Baseline Results: Zero-shot Text-to-Image Retrieval

This document records the frozen baseline results for the STATS 302 group project
"Zero-shot Image–Text Retrieval with CLIP".
All subsequent analyses and improvements are compared against this baseline.

Experimental Setup

- **Task**: Zero-shot **Text-to-Image Retrieval**
- **Dataset**: MS-COCO 2017 Validation Set
- **Images**: 5,000
- **Captions**: 25,000 (5 captions per image)
- **Model**: CLIP ViT-B/32 (`openai/clip-vit-base-patch32`)
- **Framework**: Hugging Face Transformers
- **Embedding**: L2-normalized image and text embeddings
- **Similarity**: Cosine similarity
- **Evaluation Metrics**: Recall@1, Recall@5, Recall@10
- **Random Seed**: 42

Baseline Performance

Metric	Value
Recall@1	0.3044
Recall@5	0.5477
Recall@10	0.6624

Notes

- This baseline corresponds to **Text → Image retrieval**.
- No fine-tuning or task-specific training was performed.
- Embeddings were cached to ensure reproducibility.
- These results are **frozen** and must not be modified in later stages.

Date recorded: 2026-01-23