per dataset

Distribution of Negative Pair Similarities

