

# Composed Image Retrieval for Training-FREE Domain Conversion Supplementary

Nikos Efthymiadis<sup>1\*</sup> Bill Psomas<sup>1,2</sup> Zakaria Laskar<sup>1</sup> Konstantinos Karantzalos<sup>2</sup>  
Yannis Avrithis<sup>3</sup> Ondřej Chum<sup>1</sup> Giorgos Tolias<sup>1</sup>

<sup>1</sup>VRG, FEE, Czech Technical University in Prague    <sup>2</sup>National Technical University of Athens  
<sup>3</sup>Institute of Advanced Research in Artificial Intelligence (IARAI), Austria

## 1. Method details

FREEDOM is presented in Algorithm 1 for further clarity. The multiple text inversion, lines 10 - 12, is efficiently executed by a single NN-search for a set of queries whose GPU implementations are readily available (e.g., FAISS).

## 2. Competing methods

We provide the implementation details of the literature methods used in the main paper.

**Pic2Word** [13] achieves textual inversion in the latent space of the text tokens through a three-layered MLP. In every experiment with Pic2Word, we use the officially pre-trained mapping network released by the authors. For ImageNet-R and MiniDN, the composed query has the same format as in the original paper: “a [*target domain*] of \*”, e.g. “a cartoon of \*”. For NICO++, the composed query is “a \* in [*target domain*]”, e.g. “a \* in autumn”. Finally, for the LTL dataset, the composed query “a [*target domain*] photo of \*” is used, e.g. “a today photo of \*”.

**SEARLE** [1] performs textual inversion by test-time optimization to represent query images in the latent space of the vector tokens. We opt for the optimization variant instead of their feed-forward network since it is shown to perform better. We use the official implementation for our experiments. We refer to the version with default optimization hyper-parameters as “SEARLE (default)” and to our improved hyper-parameters by “SEARLE”. Each query image is associated with different concepts retrieved from a vocabulary, which is the same as the text labels of our method. We refer to the number of those concepts by  $m$  in Table 4 of the main paper. The final composed queries are adapted for each dataset in the same way as for Pic2Word. We perform a search for learning rate in  $\{0.2, 0.02, 0.002, 0.0002\}$ , iterations in  $\{5, 10, 50, 200, 350, 500\}$ , and the number of textual labels  $m$  in  $\{1, 3, 7, 10, 15\}$ . The best results across all datasets are for  $lr = 0.0002$ ,  $iters = 350$ , and  $m = 1$ .

---

### Algorithm 1 FREEDOM.

---

```
1: procedure FREEDOM( $y, t, X, V, Z$ )
2:    $y$  : image query
3:    $t$  : text query
4:    $V, \mathcal{V}$  : textual memory (words vocabulary) and embeddings
5:    $Z, \mathcal{Z}$  : visual memory (external image set) and embeddings
6:    $X, \mathcal{X}$  : database images and embeddings
7:    $\mathbf{y} \leftarrow f(y)$  ▷ embedding of image query
8:    $\{\mathbf{y}_1, \dots, \mathbf{y}_k\} \leftarrow \text{NN}_k(\mathbf{y}; \mathcal{Z})$  ▷  $k$  nearest proxy images - including  $\mathbf{y}$ 
9:    $W^+ \leftarrow \emptyset$  ▷ collect all word inversions
10:  for  $i \in 1 \dots k$  do ▷ loop over proxy images
11:     $W^+ \leftarrow W^+ \cup \text{NN}_n(\mathbf{y}_i; \mathcal{V})$  ▷ invert proxy image -  $n$  nearest words
12:  end for
13:   $\{\hat{w}_1, \dots, \hat{w}_m\}, \{\hat{a}_1, \dots, \hat{a}_m\} \leftarrow \text{most-frequent}_m(W^+)$  ▷  $m$  most frequent words and frequencies
14:   $\mathbf{t} \leftarrow$  zero vector
15:  for  $i \in 1 \dots m$  do ▷ loop over frequent words
16:     $\mathbf{t}_i \leftarrow g(\hat{w}_i \oplus t)$  ▷ composed query (e.g. “shark origami”) embedding
17:     $\mathbf{t} \leftarrow \mathbf{t} + \hat{a}_i \mathbf{t}_i$  ▷ aggregated query - equivalent to late fusion
18:  end for
19:  Rank images  $x_j \in \mathcal{X}$  based on similarity  $\mathbf{t}^\top x_j$  ▷ execute the final query
20: end procedure
```

---

**CompoDiff** [8] is built on top of a frozen CLIP. We follow the publicly released official implementation for our experiments. We use the officially pre-trained denoising Transformer released by the authors. We do not use any masks or any mixed text condition. The query text includes only the target domain word, i.e., “[*target domain*]”.

**WeiCom** [12] is a composed image retrieval method specialized for remote sensing. It fits a normal distribution to the similarities between the text query  $g(t)$  and all the database images  $f(x)$  for  $x \in X$ , and similarly for the image query  $f(y)$ . It uses each distribution’s corresponding cumulative distribution function to transform the similarities closer to the uniform distribution. It then combines the similarities by summation.

**MagicLens** [18] is a composed image retrieval method that fine-tunes a VLM model on triplets collected from the internet, assuming that images from the same website share implicit relationships describable by textual instructions. We

(a) ImageNet-R						
METHOD	CAR	ORI	PHO	SCU	TOY	AVG
InstructPix2Pix	3.90	5.70	1.97	5.70	5.62	4.58
FREEDOM w/ img-cap	15.11	6.70	19.77	18.08	16.58	15.24
FREEDOM w/ captioners	16.68	11.74	17.44	15.68	16.94	15.70
<b>FREEDOM</b>	<b>35.97</b>	<b>11.80</b>	<b>27.97</b>	<b>36.58</b>	<b>37.21</b>	<b>29.91</b>

(b) MiniDomainNet					
METHOD	CLIP	PAINT	PHO	SKE	AVG
InstructPix2Pix	8.57	8.86	7.08	7.20	7.93
FREEDOM w/ img-cap	21.88	17.54	31.78	15.35	21.64
FREEDOM w/ captioners	27.65	17.42	33.42	17.24	23.91
<b>FREEDOM</b>	<b>41.96</b>	<b>31.65</b>	<b>41.12</b>	<b>34.36</b>	<b>37.27</b>

(c) NICO++							
METHOD	AUT	DIM	GRA	OUT	ROC	WAT	AVG
InstructPix2Pix	4.18	2.66	4.60	4.78	5.19	3.56	4.16
FREEDOM w/ img-cap	15.56	11.64	19.34	19.18	17.56	13.81	16.18
FREEDOM w/ captioners	14.07	9.54	18.67	20.86	17.34	12.37	15.48
<b>FREEDOM</b>	<b>24.35</b>	<b>24.41</b>	<b>30.06</b>	<b>30.51</b>	<b>26.92</b>	<b>20.37</b>	<b>26.10</b>

(d) LTLL			
METHOD	TODAY	ARCHIVE	AVG
InstructPix2Pix	9.83	20.02	14.92
FREEDOM w/ img-cap	<b>42.58</b>	19.16	30.87
FREEDOM w/ captioners	26.52	18.76	22.19
<b>FREEDOM</b>	<b>30.95</b>	<b>35.52</b>	<b>33.24</b>

Table 1. Evaluation of advanced baselines.

use the CLIP-L variant from the official code and evaluate its performance with two settings: the default prompt, “find this object in [target domain]”, which performed poorly across datasets, and prompts tailored per dataset. The best prompts were: “a [target domain] of this” (ImageNet-R), “a [target domain] of” (MiniDN), “in [target domain]” (NICO++), and “a [target domain] photo of” (LTLL). Results are reported as “MagicLens (original prompt)” and “MagicLens”.

### 3. Advanced baselines

In addition to the simple baselines in the main paper, we present the following more “advanced” baselines and summarize their performance in Table 1.

**InstructPix2Pix.** [4] In this baseline, InstructPix2Pix is used to generate an image from our visual and textual queries. Then, retrieval is done by image-to-image similarities. The performance of this baseline is low, indicating that the combination of the two modalities through the visual encoder is sub-optimal. We qualitatively observe that although several of the generated images are quite successful, many are completely unsuccessful.

**FREEDOM w/ img-cap.** In this baseline, we assume access to a dataset of image-caption pairs; the first 40M images and captions of LAION 400M [14] are used. This set forms a joint visual-textual memory. Proxy images are retrieved from this memory, and their captions are treated as the text labels of textual inversion. Then, they are combined with the query text, and late fusion follows with weights equal to the similarities between the query and the memory images. The hyperparameters are the same as our standard FREEDOM. Interestingly, this baseline surpasses FREEDOM on LTLL for the case of “today” as the source domain.

**FREEDOM w/ captioners.** In this baseline, two captioners are used, namely BLIP [10] and BLIP2 [9]. Each captioner captions every query image, and the results are used as the two text labels for the image. Subsequently, our stan-

dard processing pipeline is followed. The similarities of each caption are used with the query image as weights for late fusion. This baseline uses extra architectures, is 15 times slower than the standard FREEDOM, and is consistently worse.

## 4. Additional results

**Impact of FREEDOM components to different inversion methods.** The three main components of FREEDOM are text memory-based inversion, visual memory-based expansion, and late fusion. We apply the last two components on top of different inversion methods, whenever applicable, *i.e.*, with SEARLE and Pic2Word. Incorporating the two FREEDOM components (using  $m = k$ , while  $n$  is equal to 1 inherently for both methods) improves both methods, while our text memory-based inversion performs consistently the best. This experiment is summarized in Figure 1. We follow the FREEDOM workflow: A visual memory is used to enrich the query with  $k$  images. Then, inversion follows as FREEDOM, SEARLE, and Pic2Word. Finally, the combination is done by late fusion. Although the memory-based inversion is more sensitive for large  $k$  (dotted blue), our design choice of having a fixed number of final words ( $m = 7$ ) makes FREEDOM robust.

**Impact of visual memory.** We demonstrate the performance of FREEDOM with different visual memories in Table 2. Compared to no visual memory, every other option improves the performance on average. Furthermore, every visual memory is advantageous for every individual dataset except for the case of ImageNet-R with LAION 40M due to the low availability of images in specific domains such as *origami*. Therefore, the efficacy of the memory remains robust even when dealing with unstructured datasets such as the image part of LAION. Additionally, including task-relevant images, even in small proportions, proves advantageous. The best improvements are achieved using the database as memory, which is our default choice.

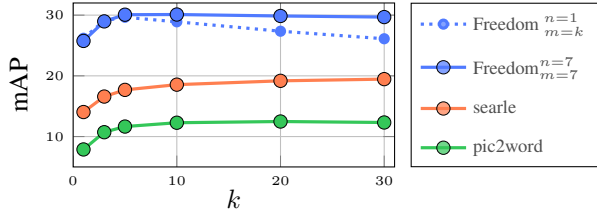


Figure 1. *Different Inversions* with visual memory expansion and late-fusion on ImageNet-R

MEMORY	AVG	IMAGENET-R	MINIDN	NICO++	LTL
NO MEMORY	27.96	25.77	32.06	23.20	30.82
LAION 40M	28.57	25.00	33.85	24.31	31.11
DATABASE + LAION 40M	29.52	26.07	34.92	24.91	32.17
DATABASE	31.63	29.91	37.27	26.10	33.24

Table 2. *Impact of the visual memory*: Comparing performance between no visual memory, the database as visual memory, a 40M-image LAION [14] visual memory, and their union.

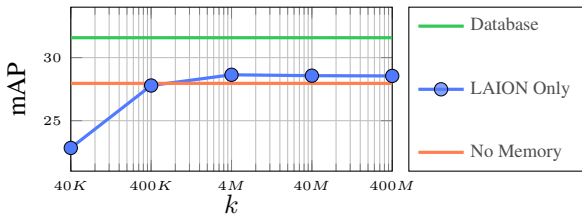


Figure 2. *Impact of the visual memory*: Performance comparison between no visual memory, the database as visual memory, and visual memory comprising LAION [14] images of various sizes.

We also study the effect of the size of the visual memory. We choose LAION subsets of size 40k, 400k, 4M, 40M, and 400M as visual memories. The average mAP of ImageNet-R, NICO++, MiniDomainNet, and LTL is reported in Figure 2. The database as a visual memory is the upper bound for this experiment, given that it is curated for the task. A performance saturation is observed for the visual memory of size 4M, which surpasses the performance of the no-visual-memory baseline FREEDOM (k=1). This supports the idea that visual memory is beneficial even when not curated. It also suggests a practical upper bound for the visual memory size. Notably, FREEDOM, with a visual memory size of 4M, has a query latency of 24.8ms for ImageNet-R.

**Query time.** Figure 3 presents the latency comparison between FREEDOM and the competitive methods: WeiCom, Pic2Word, and CompoDiff on ImageNet-R. CompoDiff achieves an mAP of 12.9 while being significantly slower, with a latency of 257.5ms. In contrast, WeiCom, Pic2Word, and FREEDOM (with m=1, n=1, and k=1) exhibit similar latencies (16.2ms, 16.3ms, and 16.6ms, respectively), but FREEDOM outperforms the rest with an mAP of 26.18 compared to 10.47 for WeiCom and 7.88 for Pic2Word. Increasing

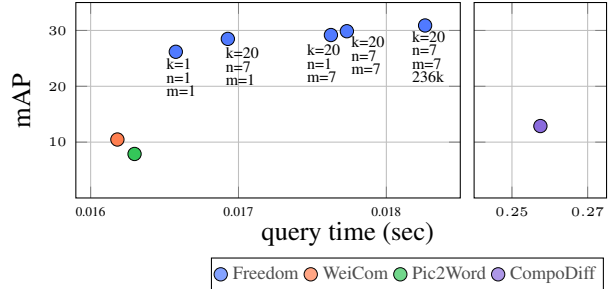


Figure 3. *Performance vs. query time*. Different variants of FREEDOM are shown by varying hyper-parameter ( $k$ ,  $m$ ,  $n$ ) values and textual memory size. FREEDOM uses textual memory with size 20k (default) or 236k (reported).

DATASET	$l$			
	1	5	10	20
IMAGENET-R	-3.48	-6.39	-7.57	-9.08
NICO++	-1.03	-2.52	-3.35	-4.42
MINIDN	-2.08	-5.32	-6.77	-8.37
LTL	-1.50	-2.43	-2.96	-4.43

Table 3. *Oracle experiment* to study the sensitivity of FREEDOM by removing the  $l$  words closest to each query’s class label from the vocabulary. We report the mAP (%) reduction.

ing FREEDOM hyperparameters to k=20 and n=7 yields an additional 2.31 mAP with a minimal latency cost of 0.3ms. The variant with k=20, m=7, n=1 shows a further latency increase of 0.7ms, leading to a 0.69 mAP gain. The default FREEDOM configuration (k=20, m=7, n=7) achieves 29.86 mAP with a total latency of 17.7ms. Expanding the text memory to 236k words [3] results in an mAP of 30.89 with a final latency of 18.3ms.

To test the scalability of FREEDOM, on top of the 236k text memory, we artificially enlarge the visual memory to 1M. The latency increases by only 1.8ms to a total of 20.1ms. By expanding the database to 1M, we get a latency of 24.5ms. Even in a database of two magnitudes larger (ImageNet-R VS our artificial 1M database), FREEDOM is more than 10 times faster than CompoDiff.

**Oracle experiment.** We demonstrate the robustness of FREEDOM with respect to the choice of textual memory. As an oracle experiment, the  $l$  words closest to each query’s class label are removed from our vocabulary, and the results with the remaining are reported. The performance reduction (mAP) for  $l = 1, 5, 10, 20$  is summarized in Table 3. Despite the performance drop, FREEDOM still outperforms the state-of-the-art on all datasets, even for  $l = 20$ . In Table 4, we present examples of excluded words for this experiment.

**Memory-based inversion examples.** In Table 5, we present some examples of the memory-based inversion of FREEDOM, including the inverted text and the corresponding frequency as weight.

IMAGENET-R					
OBJECT	NN-1	NN-2	NN-3	NN-4	NN-5
School bus	School bus	Bus	Airport bus	Minibus	Bus driver
Gullotine	Gullotine	Meat cutter	Paper cutter	Grindstone	Lock
Lawn mower	Lawn mower	Mower	Riding mower	Lawn	Walk-behind mower
African chameleon	Chameleon	Common chameleon	Lizard	Dragon lizard	Reptile
Basset	Basset hound	Basset artésien normand	Beagle	Spaniel	Bulldog
Beer glass	Beer glass	Beer	Beer	Wine glass	Pint
Collie	Collie	Australian collie	Border collie	Spaniel	Wolf
Golden retriever	Golden retriever	Retriever	Goldendoodle	Golden dream	Puppy
NICO++					
OBJECT	NN-1	NN-2	NN-3	NN-4	NN-5
Ostrich	Ostrich	Ostrich meat	Emu	Elephant	Camel
Bus	Bus	Airport bus	School bus	Bus driver	Car
Kangaroo	Kangaroo	Red kangaroo	Koala	Reindeer	Camel
Lifeboat	Lifeboat	Boat	Speedboat	Rescuer	Jollyboat
Airplane	Aircraft	Aircraft	Airliner	Air travel	Aviation
Butterfly	Butterfly	Moths and butterflies	Moth	Insect	Monarch butterfly
Crocodile	Crocodile	Alligator	Dinosaur	Crocodile	Iguana
Chair	Chair	Office chair	Folding chair	Club chair	Throne
MINIDOMAINNET					
OBJECT	NN-1	NN-2	NN-3	NN-4	NN-5
Sheep	Sheep	Wool	Shepherd	Livestock	Flock
Skateboard	Skateboard	Skateboarding	Skate	Skateboard deck	Skateboarder
Peanut	Peanut	Peanut butter	Soy nut	Bean	Biscuit
Pig	Pig	Boar	Pignolo	Ham	Rat
Rhinoceros	Rhinoceros	Indian rhinoceros	Hippopotamus	Elephant	Dinosaur
Truck	Truck	Trailer truck	Pickup truck	Truck driver	Truck racing
Carrot	Carrot	Baby carrot	Carrot cake	Vegetable	Root vegetable
Pear	Pear	Asian pear	European pear	Apple	Onion
LTLL					
OBJECT	NN-1	NN-2	NN-3	NN-4	NN-5
Notredame	Cathedral	Négociant	Jesus	Arena	Château
TempleTooth	Tooth	Temple	Mouth	Tombet	Temple Jade
BigBen	Clock	Bell	Man	Bee	Dollar
SacreCoeur	Cemetery	Tours (City)	Church	Grave	Tomb
TajMahal	Elephant	Tiger	Mahlab	Masala	Naan
Archedtrionphe	Triumphal arch	Natural arch	Arch	Tunnel	Gate
Pettaah	Mustamakkara	Pathiri	Kootu	Kozhukkatta	Poriyal
EiffelTower	Skyscraper	Tower	Mountain	Lighthouse	Windmill

Table 4. Example words removed for the robustness ablation experiment. The first column shows the class names of queries. The remaining columns (ranked in descending order) show top-ranked words from the textual memory.

**SigLIP as the backbone.** We test the transferability of FREEDOM to other backbones by using features from SigLIP [17], and the results are summarized in Table 6. We observe a significant increase in all datasets in the 4.13 to 16.36 mAP range without additional tuning.

**Datasets for general composed image retrieval.** In this work, we focus on the domain conversion task, motivated by the significance of its applications. Addressing the challenges of this task, particularly the utilization of bi-modal queries and open-world recognition across domains and objects, proves to be non-trivial. Given that our method handles these challenges well and considering that these challenges extend universally to the general composed image retrieval, we evaluate FREEDOM on benchmarks of the general task: FashionIQ [15], CIRR [11], and CIRCO [1]. The results are summarized in Table 7.

Even though FREEDOM is training-free and its scope is domain conversion, the results indicate that it is comparable with some general methods. Specifically, compared to Pic2Word, SEARLE, and CompoDiff, FREEDOM underperforms in Fashion-IQ, it performs comparably well in CIRR, and is the best approach in CIRCO.





Query Image				
NN-1	Snow leopard 1.00	Gothic architecture 1.00	Steam engine 1.00	Soccer 1.00
NN-2	Big cats 0.80	Unesco world heritage site 1.00	Locomotive 0.95	Street football 0.93
NN-3	Himalayan 0.60	Cathedral 1.00	Train 0.90	Freestyle football 0.93
NN-4	Clouded leopard 0.45	Medieval architecture 0.95	Steam 0.75	Soccer kick 0.64
NN-5	Big cat 0.35	Classical architecture 0.90	Railway 0.35	Soccer ball 0.57
NN-6	Snowball 0.25	Holy places 0.90	Railroad engineer 0.35	Kick (Sports) 0.50
NN-7	Arctic 0.25	Gothic 0.30	British rail class 81 0.35	Street sports 0.43

Table 5. Memory-based inversion. Examples of query images alongside their inverted text and corresponding weights of FREEDOM.

**Detailed results.** Following the literature [1, 8, 13], we evaluate on ImageNet-R, using only the PHOTO domain as source, and measure Recall@k. We compare with baselines and competitors in Table 8. The baselines and the SEARLE experiments are performed by us, Pic2Word performance is reported from the original paper, the rest of the Pic2Word experiments, ARTEMIS [5], CLIP4CIR [2], and CompoDiff are reported from the CompoDiff paper. FREEDOM outperforms all baselines and competitors by a large margin. CIREVL is the second best, even though it uses architectures with an estimated number of parameters of two orders of magnitude higher than FREEDOM.

Table 9 shows exhaustive results for all source-target domain combinations on ImageNet-R, NICO++, and MiniDomainNet. We compare FREEDOM with baselines and competitors on all datasets. On ImageNet-R (Table 9a), SEARLE is the second best, while CompoDiff, WeiCom, and MagicLens surpass it for specific source/target couples. On NICO++ (Table 9b), MagicLens is the second best. On MiniDomainNet (Table 9c), CompoDiff and SEARLE are the second and third-best methods, respectively.

(a) ImageNet-R							(b) MiniDomainNet					
METHOD	CAR	ORI	PHO	SCU	TOY	AVG	METHOD	CLIP	PAINT	PHO	SKE	AVG
Text	0.88	0.80	0.62	0.95	0.90	0.83	Text	0.76	0.72	0.76	0.75	0.74
Image	4.97	3.70	0.84	8.18	7.40	5.02	Image	5.07	7.53	3.68	6.15	5.61
Text × Image	6.57	4.34	4.89	6.46	7.46	5.94	Text × Image	3.00	2.60	4.34	3.18	3.28
Text + Image	7.88	5.84	3.08	13.50	12.71	8.60	Text + Image	7.79	11.33	10.80	9.02	9.74
<b>FREEDOM</b>	<b>49.46</b>	<b>27.12</b>	<b>38.11</b>	<b>47.52</b>	<b>46.90</b>	<b>41.82</b>	<b>FREEDOM</b>	<b>57.14</b>	<b>45.47</b>	<b>59.71</b>	<b>52.21</b>	<b>53.63</b>

(c) NICO++								(d) LTLL				
METHOD	AUT	DIM	GRA	OUT	ROC	WAT	AVG	METHOD	TODAY	ARCHIVE	AVG	
Text	1.08	1.13	1.04	1.26	1.10	1.11	1.12	Text	3.84	5.02	4.43	
Image	6.19	5.19	5.42	7.67	7.44	5.62	6.25	Image	10.25	28.14	19.20	
Text × Image	2.31	2.91	3.26	3.53	3.25	2.90	3.03	Text × Image	4.87	3.49	4.18	
Text + Image	8.35	7.19	8.08	11.42	10.57	8.12	8.95	Text + Image	10.16	26.73	18.44	
<b>FREEDOM</b>	<b>30.28</b>	<b>29.96</b>	<b>33.86</b>	<b>37.16</b>	<b>33.14</b>	<b>26.49</b>	<b>31.81</b>	<b>FREEDOM</b>	<b>27.45</b>	<b>47.00</b>	<b>37.22</b>	

Table 6. Domain conversion mAP (%) on four datasets, with SigLIP as a backbone. The best is denoted in bold.

(a) CIRR					(b) CIRCO				
METHOD	R@1	R@5	R@10	R@50	METHOD	mAP@5	mAP@10	mAP@25	mAP@50
Pic2Word	23.9	51.7	65.3	87.8	Pic2Word	8.7	9.5	10.7	11.3
SEARLE	<b>24.2</b>	52.5	66.3	88.8	SEARLE	11.7	12.7	14.3	15.1
CompoDiff	18.2	<b>53.1</b>	<b>70.8</b>	<b>90.3</b>	CompoDiff	12.6	13.4	15.8	16.4
<b>FREEDOM</b>	21.0	48.7	61.9	88.1	<b>FREEDOM</b>	<b>14.0</b>	<b>14.8</b>	<b>16.4</b>	<b>17.2</b>
<b>FREEDOM *</b>	23.8	52.3	65.1	88.9	<b>FREEDOM *</b>	12.0	12.8	14.4	15.0

(c) FASHION-IQ								
METHOD	Dress		Shirt		Toptee		Average	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
Pic2Word	20.0	40.2	26.2	43.6	27.9	47.4	24.7	43.7
SEARLE	20.5	43.1	26.9	45.6	29.3	50.0	25.6	46.2
CompoDiff	<b>24.8</b>	<b>44.8</b>	<b>29.5</b>	<b>47.4</b>	<b>31.4</b>	<b>53.7</b>	<b>28.6</b>	<b>48.6</b>
<b>FREEDOM</b>	16.8	36.3	23.5	38.5	24.7	43.7	21.6	39.5
<b>FREEDOM *</b>	17.2	37.8	24.9	40.8	24.8	44.7	22.3	41.1

Table 7. Composed image retrieval beyond domain conversion: We evaluate FREEDOM on the three most popular benchmarks for general composed image retrieval. We denote with \* the optimized parameters of FREEDOM obtained through hyperparameter tuning on the validation set of CIRR.

## 5. Visualizations

Figure 4 shows visualizations of the top-ranked database images of FREEDOM on ImageNet-R. We use PHOTO as the source domain and convert it to any target domain. FREEDOM can retrieve correct images in all cases. Figure 5 shows visualizations of the top-ranked database images of FREEDOM on MiniDomainNet. We perform SKETCH → PHOTO conversion, i.e., sketch-based image retrieval [6, 7, 16]. Interestingly, FREEDOM is performing well in this task, in contrast to Pic2Word [13].

Furthermore, we present challenging cases where state-of-the-art methods underperform, and the performance of FREEDOM is demonstrated. Figure 6 shows visualizations of the top-ranked database images of FREEDOM vs. competitors on the instance-level dataset LTLL. ARCHIVE →

TODAY and TODAY → ARCHIVE domain conversions are performed. We observe that the competitors confuse both domains and instances. Figure 7 shows visualizations of the top-ranked database images of FREEDOM vs. competitors on NICO++. AUTUMN → DIMLIGHT and GRASS → AUTUMN domain conversions are performed. FREEDOM has the best retrieval results, while the competitors fail almost everywhere.

In our visual examples, we excluded exact duplicates, and we performed aspect ratio changes for better presentation.

METHOD	CARTOON		ORIGAMI		TOY		SCULPTURE		AVG	
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
Text	0.15	0.95	0.87	3.73	0.71	1.77	0.36	1.89	0.52	2.09
Image	0.31	4.51	0.21	1.73	0.54	5.65	0.33	4.04	0.35	3.98
Text + Image	1.96	12.91	2.18	10.68	1.34	9.89	1.82	12.15	1.83	11.41
Pic2Word	8.00	21.90	13.50	25.60	8.70	21.60	10.00	23.80	10.05	23.23
Pic2Word (CC-3M)	7.35	18.53	12.79	25.54	10.39	22.96	10.24	23.76	10.19	22.70
Pic2Word (LAION 2B-en)	8.17	20.86	14.08	25.06	8.73	22.07	10.43	23.63	10.35	22.91
ARTEMIS w/ CompoDiff dataset	11.42	23.81	15.49	25.44	11.21	24.01	10.84	21.07	12.24	23.58
CLIP4Cir w/ CompoDiff dataset	10.90	24.12	16.08	25.60	11.01	23.57	10.45	21.86	12.11	23.79
CompoDiff (T5-XL)	8.43	20.40	15.73	25.69	11.19	22.48	9.19	18.45	11.14	21.76
CompoDiff (CLIP+T5-XL)	12.91	24.40	17.22	26.40	11.57	26.11	11.53	22.54	13.31	24.86
CompoDiff (CLIP)	13.21	24.06	17.03	26.17	11.22	26.25	11.24	22.96	13.18	24.86
KEDs	14.80	34.20	23.50	34.80	16.50	36.30	17.40	36.40	18.00	35.40
MagicLens (original prompt)	9.95	22.37	5.07	17.58	11.51	26.76	7.92	19.70	8.61	21.60
MagicLens	13.65	31.31	6.59	19.21	14.80	31.79	10.33	24.82	11.34	26.78
WeiCom	11.61	24.36	15.24	23.72	8.00	17.89	13.81	26.18	12.17	23.04
SEARLE (default)	1.49	12.38	3.78	13.88	1.99	15.34	2.18	15.34	2.36	14.24
SEARLE	10.17	30.32	17.02	32.00	8.23	9.10	11.60	32.41	11.76	30.96
CIReVL	19.20	42.80	22.2	43.10	30.20	41.30	23.40	45.00	23.75	43.05
<b>FREEDOM</b>	<b>23.77</b>	<b>48.83</b>	<b>32.84</b>	<b>42.82</b>	<b>25.70</b>	<b>47.59</b>	<b>27.86</b>	<b>48.96</b>	<b>27.54</b>	<b>47.05</b>

Table 8. *Domain conversion evaluated by Recall@k (%) on ImageNet-R. Comparison of FREEDOM with baselines and competitors. Source domain: PHOTO; target domains: CARTOON, ORIGAMI, TOY, and SCULPTURE. AVG: average Recall@10 and Recall@50 over all target domains. **Bold**: best, **magenta**: second best.*



Text						Image					Text + Image						
CART	ORI	PHO	SCU	TOY	AVG	CART	ORI	PHO	SCU	TOY	AVG	CART	ORI	PHO	SCU	TOY	AVG
0.7	1.3	0.4	0.8	0.7	0.8	0.7	1.7	1.5	2.9	3.2	4.3	5.4	5.2	11.3	4.8	5.1	6.6
0.6	0.8	0.4	1.0	0.4	0.6	2.5	5.3	2.1	2.6	3.1	4.6	3.8	1.3	5.0	1.8	1.7	2.2
0.9	1.3	0.4	0.5	0.8	0.7	3.7	1.9	13.3	4.6	5.9	7.6	9.2	8.8	12.4	6.3	6.3	9.2
0.9	1.0	0.4	0.8	0.6	0.8	4.2	1.2	11.7	3.3	5.1	8.8	10.3	6.5	10.9	6.8	8.6	8.6
0.8	1.1	0.4	0.8	0.6	0.7	3.0	1.0	10.5	1.9	2.8	3.8	7.2	5.4	9.9	4.3	4.2	6.2

(a) ImageNet-R

Text						Image					Text + Image									
AUT	DIM	GRA	OUT	ROC	WAT	AVG	AUT	DIM	GRA	OUT	ROC	WAT	AVG	AUT	DIM	GRA	OUT	ROC	WAT	AVG
1.4	1.0	0.4	0.8	1.4	1.0	1.0	6.2	13.1	4.8	4.2	3.9	6.5	11.1	14.3	4.2	5.8	7.1	8.5	8.5	
1.6	1.4	0.9	0.3	0.8	1.3	1.0	4.3	6.2	3.9	3.3	4.6	4.9	9.1	9.2	3.3	4.3	7.0	6.6	6.6	
1.6	1.4	0.9	0.3	0.8	1.6	1.2	5.9	5.5	6.2	4.8	6.0	5.7	13.0	10.7	6.9	5.4	6.7	10.3	9.2	
1.7	1.4	1.0	0.8	1.4	1.2	1.1	4.0	6.0	13.4	5.7	8.9	7.7	10.1	11.8	16.9	7.7	13.1	11.9	11.9	
1.5	1.4	0.8	0.4	1.4	1.1	1.1	5.4	6.2	12.4	7.2	7.0	7.7	12.5	11.4	14.8	6.4	10.9	11.2	11.2	
1.8	1.4	0.9	0.4	0.8	1.4	1.1	5.3	9.4	6.2	4.1	5.3	5.7	8.3	10.1	12.5	5.6	5.5	8.4	8.4	
1.7	1.4	0.9	0.4	0.8	1.4	1.1	4.7	5.8	11.3	5.7	4.4	6.1	6.3	10.6	11.0	13.6	5.0	6.0	9.7	9.3

(b) NICO++

Text						Image					Text + Image				
CLI	PAI	PHO	SKE	AVG	CLI	PAI	PHO	SKE	AVG	CLI	PAI	PHO	SKE	AVG	
0.4	0.7	0.4	0.8	0.6	2.4	10.9	8.2	7.2	7.2	4.4	9.9	14.5	9.6	9.6	
0.4	0.7	0.4	0.8	0.5	5.1	11.7	5.1	7.3	7.3	9.2	10.9	9.8	10.0	10.0	
0.4	0.7	0.4	0.8	0.6	6.1	3.2	3.8	4.4	4.4	11.3	7.4	9.0	9.2	9.2	
0.4	0.7	0.4	0.8	0.5	10.1	4.0	9.3	7.8	7.8	11.3	6.3	7.9	8.5	8.5	
0.4	0.7	0.4	0.8	0.6	7.1	3.2	10.6	5.7	6.7	10.6	6.1	9.6	11.1	9.3	

(c) MiniDomainNet

Text						Image					Text + Image				
CLI	PAI	PHO	SKE	AVG	CLI	PAI	PHO	SKE	AVG	CLI	PAI	PHO	SKE	AVG	
6.7	9.0	7.7	17.5	11.4	18.6	17.7	36.9	24.4	24.4	45.4	34.2	46.3	42.0	42.0	
11.3	12.9	6.2	15.7	9.5	12.4	12.4	28.0	17.5	17.5	25.8	30.4	38.8	31.7	31.7	
3.2	6.8	3.3	4.4	4.4	7.6	13.6	8.0	9.7	9.7	26.7	43.2	33.2	34.4	34.4	
7.1	9.6	5.7	19.1	10.4	13.0	19.0	12.7	35.6	20.1	27.6	45.0	32.6	43.9	37.3	

Table 9. Domain conversion evaluated by mAP (%) on three datasets. Comparison of FREEDOM with baselines and competitors. Across all methods, rows and columns represent the source and target domains, respectively. AVG: average mAP over respective source-target domain combinations. **Bold**: best, **magenta**: second-best.



Figure 4. *Top retrieval results* of FREEDOM. Domain conversion on ImageNet-R: (a) PHOTO → CARTOON; (b) PHOTO → ORIGAMI; (c) PHOTO → SCULPTURE; (d) PHOTO → TOY. **Orange**: image query; **green**: correctly retrieved; **red**: incorrectly retrieved.



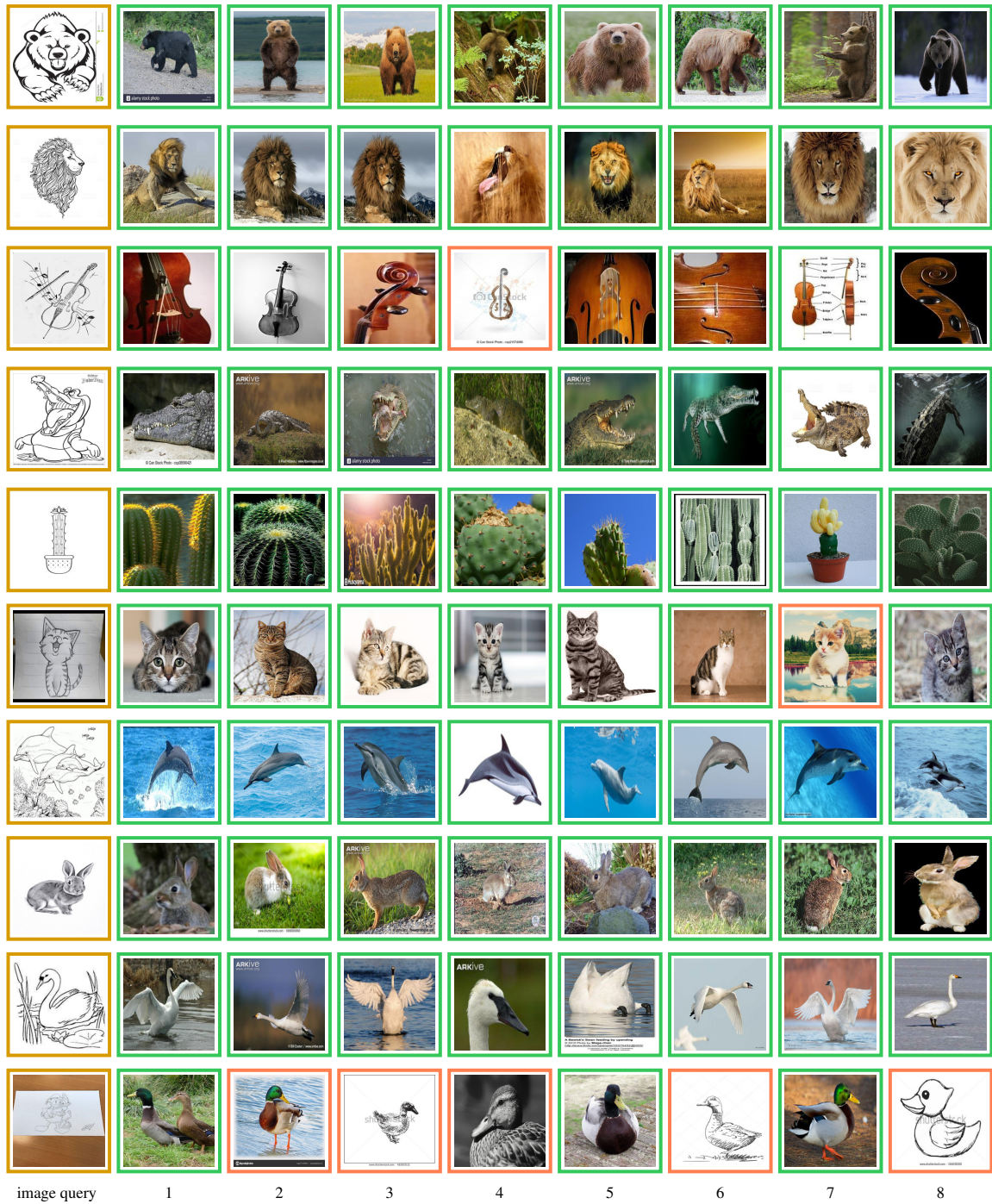


Figure 5. Top retrieval results of FREEDOM. Sketch-based image retrieval (SKETCH  $\rightarrow$  PHOTO) on MiniDomainNet. Orange: image query; green: correctly retrieved; red: incorrectly retrieved.



Figure 6. *Top retrieval results*. Competitors vs. FREEDOM. Domain conversion (ARCHIVE  $\rightarrow$  TODAY, TODAY  $\rightarrow$  ARCHIVE) on LTL. Orange: image query; green: correctly retrieved; red: incorrectly retrieved.





Figure 7. *Top retrieval results.* Competitors vs. FREEDOM. Domain conversion (AUTUMN  $\rightarrow$  DIMLIGHT, GRASS  $\rightarrow$  AUTUMN) on NICO++. Orange: image query; green: correctly retrieved; red: incorrectly retrieved.

## References

- [1] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *ICCV*, 2023. 1, 4
- [2] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *CVPR*, 2022. 4
- [3] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. ” O’Reilly Media, Inc.”, 2009. 3
- [4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 2
- [5] Ginger Delmas, Rafael Sampaio de Rezende, Gabriela Csurka, and Diane Larlus. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity. In *ICLR*, 2022. 4
- [6] Sounak Dey, Pau Riba, Anjan Dutta, Josep Lladós, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *CVPR*, 2019. 5
- [7] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*, 2019. 5
- [8] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoon Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *TMLR*, 2024. 1, 4
- [9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ICML*, 2023. 2
- [10] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2
- [11] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *ICCV*, 2021. 4
- [12] Bill Psomas, Ioannis Kakogeorgiou, Nikos Efthymiadis, Giorgos Tolias, Ondrej Chum, Yannis Avrithis, and Konstantinos Karantzas. Composed image retrieval for remote sensing. In *IGARSS*, 2024. 1
- [13] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *CVPR*, 2023. 1, 4, 5
- [14] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *NeurIPS Workshop*, 2021. 2, 3
- [15] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *CVPR*, 2021. 4
- [16] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, 2018. 5
- [17] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, 2023. 4
- [18] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhui Chen, Yu Su, and Ming-Wei Chang. MagicLens: Self-supervised image retrieval with open-ended instructions. In *ICML*, 2024. 1