INSIGHT Intelligent Systems Geometric Computing & Perception Lab

Efficient Discriminative Joint Encoders for Large Scale Vision-Language Reranking

Mitchell Keren Taraday*, Shahaf Wagner*, Chaim Baskin

*Equal contribution

BGU Ben-Gurion University of the Negev

The Reranking Bottleneck

Joint encoders like BLIP dramatically improve retrieval — but are too slow to deploy

Embedding Models (CLIP)

- Fast — independent encoding
- Precomputed embeddings

Limited cross-modal interaction

Joint Encoders (BLIP)

- Rich cross-modal reasoning
- Strong reranking accuracy

ViT runs online → Slower

Can we get the accuracy of joint encoders at the speed of embeddings?

Key Idea: Shift Vision Offline

1. Encode Offline

Run ViT once per image,
store visual features on disk

2. Compress

CrossAttention adapter:
4096 ⟶ 64 tokens

3. Rerank Online

Compact MiniLM encoder
jointly scores text + vision

Result: Online inference = only the lightweight joint encoder — no ViT at query time

Architecture

OFFLINE (per image, once)

Frozen Vision Encoder (SigLIP2)

↓ 4096 patch tokens

CrossAttention Pooling
64 learnable queries → cross-attend → project

↓ 64 compressed tokens (384-d)

Store on disk (49 kB / image)

ONLINE (per query)

Text query tokens

+ 64 compressed vision tokens

↓ concatenate

MiniLM-L12-H384
Joint Encoder (12 layers, 384-d)

↓

Matching Score

Training Pipeline

Stage 1: Pre-training

- CC3M + CC12M (15M pairs)
- ITM + MLM + ITC loss
- Frozen vision encoder throughout
- Train: adapter + joint encoder

Stage 2: Fine-tuning

- COCO / Flickr30k (task-specific)
- Same architecture, continued training
- ITM

Key design choices: Vision encoder always frozen • Learnable queries in adapter • Cross-attention compresses 4096→64 tokens • Linear projection to 384-d

Results and Comparisons to Prior Work

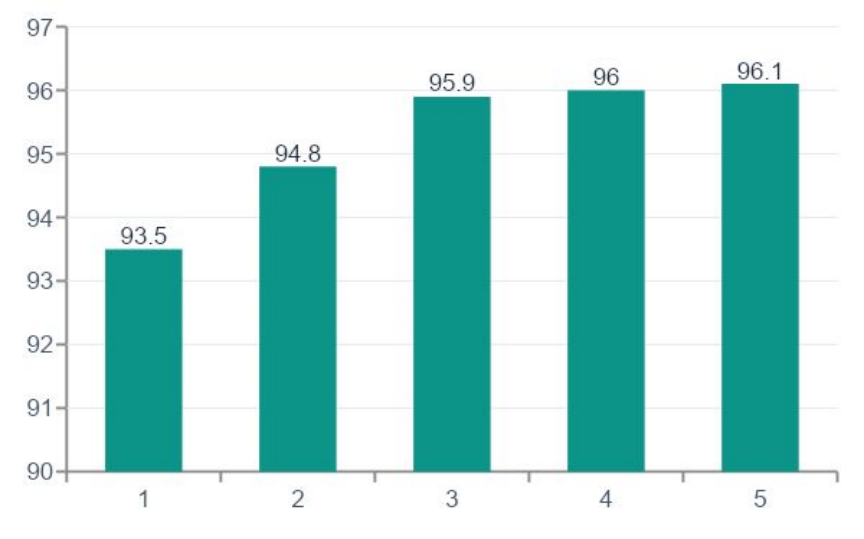
Recall@1 • Inference speed • Storage cost

Method	Train Data	Flickr-ZS T2I	Flickr-ZS I2T	COCO-FT T2I	COCO-FT I2T	Storage /image	Joint Enc. Params	Infer. (ms)
ALBEF ViT-B/16	12M	82.8	94.1	60.7	77.6	1,769 kB	147M	45.92
BLIP ViT-B/16	12M	84.9	94.8	63.1	80.6	1,769 kB	139M	83.27
BLIP ViT-L/16	129M	86.7	96.7	65.1	82.4	2,359 kB	139M	101.61
BLIP-2 ViT-L/16	400M	88.6	96.9	66.3	83.5	2,359 kB	167M	98.64
Local ViT-B/16	**12M**	**84.3**	**94.3**	**60.9**	**76.1**	**442 kB**	**33M**	**2.86**
Local ViT-L/16	**12M**	**87.8**	**96.5**	**64.9**	**81**	**442 kB**	**33M**	**4.14**
Compressed-128	**12M**	**87.1**	**96.3**	**64.6**	**81**	**98 kB**	**33M**	**2.04**
Compressed-64	**12M**	**86.9**	**96.4**	**64.6**	**80.9**	**49 kB**	**33M**	**1.91**

EDJE: competitive accuracy with 33M params, 49 kB/image, and up to 53× faster inference

Ablations & Analysis

Token Compression Trade-off



64 tokens: best efficiency/accuracy balance

Works Across Backbones

Vision Backbone	Base R@1	+EDJE
CLIP ViT-B/32	80.2	**89.1** ↑
CLIP ViT-L/14	87.4	**93.1** ↑
SigLIP2 SO	93.8	**95.9** ↑

Consistent gains

EDJE improves every backbone tested, regardless of ViT size or resolution. Acting as a universal reranker.

Summary

Efficient Joint Encoder

Shift ViT offline, compress tokens, rerank with a tiny LM — 53× faster than BLIP

Token Compression Adapter

CrossAttention pools 4096 → 64 tokens at 49 kB/image with minimal accuracy loss

Drop-in Reranker

Improves any embedding model — tested across CLIP, SigLIP2 with multiple ViT sizes

Strong Benchmarks

Matches/surpasses BLIP on Flickr30k (zero-shot) and MS-COCO (fine-tuned)

shahafwa.github.io/EDJE

Thank you! Questions?