# DS@GT at Touché: Image Search and Ranking via CLIP and Image Generation

Notebook for the Touché Lab at CLEF 2024

Benjamin Ostrower[1,†], Patcharapong Aphiwetsa[1,†]

[1]Georgia Institute of Technology, 225 North Avenue, Atlanta, 30332, United States

**Abstract**

Our team made 2 submissions in the task "Image Retrieval for Arguments", where our submission focused on retrieving images. Our two runs made use of Image Generation comparison and CLIP embeddings.

**Keywords**

Image Generation, CLIP, Image Retrieval

## 1. Introduction

The exponential growth of digital imagery has profoundly influenced various fields, ranging from social media and entertainment to scientific research and healthcare. The importance and proliferation of visual media will only accelerate as a form of efficient communication, hence the phrase "a picture is worth a thousand words". Touché offers a competition on selecting the most relevant images from a crawled corpus for a set of arguments. Therefore we attempted to enter in this touche task to improve on solutions for retrieving images related to arguments. We wanted our solutions to only focus on images and descriptions of images, to try and not use any webpage text. Our solution approaches focused around combining image descriptions and the image itself as a comprehensive unit and then for our other submission implementing one more step on top of that to add a comparison to generated images that used the arguments themselves as prompts for the generated images.

## 2. Background

### 2.1. Related work

The defining paper for retrieving images for arguments is by Kiesel [1]. In it they use natural language processing techniques found in the web text of surrounding these images to create expanded keyword searches in the web text to attempt to track the stance of an argument. Last year at Touché 2023 team Picard [2] constructed a similar solution - one of their submissions involved image generation using stable diffusion.The authors prompt the image generation with the arguments from the competition to create a benchmark image that is used to compare the competition images searching for the most similar using CLIP.

### 2.2. CLIP

CLIP [3] stands for Contrastive Language Image Pretraining. It is a model developed by OpenAI that embeds images or texts into the same vector space by training on images with their corresponding captions. It is helpful to reduce the dimensionality of text or images, but still be semantically similar to retrieve relevant results from one modality to the other.

*CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France*

† Authors contributed equally

bostrower3@gatech.edu (B. Ostrower); paphiwetsa3@gatech.edu (P. Aphiwetsa)

0000-0000-0000-0000 (B. Ostrower); 0000-0000-0000-0000 (P. Aphiwetsa)

### 2.3. Stable Diffusion

Stable Diffusion [4] is a neural network model that is capable of producing images given a prompt. By decomposing the image formation process into a sequential application of denoising autoencoders stable diffusion can achieve state-of-the-art synthesis results on image data.

## 3. System Overview

### 3.1. Embedding Pipeline

Both submissions made use of CLIP from OpenAI. The competition supplied the images and their corresponding image descriptions obtained from LLava. These modalities were embedded using CLIP into a 70-30 ratio of Image to text. These embeddings were stored in a chromaDB vector database for later retrieval.

### 3.2. Retrieval

Provided arguments (only the arguments no premises or claims used) were used as queries to be brushed against the vector database. The arguments were embedded using CLIP to keep the same dimensionality as the combined image-text embeddings of the images. These arguments were compared pairwise across each image in the database via cosine similarity keeping the top 10 for our initial submission.

### 3.3. Image generation

For the image generation submission for each topic images were generated off a set of tinyLLama generated supporting/detracting arguments depending on the stance. For example if the stance was pro the prompt would instruct to provide supporting claims if it was anti then it would prompt with claims to detract. The number of arguments generated would vary from 3-7. The prompt format for a supporting generation is found in figure 1.

```
{
    "role": "system",
    "content": "You are a student trained to think critically for each
        claim break it down into several subclaims",
},
{
    "role": "user",
    "content": f"Create some numbered prompts to give to a machine to
        create images that support the claim: '{prompt}'"
}
```

**Figure 1:** Prompt for supporting argument image generation. Using python and the transformers library pipeline method

These tinyLLama-generated supporting/detracting arguments were then fed into the stable-diffusion-2-1-base for image generation. These generated images are again embedded with CLIP and compared to the top 40 retrieved images using the method described in the prior section for a given argument. Because there was a varying number of images generated for each argument when comparing the crawled images to the generated images we take the highest average score across all generated images as our metric for most relevant images.

**Table 6**
NDCG values for the top 5, top 3, and most relevant image(s). The approaches are sorted according to the NDCG@5 score.

| Rank | Team | Approach | NDCG@5 | NDCG@3 | NDCG@1 |
|---|---|---|---|---|---|
| 1 | HTW-DIL | Ada-Summary | 0.428 | 0.409 | 0.404 |
| 2 | HTW-DIL | Moondream-Text | 0.363 | 0.355 | 0.356 |
| 3 | HTW-DIL | Moondream-Default-Image-Text | 0.293 | 0.302 | 0.317 |
| 4 | Baseline | BM25 | 0.284 | 0.273 | 0.293 |
| 5 | Baseline | SBERT | 0.232 | 0.225 | 0.221 |
| 6 | DS@GT | Generated-Image-Clip | 0.180 | 0.178 | 0.197 |
| 7 | HTW-DIL | Moondream-Image-Text-EP3 | 0.150 | 0.163 | 0.183 |
| 8 | HTW-DIL | Moondream-Image | 0.146 | 0.155 | 0.178 |
| 9 | DS@GT | Base-Clip-Submission | 0.123 | 0.111 | 0.106 |
| 10 | HTW-DIL | Moondream-Image-Text | 0.120 | 0.140 | 0.178 |

**Figure 2:** Approaches did not beat baseline

## 4. Results

Our approaches didn't beat baseline of BM25 and SBERT. We do see that the added filter of comparing the top results to images generated from the arguments does increase the accuracy of the model. It appears that Images alone

## 5. Conclusion

The image generated approach worked best, however both submissions didn't beat baseline. Future directions of work include re-ranking LLava Visual Question Answering generations - i.e. ask LLava to describe the picture in relevance to argument in question. Utilizing BM25 and webpage text to decipher keywords that might indicate the relevance of the image.

## Acknowledgments

The Data Science at Georgia Tech Club.

## References

[1] J. Kiesel, N. Reichenbach, B. Stein, M. Potthast, Image Retrieval for Arguments Using Stance-Aware Query Expansion, in: K. Al-Khatib, Y. Hou, M. Stede (Eds.), 8th Workshop on Argument Mining (ArgMining 2021) at EMNLP, Association for Computational Linguistics, 2021, pp. 36–45. doi:10.18653/v1/2021.argmining-1.4.

[2] M. Moebius, M. Enderling, S. T. Bachinger, Jean-luc picard at touch\'e 2023: Comparing image generation, stance detection and feature matching for image retrieval for arguments, arXiv preprint arXiv:2307.09172 (2023).

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.