2023 IEEE Applied Imagery Pattern Recognition Workshop (AIPR) | 979-8-3503-5952-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/AIPR60534.2023.10440710

Exploring CLIP for Real World, Text-based Image Retrieval

Manal Sultan
Computer Science
California Institute of Technology
Pasadena, California
msultan@caltech.edu

Lia Jacobs
Computer Science
Ohio State University
Columbus, Ohio
jacobs.1236@osu.edu

Abby Stylianou
Computer Science
Saint Louis University
Saint Louis, Missouri
abby.stylianou@slu.edu

Robert Pless
Computer Science
George Washington University
Washington, DC, United States
pless@gwu.edu

***Abstract*—We consider the ability of CLIP features to support text-driven image retrieval. Traditional image-based queries sometimes misalign with user intentions due to their focus on irrelevant image components. To overcome this, we explore the potential of text-based image retrieval, specifically using Contrastive Language-Image Pretraining (CLIP) models. CLIP models, trained on large datasets of image-caption pairs, offer a promising approach by allowing natural language descriptions for more targeted queries. We explore the effectiveness of text-driven image retrieval based on CLIP features by evaluating the image similarity for progressively more detailed queries. We find that there is a sweet-spot of detail in the text that gives best results and find that words describing the "tone" of a scene (such as messy, dingy) are quite important in maximizing text-image similarity.**

***Index Terms*—Deep Learning, Image Retrieval, Human Computer Interaction**

I. Introduction

Deep learning approaches to image retrieval are important in a variety of application domains, including image provenance detection in journalism, image localization in intelligence investigations, and aiding medical diagnosis through matching and retrieving relevant medical imagery. Purely image-based queries, where the input is an image and the task is to find images that are visually similar in some way, often fail to align to human-in-the-loop needs and intuitions, especially when the image retrieval models do not focus on components of the image that the human wants to focus on, or when results key on aspects of the image that the user wants to ignore.

Here we explore how well recent advances in text-based image retrieval could help. The hope is that the text interface presents a flexible interface for a user to describe exactly what they want to focus on via natural language prompts. We use Contrastive Language-Image Pretraining (CLIP) [9] to translate the prompts and our images into a common feature space. For a given prompt, the images with the most similar representations in CLIP space are returned as matches.

CLIP is optimized based on a large-scale dataset comprising billions of text-image pairs, allowing it to learn a wide range of visual concepts from natural language supervision. This model is trained using a contrastive learning approach, where it learns to closely associate the embeddings of corresponding text and image pairs while distancing the embeddings of non-matching pairs. Ideally, this means that CLIP develops a nuanced understanding of both visual and textual inputs, enabling it to map both modalities into a shared multidimensional feature space where comparisons and retrievals can be effectively performed. But the large training datasets have biases; the text part of the text-image pairs on which CLIP is trained are often shockingly terse, and/or describe only a small part of the image features. Figure 1 offers examples of the 10 best images retrieved for a simple query, as well as the text part of the text-image pair that is used in training CLIP.

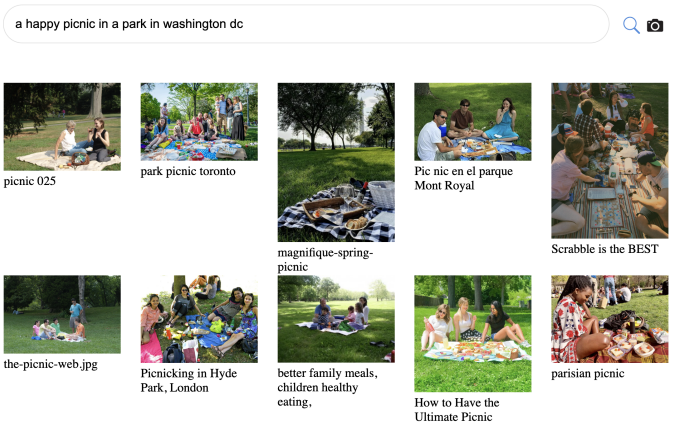


Fig. 1: Text-based image retrieval starts with a text query like: "A happy picnic in a park in Washington DC" and returns the most similar images. These images are retrieved from the set of images that CLIP was trained on, with the corresponding caption from that text image pair shown below the image. These results highlight (a) Some features, like the concept of "picnic" are cleanly captured by text-based CLIP image retrieval, while (b) some concepts (like in Washington DC) are not prioritized in the top responses, and (c) the captions often describe only one feature of the image (e.g. "picnic", or "Dcrabble".)

Because CLIP models project images and text into a common space, this approach has the potential to answer these queries. However, we find this approach often fails for queries where a human writes specific or complex queries, especially those that are longer and more detailed than the training

data. This limits the utility of CLIP for many real world applications of text-based image retrieval. For example, a query that says "find images of rooms with green carpet" may return reasonable images, but a query that says "find images of grungy rooms with stained green carpet, a dark wooden table, a striped couch, and an air conditioner on the wall" typically does not.

These types of queries may arise in image search queries in application domains like hotel classification in sex-trafficking investigations. In these investigations, there are sometimes images of hotel rooms whose identification could find potential trafficking victims or support more accurate understanding of the context of a case. Given an image of a hotel room, there is a need large scale datasets of images of hotels. These datasets arise from scraping open source datasets [14], or crowd-sourcing data via dedicated apps [15]. Datasets which have been published (e.g. [16]) support image-based image-retrieval where an image that comes up in an investigation is used as a query and the goal is to retrieve an image from the same hotel.

In many cases, images that are available for the search are very occluded, often because sensitive image regions are blocked. Figure 2 (top) shows an example image that does have relevant image features, but the main part of the image is masked off. When so much of the image is missing, even modern in-painting approaches are not able to recover enough of the image to effectively search with the image itself [3]. In such cases, we might hope that a human user could craft a text query that describes what they can see in the image in hopes that this will drive a more effective search. Figure 2 (bottom) highlights the top responses from the large CLIP training dataset (LAION 5B) to an example complex descriptive query.

In this paper we start an exploration of why CLIP might fail to support these queries. We exhibit CLIP model performance on queries with varying levels of complexity, and create qualitative evaluations of CLIP-based image generation models for image generation from prompts with varying levels of complexity.

Our specific contributions are:

- Exploring image similarities between increasingly specific text descriptions of an object and the best image responses.
- Exploring image similarities between increasingly specific text descriptions of a hotel room scene and the best image responses.
- Documentation of the image similarities between text descriptions that include mention on scene appearance other than the objects in the room (e.g. "messy").

II. Background and Prior Work

This paper explore the effectiveness of a CLIP model that is co-trained on images and text for queries with varying levels of complexity. Contrastive Language-Image Pretraining (CLIP) models [9] are co-trained on large collections of text and image data, taking pairs of related images and text (for example, an image and its caption), and using those pairs to simultaneously train a Deep Neural network to embed the image and text in similar locations in an embedding space.

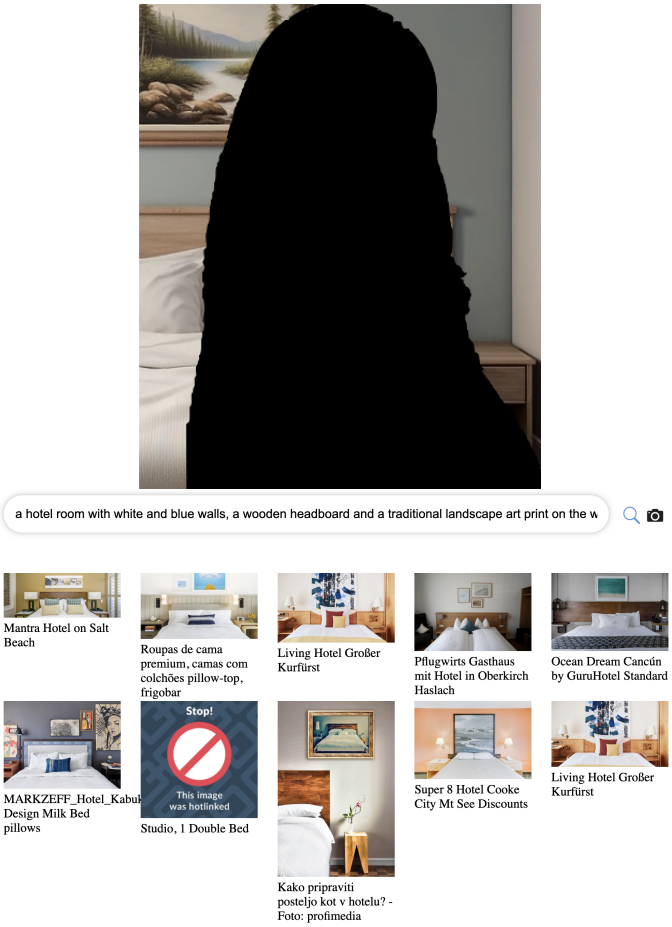


Fig. 2: For highly occluded query images (top), we hope that users might be able to construct a relevant text query, like: "a hotel room with white and blue walls, a wooden headboard and a traditional landscape art print on the wall above the bed," but the best responses often fit only one or two of the text concepts well. As in the previous figure, the poor detail in the image captions on which CLIP is trained may explain the limited effectiveness of the CLIP features at matching more complex combinations of visual features. Note, the top image was artificially generated to look substantially like a query image that arose in an active investigation. The "missing image" figure on the second row is intentionally left in to highlight challenges in using large open source databases of image links.

Studies of this CLIP space have shown impressive performance in zero-shot image classification [8], perhaps because CLIP space is composable [1] allowing it to naturally support queries that combine multiple features. It has also been shown to exhibit racial bias [18], gender bias [7], presumable inherited from the large training data sets that are needed to train it.

In this paper we explore image retrieval within one dataset that is commonly used to train CLIP, the LAION datasets, first released as a set of 400 million image text pairs [13], then as a set of 5 billion image text pairs [12]. A website support exploration and text-queries

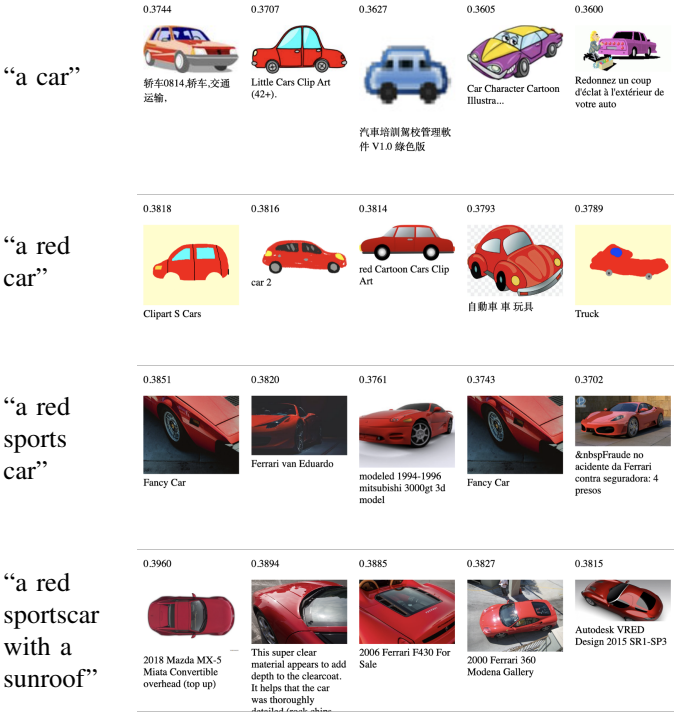


Fig. 3: A series of increasingly specific text queries for a car. The top results for these queries are good fits to the text, even though the actual similarity scores are not so high.

of this dataset, showing images both similar to the query, and the text that corresponds to the retrieved image: https://rom1504.github.io/clip-retrieval [2]

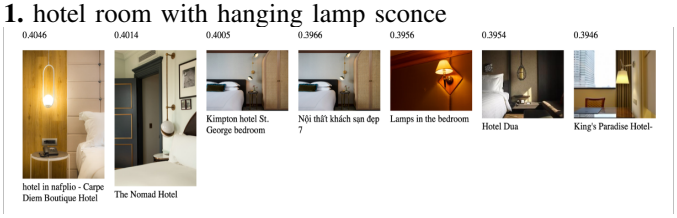
Specific studies of the relationship of images and text prompts have been most common in the design of text-prompts for image generation, rather than image retrieval.

This first commonly used such model was Dall-E [10] which created convincing, interesting and novel images from text descriptions. The ability to study these models was increased when the open-source CLIP code was used to train Stable Diffusion [11]. A recent study was carried out of 72000 text prompts that users of Stable Diffusion used to create imagery [4]. They found that these prompts almost always included keywords across multiple categories, and the categories most relevant to our image retrieval goals included the subject (or objects in the scene) in addition to the mood, the tone and the lighting. This finding drove our study of including tone words in Section III-C

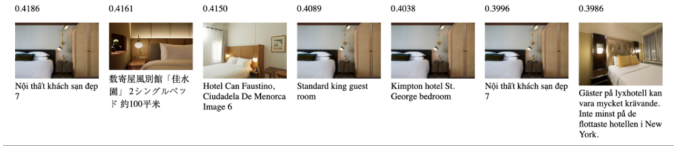
III. Experiments

In these section we explore the best results for a collection of different text queries. In all experiments, we use the CLIP Image Retrieval demo [2], which is a graphical user interface to interact with the OpenCLIP Vision Transformer model (ViT-B/32) trained on the LAION-5B dataset [5]. We first show the case of describing a specific car, and then the case of describing a specific hotel room. We show results that include the image and the similarity score between the image and the query text. The final experiment considers not just text with

1. hotel room with hanging lamp sconce

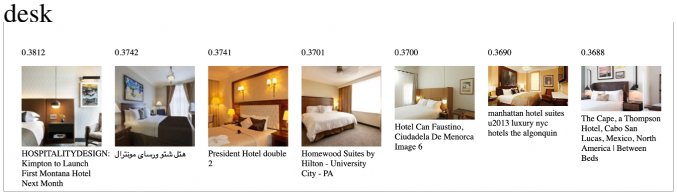


2. hotel room with hanging lamp sconce with gold body



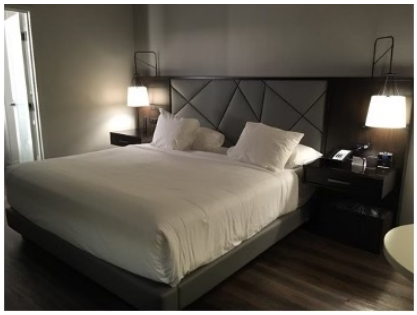
3. hotel room with hanging lamp sconce with gold body and spherical light with brown grid headboard

4. hotel room with hanging lamp sconce with gold body and spherical light with brown grid headboard and painting in gold frame and white curtain blinds and dark brown wooden desk



5. hotel room with hanging lamp sconce with gold body and spherical light with brown grid headboard and painting in gold frame and white curtain blinds and dark brown wooden desk and brown wooden nightstand with a black landline phone on top next to the bed

Fig. 4: Responses to descriptions of a hotel room with increasing levels of detail. In this case we see that the increasing level of detail makes the top similarity score decrease. Additionally, the most similar images do not have all the features listed in the text query.



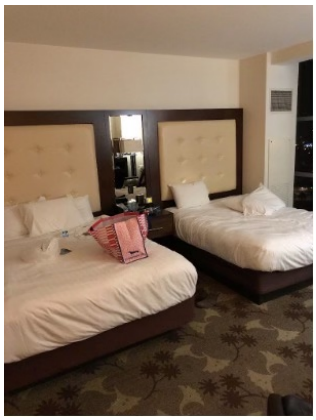
Example Case 1

L1: A hotel room with two beds with bedding and a bed skirt and a headboard. There is a mirror and a lamp sconce. There is a nightstand with a drawer. There is carpet. *Cosine Similarity: 0.369*

L2: A hotel room with two beds with white bedding and a brown bed skirt and a headboard with a cream center. There is a mirror and a lamp sconce. There is a brown nightstand with a drawer. There is brown carpet. *Cosine Similarity: 0.415*

L3: A hotel room with two beds with white striped bedding and a brown bed skirt and a wall-mounted brown headboard with a cream tufted center. There is a mirror and a wall-mounted lamp sconce with a grey metal frame and cream rectangular lampshade. There is a dark brown nightstand with a grey metal handle on the drawer. There is brown carpet with a brown and white flower pattern. *Cosine Similarity: 0.154*

L4: A hotel room with two beds with white striped bedding and a brown bed skirt and a wall-mounted dark brown wooden headboard with a cream fabric tufted center. There is a mirror and a wall-mounted lamp sconce with a grey metal frame and cream rectangular lampshade mounted to the mirror. There is a dark brown wooden nightstand with a grey metal handle on the drawer. There is brown carpet with a dark brown and white flower pattern. *Cosine Similarity: 0.060*



Example Case 2

L1: A hotel room with a king bed with bedding and a headboard. There are two wall sconces. There are two nightstands. *Cosine Similarity: 0.081*

L2: A hotel room with a king bed with white bedding and a grey headboard. There are two wall sconces with a black body. There are two brown nightstands. The wall is grey. The floor is brown. *Cosine Similarity: 0.742*

L3: A hotel room with a king bed with white bedding and a grey headboard with an abstract line pattern. There is a dark brown panel on the outside of the headboard. There are two hanging wall sconces with a black body. There are two dark brown nightstands. The wall is solid grey. The floor is dark brown. *Cosine Similarity: 0.132*

L4: A hotel room with a king bed with white bedding and a grey fabric headboard with an abstract line pattern. There is a dark brown wooden panel on the outside of the headboard. There is a hanging wall sconce with a black metal body hanging above each dark brown wooden nightstand on each side of the bed. The wall is solid grey. The floor is dark brown wood. *Cosine Similarity: 0.045*

Fig. 5: Two pairs of images and their generated captions with increasing levels of complexity.

increasingly complete descriptions of which objects are in the room, but also describing the scene overall, outside of the objects in the scene.

A. Object Retrieval with Increasingly Complex Descriptions

We first consider searching for images of a specific object as we describe that object with increasing detail. We chose a car because there are many images of cars in the LAION database, and because we feel readers may have good intuitions about how well the images match the text description. The results are summarized in Figure 3.

As the text description becomes more specific, the car images retreived match more and more specifically to the text description. In general, the similarity scores of the images become higher, as one might expect when describing the object you'd like to see more and more accurately.

The success of this trial led us to consider CLIP as a tool that might apply to hotel room search. The next section explores that direction more directly.

B. Hotel Room Retrieval with Increasingly Complex Descriptions

We then explored a similar progression in the context of a hotel room, starting with a simple prompt that highlighted just one feature, and increasing the complexity of the prompt in order to more completely defined the returned image. Figure 4

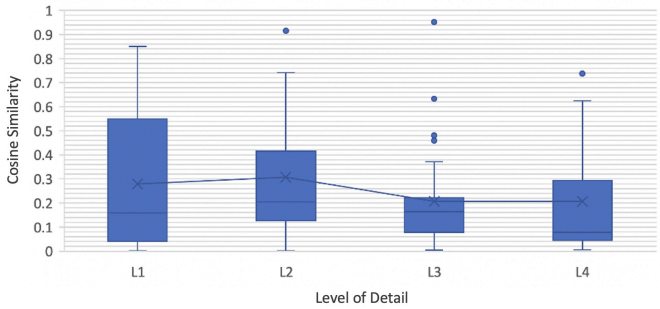


Fig. 6: We compute the similarity score for each text description with the most similar image, at all 4 levels of detail. The mean cosine similarity is highest at level 2, although the variance is high in all cases.

shows the top scoring images in the LAION dataset and the similarity scores of those image with the text. We see that the more specific descriptions do not return images with all requested features and the top scoring images have lower similarity as the level of descriptive detail increases.

We worked to formalize this across a number of possible hotel room targets and text descriptions. To evaluate the efficacy of text-based image retrieval at different levels of detail, we devised a scale to categorize the specificity of textual descriptions for a set of 46 hotel room images. The scale is defined as follows:

- **L1**: Just the object name.
- **L2**: Object type/name and color.
- **L3**: Object type/name, shape/pattern, color, and hue.
- **L4**: Object type/name, shape/pattern, color, material, hue (light/dark), relative direction of objects in relation to other objects, and specific names if possible.

In each case, this level of detail was provided for all easily visible objects in the room. Descriptions at each level of detail were generated for images of 46 different hotel rooms. The cosine similarity between CLIP embedding of each description and the CLIP embedding of the best image retrieved were computed.

We present two sets of example descriptions at varying levels of detail for our experiment in Figure 5. Each description corresponds to a unique level of detail used to query the image retrieval system.

In this case we see the level two description, that names many objects in the room and gives a detail like the color of each object gives the best scoring image matches. A second example shows the same pattern.

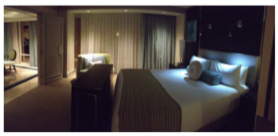
Again, the second level of detail gives the best performance. This trend carried through the analysis of all 46 image targets for which we made text descriptions. Figure 6 shows these results, while the variance is quite high, the mean similarity score is highest with the L2 score.

C. The Effect of Tone on Room Retrieval

While it is sensible to describe the objects in a scene in order to drive image retrieval, work on understanding prompting of

Query 1: "hotel room, a large bed sitting in a bedroom next to a window, 360 panorama, back room, large screen, full room view"

0.3802



Query 2: "hotel room, a large bed sitting in a bedroom next to a window, 360 panorama, back room, large screen, full room view, very elegant & complex, slick!!"

0.4028

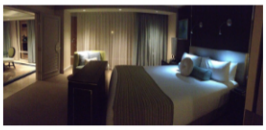


Fig. 7: Responses to text-based image retrieval queries. The left column shows the queries, while the right column shows the top retrieved image corresponding to each query. Above the image is the cosine similarity, showing improvement when tone words are added.

Query 1: "a hotel room with a bed and a desk"

0.3639

Query 2: "a hotel room with a bed and a desk, sad, ugly"

0.4161

Query 3: "a hotel room with a bed and a desk, sad, ugly, messy"

0.4173

Fig. 8: A second set of text queries that show the increased similarity score that arises from adding "tone" words in addition to object descriptions.

CLIP models highlights the importance of the description of lighting and other factors of the image appearance. We illustrate the importance of this in the text-based image retrieval context here.

We can construct object based descriptions of the room appearance, and then find the CLIP based similarity of the closest image in the database, as in the **L1** descriptions above. We can then add terms corresponding to the overall *feel* of the scene – words that don't describe specific objects in detail, but rather the overall gestalt of the scene. Examples of these tone words can be seen in Figures 7 and 8. We consistently observe that adding these tone words offers a substantial increase in

the similarity score between a caption and an image. Figure 7 shows the results on one scene, where we can see that adding tone words ("very elegant & complex", "slick") improves the similarity. In Figure 8, we show the results for another hotel room with the tone words ("sad, ugly, messy") and observe the same pattern. This increase in similarity highlights that such tone descriptors are important, although in image-retrieval situations like a hotel room search for trafficking investigations, it may not be known that the tone is of the image of the matching room in the database. Future work may explore ways to marginalize out the effect of the tone on the image matches in order to support better search algorithms.

IV. Conclusions

This study has provided observations on the capabilities of Contrastive Language-Image Pretraining (CLIP) features in the context of fine-grained, text-based image retrieval. Our experiments demonstrate that while CLIP models offer potential for understanding and mapping natural language descriptions to relevant images, there are limitations in their effectiveness, especially with increasingly complex or detailed queries. The highest performance in the hotel domain in particular was achieved with a moderate level of detail in text descriptions, emphasizing the importance of 'tone' words in enhancing the text-image similarity. However, as the complexity of queries increases, the ability of CLIP to retrieve matching images diminishes, highlighting a potential gap that is likely a result of the model's training on large-scale, yet often simplistic, image-caption pairs.

This observation opens avenues for further research, particularly in understanding the structural nuances of CLIP's feature space and developing alternative natural language interfaces for image retrieval. While we focused on the hotel retrieval domain in this work, the implications of our findings extend to diverse application domains, including journalism, intelligence, and medical diagnostics, where user-guided image retrieval is relevant. Future work might most productively explore integration of interactive, chat-based approaches to helping users build prompts [6], or more formal approaches that factor CLIP space into noun and adjective components, like the linear approaches recently proposed [17].

Acknowledgment

We gratefully acknowledge support of this research from the National Institute of Justice (2018-75-CX-0038), the Taylor Geospatial Institute Fellows Program, and the CalTech Summer Undergraduate Research Fellowship program.

References

[1] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4959–4968, 2022.

[2] R. Beaumont. Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them. https://github.com/rom1504/clip-retrieval, 2022.

[3] S. Black, S. Keshavarz, and R. Souvenir. Evaluation of image inpainting for classification and retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1060–1069, 2020.

[4] N. Dehouche and K. Dehouche. What's in a text-to-image prompt? the potential of stable diffusion in visual arts education. *Heliyon*, 2023.

[5] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt. Openclip, July 2021. If you use this software, please cite it as below.

[6] M. Levy, R. Ben-Ari, N. Darshan, and D. Lischinski. Chatting makes perfect: Chat-based image retrieval. In *Advances in Neural Information Processing Systems*, 2023.

[7] A. Mandal, S. Little, and S. Leavy. Multimodal bias: Assessing gender bias in computer vision models with nlp techniques. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 416–424, 2023.

[8] S. Pratt, I. Covert, R. Liu, and A. Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023.

[9] 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*, pages 8748–8763. PMLR, 2021.

[10] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

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

[12] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[13] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[14] A. Stylianou, A. Norling-Ruggles, R. Souvenir, and R. Pless. Indexing open imagery to create tools to fight sex trafficking. In *2015 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–6. IEEE, 2015.

[15] A. Stylianou, J. Schreier, R. Souvenir, and R. Pless. Traffickcam: Crowdsourced and computer vision based approaches to fighting sex trafficking. In *2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–8. IEEE, 2017.

[16] A. Stylianou, H. Xuan, M. Shende, J. Brandt, R. Souvenir, and R. Pless. Hotels-50k: A global hotel recognition dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 726–733, 2019.

[17] C. Tzelepis, J. Oldfield, Y. Panagakis, M. Nicolaou, and I. Patras. Parts of speech grounded subspaces in vision language models. In *Proceedings of the thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

[18] R. Wolfe and A. Caliskan. American== white in multimodal language-and-image ai. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 800–812, 2022.