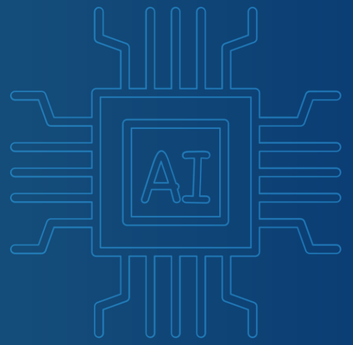
Sentinels

Zero-shot Composed Text-Image Retrieval



Introduction to CIR

- Retrieve images by leveraging a combination of reference image and textual information that illustrates desired modifications

Make it Black

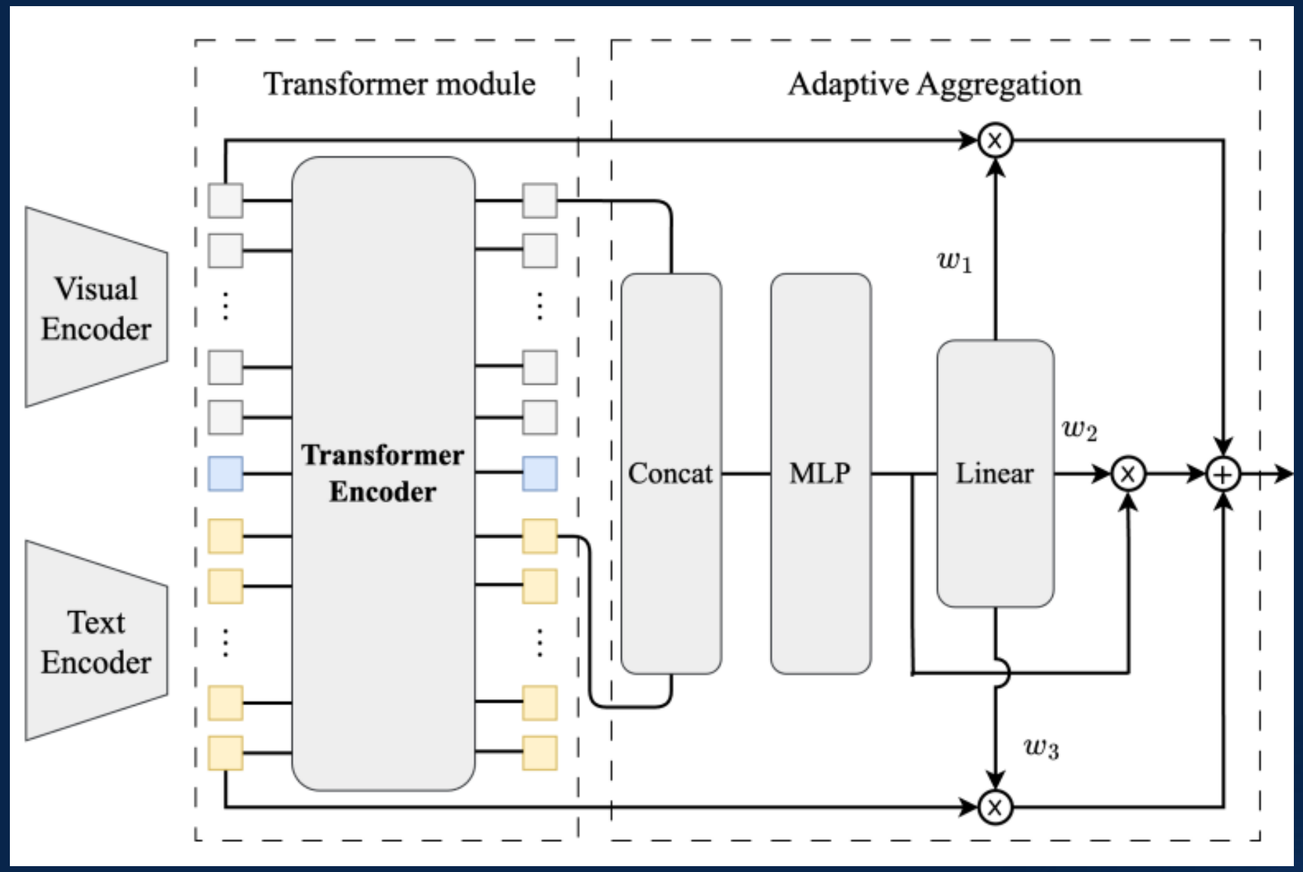
Existing Approaches and Their Problems

- Supervised models → require large annotated triplets(a reference image, a relative caption, and a target image)
- Manually constructing annotation is costly, slow and domain-limited
- Weak at generalization (new datasets, unseen domains)

- Solution proposed:
  - Automatic dataset construction from image-caption pairs(e.g., LAION-COCO)
  - New model: TransAgg with transformer fusion + adaptive aggregation

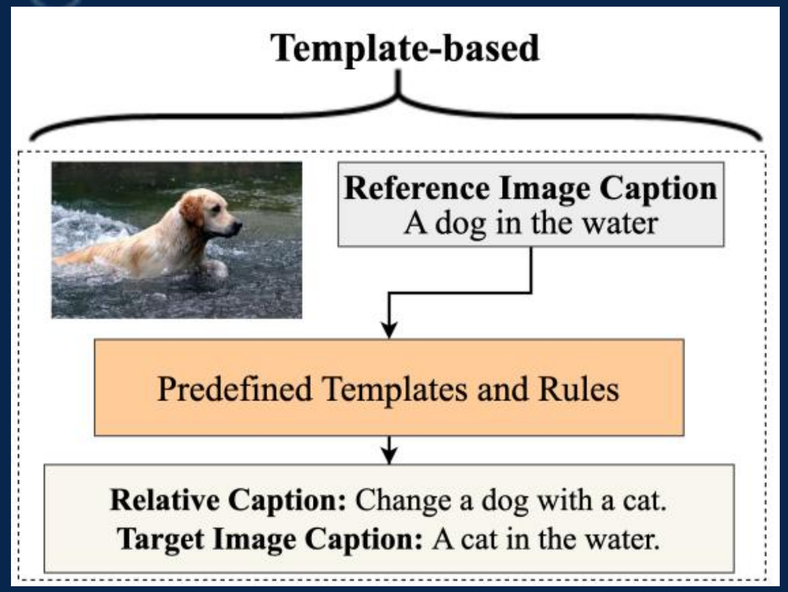
Architecture

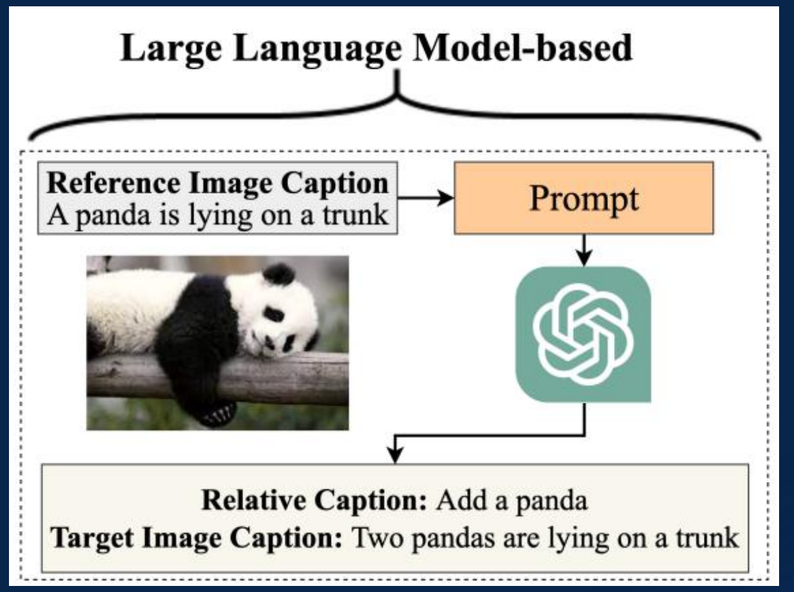
Make it Black →



Dataset Construction

- Starting with LAION-COCO image-caption pairs
- Create “relative captions” using:





- Match edited caption with real target image (sentence transformer similarity)
- Datasets: Laion-CIR-Template (16k), Laion-CIR-LLM (16k), Combined (32k)

Experimental Setup

Training Datasets:

- Laion-CIR-Template (16k)
- Laion-CIR-LLM (16k)
- Combined (32k)

The authors evaluate on two benchmarks:

- CIRR (~36k triplets, general natural images, derived from NLVR2).
- FashionIQ (~30k triplets, domain-specific fashion categories: Dress, Shirt, Toptee).

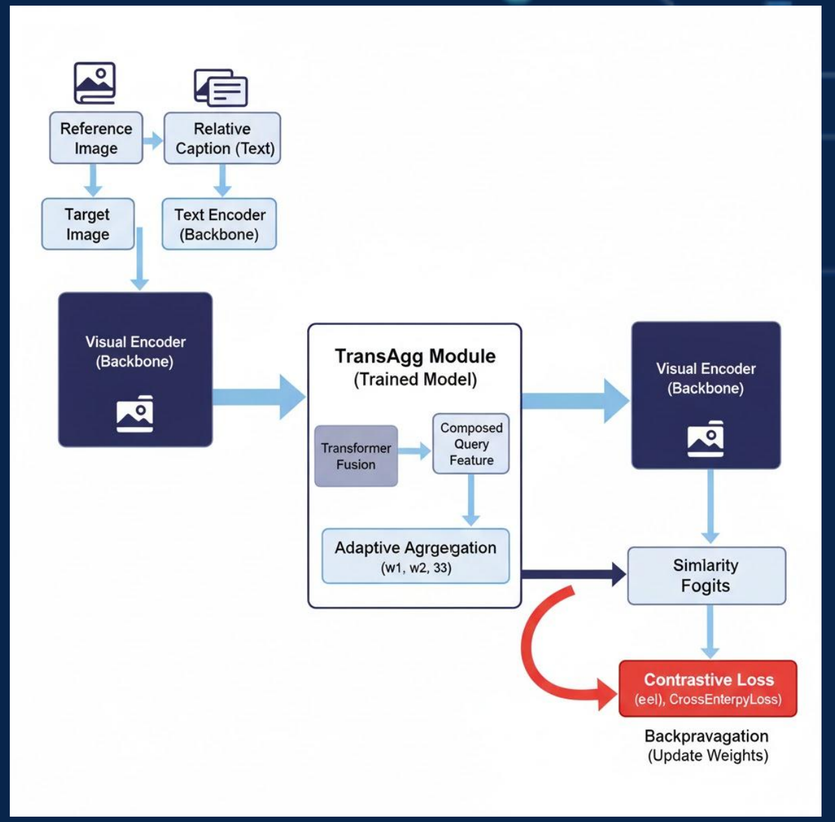
Zero-shot evaluation: The model is trained only on the automatically built Laion-CIR datasets, then directly tested on CIRR and FashionIQ—no fine-tuning.

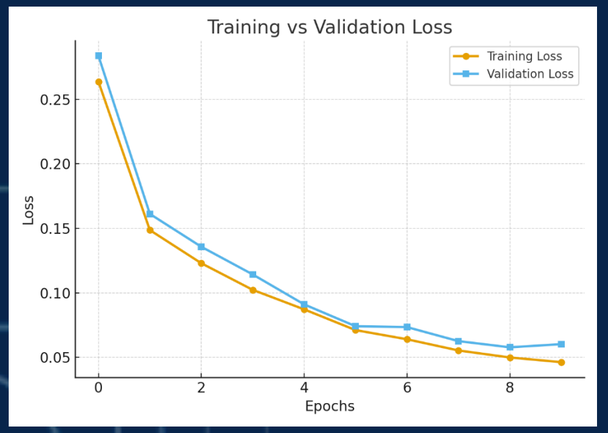
- Metrics: Recall@K (standard retrieval metric)

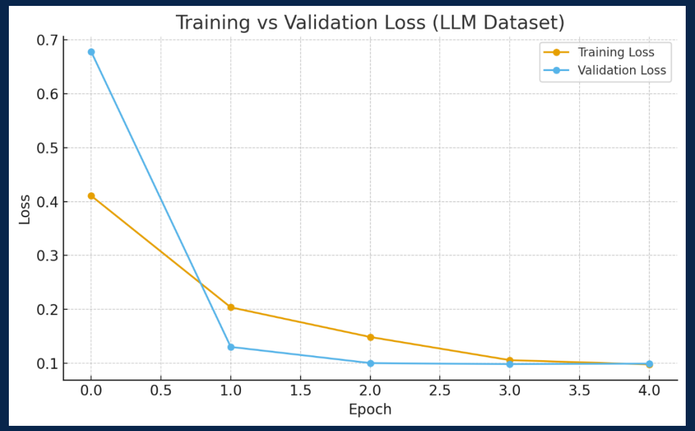
Sentinels

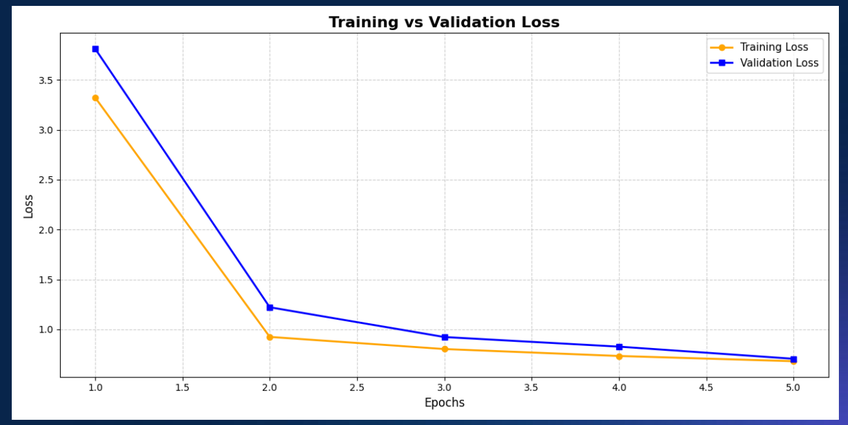
Training

- We used clip-Vit-B/32 and Blip as based models

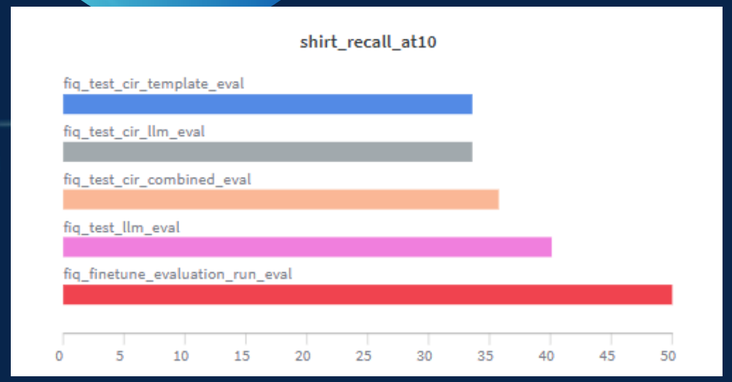


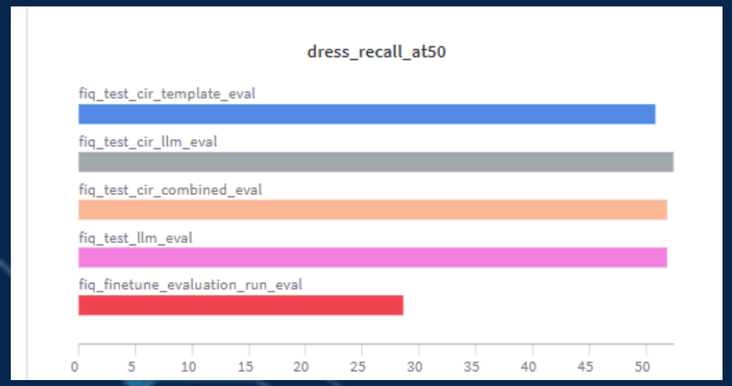


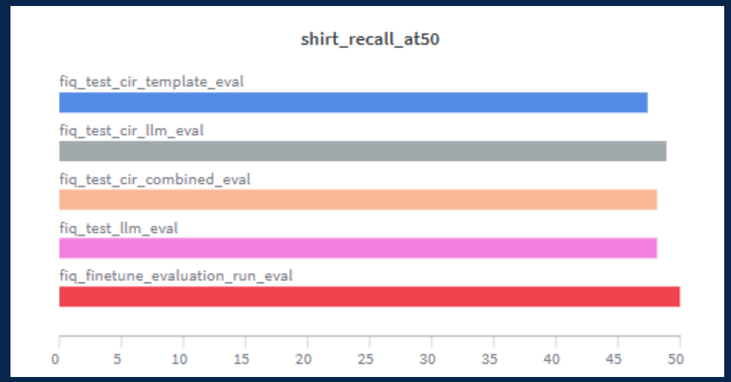




Comparison







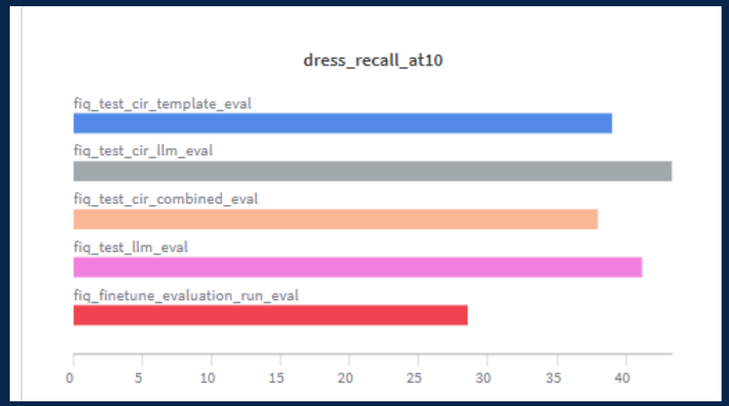


Table 1: FashionIQ Recall results for TransAgg (Laion-CIR) variants.

Method	R@10	R@50	Average
TransAgg (Laion-CIR-Template)	32.07	53.26	42.67
TransAgg (Laion-CIR-LLM)	32.77	53.44	43.11
TransAgg (Laion-CIR-Combined)	34.36	55.13	**44.75**

Table 2: Performance of CIR models trained by us on FashionIQ dataset.

Method	R@10	R@50	Average
TransAgg (Laion-CIR-Template)	36.31	49.13	**42.72**
TransAgg (Laion-CIR-LLM)	38.45	50.66	**44.56**
TransAgg (Laion-CIR-Combined)	36.88	50.02	43.45

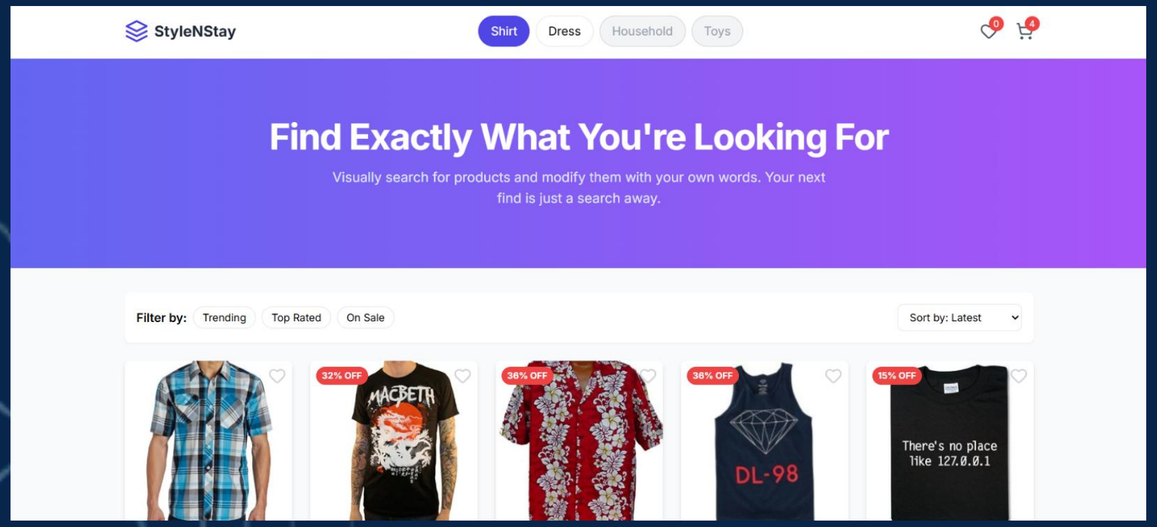
Table 3: BLIP finetuned model FashionIQ.

BLIP	**34.64**	**55.72**	**45.18**

Novel Application

An e-commerce app like *Daraz* or *AliExpress* where users can explore thousands of fashion items.

- Instead of relying purely on text searches, our system lets customers refine their product search using **both an image and natural language edits**.

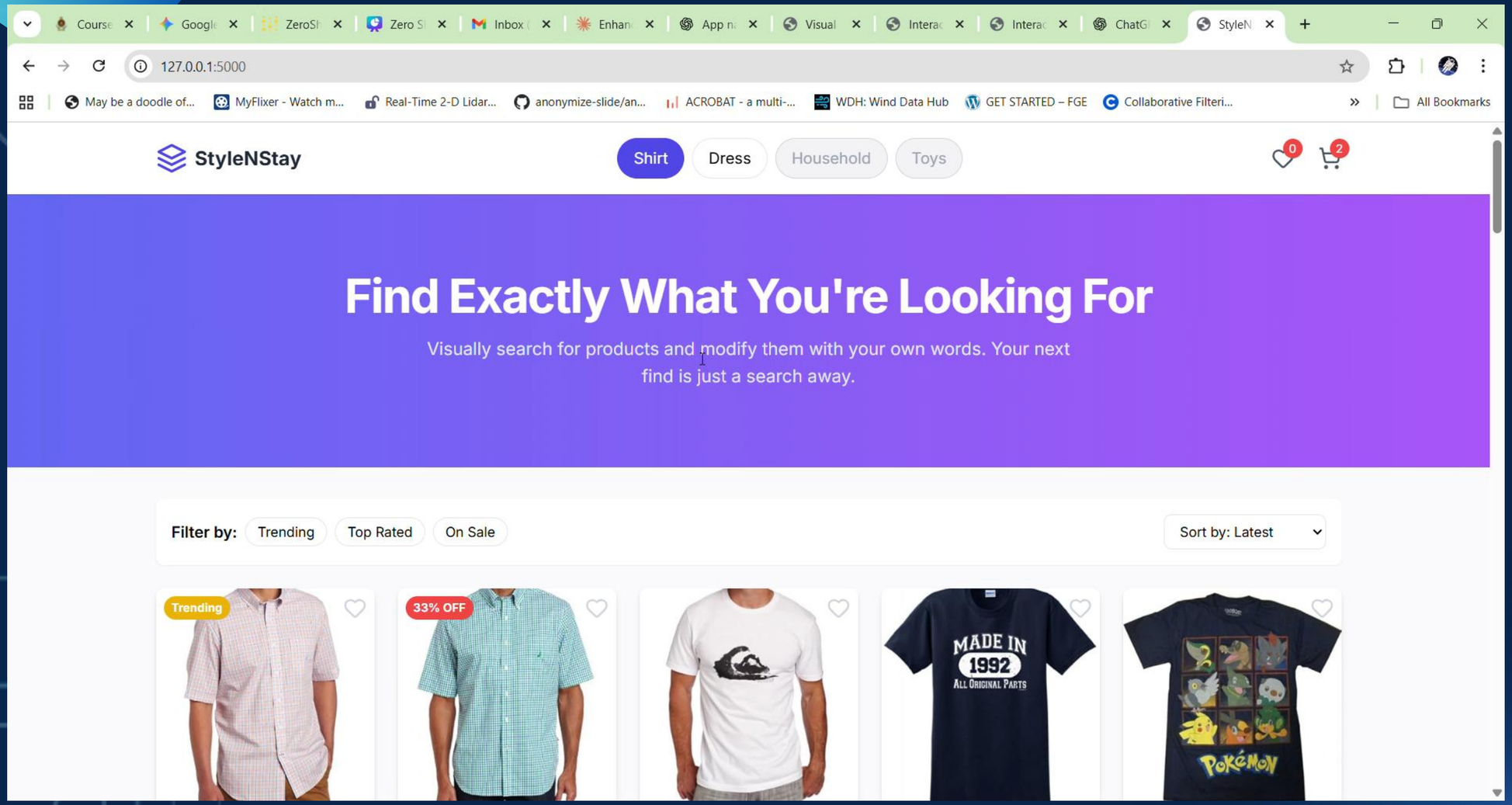


This approach enables:

- **Multimodal search** (image + text combined).
- **Fine-grained product discovery** based on visual semantics.
- **Improved personalization** since users can describe *how* they want an item changed.

Sentinels

Demonstration



Thank You !