Generalized Contrastive Learning for Universal Multimodal Retrieval

Jungsoo Lee Janghoon Cho Hyojin Park Munawar Hayat
Kyuwoong Hwang Fatih Porikli Sungha Choi[†]

Qualcomm AI Research[*]
{jungsool, janghoon, hyojinp, hayat, kyuwoong, fporikli, sunghac}@qti.qualcomm.com

Abstract

Despite their consistent performance improvements, cross-modal retrieval models (*e.g.*, CLIP) show degraded performances with retrieving keys composed of fused image-text modality (*e.g.,* Wikipedia pages with both images and text). To address this critical challenge, multimodal retrieval has been recently explored to develop a unified single retrieval model capable of retrieving keys across diverse modality combinations. A common approach involves constructing new composed sets of image-text triplets (*e.g.*, retrieving a pair of image and text given a query image). However, such an approach requires careful curation to ensure the dataset quality and fails to generalize to unseen modality combinations. To overcome these limitations, this paper proposes Generalized Contrastive Learning (GCL), a novel loss formulation that improves multimodal retrieval performance without the burdensome need for new dataset curation. Specifically, GCL operates by enforcing contrastive learning across all modalities within a mini-batch, utilizing existing image-caption paired datasets to learn a unified representation space. We demonstrate the effectiveness of GCL by showing consistent performance improvements on off-the-shelf multimodal retrieval models (*e.g.*VISTA, CLIP, and TinyCLIP) using the M-BEIR, MMEB, and CoVR benchmarks.

1 Introduction

With the growing availability of multimodal data, the ability to retrieve relevant keys across different modalities has become increasingly important. While cross-modal retrieval, retrieving images with given text or vice versa, has garnered significant attention and progress [1, 2, 3, 4, 5, 6, 7, 8], performing retrieval with fused image-text modality still remains a challenge [9, 10]. Consider, for instance, the task of finding a Wikipedia page composed of both images and text (*e.g.,* the Eiffel Tower paired with its history) in response to a query (*e.g.,* "What is the history of the Eiffel Tower?"). For such real-world multimodal retrieval scenarios, existing approaches deliver limited performance.

This performance drop stems primarily from the pervasive modality gap [11, 12, 13, 14, 15, 16, 17], a critical barrier in retrieval systems. The modality gap arises when semantically similar samples across different modalities (*e.g.,* an image and its caption) exhibit low similarity in the embedding space, while semantically dissimilar samples within the same modality appear misleadingly close [18]. For example, although an image of teddy bears is paired with the annotated caption 'a photo of teddy bears', it may be embedded closer to images of other animals in the representation space, rather than to its corresponding caption. This misalignment becomes especially acute when retrieval keys - spanning text-only, image-only, or fused text-image formats - are stored in a unified database [9, 10]. For example, a search query might ideally match a Wikipedia page with both images and text. However, without a shared representation space to bridge the three different modalities, models struggle to pinpoint the most relevant candidate, which undermines their effectiveness.

[*] Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc. [†] indicates corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

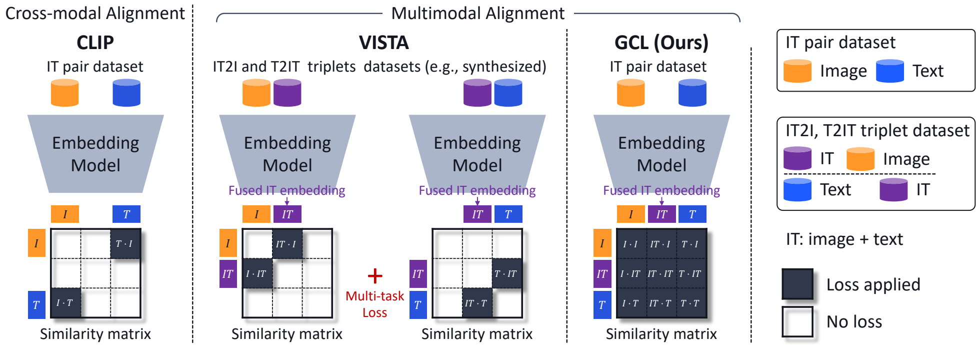


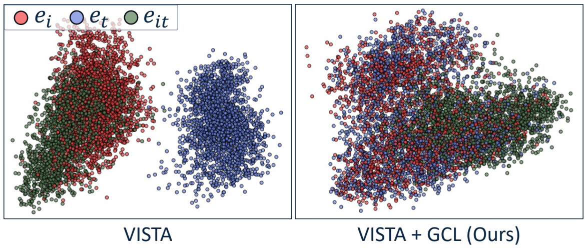
Figure 1: Overview of GCL. Given an embedding model pretrained for cross-modal alignment, previous studies (*e.g.*, VISTA [19]) constructed new triplet datasets to simulate specific multimodal retrieval scenarios. However, this approach limits generalization to unseen retrieval scenarios (white squares). In contrast, GCL improves retrieval performance across diverse scenarios (black squares). Specifically, by utilizing off-the-shelf image-caption datasets, GCL enables the learning of retrieval tasks involving nine different modality combinations.

To tackle this issue, recent research has extensively studied multimodal retrieval, which extends cross-modal retrieval by allowing searches across various modality combinations, including data samples that contain both images and text [19, 20, 9, 21, 22, 23, 24, 25, 10, 26]. As demonstrated in Fig. 1, previous studies have attempted to improve multimodal retrieval performance by generating specialized datasets tailored to specific retrieval scenarios [19, 20]. For example, VISTA [19] finetuned a pretrained cross-modal retrieval model with newly generated datasets composed of triplets: 1) IT2I dataset which includes queries of image-text pairs with candidate images and 2) T2IT dataset which includes query text with candidates of image-text pairs. While this approach can be effective for the targeted retrieval scenarios, it requires meticulous curation to ensure the quality of the generated samples (*e.g.,* verifying that the generated images accurately represent the intended content). Moreover, models trained on these composed datasets may fail to generalize to unseen modality combinations beyond those encountered in the training sets (*e.g.*, retrieving fused text-image samples when provided with corresponding text-image queries). Fig. 1 shows that the model only learns the retrieval scenarios included in the datasets (black squares), leaving the other combinations (white squares) unlearned during the finetuning phase. While such studies resort to generating new datasets for learning specific retrieval scenarios, little has been explored to fully utilize off-the-shelf image-caption paired datasets for improving multimodal retrieval performances.

To this end, we propose Generalized Contrastive Learning (GCL), a *simple yet effective* loss function that enhances multimodal retrieval performance by leveraging existing image-caption paired datasets, sidestepping the need for costly dataset construction. Specifically, GCL integrates three types of embeddings - text embeddings, image embeddings, and fused text-image embeddings - and applies contrastive loss across all modalities within a mini-batch to learn a unified representation space. As shown in Fig. 1, GCL encourages positive pairs from different modalities to be pulled closer together while pushing apart all negative pairs, regardless of modalities. This training process enables the retrieval model to learn retrieval between all combinations of modalities, which was limited to certain pairs in the previous methods that generated specific triplets. Despite its simplicity, GCL consistently improves multimodal retrieval performances across diverse tasks and datasets and outperforms a model trained with newly composed triplet datasets. The key advantage of GCL is that it does not require expensive dataset curation and can generalize well to various multimodal retrieval scenarios.

The major contributions of this paper are:

- We propose Generalized Contrastive Learning (GCL), a novel contrastive learning approach that improves multimodal retrieval by integrating different modalities (text, image, and fused text-image) within a mini-batch for building a unified representation space.
- Unlike previous methods that rely on expensive and manually curated composed datasets, GCL effectively leverages existing image-caption paired datasets, making it a cost-efficient and scalable solution for multimodal retrieval.
- We show that GCL significantly enhances multimodal retrieval performance on diverse benchmarks (*e.g.*, M-BEIR [9], MMEB [27], and CoVR [28]) and retrieval models (*e.g.*, VISTA [19], CLIP [1], and TinyCLIP [4]), showing its broad applicability and effectiveness.



	Cosine similarity between averaged vectors of each modality		
Modality Pair	$\bar{e}_i \leftrightarrow \bar{e}_t$	$\bar{e}_i \leftrightarrow \bar{e}_{it}$	$\bar{e}_t \leftrightarrow \bar{e}_{it}$
VISTA	0.8799	0.7393	0.6772
VISTA + GCL (Ours)	**0.8882**	**0.9688**	**0.8740**

Figure 2: PCA visualization of representation spaces using e_i, e_t, and e_{it}. We use MSCOCO for e_i (red) and e_t (blue) and WebQA for e_{it} (green). We sampled 2K samples from each modality, using 6K samples in total. $\bar{e}$ indicates the average embedding vector of each modality.

2 Related Work

2.1 Cross-modal Retrieval and Contrastive Learning

Cross-modal retrieval has been widely studied to enable searches across different modalities, such as retrieving images based on text queries or vice versa [1, 11, 2, 4, 3, 6, 12, 13, 8]. By leveraging contrastive learning with large-scale image-text pairs, both image and text embeddings are mapped into a shared representation space [1]. However, retrieval models trained in this manner still suffer from the modality gap, a discrepancy between image and text embeddings even with the same semantics [12, 13, 15, 16, 17, 18].

To address this issue, recent studies have proposed diverse techniques to reduce the modality gap [12, 11]. For instance, AlignCLIP introduces an intra-modality separation loss, which pushes apart samples within the same modality to improve cross-modal alignment [11]. While this approach helps mitigate the modality gap, it is mainly designed for cross-modal retrieval and does not explicitly handle data samples that contain both images and text (*e.g.*, social media pages with both images and text descriptions). As a result, its effectiveness in multimodal retrieval scenarios remains limited.

2.2 Multimodal Retrieval

Multimodal retrieval builds upon cross-modal retrieval by supporting searches across different combinations of modalities, including samples that incorporate both images and text [19, 20, 9, 29, 30, 31, 21, 22, 23, 32, 33]. UniIR demonstrates that adding image and text embeddings is effective for fusing representations for CLIP-based models, termed score-fusion (SF) [9]. The fused embeddings enable the retrieval of data samples that contain both images and text.

As mentioned previously, one common approach for multimodal retrieval is generating composed datasets containing paired image and text samples tailored for specific retrieval scenarios. For example, by using image generation models, VISTA generated the IT2T dataset, which consists of image and text queries with text-based keys, and T2IT dataset, which includes text queries with both image and text keys [19]. Similarly, MegaPairs constructed triplets consisting of two images and a descriptive text capturing the relationship between them, which are generated with multimodal large language models [20]. While dataset generation through large generative models can be effective under certain scenarios, it requires careful curation to ensure data quality, making the process labor-intensive and computationally expensive. Additionally, the retrieval model trained with such generated datasets may fail to generalize to scenarios unseen during the training phase. Therefore, a more efficient and scalable approach is needed to enhance multimodal retrieval performance without relying on expensive dataset generation.

3 Method

3.1 Problem Setup

In this paper, we define (x_i, x_t) as a pair of image and text used to train a retrieval model θ, which consists of an image encoder θ_i and a text encoder θ_t. The extracted embeddings of images and text are denoted as $e_i = \theta_i(x_i)$ and $e_t = \theta_t(x_t)$, respectively. During inference, the model retrieves candidate samples c when given a query q.

In cross-modal retrieval, the model retrieves text candidates c_t for an image query q_i ($q_i \rightarrow c_t$) and vice versa ($q_t \rightarrow c_i$). The multimodal retrieval task generalizes this setting by incorporating samples that contain both images and text, denoted as x_{it}. This results in more complex retrieval scenarios,

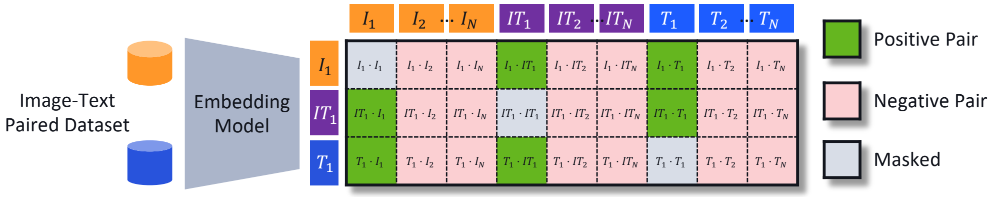


Image-Text Paired Dataset
Embedding Model
I_1 I_2 ⋯ I_N IT_1 IT_2 ⋯ IT_N T_1 T_2 ⋯ T_N
I_1
$I_1 \cdot I_1$ $I_1 \cdot I_2$ $I_1 \cdot I_N$ $I_1 \cdot IT_1$ $I_1 \cdot IT_2$ $I_1 \cdot IT_N$ $I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_N$
IT_1
$IT_1 \cdot I_1$ $IT_1 \cdot I_2$ $IT_1 \cdot I_N$ $IT_1 \cdot IT_1$ $IT_1 \cdot IT_2$ $IT_1 \cdot IT_N$ $IT_1 \cdot T_1$ $IT_1 \cdot T_2$ $IT_1 \cdot T_N$
T_1
$T_1 \cdot I_1$ $T_1 \cdot I_2$ $T_1 \cdot I_N$ $T_1 \cdot IT_1$ $T_1 \cdot IT_2$ $T_1 \cdot IT_N$ $T_1 \cdot T_1$ $T_1 \cdot T_2$ $T_1 \cdot T_N$
🟩 Positive Pair
🟥 Negative Pair
⬜ Masked

Figure 3: Training process of GCL. Given a dataset composed of image-caption pairs, we extract e_i, e_t, and e_{it}. For e_{it}, we follow the extraction method used by the retrieval model (*e.g.*, VISTA and CLIP-SF). Then, we integrate samples of the three different modalities into a single mini-batch for contrastive learning. We mask out the supervision on the positive samples with identical modalities.

where candidates of arbitrary modalities are retrieved based on queries of arbitrary modalities, such as $q_t \rightarrow c_{it}$, $q_{it} \rightarrow c_i$, and $q_{it} \rightarrow c_{it}$. Following UniIR [9] we consider two different retrieval settings: (1) global setting, retrieving candidates from a shared database regardless of modalities and tasks, and (2) local setting, retrieving candidates from a task-specific database with the same modality. We conduct experiments on both settings in this paper.

Figure 2 clearly illustrates the challenge addressed in this work. The left side of Figure 2 presents a PCA [34] visualization of the embedding spaces learned by VISTA and VISTA fine-tuned with our proposed loss function, Generalized Contrastive Learning (GCL). For this visualization, we use c_i and c_t from MSCOCO and c_{it} from WebQA. As shown, VISTA fails to construct a unified embedding space for the three different modalities. The main reason is that retrieval models trained on image-caption paired datasets [2] fail to learn a shared embedding space that incorporates e_{it}.

The table on the right further supports this observation. By using the same samples from the visualization, we compute the average embeddings of images, text, and image-text pairs, denoted as $\overline{e}_i$, $\overline{e}_t$, and $\overline{e}_{it}$, respectively. We then calculate the cosine similarity between the average embedding vectors of each modality. As demonstrated in the table, VISTA exhibits low cosine similarity across modalities, indicating a significant modality gap. Our goal is to mitigate this gap, as evidenced by the more intermixed scatter plots in the PCA visualization and the increased cosine similarities between modalities after finetuning with GCL.

3.2 Generalized Contrastive Learning

Using N samples per mini-batch, the standard contrastive learning loss is formulated as:

$$\mathcal{L}_{\text{CL}} = -\frac{1}{2N} \sum_{j=1}^{N} \sum_{(a,b)\in S} \log \frac{\exp[(e_a^j \cdot e_b^j)/\tau]}{\sum_{k=1}^{N} \exp[(e_a^j \cdot e_b^k)/\tau]}, \tag{1}$$

where j and k denote the sample indices, τ is the temperature scaling parameter, and S denotes the set of modality pairs $\{(i,t),(t,i)\}$. While this loss function has been effective in constructing a unified cross-modal representation space, multimodal retrieval performance remains limited within this embedding space. The main limitation arises because models are not explicitly trained with embeddings that represent data samples containing both images and text (*e.g.*, news articles with both pictures and textual explanations). As mentioned earlier, previous studies created new datasets consisting of triplets to simulate specific retrieval scenarios for addressing such a limitation [19, 20]. In contrast, our proposed method does not rely on these new triplet-based datasets.

Fig. 3 describes our method well. Generally, multimodal retrieval models are obtained by finetuning cross-modal retrieval models, which are pretrained with image-caption paired datasets. Instead of finetuning the cross-modal retrieval models with newly constructed triplet-based datasets, we finetune them using off-the-shelf image-caption paired datasets with our proposed Generalized Contrastive Learning (GCL) loss function. In GCL loss, negative samples are constructed from all possible combinations of embeddings (pink squares), including 1) image embeddings e_i, 2) text embeddings e_t, and 3) fused embeddings of images and text e_{it}. To obtain the fused embeddings of samples with both images and text, we follow the extraction method used in the retrieval model. For example, we either (1) use a specialized architecture pretrained for extracting fused embeddings for VISTA

[2] For the experiment, we use the checkpoint of VISTA prior to training with newly generated composed sets.

(*e.g.,* appending visual tokens alongside text tokens as input to a text encoder [19]), or (2) sum the individual image and text embeddings, $e_{it} = e_i + e_t$ for CLIP-based models following UniIR [9]. The positive pairs are then defined as samples with different modalities from the same pair (green squares). Positive pairs with the same modality correspond to the sample itself, so they are masked out during training (gray squares). Using the image embedding query as an example, the positive pair can be either the corresponding text embedding or the fused image-text embedding. GCL loss can be formulated as:

$$\mathcal{L}_{\text{GCL}} = -\frac{1}{6N} \sum_{j=1}^{N} \sum_{(a,b)\in P} \log \frac{\exp[(e_a^j \cdot e_b^j)/\tau]}{\sum_{m\in M} \sum_{k=1}^{N} \exp[(e_a^j \cdot e_m^k)/\tau]}, \tag{2}$$

where M represents the set of modalities $\{i, t, it\}$ and P denotes the set of positive modality pairs $\{(i,t), (i,it), (t,i), (t,it), (it,t), (it,i)\}$. The factor $6N$ normalizes the total loss across all six modality combinations in P, with each combination contributing losses from N training samples.

Although AlignCLIP [11] recently introduced an intra-modality separation loss to push negative image samples away from a given image query in order to reduce the modality gap, it does not incorporate contrastive learning across all possible combinations, particularly those involving fused embeddings e_{it}. In contrast, our GCL loss provides a more generalized contrastive learning framework by seamlessly integrating different modalities into a unified representation space within a single mini-batch, leading to improved multimodal retrieval performances across diverse scenarios and tasks (results shown in Table 5).

4 Experiments

4.1 Experimental Settings

Benchmarks We evaluate the effectiveness of our proposed GCL using standard multimodal retrieval benchmarks: M-BEIR[9], MMEB[27], and CoVR [28] [3]. While images are used as input for M-BEIR and MMEB, CoVR handles video input. For M-BEIR, we conduct evaluations on 10 datasets under both local and global evaluation settings. For MMEB, we perform experiments on 12 sub-datasets included in the retrieval benchmark. Regarding CoVR, following the original evaluation setting of CoVR, we sample 15 frames for each target video and average the embeddings of each frame to obtain a single visual embedding for a given video. For the image-caption paired dataset used during finetuning, we use the LLaVA Visual Instruct Pretrain LCS-558K dataset [35], in which personal information has been blurred for data sanitization. Note that our experiments are conducted in a zero-shot setting, meaning the model is not fine-tuned on the training set of the evaluation benchmark.

Models We apply the GCL loss to recent multimodal retrieval models, VISTA [19], CLIP [1], and TinyCLIP [4] [4]. For VISTA, we use the checkpoint before the second stage, the one trained without the generated dataset to demonstrate that our method can enhance multimodal retrieval performance even without newly composed triplet datasets. Throughout the paper, we refer to this checkpoint as VISTA. Following UniIR [9], we adopt score-level fusion for fused embeddings when using CLIP-based models, referred to as CLIP-SF and TinyCLIP-SF for CLIP and TinyCLIP, respectively.

Baselines We compare the effectiveness of GCL with that of pretrained model and standard contrastive learning. For VISTA, we also compare ours with VISTA finetuned using the generated datasets composed of triplets (i.e., IT2I and T2IT) used in the original paper, which we denoted as CL+Triplet.

Metrics Following prior work [9, 27], we adopt the standard retrieval evaluation metric, Recall@K. Following UniIR, we set K=5 for the local setting, except for Fashion200K and FashionIQ, where we use K=10, and set K=50 for the global setting. We set K=1 for MMEB and K=1, 5, 10, and 50 for CoVR, following the previous work [27, 28]. Further details are provided in the Supplementary.

4.2 Quantitative Evaluation

Table 1 demonstrates that applying GCL consistently improves multimodal retrieval performance in the global setting of M-BEIR across various tasks and models. Notably, the performance improvement is particularly significant for tasks involving q_{it} or c_{it}. Even without generating new data samples composed of it, GCL loss improves tasks related to it with off-the-shelf image-caption paired datasets. While VISTA trained with generated datasets (i.e., IT2I and T2IT) shows the best performance for

[3] M-BEIR and CoVR datasets are under the MIT license, and MMEB is under the Apache-2.0 license.

[4] VISTA, CLIP, and TinyCLIP are all under the MIT license.

Table 1: Comparisons on global setting of M-BEIR using Recall@50. CL and GCL indicates standard contrastive learning and our generalized contrastive learning, respectively. Triplet and Pairwise refers to training with newly composed triplet dataset and original image-text paired dataset, respectively.

Task	Dataset	VISTA [19]				CLIP-SF [9]		
		Pretrained	CL +Triplet	CL +Pairwise	**GCL (Ours) +Pairwise**	Pretrained	CL +Pairwise	**GCL (Ours) +Pairwise**
1. $q_t \to c_i$	VisualNews [36]	5.36	1.64	9.29	16.64	0.08	0.00	6.70
	MSCOCO [37]	2.72	5.60	14.42	38.85	0.00	0.00	3.25
	Fashion200K [38]	0.00	0.00	0.00	4.25	0.00	0.00	0.00
2. $q_t \to c_t$	WebQA [39]	97.07	96.90	96.86	96.25	60.29	88.55	60.24
3. $q_t \to (c_i, c_t)$	EDIS [40]	25.15	44.37	36.90	49.06	23.39	34.19	54.43
	WebQA [39]	14.22	80.88	31.74	64.00	19.87	68.42	40.62
4. $q_i \to c_t$	VisualNews [36]	1.35	0.08	1.18	4.71	0.00	0.00	2.48
	MSCOCO [37]	12.90	0.50	26.82	60.32	0.00	0.00	24.84
	Fashion200K [38]	0.02	0.00	0.00	0.72	0.00	0.00	0.16
5. $q_i \to c_i$	NIGHTS [41]	76.60	83.07	79.39	82.50	81.65	88.07	85.09
6. $(q_i, q_t) \to c_t$	OVEN [42]	5.06	1.78	3.10	8.72	0.00	0.00	3.63
	InfoSeek [43]	2.94	4.80	1.70	9.07	0.00	0.00	1.86
7. $(q_i, q_t) \to c_i$	FashionIQ [44]	6.66	16.41	6.10	10.88	11.61	0.00	4.25
	CIRR [45]	23.62	43.81	24.27	31.13	18.06	0.43	21.25
8. $(q_i, q_t) \to (c_i, c_t)$	OVEN [42]	34.31	9.67	32.83	32.92	11.04	0.58	19.47
	InfoSeek [43]	30.95	14.94	29.82	34.97	12.73	0.00	21.89
	Avg.	21.18	25.28	24.65	**34.06**	14.92	17.52	**21.89**

Table 2: Comparisons on MMEB dataset using Recall@1, following VLM2Vec [27]. Abbreviations as in Table 1.

Task	Dataset	VISTA [19]				CLIP-SF [9]		
		Pretrained	CL +Triplet	CL +Pairwise	**GCL (Ours) +Pairwise**	Pretrained	CL +Pairwise	**GCL (Ours) +Pairwise**
1. $q_t \to c_i$	VisDial [46]	10.1	17.3	17.2	16.6	22.5	27.2	31.1
	VisualNews [36]	51.7	38.4	50.7	50.5	72.4	41.1	70.5
	MSCOCO [37]	32.8	44.8	46.8	48.7	54.9	60.7	61.5
	Wiki-SS-NQ [47]	16.3	12.4	14.7	16.7	50.7	34.1	46.5
2. $q_t \to c_{it}$	WebQA [39]	65.9	83.9	73.3	79.5	61.1	73.7	62.8
	EDIS [40]	78.0	64.6	78.2	78.5	79.2	45.4	85.4
3. $q_i \to c_t$	VisualNews [36]	54.6	25.7	52.7	54.2	1.5	0.2	10.9
	MSCOCO [37]	44.0	32.9	55.3	52.8	2.0	0.1	23.1
4. $q_i \to c_i$	NIGHTS [41]	64.7	64.1	65.7	65.4	60.1	9.1	66.4
5. $q_{it} \to c_i$	CIRR [45]	8.1	14.1	9.0	11.2	10.9	46	11.6
	FashionIQ [44]	3.3	9.0	3.1	7.7	9.9	16.5	6.2
6. $q_{it} \to c_{it}$	OVEN [42]	54.3	45.4	53.6	57.3	46.1	4.7	53.8
	Avg.	40.3	37.7	43.4	**44.9**	39.3	29.9	**44.2**

tasks $q_{it}{\to}c_i$ and $q_t{\to}c_{it}$, it shows limited performance gain or performance drop with other tasks. That is, finetuning retrieval models with generated samples under certain scenarios may show promising performance for the targeted scenarios, but they may fail to generalize to the tasks unseen during the finetuning phase. It also shows degraded performance on cross-modal tasks ($q_i{\to}c_t$ and $q_t{\to}c_i$) compared to the pretrained VISTA. We conjecture that further finetuning VISTA with composed sets targeting specific retrieval scenarios may degrade its original performance on cross-modal tasks due to forgetting its initial cross-modal alignment. We want to emphasize that our goal is to perform well across a wide range of tasks and datasets, not just to excel at a specific task or dataset.

Tables 2 and 3 compare the multimodal retrieval performances under the local setting. Again, finetuning retrieval models with GCL brings further performance improvements even under the local setting. Along with the global setting, we believe that performing well in the local setting is also important since we may need databases separately divided for each task depending on the use cases we pursue in the real-world applications. Although there may exist a slight performance drop in scenario-specific retrieval tasks (*e.g.*CIRR and FashionIQ), this can largely be attributed to the nature of the fine-tuning dataset used. The LCS-558K dataset is designed for general-purpose fine-tuning, which may not fully capture the nuances of domain-specific tasks. To achieve optimal performance in these specialized applications, we believe GCL serves as an effective initial training stage, and performance can be further improved through additional fine-tuning with task-specific data.

Table 4 demonstrates that applying GCL also improves the video retrieval performance. When deploying retrieval models in real-world scenarios, visual content may be stored in video formats

Table 3: Comparisons on local setting of M-BEIR. We report the results using Recall@5 for the local setting except using Recall@10 for Fashion200K and FashionIQ, following UniIR [9, 44]. Abbreviations as in Table 1.

Task	Dataset	VISTA [19]				CLIP-SF [9]		
		Pretrained	CL +Triplet	CL +Pairwise	**GCL (Ours)** **+Pairwise**	Pretrained	CL +Pairwise	**GCL (Ours)** **+Pairwise**
1. $q_t \rightarrow c_i$	VisualNews [36]	16.04	10.01	15.78	15.42	44.34	20.97	36.71
	MSCOCO [37]	50.65	58.40	61.34	61.09	61.09	71.94	67.69
	Fashion200K [38]	9.31	8.03	9.83	9.54	6.57	8.84	7.04
2. $q_t \rightarrow c_t$	WebQA [39]	91.20	91.20	90.43	89.37	40.61	70.35	40.61
3. $q_t \rightarrow (c_i, c_t)$	EDIS [40]	36.69	40.98	35.76	45.88	43.29	34.56	48.97
	WebQA [39]	33.49	74.51	36.16	62.49	45.48	69.97	44.01
4. $q_i \rightarrow c_t$	VisualNews [36]	14.03	4.42	13.35	13.70	41.78	20.18	30.53
	MSCOCO [37]	61.66	60.44	71.98	72.56	79.00	85.78	79.04
	Fashion200K [38]	9.63	6.71	9.29	9.31	7.71	8.65	8.55
5. $q_i \rightarrow c_i$	NIGHTS [41]	26.32	26.32	28.21	28.35	26.13	30.94	30.99
6. $(q_i, q_t) \rightarrow c_t$	OVEN [42]	30.39	25.93	29.91	31.82	0.31	0.23	8.93
	InfoSeek [43]	29.87	23.16	28.47	34.26	0.29	0.00	6.78
7. $(q_i, q_t) \rightarrow c_i$	FashionIQ [44]	2.43	9.03	2.25	5.00	6.95	11.48	5.28
	CIRR [45]	10.60	21.82	11.34	14.27	13.19	37.84	15.85
8. $(q_i, q_t) \rightarrow (c_i, c_t)$	OVEN [42]	37.45	31.11	35.84	40.60	19.94	0.37	31.40
	InfoSeek [43]	23.08	28.34	23.94	35.32	19.40	0.13	24.28
	Avg.	30.18	32.53	31.49	**35.56**	28.51	29.51	**30.42**

Table 4: Comparisons on CoVR Benchmark. Following CoVR, we use the frame of middle index for the query video, while averaging 15 uniformly sampled frames for the target video. Abbreviations as in Table 1.

Rank	VISTA [19]			CLIP-SF [9]		
	Pretrained	CL +Pairwise	**GCL (Ours)** **+Pairwise**	Pretrained	CL +Pairwise	**GCL (Ours)** **+Pairwise**
R@1	31.22	33.76	**37.52**	37.32	19.68	**37.60**
R@5	58.37	59.74	**63.46**	62.60	40.30	**65.69**
R@10	68.15	69.52	**72.81**	71.99	50.67	**75.78**
R@50	88.50	88.50	**91.12**	88.18	74.92	**92.92**

(*e.g.,* detecting unexpected actions in CCTV), making video retrieval an important task due to its practicality. Consistent performance improvements in multimodal retrieval tasks even including video retrieval demonstrates that GCL is a robust and versatile approach for enhancing retrieval models across diverse scenarios.

5 Further Analysis

5.1 Ablation Studies

Table 5 compares GCL with the intra-modality separation loss proposed in AlignCLIP [11] while dissecting the contributions of the individual loss components of GCL. $\mathcal{L}_{a2b}$ indicates the loss function of GCL using a as the query modality and b as the target modality from a given positive pair. Regarding intra-modality separation loss, we added the loss term in addition to standard contrastive learning during training. For the ablation study of GCL, we excluded each of the following key loss functions: 1) cross-modal alignment terms ($\mathcal{L}_{i2t}$ and $\mathcal{L}_{t2i}$), 2) it-candidate learning terms ($\mathcal{L}_{i2it}$ and $\mathcal{L}_{t2it}$), and 3) it-query learning terms ($\mathcal{L}_{it2i}$ and $\mathcal{L}_{it2i}$). For the comparisons, we use the global setting of M-BEIR. Results on the local setting of M-BEIR and performance variance of multiple runs are included in our Supplementary.

As shown, adding the intra-modality separation loss indeed improves the multimodal retrieval performance compared to training with standard contrastive learning. However, we observe that the performance gain is limited for tasks involving retrieval with identical modalities (*e.g.*, $q_i{\rightarrow}c_i$ and $q_t{\rightarrow}c_t$) or queries with it modality (*e.g.*, $q_{it}{\rightarrow}c_i$ and $q_{it}{\rightarrow}c_t$) compared to our GCL loss. This indicates that intra-modality separation loss mitigates the modality gap but it fails to consider diverse multimodal retrieval scenarios, which are effectively addressed by GCL.

Regarding the ablation study, we observe a performance drop on the task that each module of GCL loss is responsible for. To be more specific, by excluding $\mathcal{L}_{i2t}$ and $\mathcal{L}_{t2i}$, the performance on cross-modal tasks are degraded significantly. Also, the performances on tasks of $q_t{\rightarrow}c_{it}$ are degraded after excluding $\mathcal{L}_{i2it}$ and $\mathcal{L}_{t2it}$. We want to emphasize that we did not perform an extensive hyperparameter search for finding the optimal weighing values for each loss function in GCL. While

Table 5: Ablation studies on loss functions and comparisons with intra-modality separation loss [11] using global setting of M-BEIR.

Task	Dataset	CL	Intra-modality Separation [11]	GCL w/o $\mathcal{L}_{i2t}, \mathcal{L}_{t2i}$	GCL w/o $\mathcal{L}_{i2it}, \mathcal{L}_{t2it}$	GCL w/o $\mathcal{L}_{it2i}, \mathcal{L}_{it2t}$	**GCL**
1. $q_t \rightarrow c_i$	VisualNews [36]	9.29	14.36	2.91	18.26	17.07	16.64
	MSCOCO [37]	14.42	36.67	9.77	39.43	38.99	38.85
	Fashion200K [38]	0.00	3.84	0.41	3.90	4.54	4.25
2. $q_t \rightarrow c_t$	WebQA [39]	96.86	96.17	97.68	96.13	96.21	96.25
3. $q_t \rightarrow (c_i, c_t)$	EDIS [40]	36.90	49.74	45.23	37.15	49.18	49.06
	WebQA [39]	31.74	47.59	69.53	52.93	61.65	64.00
4. $q_i \rightarrow c_t$	VisualNews [36]	1.18	2.78	0.59	5.50	4.63	4.71
	MSCOCO [37]	26.82	48.64	13.4	63.06	58.60	60.32
	Fashion200K [38]	0.00	0.61	0.08	0.76	0.63	0.72
5. $q_i \rightarrow c_i$	NIGHTS [41]	79.39	78.02	79.15	83.21	82.83	82.50
6. $(q_i, q_t) \rightarrow c_t$	OVEN [42]	3.10	5.23	6.92	9.77	7.95	8.72
	InfoSeek [43]	1.70	3.72	9.10	10.96	7.79	9.07
7. $(q_i, q_t) \rightarrow c_i$	FashionIQ [44]	6.10	6.33	9.48	11.76	10.61	10.88
	CIRR [45]	24.27	23.57	32.52	31.82	30.50	31.13
8. $(q_i, q_t) \rightarrow (c_i, c_t)$	OVEN [42]	32.83	35.74	36.75	29.20	31.22	32.92
	InfoSeek [43]	29.82	34.26	40.26	33.31	32.27	34.97
	Avg.	24.65	30.45	28.36	32.95	33.42	**34.06**

Table 6: Performance improvements on M-BEIR under global setting with TinyCLIP.

Metric	VISTA	CLIP-SF	TinyCLIP-SF	TinyCLIP-SF + GCL
Model Params.	196M	427M	120M	120M
Avg. Inference (ms)	26.06	21.58	14.67	14.67
M-BEIR	21.18	14.92	17.36	**22.71**

this may improve performance on certain tasks, it may accompany performance drops on other tasks. Since the main goal of this paper is to design a loss function that generally works well for diverse retrieval scenarios, we simply added the loss functions with identical weights. Depending on the use cases, each loss function in GCL may be weighed differently.

5.2 Empowering Lightweight Models

Deploying lightweight models for retrieval is essential for enabling fast and efficient inference in real-time or resource-constrained environments, such as mobile or edge devices (*e.g.,* retrieving personal data on mobile phones). Table 6 illustrates that applying GCL improves the retrieval performance of TinyCLIP compared to pretrained retrieval models, including VISTA and CLIP-SF, despite having fewer parameters and a lower average inference speed (ms). This result indicates that fine-tuning a small and lightweight retrieval model with GCL is a viable solution for improving retrieval performance.

5.3 Ranks of Ground Truth Candidates

Figure 4 compares the ranks of ground truth candidates between VISTA and VISTA trained with GCL. The x-axis denotes the ranks, and the y-axis indicates its frequency. For this analysis, we use q_t from MSCOCO and a candidate pool composed of c_t and c_i from MSCOCO, where the task is to retrieve the ground truth c_i given q_t. The numbers of queries and candidates are 2.5K and 29K, respectively. We visualize only candidates ranked within the top 10K.

Our findings show that when VISTA is trained with GCL, most ground truth candidates achieve high ranks, with the majority ranked within the top 500. In contrast, VISTA without GCL exhibits a non-trivial number of ground truth candidates that are ranked significantly lower. This highlights the challenge of retrieving ground truth c_t from a shared database of mixed modalities when the modality gap is not effectively reduced. GCL successfully mitigated the modality gap, as demonstrated by the high ranks of ground truth candidates.

5.4 Cosine Similarity with Candidates

Ground truth candidates Figure 5 (a) visualizes the cosine similarity between the embeddings of queries and their corresponding ground truth candidates. For this analysis, we selected one dataset from each task in M-BEIR. As shown, applying GCL to VISTA consistently increases the cosine similarities across diverse tasks and datasets, indicating improved alignment between query and ground truth representations. By improving the representation space, GCL ensures that relevant

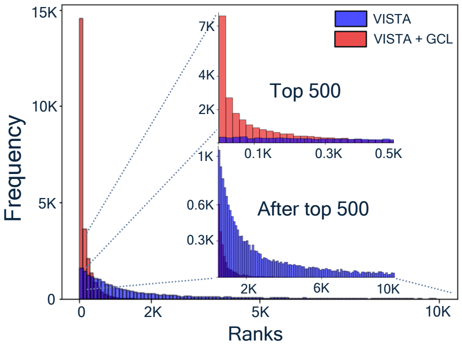
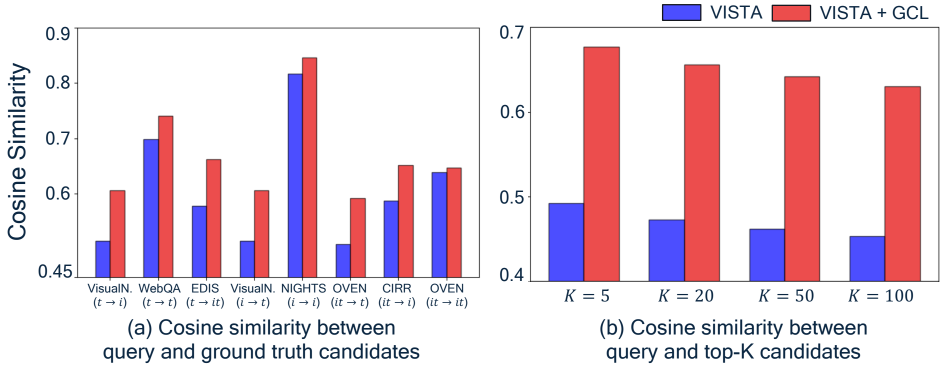


Figure 4: Rankings of ground truth candidates. The x-axis and y-axis indicate the ranks and the frequency of ranks, respectively. We use the task of $q_t \rightarrow c_i$ on the MSCOCO dataset, with a candidate pool composed of c_i and c_t from MSCOCO.



(a) Cosine similarity between query and ground truth candidates

(b) Cosine similarity between query and top-K candidates

Figure 5: (a) Cosine similarity between query and ground truth candidates. X-axis and y-axis indicates the dataset and cosine similarity, respectively. VisualN. refers to VisualNews. (b) Cosine similarity between queries and top-ranked candidates. We use MSCOCO for the task of $q_i \rightarrow c_t$.

multimodal pairs - whether text, image, or fused - are positioned closer in the embedding space, leading to more accurate retrieval.

Top-ranked candidates Figure 5 (b) further examines the cosine similarity between queries and their top-ranked retrieved candidates to evaluate the retrieval consistency. Using MSCOCO ($q_i \rightarrow c_t$), we analyze how cosine similarity trends change across different ranks. As shown, VISTA exhibits a significant drop in similarity with lower ranks, suggesting that lower-ranked candidates are less semantically relevant. In contrast, applying GCL helps maintain high cosine similarity even with lower ranks, demonstrating that relevant candidates are still retrieved even with lower ranks.

This stability across ranks highlights the ability of GCL to build a unified representation space more effectively, ensuring that even when the top candidate is not a perfect match, subsequent retrieved items remain semantically meaningful. Such improvements are crucial for real-world multimodal retrieval applications where retrieving a set of relevant candidates would be a viable solution, rather than retrieving the single best match.

6 Conclusion

In this paper, we introduced Generalized Contrastive Learning (GCL), a *simple yet effective* loss function designed to enhance multimodal retrieval performance without the need for generating triplet datasets simulating certain retrieval scenarios. By integrating text, image, and fused text-image embeddings into the contrastive learning framework, GCL mitigates the modality gap and improves multimodal retrieval performance across diverse tasks and datasets. Although not discussed in this work, one promising future work direction is integrating GCL with multimodal large language models (MLLMs) to further enhance retrieval capabilities in generative and reasoning-based tasks [23, 21, 22, 48, 49, 50]. As MLLMs continue to advance, utilizing retrieved information to generate responses would be promising, especially with databases containing mixed-modalities. By eliminating the need for labor-intensive dataset curation while improving retrieval across arbitrary modality combinations, GCL presents a scalable and effective solution for multimodal retrieval. We hope this work paves the way for future research in leveraging contrastive learning for more generalizable and robust multimodal retrieval.

Acknowledgement We would like to thank Qualcomm AI Research team for their valuable discussions.

References

[1] 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.

[2] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

[3] Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. Mobileclip: Fast image-text models through multi-modal reinforced training. In *CVPR*, 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] Young Kyun Jang, Junmo Kang, Yong Jae Lee, and Donghyun Kim. MATE: Meet at the embedding - connecting images with long texts. In *EMNLP*, 2024.

[6] Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. Clip-kd: An empirical study of clip model distillation. In *CVPR*, 2024.

[7] Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In *CVPR*, 2022.

[8] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022.

[9] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. *ECCV*, 2024.

[10] Lang Huang, Qiyu Wu, Zhongtao Miao, and Toshihiko Yamasaki. Joint fusion and encoding: Advancing multimodal retrieval from the ground up, 2025.

[11] Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Improving cross-modal alignment in CLIP. In *ICLR*, 2025.

[12] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022.

[13] Peiyang Shi, Michael C. Welle, Mårten Björkman, and Danica Kragic. Towards understanding the modality gap in CLIP. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023.

[14] Simon Schrodi, David T Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language representation learning. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024.

[15] Siddharth Srivastava and Gaurav Sharma. Omnivec: Learning robust representations with cross modal sharing. In *WACV*, 2024.

[16] Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P. Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). In *NeurIPS*, 2024.

[17] Yan Xia, Hai Huang, Jieming Zhu, and Zhou Zhao. Achieving cross modal generalization with multimodal unified representation. In *NeurIPS*, 2023.

[18] François Role, Sébastien Meyer, and Victor Amblard. Fill the gap: Quantifying and reducing the modality gap in image-text representation learning, 2025.

[19] Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. VISTA: Visualized text embedding for universal multi-modal retrieval. In *ACL*, August 2024.

[20] Junjie Zhou, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, Defu Lian, and Yongping Xiong. Megapairs: Massive data synthesis for universal multimodal retrieval, 2024.

[21] Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Gme: Improving universal multimodal retrieval by multimodal llms. In *CVPR*, 2025.

[22] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. MM-EMBED: UNIVERSAL MULTIMODAL RETRIEVAL WITH MULTIMODAL LLMS. In *ICLR*, 2025.

[23] Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-v: Universal embeddings with multimodal large language models, 2024.

[24] Woongyeong Yeo, Kangsan Kim, Soyeong Jeong, Jinheon Baek, and Sung Ju Hwang. Universalrag: Retrieval-augmented generation over multiple corpora with diverse modalities and granularities. *arXiv preprint arXiv:2504.20734*, 2025.

[25] Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, Wenhao Huang, Noura Al Moubayed, Jie Fu, and Chenghua Lin. Scimmir: Benchmarking scientific multi-modal information retrieval. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

[26] Adriel Saporta, Aahlad Manas Puli, Mark Goldstein, and Rajesh Ranganath. Contrasting with symile: Simple model-agnostic representation learning for unlimited modalities. In *NeurIPS*, 2024.

[27] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2vec: Training vision-language models for massive multimodal embedding tasks. In *ICLR*, 2025.

[28] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gül Varol. CoVR: Learning composed video retrieval from web video captions. *AAAI*, 2024.

[29] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and quality assessment for composed image retrieval. *AAAI*, 2024.

[30] Yiyang Chen, Zhedong Zheng, Wei Ji, Leigang Qu, and Tat-Seng Chua. Composed image retrieval with text feedback via multi-grained uncertainty regularization. In *ICLR*, 2024.

[31] Wei Li, Hehe Fan, Yongkang Wong, Yi Yang, and Mohan Kankanhalli. Improving context understanding in multimodal large language models via multimodal composition learning. In *ICML*, 2024.

[32] Yongqi Li, Wenjie Wang, Leigang Qu, Liqiang Nie, Wenjie Li, and Tat-Seng Chua. Generative cross-modal retrieval: Memorizing images in multimodal language models for retrieval and beyond. In *ACL*, 2024.

[33] Zhengyuan Zhu, Daniel Lee, Hong Zhang, Sai Sree Harsha, Loic Feujio, Akash Maharaj, and Yunyao Li. MuRAR: A simple and effective multimodal retrieval and answer refinement framework for multimodal question answering. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, 2025.

[34] I.T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.

[35] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] Siqi Liu, Weixi Feng, Tsu-Jui Fu, Wenhu Chen, and William Wang. EDIS: Entity-driven image search over multimodal web content. In *EMNLP*. Association for Computational Linguistics, 2023.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, 2017.

[47] Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. In *EMNLP*, 2024.

[48] Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024.

[49] Sahel Sharifymoghaddam, Shivani Upadhyay, Wenhu Chen, and Jimmy Lin. Unirag: Universal retrieval augmentation for large vision language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025.

[50] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *ICLR*, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions and the scope of the paper are well explained in the abstract and the introduction. For example, we summarized the contributions with bulletin points in the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the limitation of this work in the conclusion section by proposing possible future work direction. Specifically, while not included in our work, we believe that extending our work with multimodal large language models would be an interesting research directin.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: We do not include theoretical assumptions and proofs in our paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We included the implementation details regarding our experiments in Section 4.1 and our Supplementary. We also include the pseudocodes in our Supplementary in order to help readers reproduce our algorithm.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We are currently refactoring and cleaning the codes. We are planning to release the codes after internal review of the codes is finalized.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, we specified the training and test details in our implementation details of Section 4.1 of the main paper and Supplementary.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Since we conducted extensive experiments, it was challenging to report results of multiple runs for all experiments. However, we reported error bars for the ablation study on each loss term of GCL in our Supplementary.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We described the computing resources including the type of GPU and memory consumed for the experiments in our Supplementary.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirmed and conformed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed both potential positive societal impacts and negative societal impacts in our Supplementary.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: In Section 4.1 of the main paper, we mentioned that we blurred the facial images included in the train set for the sanitization of data.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In the main paper, we denoted the licenses of each dataset in the footnotes.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release new assets in this work.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.