Crossing Minds

UNIVERSITY OF TORONTO

Captions Are Worth a Thousand Words

Enhancing Product Retrieval with Pretrained Image-to-Text Models

Jason Tang[1], Garrin McGoldrick[2], Marie Al Ghossein[2], **Ching-Wei Chen[2]**

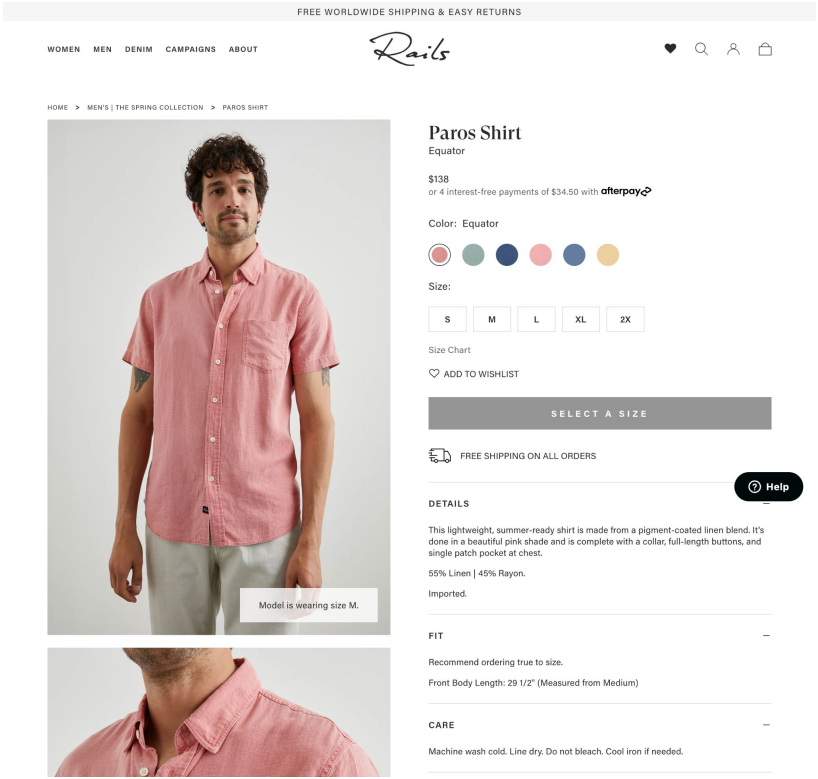
[1] University of Toronto, [2] Crossing Minds

Presented at ISIR-eCom @ WSDM 2024
Mar 8th, 2024, Mérida, Mexico

Crossing Minds

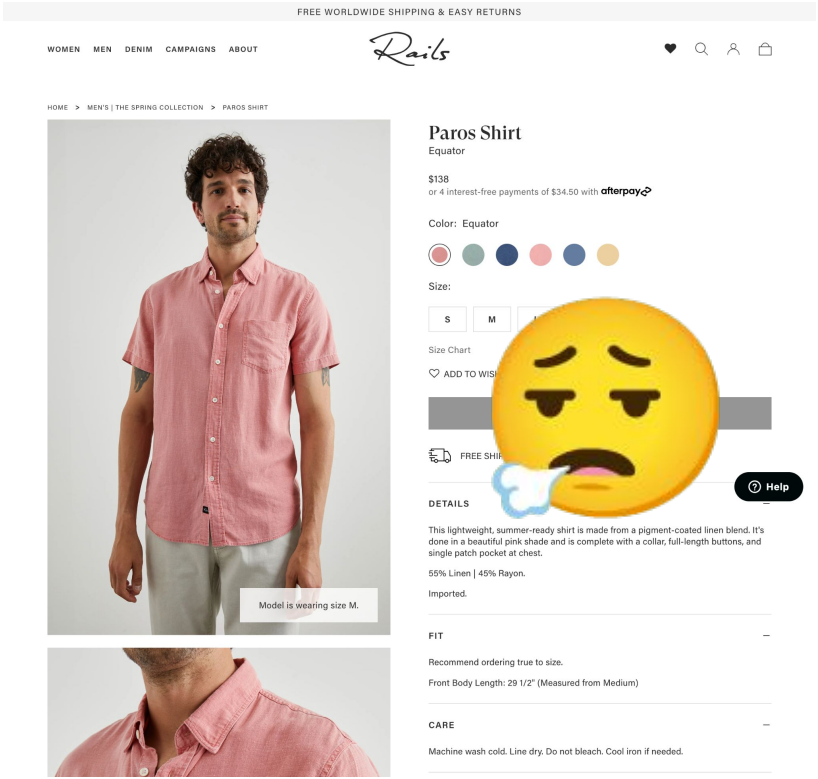
Metadata, Search and Discovery

- Metadata is essential to helping customers find and discover products in a catalog.
- Descriptive product metadata like Name, Color, Size, Description, etc, are heavily used in Search as well as Recommendations.



Metadata, Search and Discovery

- But metadata is costly to manually curate, and many eCommerce sites have varying coverage, quality, and granularity of these tags and descriptors.
- But almost all eCommerce sites will have high quality product images, and paragraph-level product descriptions.



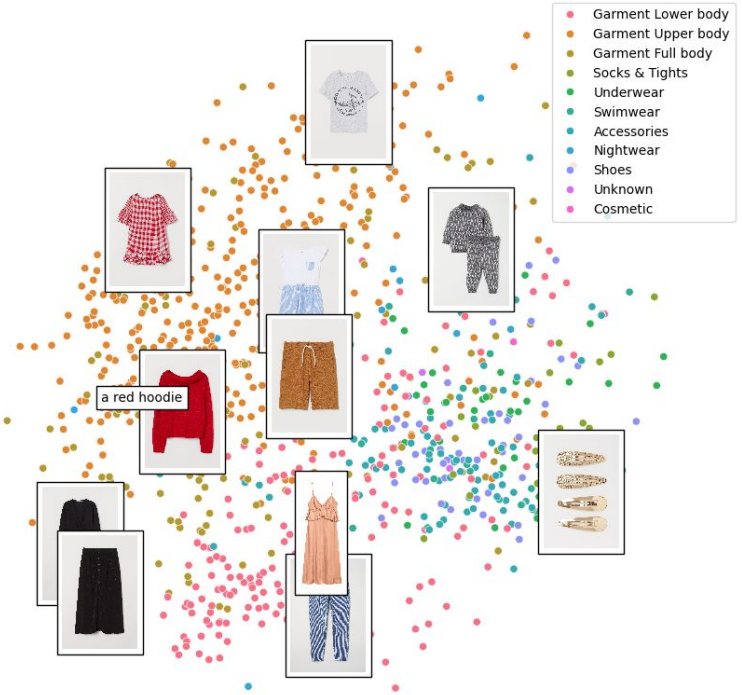
Related Work

CLIP: Image Tagging

- Discriminative image-to-text tagging
- Pre-training task: Predict (image, text) pairs that occur across a batch

FashionCLIP

- CLIP fine-tuned on fashion dataset from Farfetch



[CLIP] "Learning Transferable Visual Models From Natural Language Supervision", Radford et al., ICML'21
[FashionCLIP] "Contrastive language and vision learning of general fashion concepts", Chia et al., Scientific Reports 12 (2022)

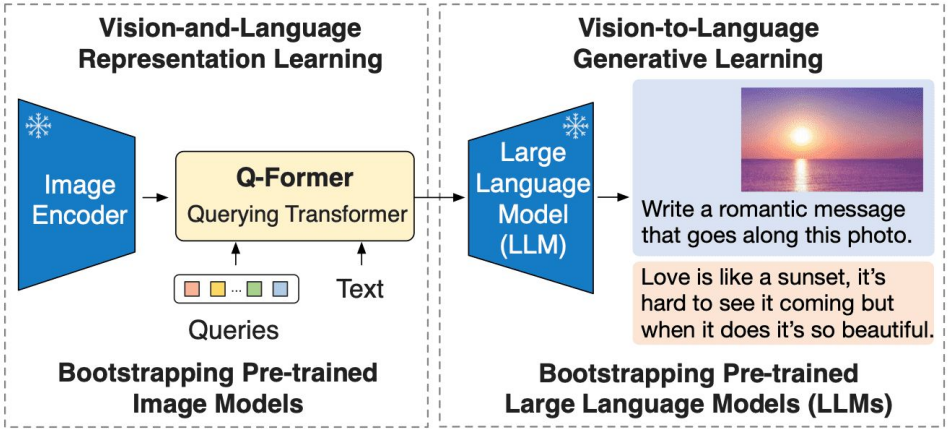
Related Work

BLIP-2: Image captioning

- Generative image-to-text captioning
- Pre-trained jointly with multiple vision-language objectives

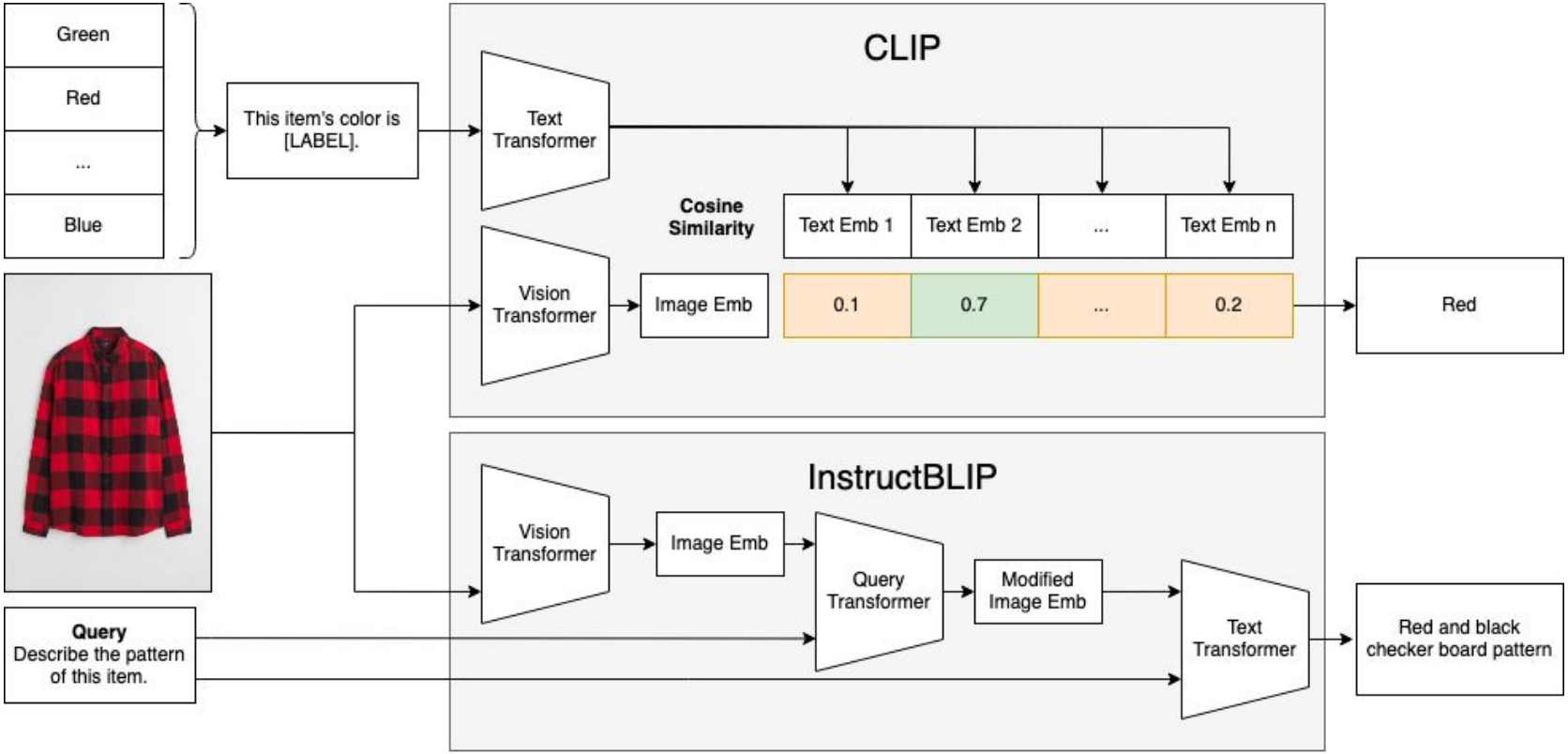
InstructBLIP

- Instruction tuning format for further pre-training



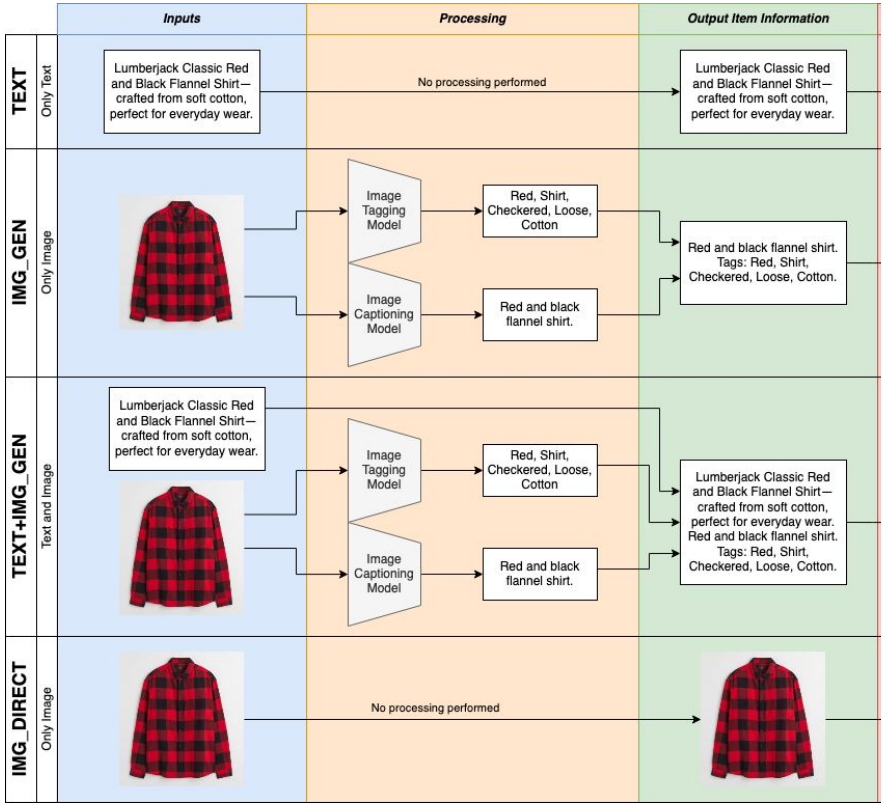
[BLIP] "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models", Li et al., ICML'23
[InstructBLIP] "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning", Dai et al., ArXiv abs/2305.06500 (2023).

Proposed Method



Crossing Minds

Proposed Method



Proposed Method

- Query preprocessing to address misspelled queries and those containing non-English terms

- Done with ChatGPT-3.5, e.g.,

"Extract at least 5 related tags or usage keywords from queries. Output in English as a comma separated list."

User Query	Processed Query
!awnmower tires without rims	lawnmower, tires, without rims
#20 paper bags without handle	paper bags, without handle, packaging, eco-friendly, retail
paws	animal, pets, claws, dogs, cats
apple iphone 11 pro unlocked	apple, iPhone, 11 pro, unlocked
자전거트레일러	bicycle trailer, bike trailer, cycling trailer, bike cart, bike carrier
眼镜框	eyeglass frames, glasses frames, eyewear, spectacle frames, glasses

Evaluation

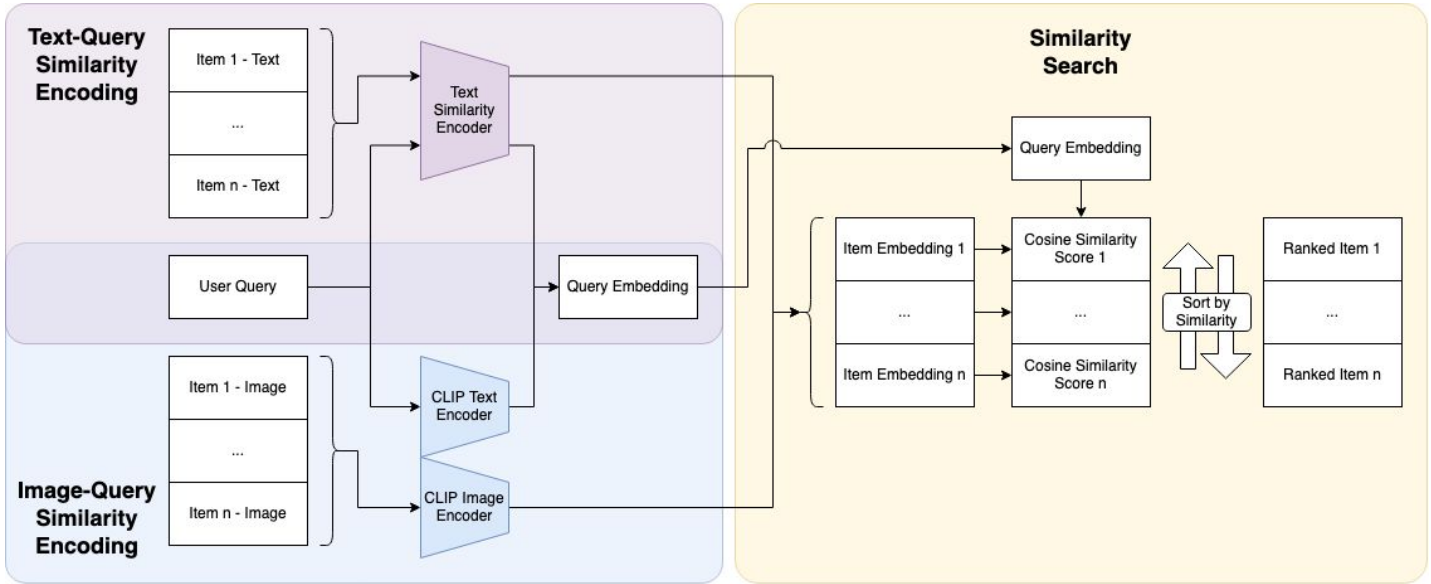
- The Amazon ESCI dataset consists of a list of query-product pairs annotated with E/S/C/I (Exact, Substitute, Complement, Irrelevant) labels
- Focus on task 1 of the KDD Cup'22: Query-Product Ranking and on the English subset of the ESCI dataset
- Use product identifiers to scrape images and select first image to constitute dataset

Dataset sampling

- Filter the dataset and consider products that appear in at least 3 queries (out of the 482k products) for efficiency reasons, given the cost of retrieving images
  - Resulting dataset: 21.6k products, 19.8k queries, 4.3 products per query
- Use padding, consisting in adding random products per query with `I` label, in order to mitigate the facts that:
  - The datasets mainly contain E and S labels per query,
  - Relevant information (e.g., product dimensions and version) might not be discernible from images

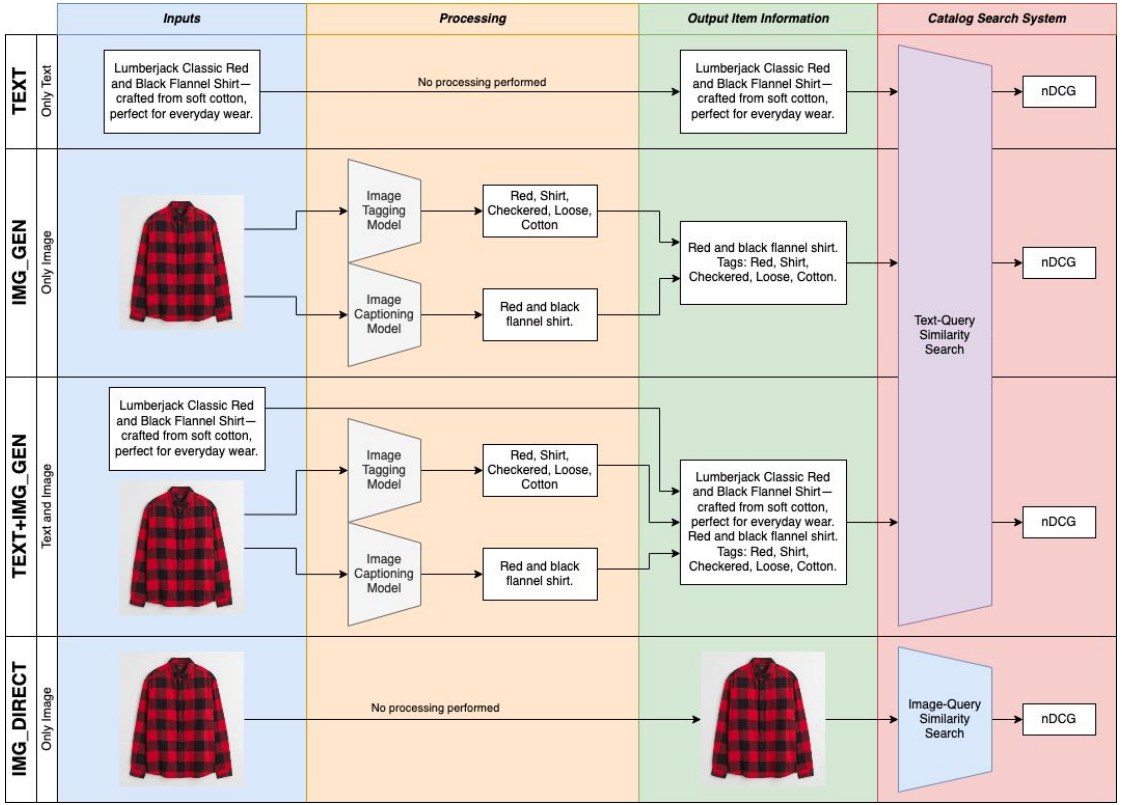
Similarity Search

- Multi-qa-mpnet-base-dot-v1
- ms-marco-MiniLM-L-2-v2



Crossing Minds

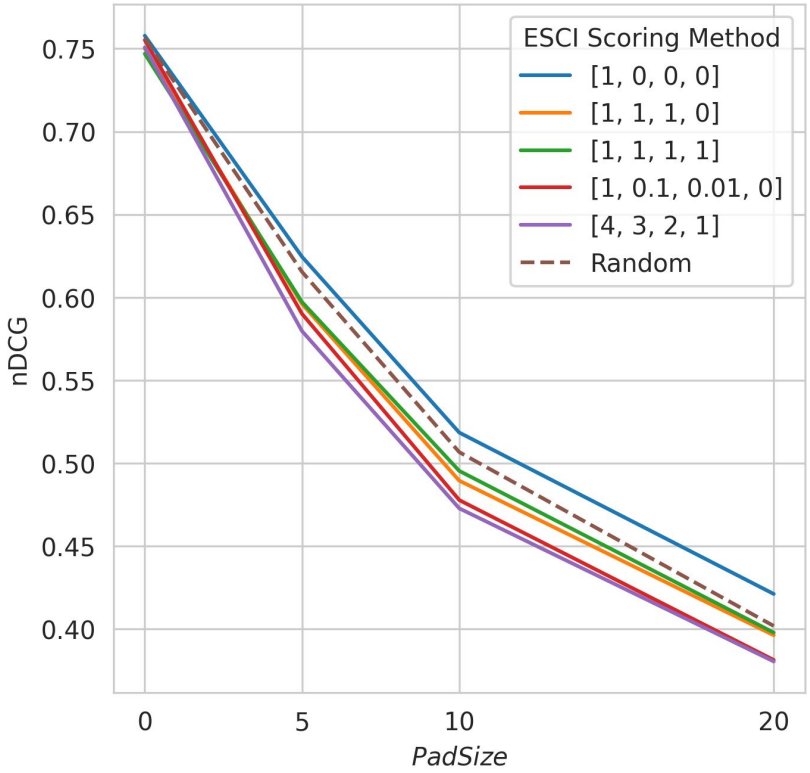
NDCG



Baselines

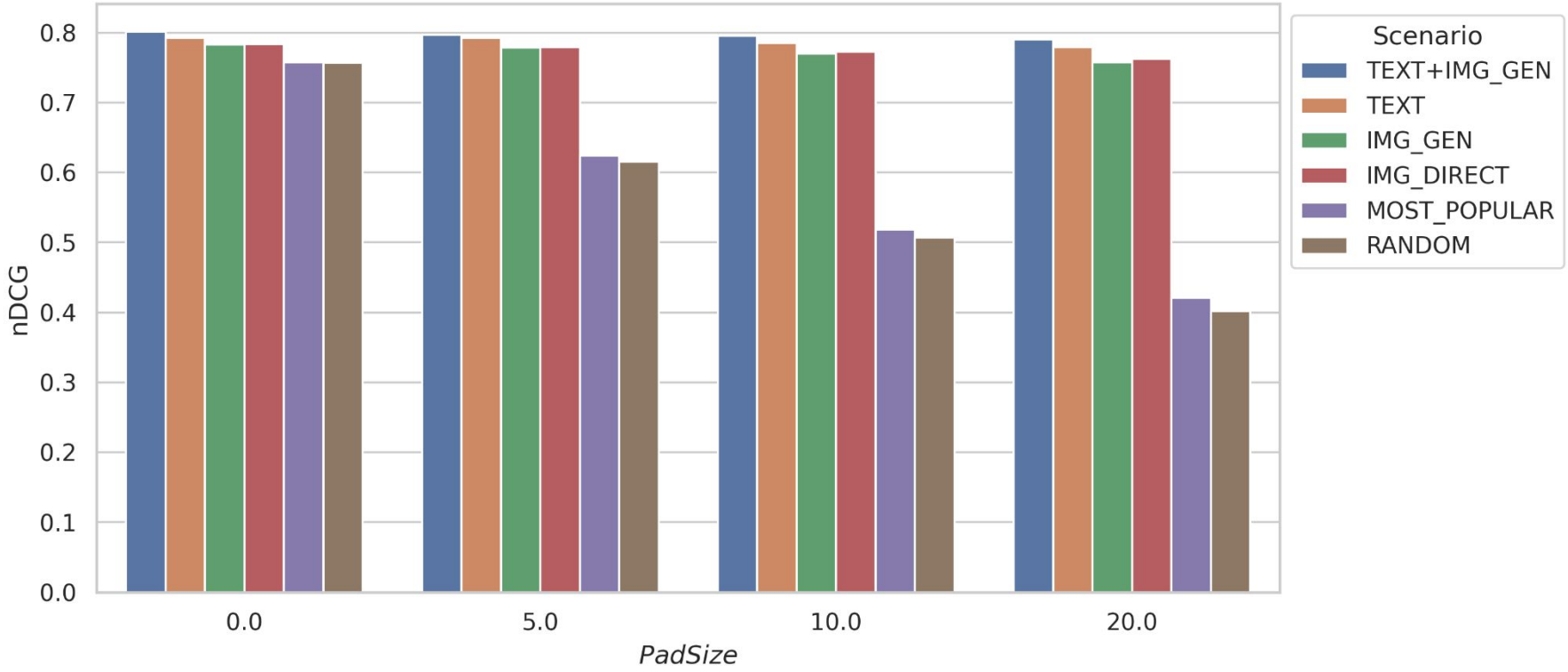
- Random
- Most popular,

  with popularity computed according to different scores attributed to the [E, S, C, I] labels



Crossing Minds

Results



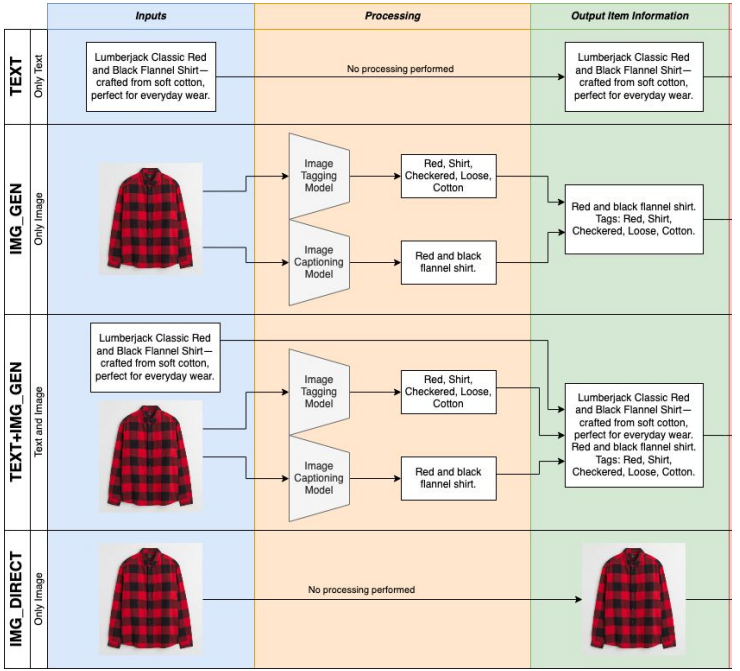
Results

- Query preprocessing showed an increase in performance across the board.

PadSize	Original Query	GPT Preprocessing
0	0.780	**0.782**
5	0.767	**0.774**
10	0.756	**0.762**
20	0.734	**0.745**

Conclusion

- We demonstrated the ability of multimodal models to generate high-performing image-derived descriptions that enable eCommerce platforms without substantial textual metadata to supplement existing text or images to improve item retrieval performance.



Future Directions

- Explore larger LLMs
- Evaluate against BM25 search engine baseline
- Apply to winning techniques from KDD Cup 2022

Crossing Minds

Thank You!