Supplementary Material for Pic2Word

1. Additional Discussion

Discussion on DreamBooth [4] and Textual Inversion [1] From the application perspective, these works generate images with the user's intent while our approach retrieves images given the intent. Besides the difference from the application perspective, there are few critical differences. First, our approach needs only unlabeled images, while their approaches rely on a set of images containing the same object. Second, our method works in real-time as it requires single forward pass of visual encoder, mapping network and text encoder at inference time. On the other hand, both Textual Inversion and DreamBooth are far from being a real-time as they require more than thousands of gradient steps to invert or finetune per set of images.

2. Experimental Details

Evaluation Dataset. Table 1 describes the details of the dataset, i.e., number of query images and candidate images used for evaluation. The evaluation datasets are preprocessed as explained in the main paper.

Dataset	Query images	Candidate images
ImageNet	10,000	16,983
COCO	4,766	4,766
CIRR (test)	4,148	2,315
Fashion (Dress)	2,017	3,817
Fashion (Shirt)	2,038	6,346
Fashion (TopTee)	1,961	5,373

Table 1. The number of images used for evaluation in each dataset.

Mapping network design. Table 2 summarizes the mapping network architecture we employ. In the next section, we give the study on the choice of the architecture.

Images used for qualitative examples. In qualitative examples, we exclude images that can recognize the identity of a person. In the third query in Fig. 8, we employ an image[1], which is not included in CC3M validation set.

[1] https://www.flickr.com/photos/enerva/9068467267 licensed with CC-BY 2.0.

Layer	Module
Output	nn.Linear(512, 768)
ReLU2	nn.ReLU
Dropout2	nn.Dropout(0.1)
FC2	nn.Linear(512, 512)
ReLU1	nn.ReLU
Dropout1	nn.Dropout(0.1)
FC1	nn.Linear(512, 512)

Table 2. Pytorch-style [3] model description of the mapping network. The output is fed into the langauge encoder.

Model Description	ImageNet R50	COCO R10	CIRR R10	Fashion R50
Default ($L = 3$, $h_d = 512$)	**23.2**	33.4	65.4	43.7
$L = 3$, $h_d = 4096$	22.4	**33.9**	**67.6**	**45.3**
$L = 5$, $h_d = 512$	22.3	32.3	65.1	42.6
Linear only ($L = 2$, $h_d = 512$)	22.1	33.4	58.8	42.4
Best zero-shot baseline	11.2	26.6	56.7	35.7

Table 3. Analysis of the design of the mapping network. The top row is the model used in the main paper. The bottom is the score of the zero-shot baseline, which performs the best of three zero-shot baselines in each dataset.

3. Additional Experiments

Analysis on the architecture design. Table 3 shows the study on the design of the mapping network. There are two observations: (1) no variants outperform the default model across all datasets, (2) removing non-linear activation can significantly reduce the performance (Linear only model). In order to faithfully predict the pseudo language token, the mapping network needs to be expressive to a certain degree. We also try other variants, e.g., varying dropout rate, but we do not see clear improvements.

Training dataset for the mapping network. Table 4 describes the comparison between the model trained on CC3M and CC12M, where CC12M is 4 times larger than CC3M approximately. We do not see the clear advantage of using CC12M, which indicates that CC3M is large enough to train the mapping network.

Comparison with CLIP fine-tuned on Conceptual Caption (CC3M). In Table 5, we show that fine-tuning CLIP on CC3M improves the baseline performance. However, these baselines still fall short to our approach.

Comparison between CLIP and BLIP. In Table 6, we

Training Dataset	ImageNet		COCO		CIRR		Fashion	
	R10	R50	R1	R10	R1	R10	R10	R50
CC3M	10.1±1.5	23.2±1.1	11.5±0.2	33.4±0.3	22.2±0.6	65.4±1.3	24.7±2.1	43.7±3.4
CC12M	8.9±1.4	21.4±0.6	12.1±1.1	33.9±1.0	22.1±2.5	64.5±3.4	25.1±0.7	43.4±1.2

Table 4. **Study on the dataset to train mapping network.** We report the results averaged over three runs and its standard deviation. Note that we report the results on validation set for CIRR.

Model	Methods	COCO		Fashion	
		R5	R10	R10	R50
Fine-tuned	Text-only	14.3	22.0	19.4	37.1
	Image+Text	22.5	29.0	20.9	36.9
Original	Text Only	15.7	23.5	17.3	32.8
	Image + Text	20.2	26.6	19.8	35.7
	Ours	24.8	33.4	24.7	43.7

Table 5. Baseline results of CLIP fine-tuned on CC3M (top two).

Model	Methods	R1	R5	R10	R50
BLIP	Image-only	7.2	25.6	36.6	62.4
	Text-only	**25.1**	52.0	62.4	82.7
	Image+Text	16.5	47.2	61.3	86.8
CLIP	Image-only	7.5	25.1	35.5	59.8
	Text-only	20.8	46.2	57.0	78.8
	Image+Text	13.2	36.6	50.5	78.1
	Pic2Word	22.6	**52.6**	**66.6**	**87.3**

Table 6. Comparison between CLIP and BLIP ViT-L/14 model [2]. Evaluation on CIRR validation set.

show the comparison between CLIP and BLIP [2] ViT-L/14 model. Text-only result of BLIP outperform that of CLIP, showing that difference in pre-training can result in the significant difference in the performance.

Additional qualitative examples. Fig. 1 shows retrieval examples of ImageNet. Both the query and candidate images are from the same synset of one category. We can see that target images are within top-4 in these examples. Fig. 2 shows retrieval results tested on the images downloaded from the web. In this evaluation, all of the top-1 images are included in the candidate set, which indicates that the composed representations can express both the characteristics of an object, specified by an image, and the attribute from language.

References

[1] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1

[2] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*. PMLR, 2022. 2

[3] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019. 1

[4] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 1

Figure 1. Image retrieval examples in the synset of n02107142 (doberman) and n02123159 (tiger cat). Note that the search is performed on 50 images from each category. Target images are highlighted with green outline.

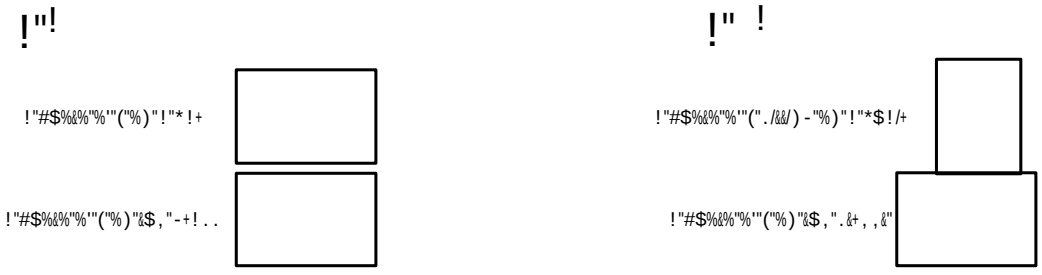

Figure 2. Qualitative retrieval results. Query images are from rawpixel and wikimedia. Top-1 images are downloaded from flickr, rawpixel, flickr, and wikimedia (top to bottom and left to right). Note that these images are licensed by either CC-BY 2.0, CC BY-SA 2.0, or CC0 1.0 Universal. We create a candidate set by these images and CC3M validation split. Target images are highlighted with green outline.