(a) Dataset Construction

Query q: Mountain Fuji

Mount Fuji is an active stratovolcano located on the Japanese island of Honshu.

Text and Image

$$\vec{e_1} = \alpha\vec{e_1^T} + (1-\alpha)\vec{e_1^I}$$

The Statue of Liberty is a colossal neoclassical sculpture on Liberty Island.

Text and Image

$$\vec{e_2} = \alpha\vec{e_2^T} + (1-\alpha)\vec{e_2^I}$$

Yellowstone National Park is a national park of the United States located in …

Text and Image

$$\vec{e_3} = \alpha\vec{e_3^T} + (1-\alpha)\vec{e_3^I}$$

(b) Results

CLIP — OVEN: 0.71, 0.77, 0.72

OpenCLIP — OVEN: 0.76, 0.82, 0.72

SigLIP — OVEN: 0.69, 0.79, 0.72

CLIP — Google WIT: 0.87, 0.79, 0.83

CLIP — MSCOCO: 0.76, 0.88, 0.85

CLIP — VisualNews: 0.82, 0.88, 0.86

Retrieval Performance (NDCG@10)

CLIP   GR-CLIP   VLM2Vec (Qwen)