

---

# Supplementary for Generalized Contrastive Learning for Universal Multimodal Retrieval

---

Jungsoo Lee Janghoon Cho Hyojin Park Munawar Hayat  
Kyuwoong Hwang Fatih Porikli Sungha Choi<sup>†</sup>

Qualcomm AI Research\*

{jungsool, janghoon, hyojinp, hayat, kyuwoong, fporikli, sunghac}@qti.qualcomm.com

## A Further Implementation Details

In addition to the implementation details provided in the main paper, we provide further implementation details. For finetuning, we utilize the GitHub repository of OpenCLIP<sup>2</sup>, using Automatic Mixed Precision BFloat16. The parameters being updated differ between models; for VISTA [1], the entire set of parameters is updated, whereas for CLIP-SF [2, 3] and TinyCLIP-SF [4], only the visual encoder parameters are updated. We utilize the cosine learning rate scheduler with 500 warmup steps and use the AdamW [5] optimizer, configured with  $\beta_1 = 0.9$  and  $\beta_2 = 0.95$ . All three models are finetuned for 5 epochs with a learning rate of  $5e-6$ . We use batch size of 128 for VISTA and TinyCLIP-SF and 32 for CLIP-SF. All finetuning experiments are conducted using 8 GPUs on a single node of A100. Regarding the evaluation, while UniIR leverages explicit task instructions during training, we do not utilize such instructions, in order to demonstrate the generalizability of our method across diverse tasks without task-specific instruction fine-tuning.

## B Pytorch-like pseudocode of GCL

Algorithm 1 further explains GCL with the Pytorch-like pseudocode to help understanding of our proposed GCL loss function. The codes of GCL are currently under internal legal review. We are planning to release the codes of GCL after the legal review is finalized.

## C GCL with Triplet-based Datasets

In the main paper, we demonstrated that training VISTA using GCL with image-text paired datasets yields superior performance compared to training VISTA with synthetically generated triplet-based datasets, which are specifically designed for certain targeted retrieval scenarios. Given the public availability of the triplet-based datasets, we further conducted a comparative analysis involving joint training of VISTA using both image-text paired dataset with GCL and the triplet-based datasets with a standard contrastive loss.

Tables 1 and 2 show the experiment results on M-BEIR under the global and local setting, respectively. As shown, additionally utilizing triplet-based datasets enhances performance in targeted retrieval tasks, specifically for the scenarios  $q_{it} \rightarrow c_i$  and  $q_t \rightarrow c_{it}$ . However, this joint training approach leads to a decline in performance across other retrieval tasks, resulting in an overall degradation in retrieval performance, a trend that was also observed in the main paper. We want to emphasize that this jointly trained VISTA still outperforms VISTA trained only with triplet-based dataset. These findings again highlight the effectiveness of leveraging off-the-shelf image-caption paired datasets with GCL for improving general multimodal retrieval performance.

---

\*Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc. <sup>†</sup> indicates corresponding author.

<sup>2</sup>[https://github.com/mlfoundations/open\\_clip](https://github.com/mlfoundations/open_clip)

## D M-BEIR Results with Diverse Ranks

Following UniIR [3], we also present the results of M-BEIR at various ranks-1, 5, 10, 20, 50-in Tables 3 and 4, corresponding to the global and local settings, respectively. The results were obtained from a separate run, independent of the experiment in the main paper, so minor differences at the decimal point level may exist.

## E Further Results of Ablation Study

Due to the page limit, we only reported the ablation study and comparison with intra-modality separation loss using the global setting of M-BEIR. Table 5 shows these results under the local setting of M-BEIR. Additionally, we show the performance variance across 3 multiple runs for each loss term in GCL in Table 6. As shown, we observe a similar tendency to that in the original manuscript, indicating that each tuple of loss term contributes meaningfully to the modality pairs.

## F Further Results of TinyCLIP-SF

In the main paper, we compared TinyCLIP-SF with VISTA and CLIP-SF to demonstrate its improved multimodal retrieval performance through GCL training, despite having fewer parameters and faster inference speed. Table 7 compares TinyCLIP-SF trained with GCL to both the pretrained model and the one fine-tuned with standard contrastive learning on M-BEIR under the global setting. As shown, training TinyCLIP-SF with GCL loss outperforms both models.

## G Analysis on Quantitative Evaluation

While CLIP has more number of parameters (427M) compared to VISTA (196M), we found that performance gains on multimodal retrieval are relatively smaller for CLIP-SF compared to VISTA. The main reason is that the architecture of VISTA is modified so that image and text are input together to extract embeddings using a single text encoder. On the other hand, CLIP-SF simply adds the extracted embeddings of images and texts, which limits creating a unified embedding space of diverse modalities. Due to this fact, the performance drop from the local setting to the global setting is more severe for CLIP-SF compared to VISTA, making it challenging to utilize CLIP-based models for multimodal retrieval tasks. Considering such an aspect, we believe that improving multimodal retrieval performance for CLIP-based models would be an interesting future research topic.

## H Broader impacts

The proposed Generalized Contrastive Learning (GCL) loss function offers several positive societal impacts. Firstly, it significantly enhances the performance of multimodal retrieval systems, making it easier for users to access relevant information across different modalities. For example, professionals fields such as healthcare, law, and research can leverage these improved systems to quickly find relevant information, combining text and images. This can lead to better patient outcomes, more informed decision-making, and accelerated research progress. Furthermore, the ability to retrieve mixed modality content seamlessly can enhance user experience on various platforms, from search engines to social media, improving satisfaction and engagement.

However, there are also negative societal impacts to consider. Information from different modalities may be wrongly combined. For instance, given that the image-text pairs are wrongly paired, a retrieval system might incorrectly pair an image of a historical monument with a text describing a completely different monument’s history. This misalignment can lead to misinformation, confusing users and potentially spreading false narratives. Such errors could be particularly harmful in educational contexts, where accurate information is crucial for learning. While the training and test sets used in academic purposes do not consider such cases, we must take into account of such a point when deploying multimodal retrieval models in the real-world applications.

---

**Algorithm 1** Pseudocode of GCL in a PyTorch-like style

---

```
for (x_i, x_t) in data_loader: # load i2t pairs
    # encode image, text, fused modality
    e_i = model.encode_image(x_i)
    e_t = model.encode_text(x_t)
    e_it = model.encode_fused((x_i, x_t))

    # similarity map for image
    i2i = sim(e_i, e_i)
    i2t = sim(e_i, e_t)
    i2it = sim(e_i, e_it)
    i_neg = concat(i2i, i2t, i2it)
    i_neg.mask_out_positives()

    # similarity map for text
    t2i = sim(e_t, e_i)
    t2t = sim(e_t, e_t)
    t2it = sim(e_t, e_it)
    t_neg = concat(t2i, t2t, t2it)
    t_neg.mask_out_positives()

    # similarity map for fused modality
    it2i = sim(e_it, e_i),
    it2t = sim(e_it, e_t)
    it2it = sim(e_it, e_it)
    it_neg = concat(it2i, it2t, it2it)
    it_neg.mask_out_positives()

    batch_size = len(x_i)
    target = torch.arange(batch_size)
    loss = 0

    # loss for i2t, i2it
    i2t_pos = i2t.mask_out_negatives()
    i2t_logits = i2t_pos + i_neg
    i2it_pos = i2it.mask_out_negatives()
    i2it_logits = i2it_pos + i_neg
    loss += ce_loss(i2t_logits, target)
    loss += ce_loss(i2it_logits, target)

    # loss for t2i, t2it
    t2i_pos = t2i.mask_out_negatives()
    t2i_logits = t2i_pos + t_neg
    t2it_pos = t2it.mask_out_negatives()
    t2it_logits = t2it_pos + t_neg
    loss += ce_loss(t2i_logits, target)
    loss += ce_loss(t2it_logits, target)

    # loss for it2i, it2t
    it2i_pos = it2i.mask_out_negatives()
    it2i_logits = it2i_pos + it_neg
    it2t_pos = it2t.mask_out_negatives()
    it2t_logits = it2t_pos + it_neg
    loss += ce_loss(it2i_logits, target)
    loss += ce_loss(it2t_logits, target)

    loss.backward()
```

---

Table 1: Comparisons on global setting of M-BEIR using VISTA [1] trained with diverse settings. GCL + Pairwise & Triplet indicates VISTA trained with both image-text paired dataset using GCL and triplet dataset using a standard contrastive learning.

Task	Dataset	Pretrained	CL +Triplet	CL +Pairwise	GCL + Pairwise & Triplet	<b>GCL (Ours) +Pairwise</b>
1. $q_t \rightarrow c_i$	VisualNews [6]	5.36	1.64	9.29	13.92	16.64
	MSCOCO [7]	2.72	5.60	14.42	35.02	38.85
	Fashion200K [8]	0.00	0.00	0.00	3.03	4.25
2. $q_t \rightarrow c_t$	WebQA [9]	97.07	96.90	96.86	96.25	96.25
3. $q_t \rightarrow (c_i, c_t)$	EDIS [10]	25.15	44.37	36.90	50.63	49.06
	WebQA [9]	14.22	80.88	31.74	78.97	64.00
4. $q_i \rightarrow c_t$	VisualNews [6]	1.35	0.08	1.18	4.30	4.71
	MSCOCO [7]	12.90	0.50	26.82	52.63	60.32
	Fashion200K [8]	0.02	0.00	0.00	0.61	0.72
5. $q_i \rightarrow c_i$	NIGHTS [11]	76.60	83.07	79.39	83.68	82.5
6. $(q_i, q_t) \rightarrow c_t$	OVEN [12]	5.06	1.78	3.10	9.24	8.72
	InfoSeek [13]	2.94	4.80	1.70	11.43	9.07
7. $(q_i, q_t) \rightarrow c_i$	FashionIQ [14]	6.66	16.41	6.10	12.41	10.88
	CIRR [15]	23.62	43.81	24.27	35.95	31.13
8. $(q_i, q_t) \rightarrow (c_i, c_t)$	OVEN [12]	34.31	9.67	32.83	24.88	32.92
	InfoSeek [13]	30.95	14.94	29.82	29.51	34.97
	Avg.	21.18	25.28	24.65	33.89	<b>34.06</b>

Table 2: Comparisons on local setting of M-BEIR using VISTA [1] trained with diverse settings. Abbreviations as in Table 1.

Task	Dataset	Pretrained	CL +Triplet	CL +Pairwise	GCL + Pairwise & Triplet	<b>GCL (Ours) +Pairwise</b>
1. $q_t \rightarrow c_i$	VisualNews [6]	16.04	10.01	15.78	13.98	15.42
	MSCOCO [7]	50.65	58.40	61.34	61.05	61.09
	Fashion200K [8]	9.31	8.03	9.83	8.67	9.54
2. $q_t \rightarrow c_t$	WebQA [9]	91.20	91.20	90.43	89.12	89.37
3. $q_t \rightarrow (c_i, c_t)$	EDIS [10]	36.69	40.98	35.76	44.62	45.88
	WebQA [9]	33.49	74.51	36.16	70.57	62.49
4. $q_i \rightarrow c_t$	VisualNews [6]	14.03	4.42	13.35	12.29	13.70
	MSCOCO [7]	61.66	60.44	71.98	71.18	72.56
	Fashion200K [8]	9.63	6.71	9.29	9.04	9.31
5. $q_i \rightarrow c_i$	NIGHTS [11]	26.32	26.32	28.21	28.49	28.35
6. $(q_i, q_t) \rightarrow c_t$	OVEN [12]	30.39	25.93	29.91	30.65	31.82
	InfoSeek [13]	29.87	23.16	28.47	32.13	34.26
7. $(q_i, q_t) \rightarrow c_i$	FashionIQ [14]	2.43	9.03	2.25	5.41	5.00
	CIRR [15]	10.60	21.82	11.34	16.81	14.27
8. $(q_i, q_t) \rightarrow (c_i, c_t)$	OVEN [12]	37.45	31.11	35.84	38.04	40.60
	InfoSeek [13]	23.08	28.34	23.94	35.31	35.32
	Avg.	30.18	32.53	31.49	35.46	<b>35.56</b>

Table 3: Comparisons on global setting of M-BEIR using ranks of 1, 5, 10, 20, and 50.

Task	Dataset	Metric	VISTA				CLIP-SF		
			Pretrained	CL +Triplet	CL +Pairwise	GCL (Ours) +Pairwise	Pretrained	CL +Pairwise	GCL (Ours) +Pairwise
1. $q_t \rightarrow c_t$	VisualNews	R@1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		R@5	0.57	0.14	1.27	3.41	0.01	0.00	0.98
		R@10	1.27	0.35	2.47	6.09	0.01	0.00	1.83
		R@20	2.42	0.72	4.45	9.86	0.02	0.00	3.33
		R@50	5.36	1.64	8.85	16.38	0.08	0.00	6.70
	MSCOCO	R@1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		R@5	0.09	0.09	0.84	7.31	0.00	0.00	0.17
		R@10	0.43	0.55	3.28	15.55	0.00	0.00	0.56
		R@20	1.08	1.66	6.83	25.02	0.00	0.00	1.38
		R@50	2.72	5.60	15.63	39.13	0.00	0.00	3.25
	Fashion200K	R@1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		R@5	0.00	0.00	0.00	0.29	0.00	0.00	0.00
		R@10	0.00	0.00	0.00	0.52	0.00	0.00	0.00
		R@20	0.00	0.00	0.00	1.57	0.00	0.00	0.00
		R@50	0.00	0.00	0.06	4.13	0.00	0.00	0.00
2. $q_t \rightarrow c_t$	WebQA	R@1	62.81	61.59	62.24	60.77	20.45	37.84	20.41
		R@5	88.31	87.09	87.41	86.03	37.80	65.58	37.76
		R@10	92.34	91.69	91.49	90.92	45.66	75.68	45.62
		R@20	95.19	94.46	94.62	94.46	52.26	82.28	52.18
		R@50	97.07	96.90	96.90	96.37	60.29	88.55	60.24
3. $q_t \rightarrow (c_i, c_t)$	EDIS	R@1	0.96	6.39	3.33	4.97	1.36	4.01	6.73
		R@5	4.07	20.52	12.90	18.64	6.48	13.42	22.31
		R@10	8.33	27.21	21.01	26.54	9.66	18.94	31.66
		R@20	14.41	34.31	30.52	35.85	14.35	25.36	42.21
		R@50	25.15	44.37	43.84	49.21	23.39	34.19	54.43
	WebQA	R@1	0.44	30.90	5.89	13.46	1.79	14.30	7.33
		R@5	2.23	55.71	18.08	31.70	5.58	33.41	17.20
		R@10	4.30	64.99	26.36	41.42	7.89	44.13	22.90
		R@20	7.53	72.32	35.28	50.82	12.47	56.07	29.51
		R@50	14.22	80.88	46.67	63.12	19.87	68.42	40.62
4. $q_i \rightarrow c_t$	VisualNews	R@1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		R@5	0.33	0.01	0.29	0.80	0.00	0.00	0.24
		R@10	0.53	0.04	0.46	1.46	0.00	0.00	0.56
		R@20	0.77	0.05	0.68	2.44	0.00	0.00	1.08
		R@50	1.35	0.08	1.21	4.69	0.00	0.00	2.48
	MSCOCO	R@1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		R@5	2.36	0.04	5.64	22.90	0.00	0.00	5.08
		R@10	3.70	0.08	9.78	32.66	0.00	0.00	9.20
		R@20	6.18	0.22	15.70	43.16	0.00	0.00	14.80
		R@50	12.90	0.50	27.24	60.12	0.00	0.00	24.84
	Fashion200K	R@1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		R@5	0.00	0.00	0.00	0.02	0.00	0.00	0.02
		R@10	0.00	0.00	0.00	0.14	0.00	0.00	0.04
		R@20	0.02	0.00	0.00	0.29	0.00	0.00	0.08
		R@50	0.02	0.00	0.06	0.65	0.00	0.00	0.16
5. $q_i \rightarrow c_i$	NIGHTS	R@1	7.03	6.32	7.26	8.02	6.60	8.82	7.74
		R@5	23.73	25.09	27.22	26.60	25.57	30.47	29.62
		R@10	40.00	41.51	42.31	42.41	42.69	49.76	48.21
		R@20	57.74	62.74	62.26	63.35	62.64	71.51	67.83
		R@50	76.6	83.07	80.38	82.74	81.65	88.07	85.09
6. $(q_i, q_t) \rightarrow c_t$	OVEN	R@1	0.23	0.12	0.19	0.41	0.00	0.00	0.05
		R@5	0.71	0.38	0.68	1.45	0.00	0.00	0.35
		R@10	1.26	0.53	1.04	2.37	0.00	0.00	0.79
		R@20	2.17	0.86	1.71	4.00	0.00	0.00	1.66
		R@50	5.06	1.78	3.92	8.49	0.00	0.00	3.63
	InfoSeek	R@1	0.04	0.01	0.07	0.19	0.00	0.00	0.01
		R@5	0.18	0.49	0.23	0.72	0.00	0.00	0.22
		R@10	0.40	1.10	0.34	1.44	0.00	0.00	0.40
		R@20	0.93	2.08	0.83	3.74	0.00	0.00	0.70
		R@50	2.94	4.80	2.74	8.70	0.00	0.00	1.86
7. $(q_i, q_t) \rightarrow c_i$	FashionIQ	R@1	0.10	0.85	0.10	0.13	0.03	0.00	0.03
		R@5	1.30	5.58	1.12	2.52	2.55	0.00	1.12
		R@10	2.30	8.16	2.22	4.08	4.46	0.00	1.98
		R@20	3.68	11.38	4.11	6.55	6.70	0.00	2.90
		R@50	6.66	16.41	6.73	10.89	11.61	0.00	4.25
	CIRR	R@1	0.82	2.04	0.82	0.82	0.62	0.05	0.65
		R@5	7.99	16.98	9.21	10.91	5.42	0.17	6.64
		R@10	11.61	23.72	13.14	15.73	8.13	0.31	10.05
		R@20	16.07	32.11	17.91	21.87	11.97	0.34	14.63
		R@50	23.62	43.81	25.80	30.86	18.06	0.43	21.25
8. $(q_i, q_t) \rightarrow (c_i, c_t)$	OVEN	R@1	4.25	0.38	4.08	2.34	0.71	0.00	1.91
		R@5	11.97	1.27	11.33	8.24	2.73	0.09	5.79
		R@10	17.58	2.32	16.97	13.53	4.39	0.16	8.72
		R@20	24.39	4.62	23.96	20.83	6.89	0.26	12.75
		R@50	34.31	9.67	33.67	32.83	11.04	0.58	19.47
	InfoSeek	R@1	1.64	0.63	1.76	1.68	0.59	0.00	1.23
		R@5	6.64	2.68	7.32	7.08	2.65	0.00	4.82
		R@10	11.98	4.70	13.23	12.82	4.48	0.00	8.59
		R@20	19.30	8.16	20.83	21.24	7.34	0.00	13.61
		R@50	30.95	14.94	32.38	34.61	12.73	0.00	21.89

Table 4: Comparisons on local setting of M-BEIR using ranks of 1, 5, 10, 20, and 50.

Task	Dataset	Metric	VISTA				CLIP-SF		
			Pretrained	CL +Triplet	CL +Pairwise	<b>GCL (Ours) +Pairwise</b>	Pretrained	CL +Pairwise	<b>GCL (Ours) +Pairwise</b>
1. $q_t \rightarrow c_i$	VisualNews	R@1	7.19	4.19	6.73	6.64	23.80	9.19	18.90
		R@5	16.04	10.01	15.34	15.09	44.34	20.97	36.71
		R@10	21.28	13.82	20.33	20.13	52.84	28.11	44.63
		R@20	27.70	18.35	25.86	25.87	60.77	35.92	52.02
		R@50	36.99	25.71	35.01	34.9	70.36	46.86	61.62
	MSCOCO	R@1	27.61	33.63	36.71	36.26	36.45	43.77	42.92
		R@5	50.65	58.40	62.30	61.26	61.09	71.94	67.69
		R@10	60.95	68.71	72.52	71.66	71.32	81.76	76.78
		R@20	71.48	78.65	81.62	81.04	80.65	89.55	85.10
		R@50	84.63	89.81	91.80	91.56	91.58	96.65	93.80
	Fashion200K	R@1	2.27	2.09	2.39	2.15	1.98	1.51	1.86
		R@5	6.52	5.64	6.86	6.46	4.71	4.94	4.25
		R@10	9.31	8.03	9.37	9.37	6.57	8.84	7.04
		R@20	14.08	10.94	13.44	14.19	10.47	12.97	10.06
		R@50	21.70	17.86	20.54	20.71	17.86	20.94	16.58
2. $q_t \rightarrow c_t$	WebQA	R@1	67.86	67.86	66.88	65.58	22.36	41.67	22.36
		R@5	91.20	91.20	90.22	89.25	40.61	70.35	40.61
		R@10	94.34	94.34	94.30	93.20	48.51	79.63	48.51
		R@20	96.99	96.99	96.66	96.09	54.75	85.74	54.75
		R@50	98.17	98.17	98.17	97.84	62.53	91.24	62.53
3. $q_t \rightarrow (c_i, c_t)$	EDIS	R@1	16.78	20.95	20.52	23.20	21.20	18.05	24.71
		R@5	36.69	40.98	41.72	45.60	43.29	34.56	48.97
		R@10	45.97	48.60	50.48	54.89	52.48	41.22	59.02
		R@20	54.49	54.86	58.32	63.04	61.00	47.89	68.37
		R@50	66.43	63.28	69.73	73.09	71.77	55.32	79.02
	WebQA	R@1	16.33	47.39	26.24	35.17	24.09	43.61	23.54
		R@5	33.49	74.51	49.26	61.77	45.48	69.97	44.01
		R@10	41.66	82.04	57.87	70.89	54.24	79.41	53.17
		R@20	51.57	87.73	67.18	78.45	62.37	86.50	62.37
		R@50	63.16	92.99	77.26	86.42	71.68	91.72	72.68
4. $q_i \rightarrow c_t$	VisualNews	R@1	5.73	1.60	5.29	5.73	22.21	8.76	15.88
		R@5	14.03	4.42	13.00	13.66	41.78	20.18	30.53
		R@10	18.95	6.48	17.72	18.47	50.20	26.31	37.18
		R@20	25.06	9.16	23.14	24.24	58.38	33.39	44.79
		R@50	34.27	14.50	31.72	32.76	68.04	43.90	54.19
	MSCOCO	R@1	37.26	34.14	47.58	47.48	55.86	62.58	56.50
		R@5	61.66	60.44	72.12	71.92	79.00	85.78	79.04
		R@10	72.58	71.58	81.76	81.30	86.60	91.36	86.72
		R@20	82.36	81.66	89.58	89.14	92.22	95.90	92.04
		R@50	91.98	91.82	95.44	95.64	97.26	99.18	97.20
	Fashion200K	R@1	2.09	1.49	1.80	2.05	1.74	1.72	1.78
		R@5	6.57	4.38	5.79	5.91	4.97	5.67	5.48
		R@10	9.63	6.71	8.80	9.06	7.71	8.65	8.55
		R@20	13.79	10.15	12.66	13.50	11.84	12.85	12.40
		R@50	21.03	16.38	20.43	21.23	19.62	21.93	19.37
5. $q_i \rightarrow c_i$	NIGHTS	R@1	7.45	6.56	7.55	8.44	6.75	8.82	8.02
		R@5	26.32	26.32	29.39	28.35	26.13	30.94	30.99
		R@10	45.80	44.58	47.88	46.70	43.49	51.23	50.66
		R@20	68.44	67.17	71.75	70.99	64.34	74.20	73.58
		R@50	88.40	87.88	90.66	90.52	83.54	90.80	90.85
6. $(q_i, q_t) \rightarrow c_t$	OVEN	R@1	15.88	12.27	15.72	16.37	0.06	0.03	3.89
		R@5	30.39	25.93	30.39	31.40	0.31	0.23	8.93
		R@10	36.43	32.63	36.80	38.07	0.54	0.44	11.52
		R@20	42.57	39.48	43.23	44.80	0.87	0.71	14.34
		R@50	50.77	48.86	51.51	53.59	1.50	1.41	18.74
	InfoSeek	R@1	15.39	10.73	15.47	17.11	0.05	0.00	2.61
		R@5	29.87	23.16	31.03	33.83	0.29	0.00	6.78
		R@10	37.31	30.46	38.88	42.15	0.65	0.23	9.49
		R@20	45.40	37.99	46.99	50.16	1.13	0.29	12.87
		R@50	55.80	47.81	57.31	60.51	2.74	0.43	18.36
7. $(q_i, q_t) \rightarrow c_i$	FashionIQ	R@1	0.07	0.82	0.07	0.10	0.07	3.60	0.07
		R@5	1.35	6.16	1.23	2.73	4.21	8.21	3.03
		R@10	2.43	9.03	2.43	4.60	6.95	11.48	5.28
		R@20	4.08	12.64	4.55	7.30	10.51	15.28	8.21
		R@50	7.50	18.84	7.70	12.53	16.91	22.96	13.86
	CIRR	R@1	0.84	2.28	0.84	0.84	0.82	17.91	0.84
		R@5	10.60	21.82	12.04	13.84	13.19	37.84	15.85
		R@10	16.19	31.03	17.77	20.89	19.88	48.30	23.91
		R@20	22.42	41.25	24.89	28.92	27.48	59.62	32.73
		R@50	32.97	55.25	36.24	41.82	40.70	73.53	48.01
8. $(q_i, q_t) \rightarrow (c_i, c_t)$	OVEN	R@1	22.31	16.65	22.27	23.83	8.60	0.04	15.97
		R@5	37.45	31.11	37.56	40.36	19.94	0.37	31.40
		R@10	43.46	37.32	43.64	46.94	25.40	0.75	38.78
		R@20	49.87	44.65	49.77	53.81	31.12	1.12	45.57
		R@50	58.43	54.68	58.27	62.99	38.81	2.04	54.64
	InfoSeek	R@1	8.62	12.83	11.16	16.54	7.26	0.08	9.27
		R@5	23.08	28.34	27.55	35.01	19.40	0.13	24.28
		R@10	31.54	35.76	35.79	43.84	26.23	0.13	32.91
		R@20	40.86	44.38	44.44	52.58	34.29	0.31	42.57
		R@50	53.82	56.44	56.44	63.78	46.95	0.49	56.11

Table 5: Ablation studies on loss functions and comparisons with intra-modality separation loss [16] under local setting of M-BEIR.

Task	Dataset	CL	Intra-modality Separation	GCL w/o $\mathcal{L}_{i2t}, \mathcal{L}_{t2i}$	GCL w/o $\mathcal{L}_{i2it}, \mathcal{L}_{t2it}$	GCL w/o $\mathcal{L}_{it2i}, \mathcal{L}_{it2t}$	<b>GCL</b>
1. $q_t \rightarrow c_i$	VisualNews	15.78	15.47	9.84	15.25	15.32	15.09
	MSCOCO	61.34	61.29	52.50	60.71	61.01	61.26
	Fashion200K	9.83	8.96	6.63	8.90	8.90	9.37
2. $q_t \rightarrow c_t$	WebQA	90.43	89.53	92.71	89.04	89.37	89.25
3. $q_t \rightarrow (c_i, c_t)$	EDIS	35.76	36.50	41.59	44.65	46.10	45.60
	WebQA	36.16	38.35	68.78	60.57	59.62	61.77
4. $q_i \rightarrow c_t$	VisualNews	13.35	13.91	8.37	13.82	13.76	13.66
	MSCOCO	71.98	71.40	63.02	72.26	71.94	71.92
	Fashion200K	9.29	9.06	6.01	9.20	8.96	9.06
5. $q_i \rightarrow c_i$	NIGHTS	28.21	30.00	27.12	27.83	28.30	28.35
6. $(q_i, q_t) \rightarrow c_t$	OVEN	29.91	30.45	22.99	31.66	31.45	31.40
	InfoSeek	28.47	30.20	27.58	34.2	33.56	33.83
7. $(q_i, q_t) \rightarrow c_i$	FashionIQ	2.25	2.45	3.76	5.35	4.46	4.60
	CIRR	11.34	11.39	15.44	14.77	13.93	13.84
8. $(q_i, q_t) \rightarrow (c_i, c_t)$	OVEN	35.84	37.87	32.64	40.02	40.29	40.36
	InfoSeek	23.94	26.72	31.89	35.28	34.43	35.01
	Avg.	31.49	32.10	31.93	35.22	35.09	<b>35.27</b>

Table 6: Performance variance on ablation studies after 3 different runs under global setting of M-BEIR.

Task	GCL w/o $\mathcal{L}_{i2t}, \mathcal{L}_{t2i}$	GCL w/o $\mathcal{L}_{i2it}, \mathcal{L}_{t2it}$	GCL w/o $\mathcal{L}_{it2i}, \mathcal{L}_{it2t}$	<b>GCL</b>
M-BEIR	28.09±0.24	32.85±0.09	33.61±0.17	34.04±0.13

Table 7: Comparisons on global setting of M-BEIR (Recall@50) using TinyCLIP.

Task	Dataset	TinyCLIP-SF [3]		
		Pretrained	CL +Pairwise	<b>GCL (Ours) +Pairwise</b>
1. $q_t \rightarrow c_i$	VisualNews [6]	0.01	0.00	0.70
	MSCOCO [7]	0.00	0.00	1.19
	Fashion200K [8]	0.00	0.00	0.06
2. $q_t \rightarrow c_t$	WebQA [9]	74.34	74.34	74.34
3. $q_t \rightarrow (c_i, c_t)$	EDIS [10]	24.68	22.12	43.88
	WebQA [9]	23.93	15.53	34.93
4. $q_i \rightarrow c_t$	VisualNews [6]	0.05	0.00	3.75
	MSCOCO [7]	0.02	0.00	36.4
	Fashion200K [8]	0.00	0.00	0.25
5. $q_i \rightarrow c_i$	NIGHTS [11]	85.28	84.20	84.62
6. $(q_i, q_t) \rightarrow c_t$	OVEN [12]	0.21	0.01	9.74
	InfoSeek [13]	0.27	0.06	4.86
7. $(q_i, q_t) \rightarrow c_i$	FashionIQ [14]	13.14	8.06	3.46
	CIRR [15]	16.14	14.99	15.95
8. $(q_i, q_t) \rightarrow (c_i, c_t)$	OVEN [12]	20.66	9.94	25.34
	InfoSeek [13]	18.96	9.47	23.92
	Avg.	17.36	14.92	<b>22.71</b>

## References

- [1] Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. VISTA: Visualized text embedding for universal multi-modal retrieval. In *ACL*, August 2024.
- [2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
- [3] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhui Chen. Uniir: Training and benchmarking universal multimodal information retrievers. *ECCV*, 2024.
- [4] Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi (Stephen) Chen, Xinggang Wang, Hongyang Chao, and Han Hu. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *ICCV*, 2023.
- [5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.
- [6] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. In *EMNLP*. Association for Computational Linguistics, 2021.
- [7] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. *CoRR*, 2014.
- [8] Xintong Han, Zuxuan Wu, Phoenix X. Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S. Davis. Automatic spatially-aware fashion concept discovery. In *ICCV*, 2017.
- [9] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *CVPR*, pages 16495–16504, June 2022.
- [10] Siqi Liu, Weixi Feng, Tsu-Jui Fu, Wenhui Chen, and William Wang. EDIS: Entity-driven image search over multimodal web content. In *EMNLP*. Association for Computational Linguistics, 2023.
- [11] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *NeurIPS*, 2023.
- [12] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. *CVPR*, 2023.
- [13] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? In *EMNLP*, 2023.
- [14] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. *CVPR*, 2021.
- [15] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4959–4968, June 2022.
- [16] Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in CLIP. In *ICLR*, 2025.