

Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models

Supplementary Material

Zheyuan Liu¹ Cristian Rodriguez-Opazo² Damien Teney^{2,3} Stephen Gould¹

¹Australian National University

²Australian Institute for Machine Learning, University of Adelaide ³Iidiap Research Institute

{zheyuan.liu, stephen.gould}@anu.edu.au

cristian.rodriguezopazo@adelaide.edu.au, damien.teney@idiap.ch

A. Implementation Details

Existing methods on CIRR. As discussed in Sec. 5, we adopt the default configurations for state-of-the-art (SoTA) methods when testing on our proposed dataset CIRR.

Specifically, for TIRG [9] and its corresponding baselines (incl. Random, Image/text-only, Random Image+Text and Concatenation), we use ResNet18 pretrained on ImageNet [7] as the image encoder, and a randomly initialized LSTM as the text encoder. We note that the above methods do not benefit from more complex ResNet features (*e.g.*, ResNet152) or word embedding initializations on CIRR. We train the models using soft-triplet based loss [9], as we discover that the batch-based classification loss introduces serious overfitting for TIRG on CIRR. For MAAF, following Dodds et al. [1], we use a pretrained ResNet50 along with an LSTM, while training the model with batch-based classification loss.

Note that the implementations of MAAF and TIRG share the same codebase. Hence, all methods above use a hidden size of 512, and models are optimized with vanilla stochastic gradient descent (SGD) as in Vo *et al.* [9].

CIRPLANT on Fashion-IQ. We do not perform hyper-parameter tuning for our model on Fashion-IQ (*i.e.*, the setup is kept the same as on CIRR, see Sec. 5 for details). Additionally, since the three subtypes in Fashion-IQ distinct greatly from each other, we sample each minibatch from a single subtype during training, as in Dodds et al. [1].

B. Additional Metrics

See Table S1 on performance in mAP@ K , where the comparisons are similar to Recall (Table 3, Sec 5.1). Note that since our task has only one true-positive for each query, Precision@ K and Recall@ K are the same (hence P@ K not shown).

Methods	mAP@ K				mAP _{Subset} @ K		
	$K = 1$	$K = 5$	$K = 10$	$K = 50$	$K = 1$	$K = 2$	$K = 3$
8 TIRG [9]	14.61	25.05	27.25	28.63	22.67	33.25	39.92
10 MAAF [1]	10.31	15.77	17.72	19.43	21.05	29.84	36.80
14 Ours (no init.)	15.18	25.06	27.19	28.65	33.81	45.50	51.60
15 Ours (init.)	19.55	30.40	32.55	33.85	39.20	50.68	56.28

Table S1. mAP scores for SoTA methods (and ours) on CIRR. See Table 3 (corresp. row numbers) for comparison with Recall.

C. Auxiliary Annotations in CIRR

As discussed in Sec. 4.1, following the collection of the modification sentences (main annotation), we additionally collect auxiliary annotations for each image pair. The auxiliary annotations are meant to provide explicit training signals that address the ambiguities caused by implicit human-agreements. Although we do not use such annotations in this work, we believe that they can benefit future work for clarifying and interpreting such ambiguities.

Collection. For each pair of reference-target image, we collect the answers to the following four questions from Amazon Mechanical Turk (AMT) workers, which tangibly address the implicit ambiguities mentioned above:

- Q1** What characteristics of the objects are preserved across images?
- Q2** What objects were changed, but not relevant to the modifying sentence?
- Q3** Is there any change in camera angle/focus/viewpoint?
- Q4** Is there any change in the background/lighting across the images?

We provide the AMT workers with the reference-target image pair along with the collected modification sentence (main annotation). For each question, workers can choose to answer with a sentence or mark as not applicable (*e.g.*, nothing worth mentioning or already covered by the main

annotation). Statistics are shown in Table S2. Collection interface is shown in Fig. S8 (bottom), see examples in Fig. S7.

	Nb. image subsets	Nb. pairs	Nb. pairs per subset	Nb. images	Pairs with auxiliary (%)			
					Q1	Q2	Q3	Q4
Train	3,345	28,225	7.54	16,939	66.72	68.09	48.06	58.45
Val.	503	4,184	8.32	2,297	71.87	67.67	49.43	64.66
Test	503	4,148	8.25	2,316	69.62	69.01	46.44	63.00
Total	4,351	36,554	8.40	21,552	67.65	68.15	48.02	59.69

Table S2. Statistics of CIRR with auxiliary annotations. The visual contents and the (main) annotation determine whether a pair also has auxiliary annotations for Q1–4.

D. Collection Details on CIRR

We provide additional details about our data collection procedure (Sec. 4.1) including examples of each step (excl. the auxiliary annotations, which is discussed in Sec. C).

Image subsets. Fig. S1 shows the procedure for constructing an image subset of six elements, noted as $\mathcal{S} = \{I_1, \dots, I_6\}$ in Sec. 3.1. We specifically demonstrate cases where images are removed. The process was designed to ensure that images in a given subset are visually similar to one another while exhibiting some appreciable differences.

Image pairs. As explained in Sec. 3.1, we draw nine pairs from each subset. Fig. S2 demonstrates how we form consecutive modifications among pairs, which could facilitate the training and evaluation of dialogue systems in the future. Fig. S3 shows that one reference image leads to multiple targets in each subset. This should allow the study of the impact of language modality in the future.

We point out that the length of dialogue paths can vary for two reasons. First, we allow a slight overlap between the images of two subsets. Therefore, it is possible to form dialogue paths across subsets with variable lengths, as shown in Fig. S4. Second, AMT workers can mark pairs of poor quality and choose not to annotate them (see below). Such pairs will be removed from the dataset, thus rendering the dialogue incomplete. In total, 71.1% of the subsets have closed-loop dialogue paths (see Table S2 for detailed statistics).

Annotation collection on AMT. Table S4 demonstrates our guideline to AMT workers, specifying types of annotations to avoid.

Fig. S8 shows our collection interface. (top) For main annotations, we require AMT workers to write sentences that only lead to the true target image, thus removing false-negatives in each subset. We also allow them to mark image pairs of poor quality for removal. (middle) For auxiliary annotations, we ask four detailed questions to clarify ambiguities within the given pair. (bottom) We evaluate human retrieval performance in $\text{Recall}_{\text{Subset}}$ using the test-split.

Quality control. We conduct a pre-selection process to manually whitelist workers with good annotation quality. Our pre-selection procedure plays a critical role in quality assurance, where we filter out over 60% of the submitted workers. Workers who have passed the selection process produce annotations with over 90% acceptance rate.

For annotations submitted by workers in the whitelist, we manually review $\sim 30\%$ of the annotations from each worker. The remaining are examined with an automated script to check for potential abuse of the use of checkboxes, irresponsible contents (e.g., very short sentences), and annotations that violate our guidelines.

Human performance. Table 3 (row 7) lists the human retrieval performance of $\text{Recall}_{\text{Subset}}@1$ (see Sec. 5 for details of the $\text{Recall}_{\text{Subset}}$ metric) on test-split. Here, we present the collection procedures of this score.

Fig. S8 (bottom) shows the collection interface. Specifically, we ask AMT workers to choose the most probable target image for a given text-image query. We employ three different AMT workers for each pair in the test-split, our final score is calculated by averaging over all submitted results.

E. Additional Analysis on CIRR

Image synsets. We analyze the image contents using the synset information in NLVR² [8]. CIRR includes 124 out of the 1000 synsets in NLVR². Each synset is associated with 136.6 ± 73.1 ($\mu \pm \sigma$) images. The five most common synsets are bookcase, bookshop, dobreman, timber wolf and pug. The five least common synsets are acorn, skunk, orange, ox, broccoli and padlock. Distributions of samples are shown in Fig. S5.

Note that we do not distinguish synsets of similar concepts (e.g., dobreman and French bulldog) when forming image pairs, instead, we choose by visual similarity. Additionally, we point out that for composed image retrieval, synset may not fully characterize an image, as the annotations focus on fine-grained visual comparisons.

Comparison to existing datasets. Table S3 compares CIRR with existing datasets used for composed image retrieval. We demonstrate that CIRR is comparable in size with existing datasets. Additionally, it provides rich auxiliary annotations for open-domain images.

False-negative analysis. Fig. S6 demonstrates the presence of false-negatives in Fashion-IQ [3], as explained in Sec. 5. For comparison, our data collection procedures ensure that no false-negatives are present within each image subset, as discussed in Sec. D. Examples of CIRR are shown in Fig. S2, Fig. S3, and Fig. S7.

(a) Randomly pick an image as I_1 (leftmost), sort the remaining images in the large image corpus \mathcal{D} by their cosine similarity to I_1 using ResNet features pre-trained on ImageNet, noted as κ_i for I_i . Images are ranked from left to right.

I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	\dots
									\dots
$\kappa_i = 1.0$	0.9981	0.8691	0.8663	0.8603	0.8490	0.8488	0.8456	0.8435	\dots

(b) Remove near-identical images with $\kappa_i \geq 0.94$.

I_1	Removed	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	\dots
									\dots	
$\kappa_i = 1.0$	0.9981	0.8691	0.8663	0.8603	0.8490	0.8488	0.8456	0.8435	\dots	

(c) Select the next top-20 ranked images (not fully shown below).

I_1	\Leftarrow Shifted	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	\Rightarrow top-20
	\Leftarrow Shifted									\dots
$\kappa_i = 1.0$		0.8691	0.8663	0.8603	0.8490	0.8488	0.8456	0.8435	0.8421	\dots

(d) Greedily add each image as ranked. Meanwhile, to ensure sufficient variations between images, skip an image if its κ_i is within 0.002 of the last added image. We demonstrate the greedy process as below. In each step, curved arrow suggests a comparison of κ_i and κ_{i+1} , added image is marked with a tick while skipped is crossed out.

I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	\dots
									\dots
$\kappa_i = 1.0$	0.8691								
$1.0 - 0.8691 > 0.002$									
I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	\dots
									\dots
$\kappa_i = 1.0$	0.8691	0.8663							
$0.8691 - 0.8663 > 0.002$									
...									
I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	\dots
									\dots
$\kappa_i = 1.0$				0.8490	0.8488				
$0.8490 - 0.8488 \leq 0.002$									
I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	\dots
									\dots
$\kappa_i = 1.0$				0.8490	0.8456				
$0.8490 - 0.8456 > 0.002$									

(e) Form an image subset $\mathcal{S} = \{I_1, \dots, I_6\}$ if 6 images can be greedily added (true for this example), otherwise discard the entire set and restart at (a).

I_1	I_2	I_3	I_4	I_5	I_6

Figure S1. The procedure of forming an image subset as described in Section 4.1. We specifically show cases where images are removed or skipped. Note that after forming the subsets, we further filter them to avoid heavy overlaps (see Section 4.1).

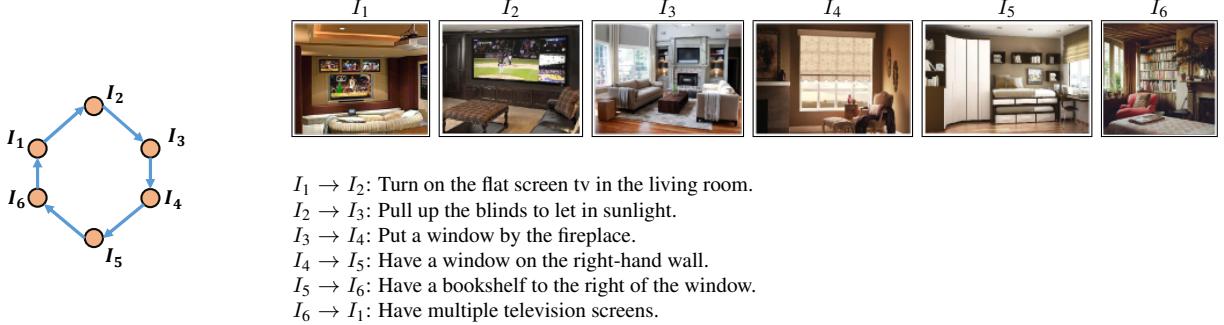


Figure S2. Left: The six pairs we draw from a subset (in total we draw nine) that form a closed-loop dialogue. Each arrow represents a reference-to-target image pair with modification sentences. Right: An example of consecutive modification sentences that forms a dialogue.

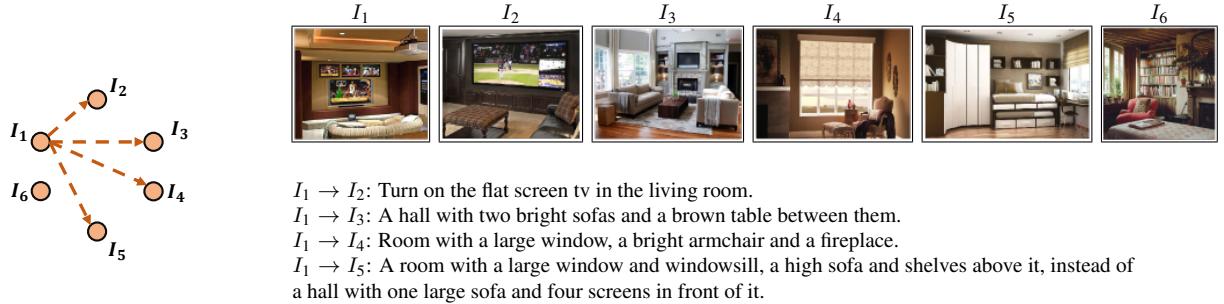


Figure S3. Left: The four pairs we draw from a subset to have multiple outcomes from the same reference image. Each arrow represents a reference-to-target image pair with modification sentences. Right: An example of the four pairs with the same reference image.

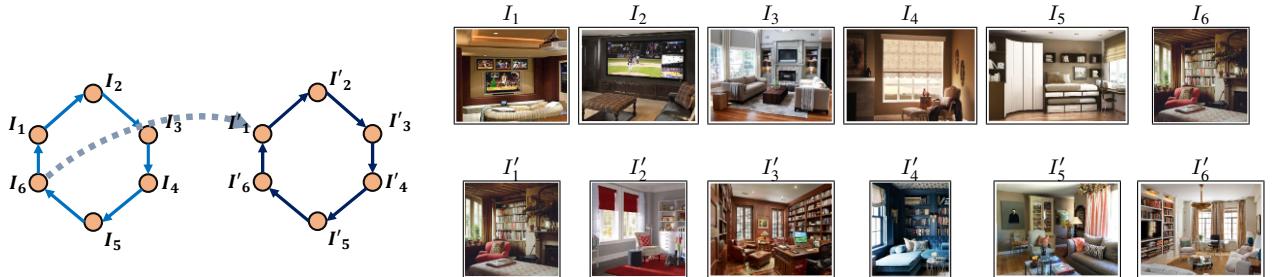


Figure S4. An example of connecting pairs from two subsets to form longer dialogue paths. Note that in this example, $I_6 \equiv I'_1$.

F. Additional Examples of CIRR

We provide additional examples from the dataset in Fig. S7, where we demonstrate negative retrieval results from both TIRG and CIRPLANT. We point out that CIRR focuses more on the challenging task of distinguishing among visually similar images. Let us note that the auxiliary annotations provide explicit interpretations of errors, particularly regarding the implicit human-agreements in visual and language modalities. This suggests that the annotations can be used for fine-grained analysis or as training signals in future research on composed image retrieval.

G. Dataset File Description

Table S5 summarizes information we provide for each image pair.

References

- [1] E. Dodds, J. Culpepper, S. Herdade, Y. Zhang, and K. Boakye. Modality-agnostic attention fusion for visual search with text feedback. *ArXiv*, abs/2007.00145, 2020. [1](#), [5](#)
- [2] M. Forbes, C. Kaeser-Chen, P. Sharma, and S. J. Belongie. Neural Naturalist: Generating fine-grained image comparisons. In *Conference on Empirical Methods in Natural Language Processing*, 2019. [5](#)

Datasets	Statistics			Images			Annotations		Non-repurposed	Additional annotation	
	Nb. pairs	Nb. images	Domain	Natural images	Pairing strategy	Natural language	Dialogue paths	Examples			
1 CSS [9]	–	38,069*	–	–			<i>Make top-center blue object yellow.</i>			✓	
2 MIT-States [5, 9]	–	63,440*	Entity states	✓	by entity class			<i>(Change state) to melted.</i>			
3 Fashion200k [4, 9]	–	338,372* [†]	Fashion	✓	by similar attributes			<i>(Replace with) beige.</i>			
4 Fashion-IQ [3]	30,122 [§] 46,609 [‡]	Fashion	✓	by product category	✓	<i>Is short sleeved and has a collar.</i>			✓	Product attribute (partial)	
5 Birds-to-Words [1, 2]	3,347	–	Birds	✓	by visual similarity	✓	<i>Animal1 is white with dark brown and white wings and a golden head . Animal2 is brown-gold with dark solid-colored brown wings and a dark head.</i>				
6 Spot-the-Diff [1, 6]	23,089	–	Surveillance footage	✓	by video frame	✓	<i>A white truck has appeared in the after image. A person is now walking on the footpath.</i>				
7 CIRR	36,554 21,552	Open	✓	by visual similarity	✓	✓	<i>Room with a large window, a bright armchair and a fireplace.</i>			Auxiliary annotation clarifying ambiguities	

* Nb. pairs not pre-defined, pairs are generated on-the-fly.

† Approx. 100,000 images have low detection score, thus could be removed [4]. Here, we show the available nb. images in total.

‡ Each pair has two sentences.

§ Combining all three subtypes. Note that pairs and images overlap between subtypes.

Table S3. Comparison between CIRR (bolded) and existing datasets for composed image retrieval. CIRR is comparable in size (nb. pairs) while containing richer annotations of open-domain images.

To avoid	Examples
1 Mentioning text/numbers	<i>Text on the pillow says “LOVE”.</i>
2 Discussing photo editing properties	<i>Crop the staircases out of the photo.</i>
3 Subjective opinions	<i>The dogs look very cute.</i>
4 Simply describing the target image, not comparing the pair	<i>Having a large table in the center of the room.</i>
5 Simple side-by-side comparison	<i>The left image shows a laptop on the wooden table, the right image has a flatscreen.</i>
6 Writing sentences that are not unique for the given image pair in the subset	–

Table S4. Types of annotations we discourage workers from writing. Rows 4 and 5 might be admissible if the annotation contains implicit comparisons. (see Fig. S8 (top)).

- [3] X. Guo, H. Wu, Y. Gao, S. J. Rennie, and R. Feris. The Fashion IQ Dataset: Retrieving images by combining side information and relative natural language feedback. *ArXiv*, abs/1905.12794, 2019. [2](#), [5](#), [8](#)
- [4] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis. Automatic spatially-aware fashion concept discovery. In *IEEE International Conference on Computer Vision*, 2017. [5](#)
- [5] P. Isola, J. J. Lim, and E. H. Adelson. Discovering states and transformations in image collections. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. [5](#)
- [6] H. Jhamtani and T. Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *Conference on Empirical Methods in Natural Language Processing*, 2018. [5](#)
- [7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Association for Computing Machinery*, 2017. [1](#)
- [8] A. Suhr, S. Zhou, A. Zhang, I. Zhang, H. Bai, and Y. Artzi. A corpus for reasoning about natural language grounded in photographs. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. [2](#), [6](#)
- [9] N. Vo, L. Jiang, C. Sun, K. Murphy, L.-J. Li, L. Fei-Fei, and

Identifiers (keys)	Explanations	Content Details (values)	Examples
1 pairid	Unique pair id*		12554
2 reference	Reference image		"dev-147-2-img0"
3 target_hard	Target image [§]		"dev-846-2-img0"
4 target_soft	Target image with additional labeling (if exists) ^{§†}	Follow NLVR ² [8] image naming conventions.	{dev-846-2-img0": 1.0, "dev-743-3-img0": -1.0}
5 caption	(Main) annotation		"Catch the crab in the circular ring and place them on the metal table."
6 caption_extend	Auxiliary annotation [‡]		
7 0	Q1		"[c] None existed"
8 1	Q2	Begin with [c] if N/A.	"We don't see the gloved hands of the fisherman"
9 2	Q3		"Focus on the net full of crabs"
10 3	Q4	Begin with [cr0] if Nothing worth mentioning, begin with [cr1] if Covered in brief annotation.	"[cr0] Nothing worth mentioning"
11 img_set	Subset information		
12 id	Unique subset id		106
13 members	Images within subset	Follow NLVR ² [8] image naming conventions.	["dev-147-2-img0", "dev-224-1-img1", "dev-410-2-img0", "dev-743-3-img0", "dev-846-2-img0", "dev-998-1-img0"]
14 reference_rank	Sequence identifier		0
15 target_rank	as in Fig. S2 [§]	Range from 0 to 5, correspond to I_1 - I_6 .	1

* Used for cross-referencing image pairs between `cap.sample.json` and `cap.ext.sample.json`.

† See Fig. S8 (a) for the three possible labels. When constructing `target_soft`, images labelled as [The same image] is added as 1.0, [No differences worth mentioning] is added as 0.5, [Images that are too different] is added as -1.0.

‡ See Fig. S8 (b) for the options we provide for AMT workers.

§ Not public for test-split. Instead, see our project website for the test-split evaluation server.

Table S5. Data structure as in the data files. For details please refer to our project website.

J. Hays. Composing text and image for image retrieval - an empirical odyssey. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1, 5

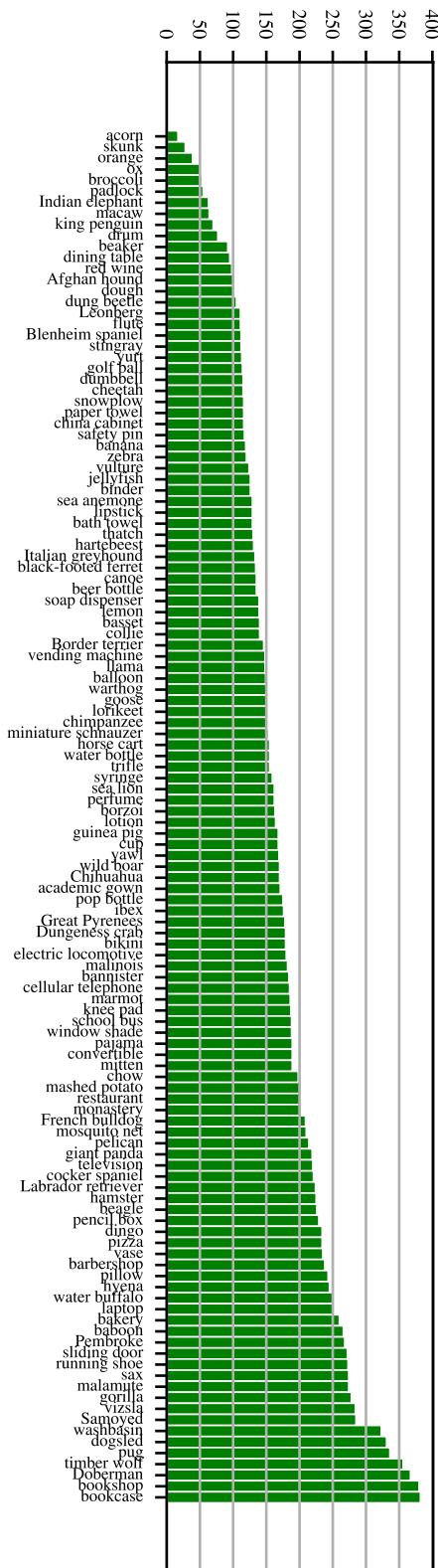


Figure S5. Number of examples per synset (sorted in ascending).

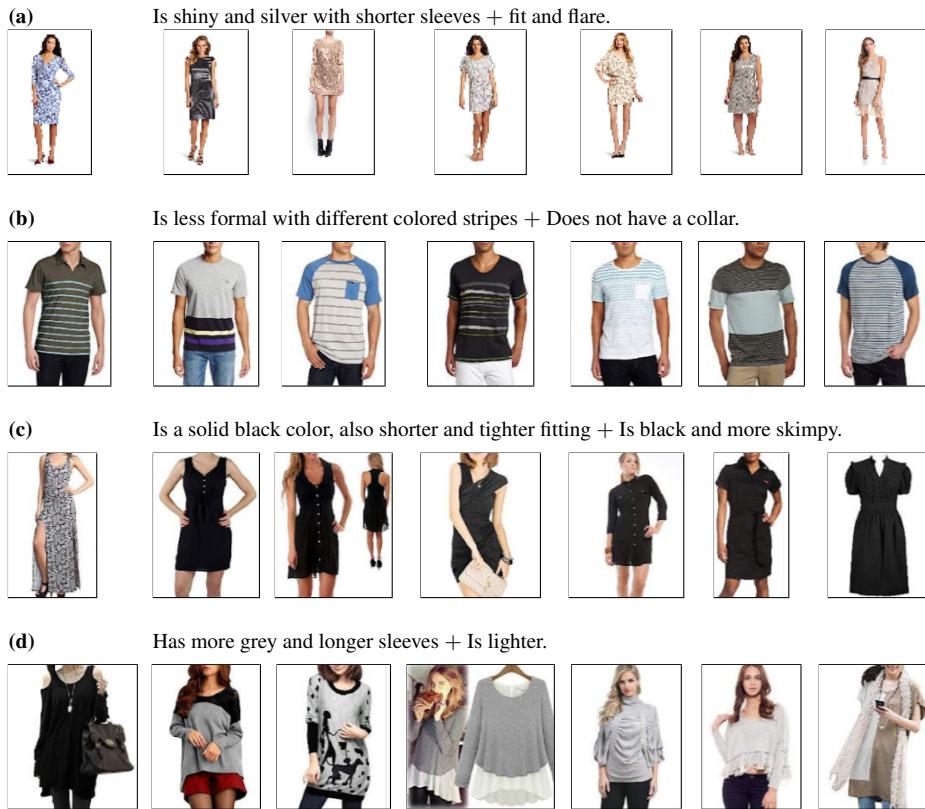


Figure S6. Examples of false-negatives in Fashion-IQ [3]. First column shows the reference image. Each sample contains two modification sentences. For each query set (reference image + modification sentences), only one candidate image is labeled as the target. Thus, rendering the remaining valid predictions as false-negatives.

(a)



Main – Goes from a black and white dog running to two dogs running.

Q1 – [N/A] Nothing worth mentioning

Q2 – Change to a brown-and-white dog and a black-and-white dog.

Q3 – [N/A] Nothing worth mentioning

Q4 – Make the grass a darer green.

(b)



Main – Remove the concrete to the right.

Q1 – Has marine animal in similar blue backdrop.

Q2 – Remove the blue thing on right.

Q3 – [N/A] Nothing worth mentioning

Q4 – [N/A] Nothing worth mentioning

(c)



Main – Remove the seashells and make the water green.

Q1 – Shows manta rays.

Q2 – Make the rays older, spread the rays further apart.

Q3 – View straight on.

Q4 – [N/A] Covered in main annotation

(d)



Main – More monkeys

Q1 – A group of monkeys side by side in same color.

Q2 – [N/A] Nothing worth mentioning

Q3 – More focused on the animals.

Q4 – [N/A] Nothing worth mentioning

Figure S7. Negative results of TIRG and CIRPLANT on CIRR. Here, we show the $\text{Recall}_{\text{Subset}}$ rankings where we consider candidates from corresponding image subsets (see Sec. 5). First column shows the reference images. True targets are in green boxes. Each pair contains a main annotation (Main) and four auxiliary annotations (Q1–4) as explained in Section 4.1 and Sec. C. We demonstrate errors of the models where: (a) fails to associate text with both reference and target image; and (b–d) fails to identify and preserve implicit global visual similarity. We show that CIRR focuses on the challenging task of distinguishing harder negatives that require fine-grained visual reasoning. Let us note that the errors can be explicitly interpreted with our auxiliary annotations (bolded), which previous datasets cannot. This suggests that future work can leverage the auxiliary annotations for analysis of methods, and possibly training of models that account for implicit human ambiguities.

Write a sentence so that ...


→


Reference Image ⇒ Target Image

Meanwhile, (Note!) your sentence do not lead to ...




⇒ Other similar images

Use the prompt below as a guide, complete the rest of the sentence...

Unlike the Reference Image, I want the Target Image to ... / the Target Image (is/has/shows) ...

Alternatively, the Target Image is...

the same image (or a cropped version) as the Reference Image.

not the same, but has no differences you can pick up to describe.

COMPLETELY irrelevant contents, such as very different type of objects or scenes.


→


"Change the angle to side, add steel handrail."

Examine the above image pair and the modifying sentence, answer each question by either writing a sentence, or clicking (one of) the button(s) below:

• Q1 What characteristics of the objects are preserved across images, but hasn't been mentioned in the modifying sentence?

+ Be more specific, write e.g., "The wooden table", instead of "The table".

+ Do NOT talk about background content in this question.

+ Do NOT talk about things that are changed; focus on the PRESERVED aspects.

e.g., Having two ducks of the same breed / Being a photo of a bathroom ...

Nothing is preserved, or everything is already covered in the given text. (Only use in rare occasions!)

• Q2 What objects were changed, but not relevant to the modifying sentence (we are looking for "negligible changes" here)?

+ Do NOT simply list the objects. Instead, write down those changes that took place in details.

+ Do NOT talk about background content in this question.

+ ! "Change" means change from the left(top) image to the right(bottom) image.

(! Follow these examples !) e.g., Add a collar to the dog, and change the color of its eye to black ...

There is nothing to be described.

• Q3 Is there any change in camera angle / focus / viewpoint?

+ Do NOT compare the images side-by-side.

+ Instead, write the sentence that describes the changes from the left(top) to the right(bottom) image.

No, there isn't any change worth mentioning.

Yes, and it has already been covered in the modifying sentence.

Yes, BUT it hasn't been mentioned yet (please write them down):

e.g., More focused on the pillows / change to a lower angle of shot ...

• Q4 Is there any change in the background across the images?

+ Do NOT compare the images side-by-side.

+ Instead, write the sentence that describes the changes from the left(top) to the right(bottom) image.

No, there isn't any change worth mentioning.

Yes, and it has already been covered in the modifying sentence.

Yes, BUT it hasn't been mentioned yet (please write them down):

e.g., Background should contain grass / add a window in the background ...


→ ?

"remove all but one dog and add a woman hugging it"

Choose the most appropriate target image: (It should contains the described changes, and still resembles the Reference Image in other ways)

Figure S8. Snapshots of our collection interface (recommend viewing digitally by zooming in). (top) (Main) annotation, where we specifically require unique sentences within each subset. (middle) Auxiliary annotation, where we ask Amazon Mechanical Turk (AMT) workers of four detailed questions per pair. (bottom) Human performance (Recall_{Subset}@1) evaluation on test-split, where we ask AMT workers to choose the most probable target image within the subset.