Assessing Effective Token Length of Multimodal Models for Text-to-Image Retrieval

Le Nguyen, Preet Jain, Krutik Panchal, Md Tanvirul Alam, Nidhi Rastogi

RIT
SIGIR 2025
Padova
ITALY
Software Engineering
Rochester Institute of Technology

Scan QR for GitHub

Introduction

Multimodal models revolutionize text to image retrieval by mapping text and image embeddings into a shared vector space.

We systematically benchmark current state of the art multimodal models across diverse datasets to quantify effective token lengths and domain-specific robustness.

Our open-source, reproducible framework guides optimal query design and establishes standard benchmarks for long-text image retrieval.

Research Questions

- **RQ1 :** What is the effective token length for CLIP, BLIP-2, ALIGN, OpenCLIP, and Long-CLIP?
- **RQ2 :** How does domain-specific language (medical, news, AI-generated, urban scenes) affect effective token length?
- **RQ3 :** Do chunking and pooling strategies that use all tokens available in a document affect the effective token length?

Methodology

1.Progressive truncation

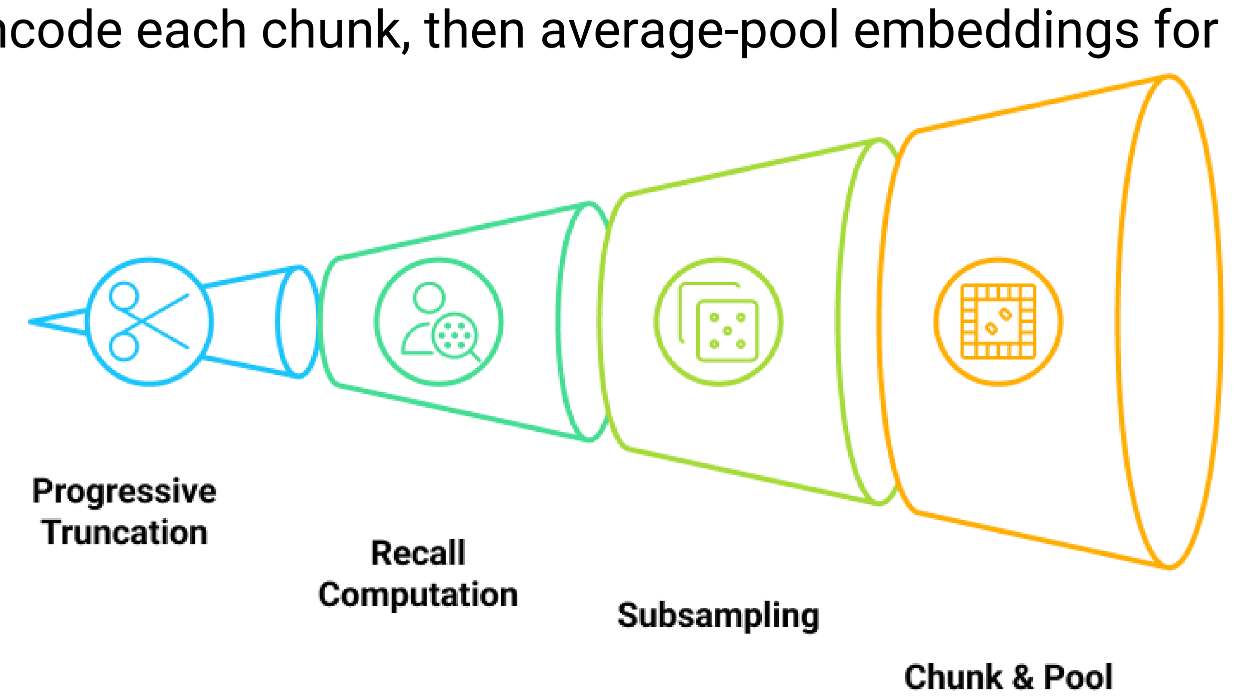
- Truncate each caption at increasing token lengths.
- Compute Recall@1 with FAISS retrieval.

2.Subsampling

- Draw 10 random 1000-item subsets per dataset.
- Repeat truncation experiment to build confidence intervals.

3.Chunk & pooling

- Split texts exceeding the model's token limit into equal sized chunks before processing.
- Encode each chunk, then average-pool embeddings for retrieval.

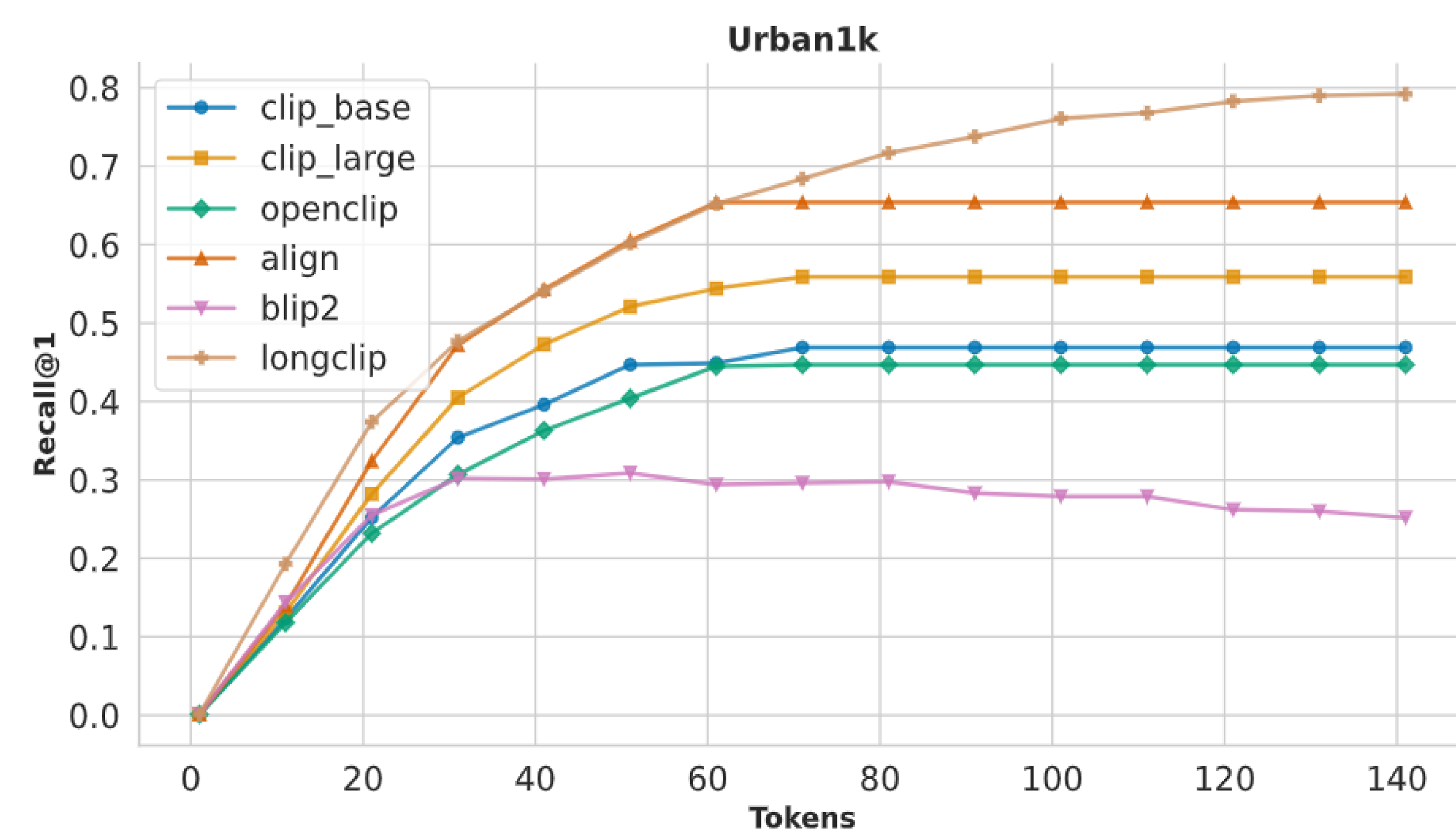


Dataset

Dataset	Domain	Avg. Caption Length
Urban1k	Urban scenes	101 tokens
ROCO	Medical imaging	25 tokens
ShareGPT4V	AI-generated	160 tokens
Factify2	News reports	1736 tokens

Results

RQ1. Effective Token Length



Recall@1 by caption length on the Urban1k Dataset

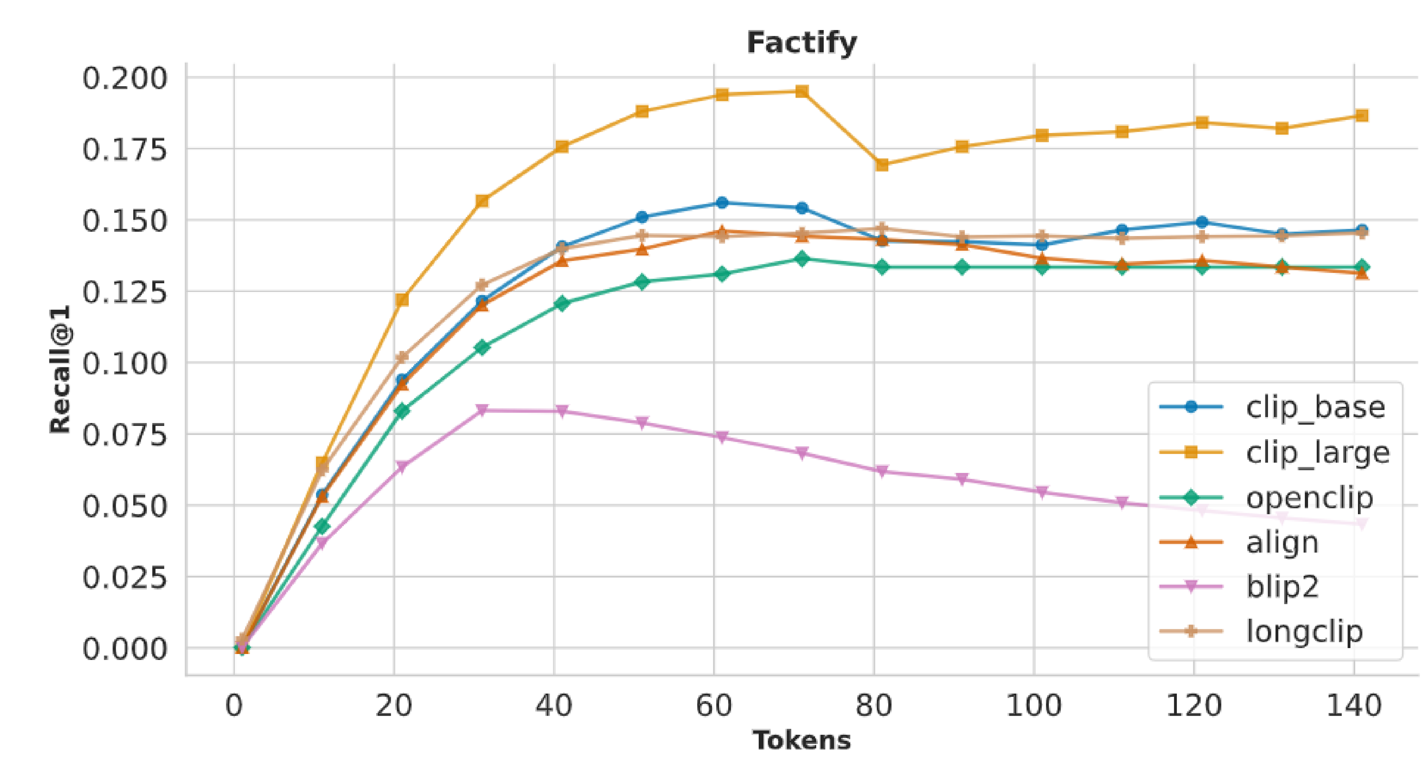
- **CLIP-Base** plateaus (~0.47) by **40 tokens** → 95 % of max recall.
- **CLIP-Large** reaches its plateau (~0.56) by **50 tokens**.
- **OpenCLIP** and **ALIGN** both hit ~95 % recall by **50–60 tokens.**
- **BLIP-2** (512 token limit) tops out earlier (~0.30) around **30 tokens**, then slightly declines.
- **Long-CLIP** exhibits the highest plateau (~0.79) but only after **90 tokens**—well below its 248-token input limit.

All models achieve near-maximum retrieval performance at **40–90 tokens**, significantly below their architectural token limits **(77–512 tokens)**, confirming each model's "effective token length".

RQ2. Domain Specific Language

- **ROCO (Medical Imaging)** and **Factify2 (News Reports)** both show lower overall Recall@1 (max ≤ 0.1 on ROCO) and **more varied effective lengths** across models.
- **Long-CLIP's** effective length on **Factify2** drops to **30 tokens** at 95% recall, **half of its Urban1k performance,** emphasizing how technical or verbose text can limit token utility.
- **ShareGPT4V** and **Urban1k** exhibit higher and more consistent effective lengths **(~50 tokens)** across all models.
- **OpenCLIP**'s massive web-scale training yields a **consistent 50-token limit**—highlighting broad-corpus benefits.

RQ3. Chunking and Pooling



Recall@1 with Extended Text Chunking and Pooling on the Factify2 Dataset

- For RQ3 (performed on Factify2), splitting texts into chunks and averaging their embeddings **did not improve Recall@1** or shift the effective token length beyond each model's native limit.
- Simple chunk-and-pool strategies yield **no significant gains** on image retrieval performance (Recall@1 curves remain flat past the model limit), suggesting that embedding models heavily prioritize initial tokens.

Conclusions

- **Early Plateau :** Models reach ≥95% Recall@1 by 40–60 tokens, far below their input limits.
- **Domain Impact :** ROCO and Factify2 show lower, variable recall; Urban1k and ShareGPT4V are more stable.
- **Chunking Ineffective :** Chunking and Pooling tokens doesn't rescue performance; models prioritize initial tokens.