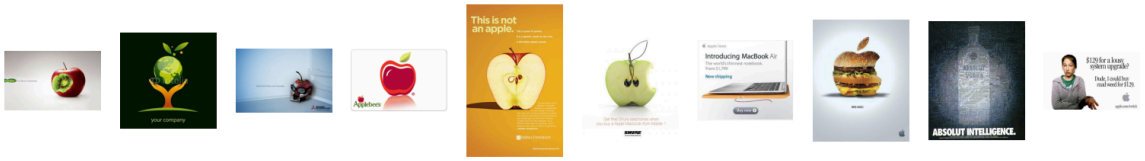
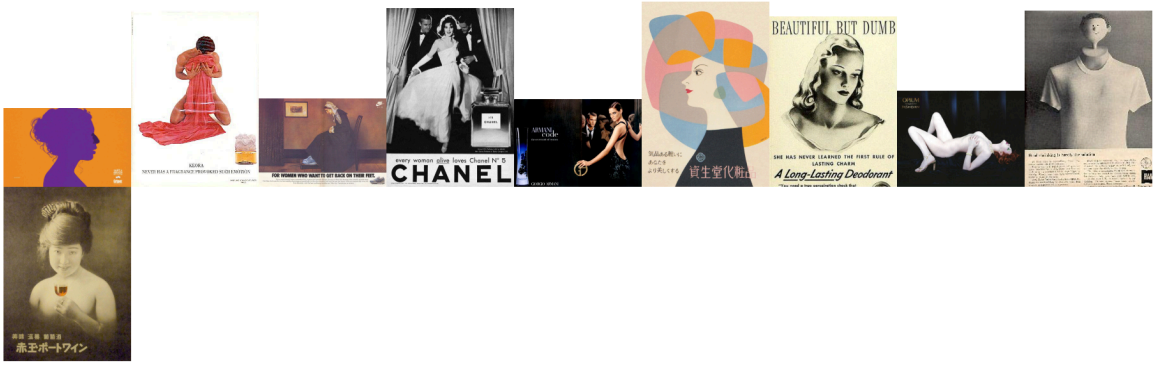
## Query Evaluation

apple



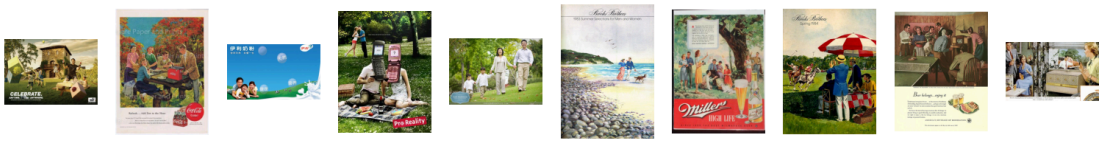

Woman



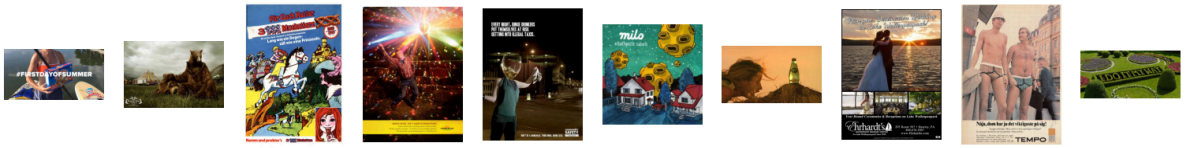

As it is apparent, searching for apple shows mostly images that contain apples, and searching for women shows mostly images that include women. However, it is also apparent that some of the images are not what is searched for. For instance, when searching for women, the image on the most right is not a woman. Furthermore, there is the issue of the inclusion of the word “woman” in an image and an image of a woman. The algorithm cannot separate these unfortunately as seen in the second image from left

Identifying queries that do not perform well in a text-to-image search system can help pinpoint areas for improvement. These examples often involve ambiguity, abstract concepts, or insufficient training data. Below are some examples of queries that might not yield accurate or relevant results, along with explanations for their shortcomings. Another point where the system falls short is when the query is more detailed, like “A family having a picnic under a cherry blossom tree in spring”. In this case it returns everything that may be related to the query which results in an image like the sixth from left which shows a family, but they are not having dinner. This shows that the system is not specific enough and the more

detailed the description gets the more apparent this issue gets.



like in this query” midsummer festival sweden”



The system gives all these images, but none of these are really what the query is. If there is no image related to what the query is, then it should give no output at all.

Suggest a method of quantitative evaluation of retrieval accuracy. (e.g. how to label dataset and prepare queries?)

For the quantitative evaluation of retrieval accuracy, it is first priority to have a thoroughly labeled dataset with which the accuracy can be compared. The dataset should be as large and as multifaceted as possible. Furthermore, the processing of the processing of the queries should be evaluated. As an example, if “family dinner in Sweden” gives the same results as “family dinner” accordingly adjustments have to be done on the processing of the queries. This could be for instance be done by saying on what the results are based on and probably a safety score that shows how safe the system is. As general statistical evaluation here precision, recall, f1, and mean average precision are important. The system has to be adjusted accordingly.