SCOT: Self-Supervised Contrastive Pretraining For Zero-Shot Compositional Retrieval

Supplemental Material

In the following sections, we provide details about the composition function f_c along with additional quantitative and qualitative results.

6. Composition Function Architecture

As mentioned in Section 3, we use the *Combiner* architecture [3] as the composition operation f_c in our pretraining strategy. The Combiner architecture is illustrated visually and through pseudocode in Figs. 8 and 9 respectively. In brief, textual and visual features are first linearly projected, followed by a ReLU activation and then concatenated together. This concatenated representation is used in two ways: (i) it is transformed by another series of nonlinear projections to contribute directly to the output, and (ii) it is used to determine coefficients that control the influence of the original textual and visual features on the output.

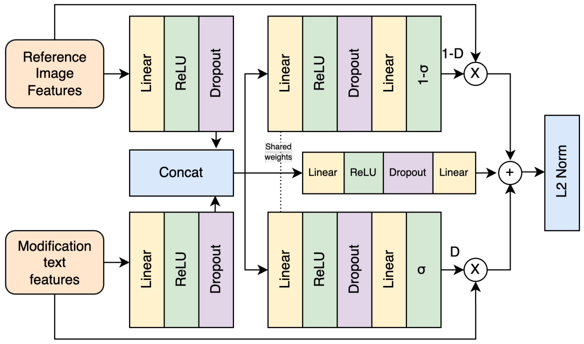

Figure 8. Overview of the Combiner architecture [3].

```
t_f  = Dropout(Relu(Linear(txt_feat)))
i_f  = Dropout(Relu(Linear(img_feat)))
c_f  = Concat((t_f, i_f), dim=-1)
F    = Linear(Dropout(Relu(Linear(c_f)))
    )
D    = Sigmoid(Linear(Dropout(ReLU(
    Linear(c_f)))))
out  = D*txt_feat + (1 - D)*img_feat + F
out  = L2_normalize(out, dim=-1)
```

Figure 9. Pytorch-like pseudocode for the Combiner architecture [3]. Here, D denotes the learned coefficient to dynamically modulate the contribution of image and text features in the output.

Specifically, the concatenated features are passed through learned projections with ReLU activations and a sigmoid output layer to learn a coefficient $D \in [0, 1]$. The final output is then determined as a weighted combination of the projected concatenated features along with the original textual features (weighted by D) and original visual features (weighted by $1 - D$). This ensures that the Combiner output remains in the same space as the contrastively-paired image and text encodings, while enabling dynamic control over the influence of each modality.

7. Results on CIRCO Dataset

In Table 4, we present results on CIRCO's test set [2]. The CIRCO dataset consists of real-world images from the COCO 2017 unlabelled image set. We note that this set of images used in CIRCO has no intersection with the MSCOCO images used in our pre-training strategy. Following [2], we present Recall@K and mAP metrics on CIRCO. The Recall@K metric evaluates using a single ground truth per sample whereas mAP is computed using multiple ground truths per test example. With respect to published approaches, SCOT performs significantly better, with our best performing model providing a 3.75% gain over SEARLE on Recall@5 and 1.56% gain on mAP@K. We also note that the choice of backbone has a large influence on the relative performance of SCOT and SEARLE. While SEARLE does marginally better than SCOT when both use CLIP B/32, when using CLIP L/14, SCOT performs on par with SEARLE. Most importantly, SCOT outperforms SEARLE by a margin of 2-3% in mAP@K when using the BLIP backbone. This is in line with our previous results and in-depth analysis on performance across backbones that was presented in Section 4.4.

We include the results on CIRCO in this appendix instead of the main paper due to space constraints. In the main paper, we present results on CIRR and FashionIQ which almost all prior works compare against, allowing for more points of comparison. While CIRCO involves more careful labeling than CIRR, the distribution of images in both datasets are similar as they are both open-world compositional retrieval datasets.

8. Extended Quantitative Comparison

The state-of-the-art comparison in Section 4 is predominantly confined to recent work that has been formally published. One reason for this is that recent preprints have not yet been peer reviewed, leaving their quantitative results subject to change. However, recognizing the importance of a broader comparison with concurrent work in composi-

Backbone	Method	Recall@K				mAP@K			
		$K = 5$	10	25	50	$K = 5$	10	25	50
CLIP B/32	Image-only	3.88	6.63	14.13	22.00	1.34	1.60	2.12	2.41
	Text-only	4.75	6.63	9.50	13.50	2.56	2.67	2.98	3.18
	Image+Text	8.25	14.13	25.50	34.75	2.65	3.25	4.14	4.54
	Captioning	10.25	14.33	21.38	29.00	5.48	5.77	6.44	6.85
	PALAVRA [5]	12.63	20.63	32.00	41.75	4.61	5.32	6.33	6.80
	SEARLE-OTI [2]	16.88	25.00	37.00	46.38	7.14	7.83	8.99	9.60
	SEARLE [2]	**19.75**	**28.00**	39.50	**50.63**	**9.35**	**9.94**	**11.13**	**11.84**
	SCOT (Ours)	17.25	25.75	**39.62**	50.38	7.58	8.24	9.46	10.14
CLIP L/14	Text Only	5.38	7.25	12.50	18.38	3.01	3.18	3.68	3.93
	Image Only	5.75	12.00	20.25	30.12	1.80	2.44	3.05	3.46
	Image+Text	14.50	21.75	36.12	46.00	3.92	4.79	5.93	6.48
	Pic2Word [29]	16.13	24.38	37.25	46.50	8.72	9.51	10.64	11.29
	SEARLE-XL-OTI [2]	22.75	32.00	45.13	**58.00**	10.18	11.03	12.72	13.67
	SEARLE-XL [2]	23.50	32.63	45.25	55.63	**11.68**	**12.73**	**14.33**	**15.12**
	SCOT (Ours)	**23.62**	**34.50**	**46.25**	56.25	10.74	11.95	13.62	14.46
BLIP	Text Only	9.12	13.50	20.50	27.50	5.35	5.74	6.49	6.87
	Image Only	6.38	11.88	20.50	29.25	1.87	2.46	3.20	3.36
	Image+Text	7.88	14.25	24.88	33.38	2.33	3.18	4.00	4.47
	Pic2Word† [29]	-	-	-	-	8.69	9.36	10.40	10.99
	SEARLE† [2]	-	-	-	-	10.65	11.34	12.40	13.02
	SCOT (Ours)	**26.75**	**38.62**	**54.5**	**64.88**	**12.45**	**13.58**	**15.41**	**16.25**
BLIP-2	Text Only	8.12	13.25	22.62	31.87	2.24	2.88	3.71	4.23
	Image Only	5.00	7.38	13.88	19.62	3.07	3.22	3.80	4.13
	Image+Text	21.62	33.62	47.88	63.50	8.39	9.66	11.31	12.22
	SCOT (Ours)	**27.25**	**37.88**	**54.12**	**64.88**	**13.24**	**14.24**	**16.05**	**17.05**

Table 4. **Results on CIRCO.** Zero-shot results on the CIRCO test set with Recall@K and mAP@K metrics. †Denotes results for Pic2Word and SEARLE with the BLIP backbone, which were taken from the results section of a recent preprint.[6]

tional image retrieval, we include an expanded comparison here, which covers both published work and unpublished preprints. Some very recent preprints were not included among the references in the main paper, corresponding to the approaches named LinCIR[4], CoVR[5], ISA[6] and GRB[7], whose results are included within Tables 5 and 6.

Retrieval metrics on the FashionIQ dataset [37] are presented in Table 5. The concurrent LinCIR approach achieves the best overall performance by a significant margin when using the CLIP G/14 backbone. At the same time, its performance using the CLIP L/14 backbone drops significantly, falling behind that of other approaches. Thus, much like SCOT, LinCIR seems to be dependent on the choice of backbone. LinCIR is an inversion-based approach, which uses a projection model ϕ to map image embeddings into text token embeddings.[8] Those token embeddings are then combined with the token embeddings corresponding to the input modification text, and then passed on to the text encoder, whose output will be a composed embedding that can be used for retrieval. LinCIR proposes an efficient procedure to train the projection model ϕ by using solely text embeddings. At training time, the model's input, which would normally be a reference image embedding, is instead replaced by a reference caption embedding, which results in their efficiency gains. By virtue of using caption embeddings as replacements for image embeddings, LinCIR becomes dependent on backbones with strong image-text alignment.

As observed in Tables 5 and 6, while LinCIR shows commendable performance on FashionIQ, its performance on CIRR is lower than SCOT BLIP B/16 and SCOT BLIP-2. Furthermore, it is noteworthy that LinCIR uses 5.5M captions, in contrast to the 290K captions used by SCOT. Similar to Pic2Word [29], LinCIR[4] is a textual inversion-based approach and does not allow fine-tuning backbones nor facilitate the potential for early fusion. In contrast, SCOT performance can be further improved by finetuning the backbones during training as well as performing early fusion of

[4] Gu, G., Chun, S., Kim, W., Kang, Y., Yun, S.: Language-only Efficient Training of Zero-shot Composed Image Retrieval. arXiv preprint arXiv:2312.01998 (2023).

[5] Ventura, L., Yang, A., Schmid, C., Varol, G.: CoVR: Learning composed video retrieval from web video captions. arXiv preprint arXiv:2308.14746 (2023).

[6] Du, Y., Wang, M., Zhou, W., Hui, S., Li, H.: Image2Sentence based Asymmetrical Zero-shot Composed Image Retrieval. arXiv preprint arXiv:2403.01431 (2024).

[7] Sun, S., Ye, F., Gong, S.: Training-free Zero-shot Composed Image Retrieval with Local Concept Reranking. arXiv preprint arXiv:2312.08924 (2023).

[8] See Section 2 for references to other inversion-based methods and related discussion.

Backbone	Method	Average		Dress		Shirt		Top/Tee	
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
CLIP B/32	Image-Only	5.88	13.19	6.96	14.08	4.46	11.89	6.22	13.61
	Text-Only	18.41	36.28	14.92	33.81	19.77	34.69	20.55	40.33
	Image+Text	13.36	27.51	12.44	28.55	12.61	24.82	15.04	29.16
	PALAVRA [5]	19.76	37.25	17.25	35.94	21.49	37.05	20.55	38.76
	SEARLE [2]	22.89	42.53	18.54	39.51	24.44	41.61	18.54	39.51
	SEARLE-OTI [2]	22.44	42.34	17.85	39.91	25.37	41.32	24.12	45.79
	TransAgg [23]	23.91	44.68	19.44	42.04	25.37	42.69	26.93	49.31
	TransAgg$_{\text{FT}}$ [23]	25.15	46.10	20.58	43.28	27.48	46.52	27.38	48.50
	*CIReVL [16]	28.29	49.35	25.29	46.36	28.36	47.84	31.21	53.85
	*Chen and Lai [4]	31.31	53.24	25.71	47.81	33.36	53.47	34.87	58.44
	SCOT (Ours)	24.14	43.44	19.73	41.24	25.51	42.93	27.18	46.14
CLIP L/14	Image-Only	7.97	17.43	5.25	13.63	10.54	20.65	8.10	18.01
	Text-Only	19.01	35.26	15.22	33.01	19.82	33.31	21.87	39.46
	Image+Text	18.12	33.17	14.27	31.33	19.13	32.28	20.95	35.90
	Pic2Word [29]	24.7	43.7	20.0	40.2	26.2	43.6	27.9	47.4
	SEARLE-XL [2]	25.56	46.23	20.48	43.13	26.89	45.58	29.32	49.97
	SEARLE-XL-OTI [2]	27.61	47.90	21.57	44.47	30.37	47.49	30.90	51.76
	TransAgg [23]	28.57	48.29	23.85	44.57	29.54	47.79	32.33	52.52
	TransAgg$_{\text{FT}}$ [23]	32.63	53.65	27.71	49.68	34.79	53.39	35.39	57.88
	*Context-I2W [33]	27.8	48.9	23.1	45.3	29.7	48.6	30.6	52.9
	*CIReVL [16]	28.55	48.57	24.79	44.76	29.49	47.40	31.36	53.65
	*Chen and Lai [4]	35.39	57.44	28.11	51.12	38.63	58.51	39.42	62.68
	*CompoDiff [12]	37.36	50.85	33.91	47.85	38.10	52.48	40.07	52.22
	*LinCIR[4]	26.28	46.49	20.92	42.44	29.10	46.81	28.81	50.18
	SCOT (Ours)	28.27	47.44	23.69	45.06	29.09	47.01	32.02	50.33
CLIP G/14	*CIReVL [16]	32.19	52.36	27.07	49.53	33.71	51.42	35.80	56.14
	*CompoDiff [12]	39.02	51.71	37.78	49.10	41.31	55.17	44.26	56.41
	*LinCIR[4]	**45.11**	**65.69**	**38.08**	**60.88**	**46.76**	**65.11**	**50.48**	**71.09**
BLIP B/16	Image-Only	6.65	15.40	5.05	12.19	7.55	17.76	7.34	16.26
	Text-Only	24.01	42.73	20.03	39.96	24.63	41.02	27.38	47.22
	Image+Text	8.06	18.16	6.14	19.78	9.37	19.87	8.66	19.78
	Pic2Word[6] [29]	25.99	46.00	21.35	42.68	27.51	46.01	29.12	49.33
	SEARLE[6] [2]	27.62	47.56	22.11	41.79	29.72	48.53	31.03	52.37
	TransAgg [23]	26.95	46.10	21.67	41.89	28.07	45.63	31.11	50.79
	TransAgg$_{\text{FT}}$ [23]	34.64	55.72	31.28	52.75	34.84	53.93	37.79	60.48
	*ISA[6]	29.79	49.19	24.69	43.88	30.79	50.05	33.91	53.65
	*ISA Eff-Net B2[6]	29.60	49.54	25.33	46.26	30.03	48.58	33.45	53.80
	*ISA Eff-ViT M2[6]	29.35	49.50	25.48	45.51	29.64	48.68	32.94	54.31
	SCOT (Ours)	30.68	51.33	26.42	49.23	30.91	49.65	34.72	55.12
BLIP L/16	*CoVR[5]	27.70	44.63	21.95	39.05	30.37	46.12	30.78	48.73
BLIP-2	Image-Only	7.53	17.93	4.21	11.89	10.59	23.51	7.81	18.41
	Text Only	24.68	43.59	20.77	41.64	25.95	42.83	27.33	46.31
	Image+Text	29.21	50.05	23.30	45.61	32.82	53.09	31.51	51.45
	*GRB[7]	30.74	51.44	24.14	45.56	34.54	55.15	33.55	53.60
	SCOT (Ours)	38.45	60.03	32.78	55.91	41.42	61.09	41.15	63.10

Table 5. **Expanded results on FashionIQ.** Zero-shot results on the FashionIQ validation set, including results from concurrent work. The **best**, second-best and third-best results are correspondingly highlighted. SCOT consistently achieves a top-3 result on every metric. Methods using CLIP G/14 adopt the OpenCLIP implementation. TransAgg$_{\text{FT}}$ denotes TransAgg with backbone fine-tuning. *Methods preceded by an asterisk are from recent preprints.

visual and textual features. We also wish to emphasize the complementary aspects of LinCIR and SCOT. The significant gains of LinCIR can be credited to their approach of incorporating random noise into textual embeddings during training, so that the resulting noisy text embedding distribution better resembles that of image embeddings. Given that SCOT employs textual embeddings as a substitute for visual embeddings in its training targets, applying LinCIR's noise addition strategy to our textual representations during training could potentially enhance ZS-CIR performance.

On FashionIQ (Table 5), LinCIR is followed by SCOT and CompoDiff [12]. While CompoDiff achieves strong results on Recall@10, SCOT significantly surpasses CompoDiff in Recall@50 metrics, both within categories and on average. Notably, the training procedure for CompoDiff utilizes a substantially larger dataset including 2 billion

Backbone	**Method**	**Recall@K**				**Recall$_{\text{subset}}$@K**		
		$K=1$	$K=5$	$K=10$	$K=50$	$K=1$	$K=2$	$K=3$
CLIP B/32	Image-only	6.94	22.94	33.71	59.18	21.06	41.01	60.34
	Text-only	21.16	45.35	57.40	81.06	62.26	81.08	90.75
	Image+Text	10.46	32.41	46.39	75.11	30.09	54.24	73.20
	PALAVRA [5]	16.62	43.49	58.51	83.95	41.61	65.30	80.94
	SEARLE [2]	24.00	53.42	66.82	89.78	54.89	76.60	88.19
	SEARLE-OTI [2]	24.27	53.25	66.10	88.84	54.10	75.81	87.33
	TransAgg [23]	24.46	53.61	67.54	89.81	57.81	78.17	89.54
	TransAgg$_{FT}$ [23]	29.30	60.48	73.25	92.31	63.57	82.31	91.95
	*CIReVL [16]	23.94	52.51	66.0	86.95	60.17	80.05	90.19
	*Chen and Lai [4]	18.80	46.07	60.75	86.41	44.29	68.10	83.42
	SCOT (Ours)	22.80	53.18	66.22	89.64	53.25	75.45	88.31
CLIP L/14	Image-only	7.47	23.88	34.07	57.57	20.87	41.95	61.13
	Text-only	22.00	45.79	57.57	79.59	61.71	80.26	90.43
	Image+Text	10.55	32.70	45.71	74.26	31.06	55.69	73.93
	Pic2Word [29]	23.9	51.7	65.3	87.8	-	-	-
	SEARLE-XL [2]	24.24	52.48	66.29	88.84	53.76	75.01	88.19
	SEARLE-XL-OTI [2]	24.87	52.31	66.29	88.58	53.80	74.31	86.94
	TransAgg [23]	25.04	53.98	67.59	88.94	55.33	76.82	88.94
	TransAgg$_{FT}$ [23]	33.04	64.39	76.27	**93.45**	63.37	82.27	92.22
	*Context-I2W [33]	25.6	55.1	68.5	89.8	-	-	-
	*CIReVL [16]	24.55	52.31	64.92	86.34	59.54	79.88	89.69
	*Chen and Lai [4]	25.52	54.58	67.59	88.70	55.64	77.54	89.47
	*CompoDiff [12]	19.37	53.81	72.02	90.85	59.13	78.81	89.33
	*LinCIR[4]	25.04	53.25	66.68	-	57.11	77.37	88.89
	SCOT (Ours)	24.36	53.52	67.37	89.35	51.47	74.24	87.90
CLIP G/14	*CIReVL [16]	34.65	64.29	75.06	91.66	67.95	84.87	93.21
	*CompoDiff [12]	26.71	55.14	74.52	92.01	64.54	82.39	91.81
	*LinCIR[4]	35.25	64.72	76.05	-	63.35	82.22	91.98
BLIP B/16	Image-only	7.23	25.78	37.35	62.34	20.60	40.96	61.35
	Text-only	34.19	61.68	71.74	87.83	72.34	87.97	94.79
	Image+Text	8.24	28.96	41.23	68.07	23.64	45.35	66.29
	Pic2Word[6] [29]	26.70	53.16	64.10	84.36	-	-	-
	SEARLE[6] [2]	29.27	54.86	66.57	86.16	-	-	-
	TransAgg [23]	34.89	64.75	76.24	92.22	66.34	83.76	92.92
	TransAgg$_{FT}$ [23]	37.18	**67.21**	**77.92**	93.43	69.34	85.68	93.62
	*ISA[6]	29.68	58.72	70.79	90.33	-	-	-
	*ISA Eff-Net B2[6]	30.84	61.06	73.57	92.43	-	-	-
	*ISA Eff-ViT M2[6]	29.63	58.99	71.37	91.47	-	-	-
	SCOT (Ours)	36.31	66.19	77.37	92.96	64.73	83.20	92.15
BLIP L/16	*CoVR	**38.48**	66.70	77.25	91.47	69.28	83.76	91.11
BLIP-2	Image-only	7.59	24.43	35.56	61.42	20.74	40.67	61.08
	Text-only	33.52	61.50	71.35	88.31	72.53	88.02	**94.87**
	Image+Text	19.69	49.98	64.39	90.01	45.69	71.18	85.83
	*GRB+LCR[7]	30.92	56.99	68.58	85.28	66.67	78.68	82.60
	SCOT (Ours)	36.82	64.34	74.48	93.42	**75.73**	**88.70**	94.84

Table 6. **Expanded results on CIRR.** Zero-shot results on the CIRR test set, including results from concurrent work. The **best**, second-best and third-best results are correspondingly highlighted. SCOT is the only approach that consistently achieves a top-3 result on every metric. Methods using CLIP G/14 adopt the OpenCLIP implementation. TransAgg$_{FT}$ denotes TransAgg with backbone fine-tuning. *Methods preceded by an asterisk are from recent preprints.

captioned images from LAION and 18 million synthetically generated triplets; in contrast, SCOT is trained on only 290K samples. The synthetic image-text-image triplets generated by CompoDiff's approach are complementary to our proposed use of image-text-text triplets, and we hypothesize that a combination of both pretraining schemes may yield stronger results than each.

Table 6 presents different recall metrics over the CIRR dataset [24]. We observe that in subset evaluations, SCOT significantly outperforms all prior and concurrent ZS-CIR methods, while being competitive with the best *supervised* approach from Table 2. On the full set recall metrics, TransAgg$_{FT}$ presents the best results among ZS-CIR methods. This version of TransAgg fine-tunes its BLIP backbone during training. Note that when using the same backbone, SCOT presents better performance than the version of TransAgg which does not fine-tune its backbone. Thus, it would be reasonable to expect that SCOT with fine-tuned backbones would outperform TransAgg$_{FT}$, resulting in state-of-the-art performance on the full set recall metrics as well.

9. Ablations of Encoder Backbones

In Section 4.4, an ablation was conducted to examine the effect of varying the contrastively-trained encoders on zero-shot compositional retrieval performance on the FashionIQ and CIRR datasets. Here, we present additional observations around the performance of SCOT relative to the simple baselines (Image-Only, Text-Only, and Image+Text) as the backbones are changed.

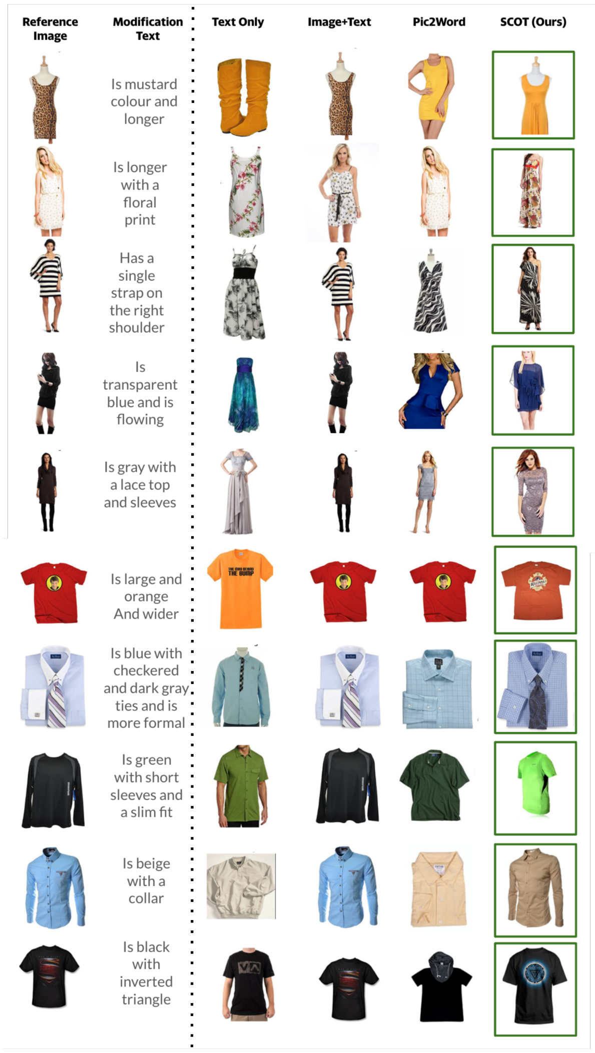
Reference Image
Modification Text
Text Only
Image+Text
Pic2Word
SCOT (Ours)
Is mustard colour and longer
Is longer with a floral print
Has a single strap on the right shoulder
Is transparent blue and is flowing
Is gray with a lace top and sleeves
Is large and orange And wider
Is blue with checkered and dark gray ties and is more formal
Is green with short sleeves and a slim fit
Is beige with a collar
Is black with inverted triangle

Figure 10. **Additional qualitative retrieval results on FashionIQ [37] validation set.** A green box indicates the correctly retrieved image. Here we present the closest (Rank 1) matches for all methods.

Table 1 provides comprehensive results on the FashionIQ dataset when the image-text encoder backbone is varied, including retrieval performance across the Dress, Shirt, and Top/Tee categories. We observe that performance is fairly correlated across the categories, with BLIP-2 encoders improving dramatically over CLIP and BLIP. Interestingly, we note that Text-Only performance significantly exceeds that of Image+Text for both CLIP B/32 and BLIP encoders, while the gap shrinks for the larger CLIP L/14 model and reverses with BLIP-2. We hypothesize that the performance of Text-Only relative to Image+Text may reflect the quality of embeddings produced by pretrained image-text encoders for a given dataset.

Table 2 contains results on the CIRR dataset, also with varied image-text encoders. We observe that, in subset evaluations on CIRR, when models use CLIP B/32, they demonstrate marginally better performance than when using CLIP L/14; however, this gets reversed in full-set evaluations. The subset evaluations are also significantly dominated by Text-Only retrieval for the CLIP variants and for BLIP, unlike what happens in the full-set evaluations. Some CIRR triplets are known to have target images which are fully described by the modification text—rendering the reference image irrelevant—which partly explains these results, as also previously observed by Baldrati et al. [2]. Using the stronger BLIP-2 encoder improves the performance of the Text-Only retrieval baseline, while at the same time leading to larger gains when using SCOT.

10. Additional Qualitative Results

Qualitative retrieval results on samples from FashionIQ are presented in Fig. 10. The closest (Rank 1) match is shown for all the presented methods. The examples highlight the central challenge of compositional retrieval, namely the merging of visual information from the original image with the modification text to retrieve the relevant target image. For example, in row one, text-only retrieval yields mustard-colored long shoes, the simple image+text method retrieves the original image, and Pic2Word retrieves a similar yet shorter mustard-colored dress. In contrast, SCOT successfully retrieves a long mustard-colored dress with a neckline similar to that of the original image.

In Fig. 11, we present some Recall@1 failure cases where SCOT did not retrieve the ground truth image as the closest match. We present the top four samples retrieved by SCOT and highlight the annotated ground truth sample. We often see that the top result retrieved by SCOT is *also* a relevant product that matches the reference image with the described modification applied. However, because the annotations are not exhaustive and only one image is labeled as the ground truth, the retrieved result is not judged as correct. This can be clearly observed in the second row: whereas the ground-truth image featuring a black t-shirt with a Godzilla

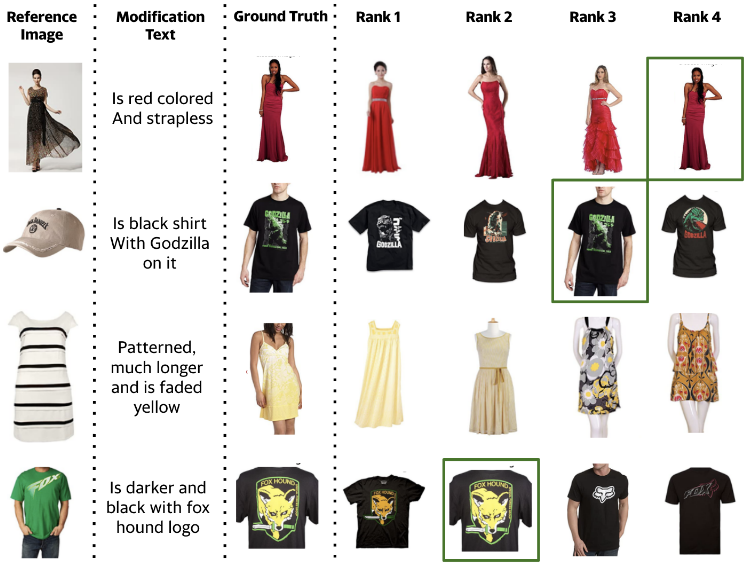
Reference Image
Modification Text
Ground Truth
Rank 1
Rank 2
Rank 3
Rank 4
Is red colored And strapless
Is black shirt With Godzilla on it
Patterned, much longer and is faded yellow
Is darker and black with fox hound logo

Figure 11. **Additional qualitative retrieval results on the FashionIQ [37] validation set.** Samples where SCOT failed to retrieve the annotated ground truth (green box) as the closest match.

graphic shows up in the third position, the first two retrieved images are *also* black t-shirts with Godzilla graphics on them and could thus also have reasonably been judged as correct retrievals.

11. Visualization of Composed Embeddings

In Section 3, we present a method for learning a composed image embedding $\mathcal{V}_c$ which is utilized for retrieval at inference time. In Fig. 12, we attempt to visualize these latent composed image embeddings $\mathcal{V}_c$ in visual space by using the CLIP L/14 backbone along with unCLIP.[9]

The visualization illustrates how SCOT successfully retains the essential elements of the original image while applying the necessary modifications to generate the composed representation. The second row of the visualization provides a notable example: the modification text explicitly requests the replacement of birds for rabbits, and the resulting composed image contains not only rabbits but also additional elements—a man and cages–which were not in the modification text. These elements appear to be inferred from the original image and then incorporated by the model into the resulting composed embedding.

Similarly, in the first row the blue and white costume was likely drawn from "Alice costume" in the modification text, but the presence of multiple women and the dancing poses were likely inferred from the original image. In the fourth row, the deer were mentioned in the modification text

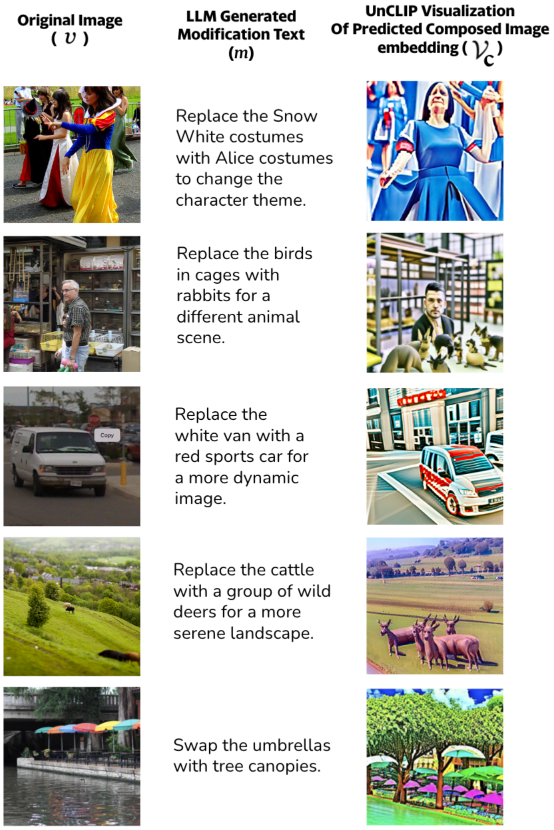
Original Image (v)
LLM Generated Modification Text (m)
UnCLIP Visualization Of Predicted Composed Image embedding ($\mathcal{V}_c$)
Replace the Snow White costumes with Alice costumes to change the character theme.
Replace the birds in cages with rabbits for a different animal scene.
Replace the white van with a red sports car for a more dynamic image.
Replace the cattle with a group of wild deers for a more serene landscape.
Swap the umbrellas with tree canopies.

Figure 12. **UnCLIP visualizations** of the composed image embedding $\mathcal{V}_c$ (using CLIP ViT-L/14) for selected samples.

[9] We use the karlo-v1-alpha checkpoint for unCLIP, which reimplements Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022).

but the wide shot of grassland scenery was likely inferred from the image. The orientation of the vehicle and tables in the third and fifth images respectively appear to have been drawn from the images as well. These examples underscore the model's ability to preserve contextually relevant information while making specific alterations as per the given modification text.