Qualcomm

NEURAL INFORMATION PROCESSING SYSTEMS

Dec 02-07, 2025

Generalized Contrastive Learning for Universal Multimodal Retrieval

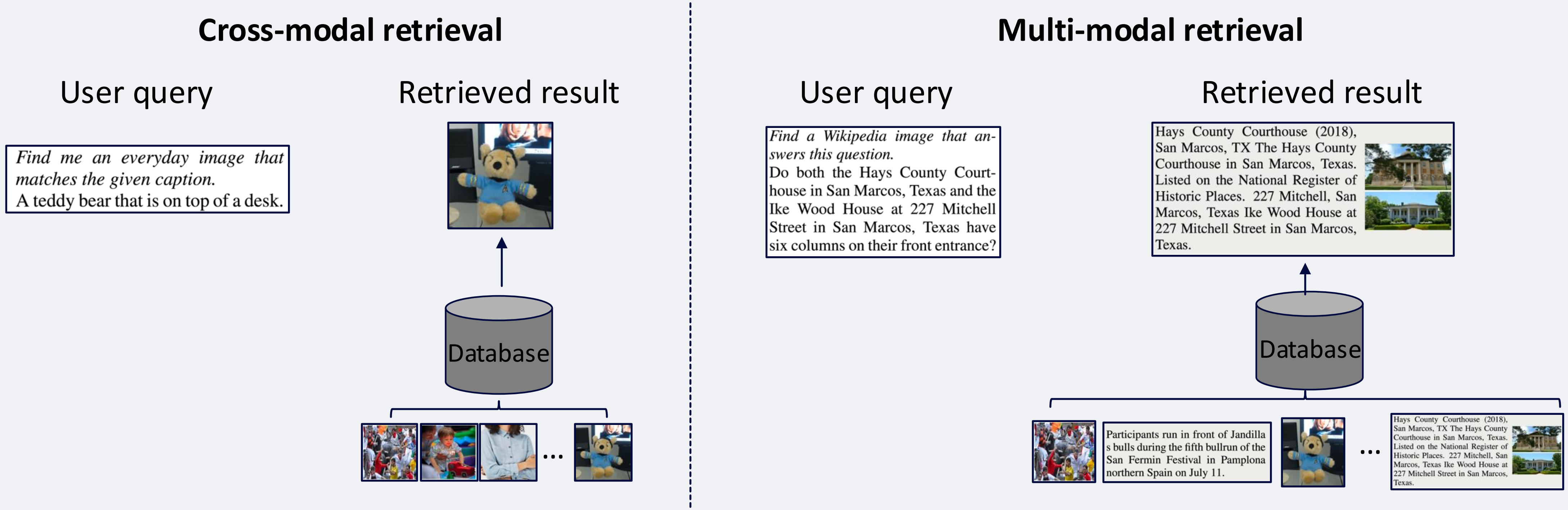
Jungsoo Lee, Janghoon Cho, Hyojin Park, Munawar Hayat, Kyuwoong Hwang, Fatih Porikli, and Sungha Choi

Qualcomm AI Research*

*Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

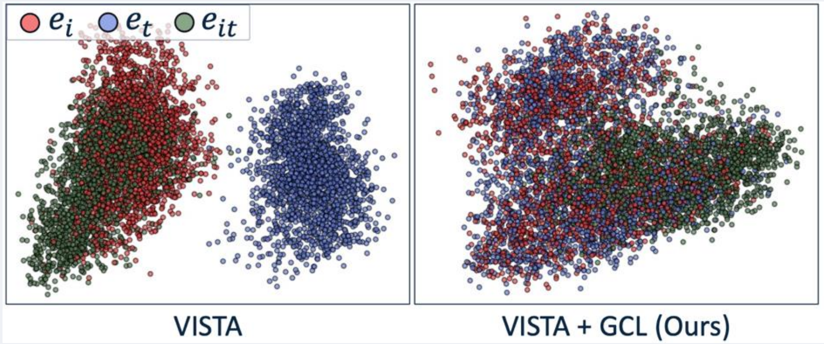
Introduction

- While cross-modal retrieval (e.g., T2I retrieval) has been widely explored, **multimodal retrieval** (e.g., T2IT retrieval) has been relatively underexplored
- Especially, retrieving samples in a **database with mixed modalities**, where image, text, and image + text (e.g., Wikipedia images with texts) are included together, remains challenging



Motivation

- The main reason behind such a performance degradation is the **modality gap** between images and texts
  - Modality gap: phenomenon where data samples that share similar semantics but belong to different modalities exhibiting low similarity
- Analyzed modality gap using recent multimodal embedding model, VISTA [1]
  - PCA visualization using embeddings of image, text, and image + text -> **clear discrepancy exists among three modalities**
  - **Cosine similarity** between averaged vectors of each modality is also **low**



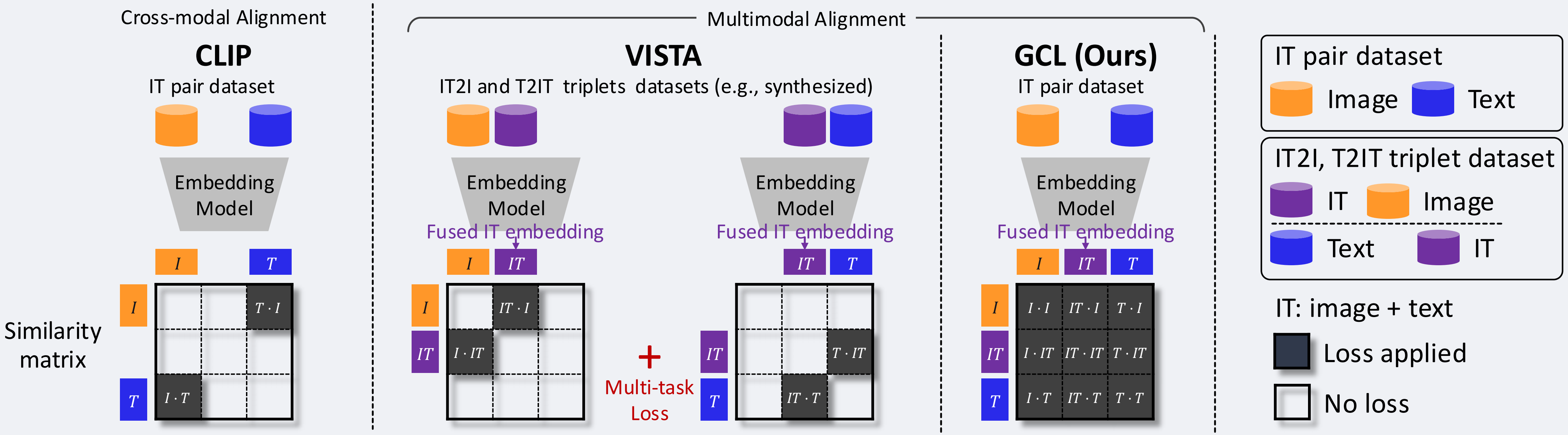
	Cosine similarity between averaged vectors of each modality		
Modality Pair	$\bar{e}_i \leftrightarrow \bar{e}_t$	$\bar{e}_i \leftrightarrow \bar{e}_{it}$	$\bar{e}_t \leftrightarrow \bar{e}_{it}$
VISTA	0.8799	0.7393	0.6772
VISTA + GCL (Ours)	**0.8882**	**0.9688**	**0.8740**

- Goal: minimize discrepancy between embeddings of each modality our proposed method

[1] VISTA: Visualized Text Embedding For Universal Multi-Modal Retrieval (ACL 2024)

Motivation

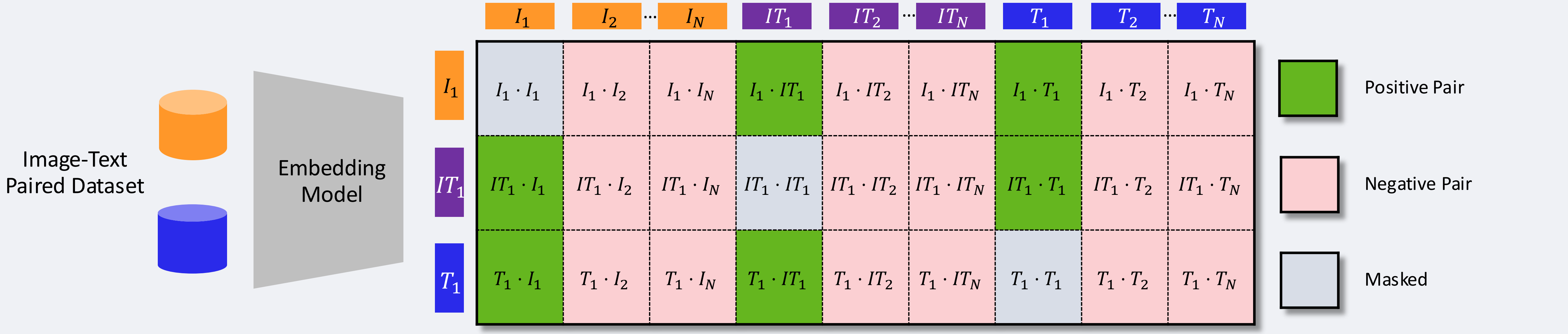
- Improving multimodal retrieval performance is also challenging due to the **scarcity of datasets including composed sets of image and texts**
  - Previous studies resorted to **generating such triplets**, which require a non-trivial amount of labor and time.
  - Additionally, models trained with these triplets might only be capable of performing **multimodal retrieval in scenarios that were seen during their training**, while showing **poor performance for unseen scenarios**
- We need a method that enables to 1) **learn retrieving any combination of modalities** 2) **without generating such datasets**



Proposed Method: Generalized Contrastive Learning

Generalized Contrastive Learning

- We propose Generalized Contrastive Learning (GCL), a loss function that improves the multimodal retrieval performance **without generating datasets with composed sets of image and texts.**
- Given that multimodal retrieval models are further finetuned from cross-modal retrieval models, we utilize **original image-to-text caption** dataset (e.g., LCS-558K) used for training cross-modal retrieval models
- Simply by imposing contrastive loss on samples with all modalities in a mini-batch, we can improve multimodal retrieval performance
- For image-text (IT) embeddings, we can either simply add the embeddings of image and texts or utilize an embedding vector extracted from a model fusing both embeddings



Proposed Method: Generalized Contrastive Learning

Generalized Contrastive Learning

- GCL loss can be formulated by extending the types of modalities in the standard contrastive learning loss function
- Notations
  - Set of modalities: $M = \{i, t, it\}$
  - Set of positive modality pairs: $P = \{(i,t), (i,it), (t,i), (t,it), (it,t), (it,i)\}$

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

Experimental settings

- Benchmarks
  - We evaluate the effectiveness of our proposed method on established multimodal retrieval benchmarks, including M-BEIR [2], MMEB [3], and the video retrieval benchmark CoVR [4]
  - These benchmarks include both cross-modal retrieval and multi-modal retrieval tasks
  - M-BEIR is composed of both global and local evaluation settings
    - Global: **All candidates across different tasks and datasets** are used
    - Local: **Only candidates from the specific dataset** are used

- Models
  - We apply GCL loss on CLIP-SF [2] and VISTA [1] and show consistent performance improvements under various benchmarks and tasks

[1] VISTA: Visualized Text Embedding For Universal Multi-Modal Retrieval (ACL 2024)
[2] UniIR: Training and Benchmarking Universal Multimodal Information Retrievers" (ECCV 2024)
[3] VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks (ICLR 2025)
[4] CoVR: Learning Composed Video Retrieval from Web Video Captions (AAAI 2024)

Quantitative Evaluation

- Cross-/multimodal Retrieval performance on MMEB and M-BEIR
  - Our framework consistently improves cross-/multi-modal retrieval performances on MMEB and M-BEIR benchmarks using CLIP-SF and VISTA

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**

M-BEIR (Global)

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**

MMEB

Quantitative Evaluation

- Improves video retrieval performance of VISTA and CLIP-SF

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**

- Apply GCL to TinyCLIP-SF even outperforms pretrained VISTA and CLIP-SF
  - This shows that GCL enables multimodal retrieval even with small number of model parameters and fast inference

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**

Thank you

Nothing in these materials is an offer to sell any of the components or devices referenced herein.

References in this presentation to “Qualcomm” may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable. Qualcomm Incorporated includes our licensing business, QTL, and the vast majority of our patent portfolio. Qualcomm Technologies, Inc., a subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of our engineering, research and development functions, and substantially all of our products and services businesses, including our QCT semiconductor business.

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries. Qualcomm patented technologies are licensed by Qualcomm Incorporated.

Follow us on:

For more information, visit us at qualcomm.com & qualcomm.com/blog