Does the performance of Text-to-Image retrieval models generalize beyond captions-as-a-query?

Juan Manuel Rodriguez[1,2,5][0000−0002−1130−8065], Nima Tavassoli[1], Eliezer Levy[3], Gil Lederman[3], Dima Sivov[3], Matteo Lissandrini[2,4][0000−0001−7922−5998], and Davide Mottin[1][0000−0001−8256−2258]

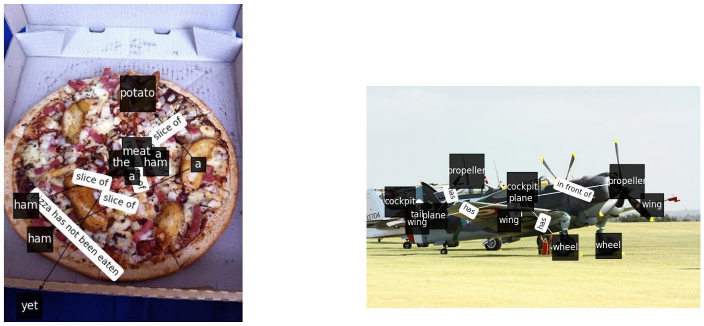
[1] Aarhus University nima.tavassoli2@gmail.com, davide@cs.au.dk
[2] Aalborg University jmro@cs.aau.dk
[3] Pnueli Lab, Tel Aviv Research Center Huawei Technologies
{eliezer.levy, gil.lederman, dima.sivov}@huawei.com
[4] University of Verona matteo.lissandrini@univr.it
[5] UNCPBA-CONICET

Abstract. Text-image retrieval (T2I) refers to the task of recovering all images relevant to a keyword query. Popular datasets for text-image retrieval, such as Flickr30k, VG, or MS-COCO, utilize annotated image captions, e.g., "a man playing with a kid", as a surrogate for queries. With such surrogate queries, current multi-modal machine learning models, such as CLIP or BLIP, perform remarkably well. The main reason is the descriptive nature of captions, which detail the content of an image. Yet, T2I queries go beyond the mere descriptions in image-caption pairs. Thus, these datasets are ill-suited to test methods on more abstract or *conceptual queries*, e.g., "family vacations". In such queries, the image content is implied rather than explicitly described. In this paper, we replicate the T2I results on descriptive queries and generalize them to conceptual queries. To this end, we perform new experiments on a novel T2I benchmark for the task of conceptual query answering, called ConQA. ConQA comprises 30 descriptive and 50 conceptual queries on 43k images with more than 100 manually annotated images per query. Our results on established measures show that both large pretrained models (e.g., CLIP, BLIP, and BLIP2) and small models (e.g., SGRAF and NAAF), perform up to 4× better on descriptive rather than conceptual queries. We also find that the models perform better on queries with more than 6 keywords as in MS-COCO captions.

1 Introduction

Text-to-Image retrieval (T2I) aims to retrieve images that answer a user keyword query [3]. Recent methods based mainly on deep learning models such as CLIP [25], BLIP [19], and BLIP2 [18] achieve state-of-the-art performance on such a task by retrieving the relevant image 68% of the time. These models jointly learn vector embeddings from image-caption pairs, so the embedding of the image should be close to that of the caption describing that image.

Since the most adopted datasets are Flickr30k [33], VG [16], and MS-COCO [20], with only image-caption pairs, all evaluations implicitly treat captions as surrogate queries, overlooking that *a user query might not describe so explicitly the image content*. Moreover, these datasets contain at best N(captions)-to-1(image) T2I mapping but miss any 1-to-N mapping. This is again problematic for real applications in which a query usually aims to retrieve several images.



(a) *A pizza shown in an open box.* (b) *A small propeller plane sitting on a field.*
Fig. 1: Examples of images, captions, and annotations in VG dataset.

For instance, existing T2I evaluations for Figure 1b assume that image to be the only relevant image for the caption. Further, since they are trained only on mere descriptions of the images, they are unable to understand *conceptual queries* like “symbols of wars from the past” and potentially missing that image.

This paper focuses on **replicating [1] results in T2I** [25,19,18,10,35] under the lens of descriptive vs. conceptual queries. To this end, **we introduce a new dataset** for Conceptual Query Answering, ConQA, comprising both descriptive and conceptual queries with a length of three or four words. This length is similar to the average queries in Web search engines [14,22], and in the Google Trends for Google Images[1]. ConQA[2] comprises 80 queries on $43k$ images and over 100 annotated images per query using Amazon Mechanical Turk; 30 queries are descriptive, similar but shorter than those in previous datasets, and 50 queries are conceptual, an underrepresented type of queries in existing datasets.

Our extensive results highlight that state-of-the-art T2I models perform well when the text is a long image description, while the performance declines when the text is short or conceptual. These results confirm and extend those in a previous study [11] that highlighted deficiencies of T2I models, such as CLIP, in retrieving images containing a single object (e.g., a bird).

Contributions. In summary, we contribute with (i) A thorough replicability study of state-of-the-art T2I models validating previously reported results; (ii) A new dataset, ConQA, that extends MS-COCO and VG annotations with 1-to-N query-image pairs, providing a new benchmark for the T2I for both conceptual

[1] Google Trends for Image search in 2023
[2] The code and ConQA: https://github.com/AU-DIS/ConQA

and descriptive queries; (iii) A reproducibility study on T2I models on ConQA short descriptive and conceptual queries. (iv) Important experimental findings showing the limitations of current T2I models those queries.

2 Related Work

Table 1: Main datasets for T2I and their characteristics. Annotation type: Caption (**C**), Object (**O**), Relationship (**R**), Segment (**S**), Attribute (**A**).

Dataset	**Images**	**Relevant per Query**	**Conc. Queries**	**Annotation**	**Task**
Flickr30k	31K	1	✘	C	Captioning
VG	108K	1	✘	O, A, R	Relationship Detection
MS-COCO	328K	1	✘	O, C, S	Obj. Det., Segmentation
GCC	3 300K	1	✘	C	Captioning, Retrieval
ConQA	43K	24.2	✔	O, C, A, R	Retrieval

Image retrieval aims to find images that match a query [3], whereby the query type reflects the task to solve. *Image-to-image* retrieval [32] returns images visually or semantically similar to a query image. *T2I* [13,31,34] returns images that match text queries or descriptions. *Hybrid* [36] retrieval finds images based on a combination of text, images, and additional annotations (e.g., semantic annotations). Here, we study the T2I task and describe current benchmarks' limitations to evaluate the T2I task.

Common benchmarks in T2I. Two of the most popular datasets for evaluating image retrieval tasks are Flickr30k [33] and MS-COCO [20]. MS-COCO [20] consists of more than 328K images classified into 11 super-categories, e.g., "vehicle," divided into 91 categories e.g., "bicycle". MS-COCO is intended for object detection, segmentation, and captioning, as such images are paired with captions and annotated objects and segments. Similarly, Flickr30k [33] provides about 31K images crawled from Flickr; each image is associated with five captions. Flickr30k aims to analyze the semantic relationships and similarities between different captions. As such, MS-COCO and Flickr30k feature only one known correct image for each piece of text (caption or description). A more recent dataset for image captions is the Google Conceptual Caption (GCC) [29] that includes conceptual relations through synonyms and hypernyms. GCC consists of 3.3M images from the web, each with multiple captions, with 10.3 words on average. Yet, the image captions are quite detailed and descriptive. VG [16] is a dataset associating semantic annotations to objects in the image with a collection of 108*k* images from MS-COCO and YFCC100M [15] annotated with *scene graphs*. A *scene graph* is a graph with objects in the image as nodes and the relationships among them as edges, describing the image contents from the semantic point of view [21]. Images are manually annotated by more than 33 000 workers using crowdsourcing. VG images are still associated only with descriptive captions. Therefore, *none of the existing datasets are intended (nor appropriate) as a benchmark for the T2I task, since for a given textual clue we only have one image annotated as relevant (see Table 1).* Furthermore, these datasets implicitly encourage descriptive texts.

State-of-the-art models in T2I. The current best-performing methods for T2I are *vision-language models*, such as CLIP [25], BLIP [19], or BLIP2 [18]. Such models are large models with up to billions of parameters that require a large amount of training data and computational resources, e.g., some versions of CLIP require 500 V100 GPUs and 18 days of training. CLIP [25] employs a contrastive loss that promotes high similarity only for true image-caption pairs. BLIP [19] and BLIP-2 [18] are flexible models trained to perform several tasks. As CLIP, they are trained for the Image-Text Constastive (ITC) task, which allows fast image and text search by kNN search. Both models can also perform a fusion image-caption to estimate the probability that both are related, called Image-Text Matching (ITM). ITM is much slower than ITC, but it is more precise due to an attention mechanism that aligns the image and the text.

Besides large vision-language models, small *specialized models* [17,10,35] explicitly target T2I, often employing cross-attention mechanisms [7] to align words and image regions. SGRAF [10] leverages graph convolutional networks to model relationships between words and regions. NAAF [35] considers both matched and mismatched word-region pairs. These lightweight models train on consumer hardware. Yet, since such models are trained on MS-COCO or Flickr30k, it is unclear whether they generalize to more complex queries.

Diffusion models. Diffusion models [27] are an orthogonal line of work aiming to solve a different task, namely generating images given textual descriptions. T2I retrieval can improve the results of such generative models [5] by including T2I models for similarity search as an initial filtering for image generation.

3 ConQA creation

Despite recent advances in multi-modal representation learning, the replicability of the results and their generalizability when these methods are employed explicitly for the task of T2I have not been studied before. To provide a more robust test dataset for the T2I task, in this work, we design and collect new annotations for the popular VG [16] dataset to benchmark T2I methods. Then, we identify two query sets: descriptive and conceptual.

Image selection. ConQA is a curated subset of VG 1.4. We select VG as it enriches MS-COCO with scene-graphs, which we use to select images with rich content. We filter images according to the following annotation quality criteria:

1. **Description**: The image should have one or more captions. Hence, we discarded the YFCC100M images with no caption, obtaining 91 524 images from the MS-COCO subset of VG.

2. **Significance**: The image should be a non-trivial selection of scenes with at least two objects and one interaction. Notice that a picture of a single object, such as a car, can pass this filter if the graph annotates the parts of the objects and their relations, e.g., "license plates-on→bumper". We kept only images with *scene graphs* containing at least one edge, thus retaining 67 609 images.
3. **Coherence**: some VG edges contain noisy, erroneous, or meaningless annotations. In our case, we enforce that all relationships should be verbs and not contain nouns or pronouns. To detect this, we generated sentences for each edge as a concatenation of the words in the node labels and the relationship and applied Part of Speech tagging[3]. We remove all images with scene graphs that contain an edge not tagged as a verb or where the tag is not in an ad-hoc list of allowed non-verb keywords[4]. After filtering, we obtain 43 413 images.

Figure 1a depicts a discarded image as the *scene graph* has the edge "pizza has not been eaten yet", which contains the noun *pizza*. Figure 1b shows an image with an acceptable *scene graph*.

Query generation. To ensure high-quality query-image relevance annotations, we devise a two-step approach.

The first step consists of generating queries for the dataset. Since there are no publicly available T2I query logs, the researchers created them. Six researchers, divided into three pairs, served as annotators to create both *conceptual* and *descriptive* queries. A descriptive query mentions some objects or actions in the image, such as "people chopping vegetables". While, a conceptual query does not mention specific objects or actions, but refers to a generic abstraction, e.g., "working class life". As a result, images with different objects can fit the query, e.g., a picture of an office or a factory ground can both be considered relevant. Notice that unlike GCC [29] that creates its "conceptual captions" by removing details of the original caption, e.g., removing proper names and object counts, or replacing a word by its hypernym, our conceptual queries are created independently of the images. As a result, while the GCC conceptual captions still describe the scene objects and relationships with fewer details than the original captions, *our conceptual queries do not prescribe a particular type of object or action in the scene.* The annotators also ensured that the queries are as realistic as possible. A post-hoc analysis confirms that the average number of words per query (3.4 words) is closer to the number of words reported for real query logs from Web search engines (2.56 words) [14] than the length of the average MS-COCO caption (11.26 words). The annotators generated 30 descriptive queries and 50 conceptual queries. After generating the queries, the annotators were tasked to search the dataset for images relevant to their query. To this end, the annotators used a prototype search engine using the "ViT-B/32" CLIP model [25]. The annotators could use ad-interim proxy or reformulated queries

[3] `en_core_web_sm` model in SpaCy [12]

[4] Allowed keywords: {top, front, end, side, edge, middle, rear, part, bottom, under, next, left, right, center, background, back, parallel}

to find such relevant images. Anecdotally, we report that the annotators had to reformulate their queries several times before finding at least one relevant image. The annotators kept the queries for which they could define a non-empty initial relevant set.

Data augmentation. We augment the initial set of images by adding the top-100 closest images in VG according to a pretrained ResNet152 model [9], meaning that the new images share visual features with the images in the original set. The purpose of this step is twofold, (1) evaluating a large amount of images, and (2) discriminating human- from machine-retrieved images.

Human annotation. Finally, we set up a set of Human Intelligence Tasks (HITs) on Mechanical Turk (MTurk) [2], where each MTurk task consists of a query and 5 candidate images sampled from the augmented set of images per query. The workers are instructed to mark each image "Relevant" or "Irrelevant" for the given query[5]. They are also allowed to report "Unsure" when undecided. To reduce presentation bias, we randomize the order of images in each query. We also include validation tasks with control images to ensure a minimum quality in the annotation process, so workers failing 70% or more of validation tasks are excluded. Hence, we purposely present the worker images that the initial annotators manually tagged as completely irrelevant and only for descriptive queries. As a result, we employ 2 190 workers to tag 6.9k images with at least 3 workers annotating the same query-image pair. Each worker answered on average 2.5 HITs, 77% of workers performed at most 3 and only one performed 16 HITs.

Annotation cost. We tag 100 images per query for a total of 80 queries. We show 5 images per HIT, and decided to obtain three judgments per query-image pair. Every time we show a descriptive query, one of the 5 images is replaced with a control image for that query. Therefore, to tag all the images for conceptual query we need 60 HITs and 75 for descriptive query. Since we pay USD 0.2 per HIT and Amazon charges us 10% of that, in total, the entire dataset tagging costs USD 1 155. Yet, since we run several preliminary versions of the HITs to validate the interface and the process before running the final task, the cost is about USD 1 500. Notice that this cost is only for Amazon Mechanical Turk, and it does not consider the time expended by the researchers defining the queries, analyzing the results of the tagging, and re-running HITs when needed.

4 ConQA Annotations and Content

After filtering the most inaccurate users, we obtain 6 941 images annotated as relevant/irrelevant for at least one query, and on average 100 images per query tagged as either relevant, non-relevant, or unsure by at least three workers. The queries have between 1 and 7 words with a median length of 3 words, and each has 100 annotated images as relevant/irrelevant by the workers. Finally, Figure 2 shows that the images span several MS-COCO categories uniformly.

[5] An example is shown at `https://benevolent-sawine-669286.netlify.app/`

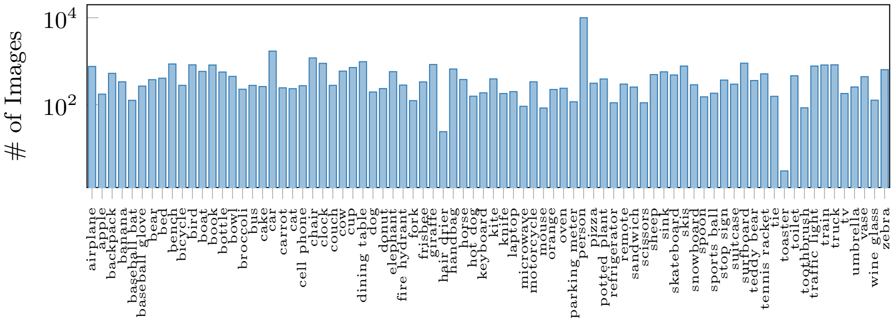


Fig. 2: Number of images per category in ConQA.

Table 2: Query depth in Wordnet hyponymy/hypernymy

	Depth		
Queries (#)	**Min**	**Max**	**Average**
ConQA Conceptual (50)	2.96±1.94	6.73±1.79	4.88±1.57
ConQA Descriptive (30)	3.56±2.30	8.03±1.43	5.99±1.35
MS-COCO captions (1.5M)	1.53±0.88	9.72±1.52	5.75±0.93

Conceptual vs descriptive queries. We analyze the different abstraction levels of words in conceptual and descriptive queries. Hence, we measure the depth of the query words in the WordNet hyponym structure [23], which is a hierarchy where depth is related to concreteness as it goes from the most generic term [6] to the most concrete. We expect that the words in the conceptual queries have fewer nodes between them and a root concept, i.e. the most generic. Since queries have more than one word, Table 2 presents the the average/minimum/maximum word depth averaged per query. The results show that conceptual query words are closer to the root node than descriptive query words. MS-COCO captions are similar to descriptive queries on average while the minimum and maximum values are, respectively, lower and higher than ConQA queries.

Validation task. We study the annotation quality by analyzing the performance of the validation task. Figure 3 shows the correct/incorrect (unsure means the worker selected unsure as the answer) number of validation HITs with respect to the number of times a worker has been evaluated. Additionally, about one-third of the workers are not assigned any task with validation, but these are mostly workers who completed only 1 or 2 hits. Thus, it is reasonable to conclude that the vast majority of labels obtained are of high quality. To assess the inter-rater agreement between our evaluation and the workers' output, we computed Cohen's Kappa, which yields 0.676, suggesting a strong agreement.

Image relevance. We further evaluate the reliability of obtained labels by measuring the inter-judgment agreement between workers. We classify each image into 5 categories in our analysis: "Fully Relevant" meaning all workers selected relevant, "Fully Non-relevant" meaning all workers selected non-relevant,

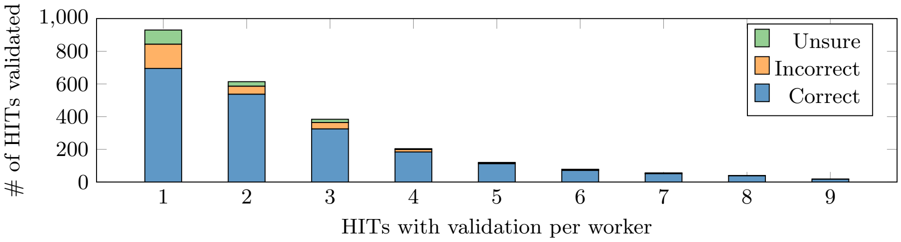


Fig. 3: Number of HITs with validation, only few control tasks failed.

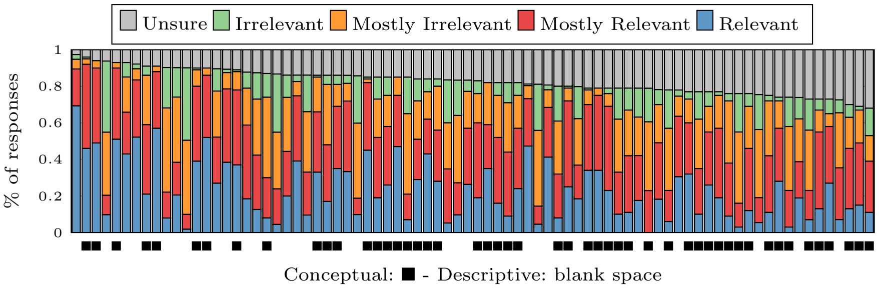


Fig. 4: Distribution of workers' responses per query.

"Majority Relevant" meaning most workers selected relevant, "Majority Non-relevant" meaning most workers selected non-relevant, and "Unsure" otherwise. Figure 4 illustrates the distribution of different mentioned categories. The frequency with which workers selected the "Unsure" category for images in conceptual and descriptive queries has a median of 20.0% and 13.97%, respectively. Moreover, the maximum frequency of "Unsure" is 32% for a conceptual query and 27.43% for a query in the descriptive category, showing that even humans tend to have difficulties when assigning relevance to some conceptual queries.

5 Replicating and generalizing T2I experiments

We present the results of our experiments in two parts. First, we *reproduce* the T2I over MS-COCO-5K experiments as reported in the papers CLIP [25], BLIP [19], BLIP2 [18], SGRAF [10], and NAAF [35]. The second experiment aims at replicating the results in a zero-shot regime on the portion of the MS-COCO dataset for which we have annotations in ConQA. We additionally evaluate the models on ConQA conceptual and descriptive queries. We define reproducibility and replicability as they are defined by ACM in [1].

Models and configurations. We evaluate three pretrained large vision-language models (LVMs) CLIP [25], BLIP [19], and BLIP2 [18], and two specialized T2I models, NAAF [35] and SGRAF [10]. In all the cases, we use the original implementations and pretrained weights. CLIP is trained on the WebImageText dataset (∼400M images), while BLIP and BLIP2 are trained using a combination of datasets, including MS-COCO [20], VG [16], GCC [29], Conceptual [8], and SBU captions [24], resulting in more than 14M images. Further, LAION [28] is used as a source of noisy captions adding 100M images. For CLIP, we report the results for the ViT-L_14@336px model that attains the best results in the original paper [25]. For BLIP [19] and BLIP2 [18] we both use a version fine-tuned on MS-COCO for the reproducibility study (Section 5.1) and the pretrained version for the replicability experiment (Section 5.2). For BLIP and BLIP2, we use the ITC modality for obtaining an initial set of 128 images and then use ITM modality to re-rank them, as described in [19,18]. For SGRAF we use the pretrained model on MS-COCO for the reproducibility experiment (Section 5.1) and the model trained on Flickr30k for replicability (Section 5.2). For NAAF the only available pretrained model is on Flickr30k; for this reason we do not report results on reproducibility. The subscript FT next to each model's name indicates the version fine-tuned on MS-COCO.

Datasets. In the reproducibility analysis, we use MS-COCO-5K which comprises 5k images with 5 captions each used to test the model in the original papers [25,19,18,10,35]. For the replicability experiments, we report results on five datasets: Conc, Desc, MS-COCO-VG, GPT-Conc, and GPT-Desc. All datasets consist of the $6.3K$ images tagged by the MTurkers excluding the seed image annotated by the researchers. Conc and Desc are queries in ConQA in which an image is relevant for the query only if three MTurkers deem it relevant. MS-COCO-VG queries are the 5 MS-COCO captions for each image in the ConQA dataset. Note that while LVMs are trained with images from various sources, to the best of our knowledge they have not been exposed to the MS-COCO-VG queries tested in this work as we use the original hold-out set to test the models. The results in our reproducibility and replicability analyses seem to confirm this assumption. Finally, GPT-Conc and GPT-Desc are respectively the top-10 rephrasings of the Conc and Desc queries obtained using the GPT-J6B model [30] with prompt *"QUERY" can be rephrased as* to task the model to rephrase the query.

Measures. We report NDCG, R-precision, and Recall at 1, 5, and 10. NDCG evaluates the entire ranking by assigning a higher score to relevant images in the top positions. R-precision is the percentage of relevant images in the first R positions, where R is the number of relevant images for the query. For consistency with the T2I literature [17,35,10,25,19,18], here, we define Recall@K as the percentage of queries that return at least one relevant image within the top-K ranked images. We note that this definition assumes explicitly that each query (caption) can retrieve at most 1 relevant image. Our evaluation code adopts the ranx library [4], which follows TREC definitions, hence what we report as Recall@K here, it is called hit-rate@K in ranx.

5.1 Reproducing T2I results

We first compare the experiments in the original papers and in the previous study [11]and highlight some differences among them. In the case of CLIP, the original paper presents a zero-shot experiment; while in the other cases the models (including BLIP and BLIP2) are fine-tuned on MS-COCO. Table 3 presents

Table 3: Results for T2I in MS-COCO-5K: our results vs. originally reported

	Replication			**Originally reported**		
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
BLIP$_{\text{FT}}$ [19]	65.1	86.3	91.8	65.1	86.3	91.8
BLIP2$_{\text{FT}}$ [18]	67.9	87.5	92.4	68.3	87.7	92.6
CLIP [25]	37.1	61.6	71.5	37.8	62.4	72.2
ECIR'23 Rep. CLIP [11]	21.9	40.2	49.8			
CLIP ViT-L/14	36.5	61.0	71.1			
SGRAF$_{\text{FT}}$ [10]	40.5	69.6	80.3	40.2	-	79.8

Table 4: Replicating results from ECIR'23 Rep. CLIP [11]

	Replication			**ECIR'23 Rep.**		
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
CLIP ViT-L/14 [25]	21.9	40.7	49.9	21.9	40.2	49.8
ST-CLIP ViT-L/14 [26]	22.1	40.7	49.9			

our reproduction of the T2I experiments as presented in the original works. The results on CLIP, BLIP$_{\text{FT}}$ and BLIP2$_{\text{FT}}$ are mostly consistent to those in the original papers [25,19,18]. Minor differences can be ascribed to technical differences in the platform as stated in the PyTorch documentation[6]. The most apparent difference is between our results and that of a previous evaluation [11] for the CLIP model that is depicted in the row "ECIR'23 Rep. CLIP" of Table 3. That work [11] reports on average 20% less Recall that originally reported for CLIP [25]. An inspection revealed that the ECIR'23 evaluation was conducted using the ViT-L_14 in opposition to the ViT-L_14@336px used in the original CLIP paper and in our evaluation. Moreover, Sentence Transformer library [26] implementation was used in the ECIR'23 [11], while our experiments use the CLIP original implementation [25]. Therefore, we performed the experiment with ViT-L_14 using both implementations. Our results show that the model and implementation differences do not account for the reported difference. Further analysis of the "ECIR'23 Rep. CLIP" code, shows that the authors used an ad-hoc list of 20 252 images/captions from MS-COCO with only one caption per image instead of five captions per images, despite stating that they used the standard MS-COCO-5K [11]. Table 4 shows that we could replicate the results reported in [11] using the original CLIP implementation and the Sentence Transformer, called ST-CLIP in the table. Notice that, the MS-COCO based dataset used in [11] (mistakenly labeled there as MS-COCO-5K), has 4x more images than MS-COCO-5K. Hence, the task is more challenging than in MS-COCO-5K as the task consist in ranking 20 252 images instead of only 5 000, which explains the difference in the results.

[6] `https://pytorch.org/docs/stable/notes/randomness.html`

Table 5: Results for T2I Retrieval on Descriptive vs. Conceptual queries. The results with MS-COCO differ from those reported in previous work as we use 25% more testing examples and we do not fine-tune the models.

Method	Queries	NDCG	R-precision	Recall@1	Recall@5	Recall@10
BLIP2 [18]	MS-COCO-VG	**70.8**	**51.6**	**51.6**	**76.1**	**84.5**
	Desc	48.0	15.3	20.7	51.7	62.1
	GPT-Desc	40.2	9.5	12.8	32.4	43.4
	Conc	36.4	5.4	8.2	28.6	36.7
	GPT-Conc	32.1	3.4	3.9	16.1	25.3
BLIP [19]	MS-COCO-VG	**66.9**	**46.3**	**46.3**	**71.8**	**80.9**
	Desc	46.2	15.3	20.7	58.3	62.1
	GPT-Desc	39.3	10.4	13.4	32.8	48.6
	Conc	35.0	5.4	4.1	28.6	40.8
	GPT-Conc	31.6	3.8	3.1	14.5	22.0
CLIP [25]	MS-COCO-VG	**50.5**	**28.0**	**28.0**	49.8	60.3
	Desc	47.8	16.5	20.7	**58.6**	**65.5**
	GPT-Desc	40.2	10.1	12.1	34.5	45.2
	Conc	37.9	6.8	12.2	30.6	36.7
	GPT-Conc	32.4	3.7	5.3	15.5	24.3
NAAF [35]	MS-COCO-VG	41.0	**17.7**	**17.7**	**37.4**	**48.0**
	Desc	**41.5**	10.6	13.8	34.5	44.8
	GPT-Desc	36.7	7.3	9.0	26.6	33.8
	Conc	30.7	2.4	4.1	12.2	16.3
	GPT-Conc	29.2	1.9	3.7	8.2	12.2
SGRAF [10]	MS-COCO-VG	**39.6**	**16.5**	**16.5**	**35.6**	**46.1**
	Desc	36.3	7.9	6.9	24.1	34.5
	GPT-Desc	33.9	5.8	5.9	18.3	25.2
	Conc	28.4	1.3	0.0	8.2	10.2
	GPT-Conc	27.8	1.3	0.8	6.1	9.6

Finally, we obtain results with SGRAF_{FT} slightly better than those reported in the original paper; our results refer to an improved model released by the authors in their code repository. As mentioned above, it is not possible to evaluate the NAAF model under the same conditions described in the paper [35].

5.2 Replicating and generalizing T2I results

We study the performance of T2I models under different query regimes. These results extend those of previous experimental evaluations [25,19,18,10], and pave the way to novel research in this field. Table 5 shows that the results vary significantly across query types.

Finding 1: The LVMs perform well in zero-shot MS-COCO-VG queries as opposed to specialized models. In zero-shot, the LVMs perform better than the small fine tuned NAAF and SGRAF. Yet, we see that the performance of the LVMs decreases in the ConQA queries compared to their original performance on MS-COCO-5K. This can be seen in BLIP, BLIP2 and CLIP Recall@1, 5, 10 in Table 3 with Table 5, e.g., BLIP2 Recall@1 on MS-COCO-5K is 67.9% and on MS-COCO-VG is 51.6%. Hence, while we conclude BLIP2 to be the best off-the-shelf method, and with superior zero-shot capabilities than small fine-tuned models, we still highlight difficulties in generalizing to other query sets.

Finding 2: The models are challenged by shorter queries. The models perform best on MS-COCO-VG queries that are similar to those employed in the model training. Yet, we observe a consistent performance drop on ConQA's queries, even those descriptive. Recall that ConQA has on average 24 relevant images per query and, despite such an easier case, most models do not identify relevant images in the top-5. This confirms our hypothesis that current models are less suited for shorter queries commonly found in web-search.

Finding 3: Current models struggle on conceptual queries. On conceptual queries that describe the content of the image in an abstract manner only, all models experience a significant setback up to 30% on all metrics compared to MS-COCO-VG queries. For example, Table 5 shows that BLIP2 Recall@1 on MS-COCO-VG is 51.6% and on conceptual queries is 8.2%

Table 6: Relative percentage improvement with p-value< 0.05 among pairs of queries and difference measures; values in bold: p-value< 0.01; X: not statistically significant; inf: the measure is 0 for the set of queries.

Method	Metrics	CONC DESC	CONC MS-COCO-VG	DESC MS-COCO-VG	GPT-CONC GPT-DESC	GPT-CONC MS-COCO-VG	GPT-DESC MS-COCO-VG
BLIP2 [18]	NDCG	**31.9**	**94.2**	**47.3**	**25.1**	**120.4**	**76.2**
	R-precision	**186.1**	**862.2**	236.3	**177.1**	**1412.1**	**445.7**
	Recall@1	X	**531.8**	**149.3**	**229.0**	**1230.1**	**304.2**
	Recall@5	81.0	**166.3**	**47.1**	**101.0**	**371.9**	**134.7**
	Recall@10	69	**130.0**	**36.1**	**71.7**	**233.9**	**94.5**
BLIP [19]	NDCG	**32.1**	**91.1**	**44.7**	**24.5**	**111.6**	**70.0**
	R-precision	181.5	**752.0**	X	**174.0**	**1117.7**	**344.3**
	Recall@1	406.9	**1034.3**	**123.8**	**339.3**	**1412.4**	**244.3**
	Recall@5	69	**151.2**	**48.7**	**126.1**	**395.4**	**119.1**
	Recall@10	52.1	98.3	**30.4**	**120.6**	**267.2**	**66.5**
CLIP [25]	NDCG	**26.2**	X	X	**23.9**	**55.9**	25.7
	R-precision	**142.9**	X	X	**173.5**	X	X
	Recall@1	X	**128.3**	X	**127.5**	**426.9**	**131.7**
	Recall@5	**91.5**	**62.5**	X	**122.3**	**220.8**	**44.3**
	Recall@10	**78.4**	**64.3**	X	**86.0**	**148.5**	**33.6**
NAAF [35]	NDCG	**35.3**	X	X	**25.6**	X	X
	R-precision	**333.4**	X	X	**295.1**	X	X
	Recall@1	X	**333.5**	X	**144.1**	**381.7**	**97.4**
	Recall@5	**181.6**	**205.1**	X	**225.3**	**357.6**	**40.7**
	Recall@10	**174.6**	**193.8**	X	**176.0**	**291.8**	**42.0**
SGRAFP [10]	NDCG	27.7	X	X	**22.1**	X	X
	R-precision	**504.7**	X	X	**335.8**	**1146.4**	X
	Recall@1	inf	inf	X	**618.1**	**1917.3**	**180.9**
	Recall@5	195.7	**335.7**	X	**198.5**	**480.9**	**94.6**
	Recall@10	**237.9**	**352.0**	X	**162.4**	**380.8**	**83.2**

Finding 4: Most of the results are statistically significant and exhibit a noticeable difference from MS-COCO-VG. We further report an experiment on the differences observed in Table 5. We perform a Mann-Whitney U test , a form of non parametric test to assess statistical significance of the alternative hypothesis that the model when retrieving queries in X got a lower score than those in Y. In our case, we compute differences for the following (X, Y) dataset pairs (Conc, MS-COCO-VG), (Desc, MS-COCO-VG), (GPT-Conc, GPT-Desc), (GPT-Conc, MS-COCO-VG), (GPT-Desc, MS-COCO-VG). We repeat the test for all measures and report the relative percentage improvement calculated as $(m_y - m_x)/m_x$ for a pair of measures (m_x, m_y). We report only the relative improvements for which the p-value< 0.05. The results in Table 5 confirm our findings. In particular, all models exhibit statistically significant differences in at least two metrics when comparing Conc with Desc and MS-COCO-VG queries. The Desc queries are also significantly different from those in MS-COCO-VG in the two largest models, BLIP and BLIP2. We evince that LVMs have a bias towards longer queries or image descriptions. Finally, we experience significant differences among GPT-paraphrased queries and MS-COCO-VG queries, showing that the difference is due to the content rather than the language.

6 Conclusions

We successfully reproduced and replicated results in the T2I task using LVMs, such as CLIP, BLIP and BLIP2, and specialized models trained on a specific dataset, such as SGRAF and NAAF. In particular, we obtained results within 1% of those reported for CLIP, BLIP, BLIP2, and SGRAF while, we could not reproduce NAAF due to the lack of pretrained model on MS-COCO. Moreover, we point out a limitation of a previous CLIP reproducibility study [11]. To perform a more systematic evaluation, we introduced ConQA, a dataset built on top of VG and MS-COCO datasets, that enriches the T2I task by adding conceptual queries that express the content of images in abstract terms. We found that small models, such as SGRAF and NAAF, are not able to generalize, even when training on Flickr30k and testing on MS-COCO. In contrast, BLIP and BLIP2 outperform CLIP even without fine-tuning. Furthermore, BLIP and BLIP2 perform 1.5× better on MS-COCO-VG queries than ConQA descriptive queries. Hence, these models perform better on larger and more descriptive queries. Overall, we found out that all the models tend to perform better on descriptive queries rather than conceptual ones. Particularly, we found that some measures are 5× lower for conceptual than for ConQA's descriptive queries, and up to 10× than MS-COCO-VG queries. This shows a limitation of state-of-the-art models to generalize on conceptual queries. Our study provides evidence of unknown issues of current T2I approaches and paves the way to novel models that incorporate higher level abstractions to properly answer conceptual queries.

References

1. ACM (2020) artifact review and badging - current. https://www.acm.org/publications/policies/artifact-review-and-badging-current (2020)
2. Aguinis, H., Villamor, I., Ramani, R.S.: Mturk research: Review and recommendations. Journal of Management **47**(4), 823–837 (2021)
3. Alemu, Y., Koh, J.b., Ikram, M., Kim, D.K.: Image retrieval in multimedia databases: A survey. In: 2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing. pp. 681–689 (2009)
4. Bassani, E.: ranx: A blazing-fast python library for ranking evaluation and comparison. In: ECIR (2). Lecture Notes in Computer Science, vol. 13186, pp. 259–264 (2022)
5. Blattmann, A., Rombach, R., Oktay, K., Müller, J., Ommer, B.: Retrieval-augmented diffusion models. In: Advances in Neural Information Processing Systems. vol. 35, pp. 15309–15324 (2022)
6. Bouras, C., Tsogkas, V.: W-kmeans: Clustering news articles using wordnet. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) Knowledge-Based and Intelligent Information and Engineering Systems. pp. 379–388 (2010)
7. Cao, M., Li, S., Li, J., Nie, L., Zhang, M.: Image-text retrieval: A survey on recent research and development. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22. pp. 5410–5417 (7 2022), survey Track
8. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3558–3568 (June 2021)
9. Chen, P., Liu, S., Jia, J.: Jigsaw clustering for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11526–11535 (June 2021)
10. Diao, H., Zhang, Y., Ma, L., Lu, H.: Similarity reasoning and filtration for image-text matching. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 1218–1226 (2021)
11. Hendriksen, M., Vakulenko, S., Kuiper, E., de Rijke, M.: Scene-centric vs. object-centric image-text cross-modal retrieval: A reproducibility study. In: Advances in Information Retrieval. pp. 68–85 (2023)
12. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python (Jan 2022)
13. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Image retrieval using scene graphs. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3668–3678 (2015)
14. Kacprzak, E., Koesten, L.M., Ibáñez, L.D., Simperl, E., Tennison, J.: A query log analysis of dataset search. In: Web Engineering. pp. 429–436 (2017)
15. Kalkowski, S., Schulze, C., Dengel, A., Borth, D.: Real-time analysis and visualization of the yfcc100m dataset. In: Workshop on Community-Organized Multimodal Mining: Opportunities for Novel Solutions. p. 25–30. New York, NY, USA (2015)
16. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision pp. 32 – 73 (2017)

17. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
18. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models (2023)
19. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 12888–12900 (2022)
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 740–755 (2014)
21. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: Computer Vision – ECCV 2016. pp. 852–869. Springer International Publishing (2016)
22. Luo, C., Headden, W., Avudaiappan, N., Jiang, H., Cao, T., Yin, Q., Gao, Y., Li, Z., Goutam, R., Zhang, H., Yin, B.: Query attribute recommendation at amazon search. In: Proceedings of the 16th ACM Conference on Recommender Systems. p. 506–508. RecSys '22 (2022). `https://doi.org/10.1145/3523227.3547395`
23. Miller, G.A.: Wordnet: A lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
24. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. In: Advances in Neural Information Processing Systems. vol. 24. Curran Associates, Inc. (2011)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. PMLR, vol. 139, pp. 8748–8763 (2021)
26. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (11 2019)
27. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
28. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs (2021)
29. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. pp. 2556–2565 (2018)
30. Wang, B.: Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. `https://github.com/kingoflolz/mesh-transformer-jax` (May 2021)
31. Wang, S., Wang, R., Yao, Z., Shan, S., Chen, X.: Cross-modal scene graph matching for relationship-aware image-text retrieval. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2020)

32. Yoon, S., Kang, W.Y., Jeon, S., Lee, S., Han, C., Park, J., Kim, E.S.: Image-to-image retrieval by learning similarity between scene graphs. AAAI **35**(12), 10718–10726 (2021)
33. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)
34. Yu, T., Fei, H., Li, P.: U-bert for fast and scalable text-image retrieval. In: Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval. p. 193–203 (2022)
35. Zhang, K., Mao, Z., Wang, Q., Zhang, Y.: Negative-aware attention framework for image-text matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15661–15670 (2022)
36. Zhao, Y., Song, Y., Jin, Q.: Progressive learning for image retrieval with hybrid-modality queries. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1012–1021 (2022)