Simple but Effective Raw-Data Level Multimodal Fusion for Composed Image Retrieval

Haokun Wen[1], Xuemeng Song[2*], Xiaolin Chen[2], Yinwei Wei[3], Liqiang Nie[1*], and Tat-Seng Chua[4]

1 Harbin Institute of Technology (Shenzhen), Shenzhen, China
2 Shandong University, Qingdao, China
3 Monash University, Melbourne, Australia
4 National University of Singapore, Singapore

whenhaokun@gmail.com

Outline

- Background
- Related Work
- Motivation
- Framework
- Experiment
- Conclusion

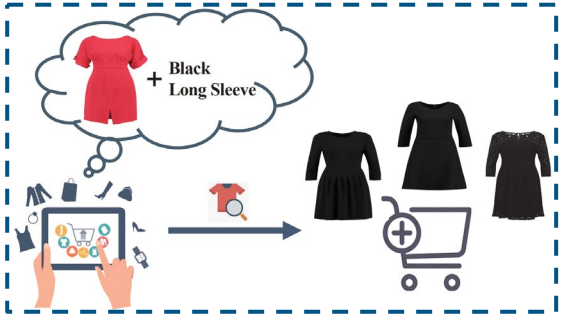
Background

Traditional single-model query-based image retrieval system cannot well deliver the user's sophisticated search intention. Composed image retrieval (CIR) allows users using the multimodal query to express the search intentions more flexibly.

- Extending the retrieval paradigm of the image retrieval systems.
- Enhancing the interaction ability of the retrieval system.
- Commercial product search.
- Interactive intelligent robot.

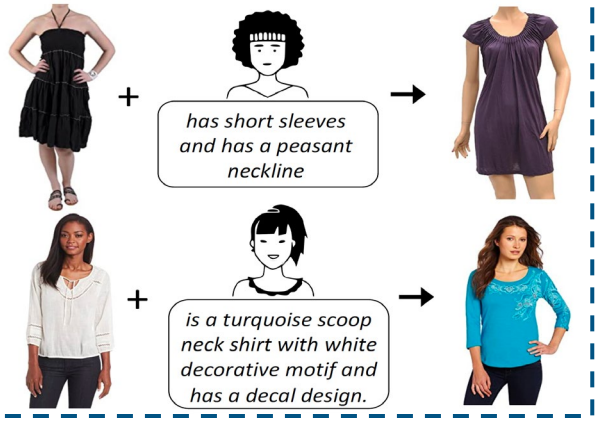
Related Work

The modification is limited to pre-defined attributes.

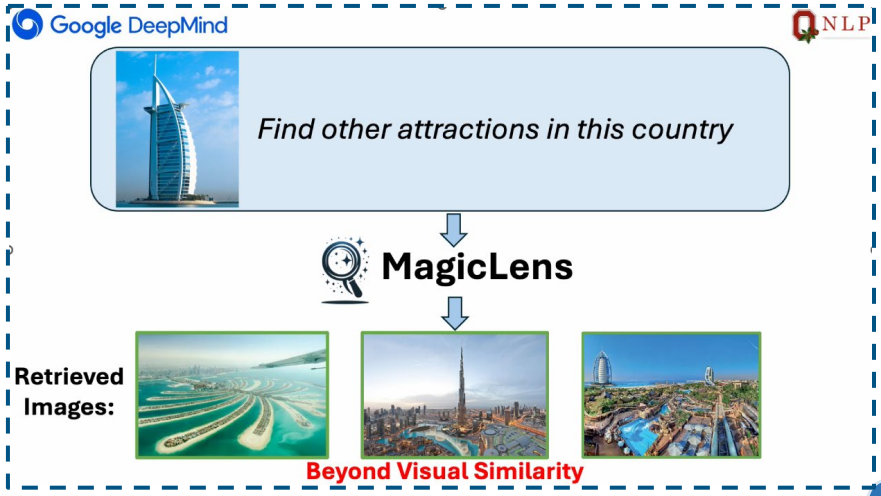


AMNet@CVPR'17

The multimodal feature extraction and cross-modal retrieval capabilities are limited.



TIRG@CVPR'19



MagicLens@ICML'24

CLIP4CIR@CVPR'22

VLP model-based method

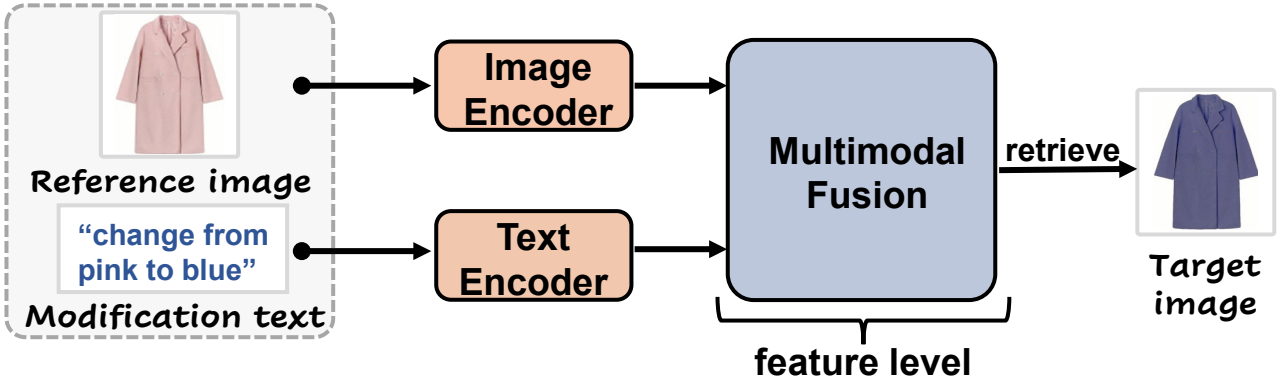
- CLIP
- BLIP
- BLIP2
- ...

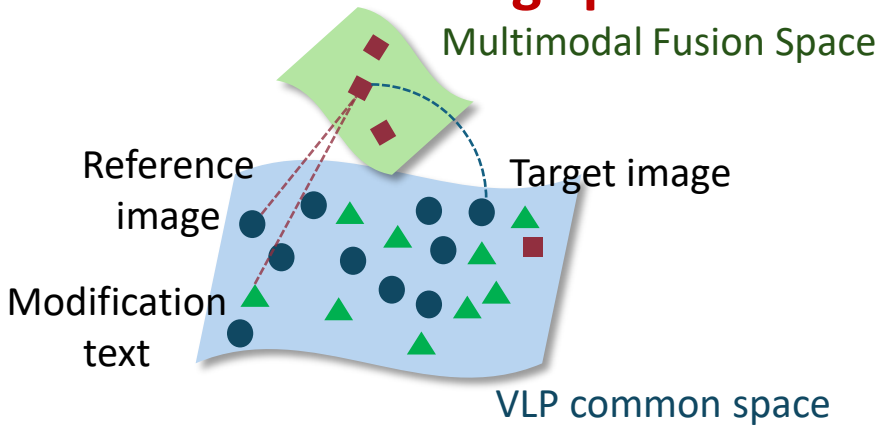
Traditional model-based method

Attribute-based CIR → **Natural language modification-based CIR** → **Open-ended instruction-based CIR**

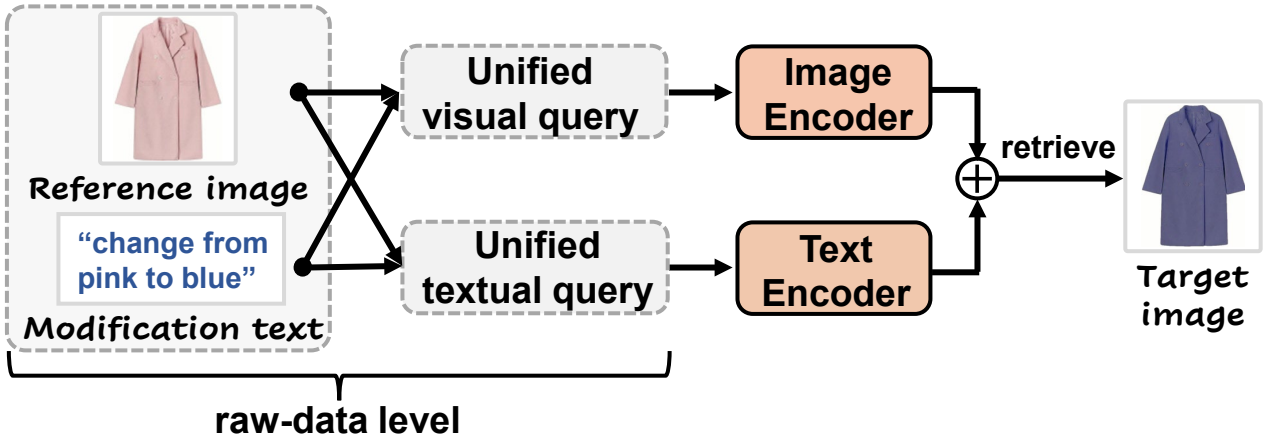
Motivation

- **Existing methods**: the nonlinear multimodal fusion function may potentially cause the fused multimodal query feature to **deviate from the original common embedding space**





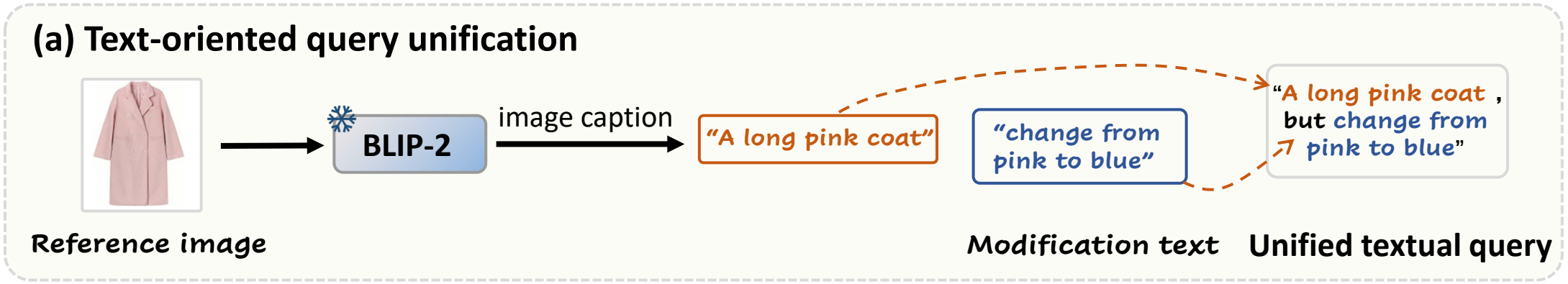
- **Our proposal**: shift the multimodal fusion **from the feature level to the raw-data level**, which can fully leverage VLP model's multimodal encoding and cross-modal retrieval capabilities.



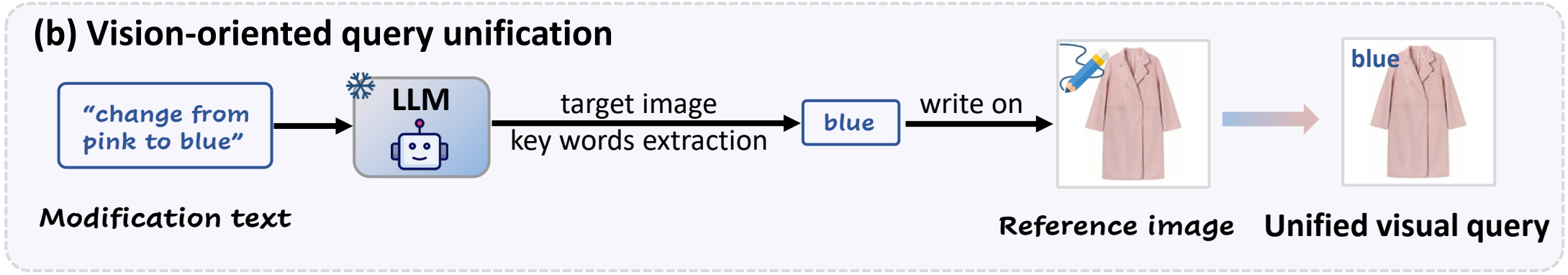
Framework

❑ Dual Query Unification-based Composed Image Retrieval framework (DQU-CIR)

Unifying the multimodal query into a pure text query:

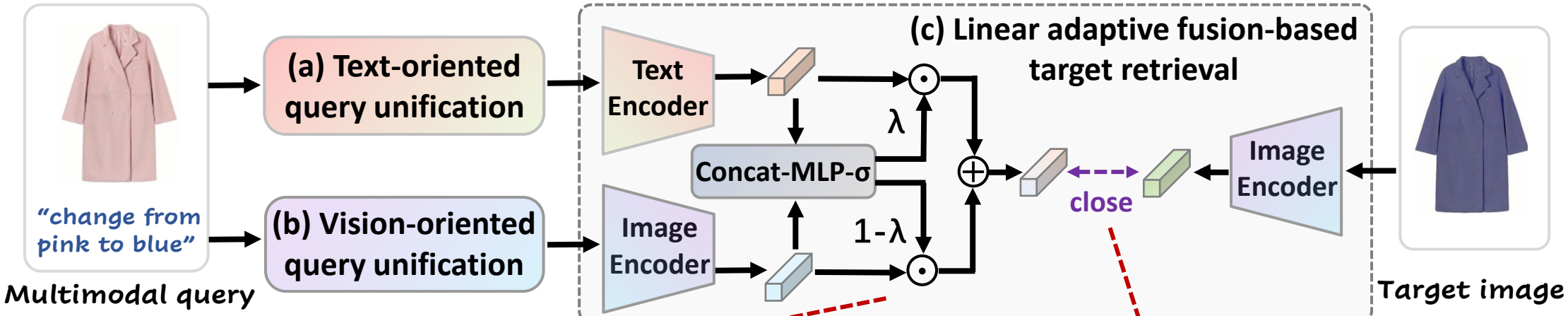


Unifying the multimodal query into an image query:



Framework

❑ Dual Query Unification-based Composed Image Retrieval framework (DQU-CIR)



$$\begin{cases} \mathbf{f}_q = \lambda * \mathbf{f}_{textual} + (1-\lambda) * \mathbf{f}_{visual}, \\ \lambda = \sigma\left(\mathrm{MLP}\left(\left[\mathbf{f}_{textual} \| \mathbf{f}_{visual}\right]\right)\right), \end{cases}$$

$$\mathcal{L} = \frac{1}{B}\sum_{i=1}^{B} -\log\left\{\frac{\exp\left\{\cos\left(\mathbf{f}_{qi}, \mathbf{f}_{ti}\right)/\tau\right\}}{\sum_{j=1}^{B}\exp\left\{\cos\left(\mathbf{f}_{qi}, \mathbf{f}_{tj}\right)/\tau\right\}}\right\}$$

Experiment

SIGIR 2024 Washington, D.C.

❑ **Performance Comparison on fashion domain: FashionIQ, Shoes, and Fashion200K**

Split	Method	Dresses		Shirts		Tops&Tees		Average		Avg.
		R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50	
VAL-Split	*Traditional Model-Based Methods*									
	TIRG [34] (CVPR'19)	14.87	34.66	18.26	37.89	19.08	39.62	17.40	37.39	12.60
	VAL [5] (CVPR'20)	21.12	42.19	21.03	43.44	25.64	49.49	22.60	45.04	16.49
	CIRPLANT [27] (ICCV'21)	17.45	40.41	17.53	38.81	21.64	45.38	18.87	41.53	30.20
	CLVC-Net [37] (SIGIR'21)	29.85	56.47	28.75	54.76	33.50	64.00	30.70	58.41	44.56
	ARTEMIS [7] (ICLR'22)	27.16	52.40	21.78	43.64	29.20	54.83	26.05	50.29	38.17
	EER [49] (TIP'22)	30.02	55.44	25.32	49.87	33.20	60.34	29.51	55.22	42.37
	CRR [48] (MM'22)	30.41	57.11	30.73	58.02	33.67	64.48	31.60	59.87	45.74
	AMC [53] (TOMM'23)	31.73	59.25	30.67	59.08	36.21	66.60	32.87	61.64	47.26
	CRN [44] (TIP'23)	32.67	59.30	30.27	56.97	37.74	65.94	33.56	60.74	47.15
	CMAP [21] (TOMM'24)	36.44	64.25	34.83	60.06	41.79	69.12	37.64	64.42	51.03
	VLP Model-Based Methods									
	Prog. Lrn. [51] (SIGIR'22)	38.18	64.50	48.63	71.54	52.32	76.90	46.37	70.98	58.68
	TG-CIR [39] (MM'23)	45.22	69.66	52.60	72.52	56.14	77.10	51.32	73.09	62.21
	LIMN+ [38] (TPAMI'24)	52.11	75.21	57.51	77.92	62.67	82.66	57.43	78.60	68.02
	SPIRIT [6] (TOMM'24)	43.83	68.86	52.50	74.19	56.60	79.25	50.98	74.10	62.54
	DQU-CIR	$57.63^{\pm 0.24}$	$78.56^{\pm 0.50}$	$62.14^{\pm 0.66}$	$80.38^{\pm 0.15}$	$66.15^{\pm 0.50}$	$85.73^{\pm 0.25}$	$61.97^{\pm 0.28}$	$81.56^{\pm 0.22}$	$71.77^{\pm 0.17}$
Original-Split	*Traditional Model-Based Methods*									
	TIRG [34] (CVPR'19)	14.13	34.61	13.10	30.91	14.79	34.37	14.01	33.30	23.66
	ARTEMIS [7] (ICLR'22)	25.68	51.05	21.57	44.13	28.59	55.06	25.28	50.08	37.68
	VLP Model-Based Methods									
	CLIP4CIR [1] (CVPR'22)	31.63	56.67	36.36	58.00	38.19	62.42	35.39	59.03	47.21
	Prog. Lrn. [51] (SIGIR'22)	33.60	58.90	39.45	61.78	43.96	68.33	39.02	63.00	51.01
	FAME-ViL [11] (CVPR'23)	42.19	67.38	47.64	68.79	50.69	73.07	46.84	69.75	58.30
	SPIRIT [6] (TOMM'24)	39.86	64.30	44.11	65.60	47.68	71.70	43.88	67.20	55.54
	BLIP4CIR+Bi [28] (WACV'24)	42.09	67.33	41.76	64.28	46.61	70.32	43.49	67.31	55.40
	DQU-CIR	$51.90^{\pm 0.64}$	$74.37^{\pm 0.39}$	$53.57^{\pm 0.27}$	$73.21^{\pm 0.34}$	$58.48^{\pm 0.46}$	$79.23^{\pm 0.29}$	$54.65^{\pm 0.38}$	$75.60^{\pm 0.18}$	$65.13^{\pm 0.14}$

Experiment

SIGIR 2024 Washington, D.C.

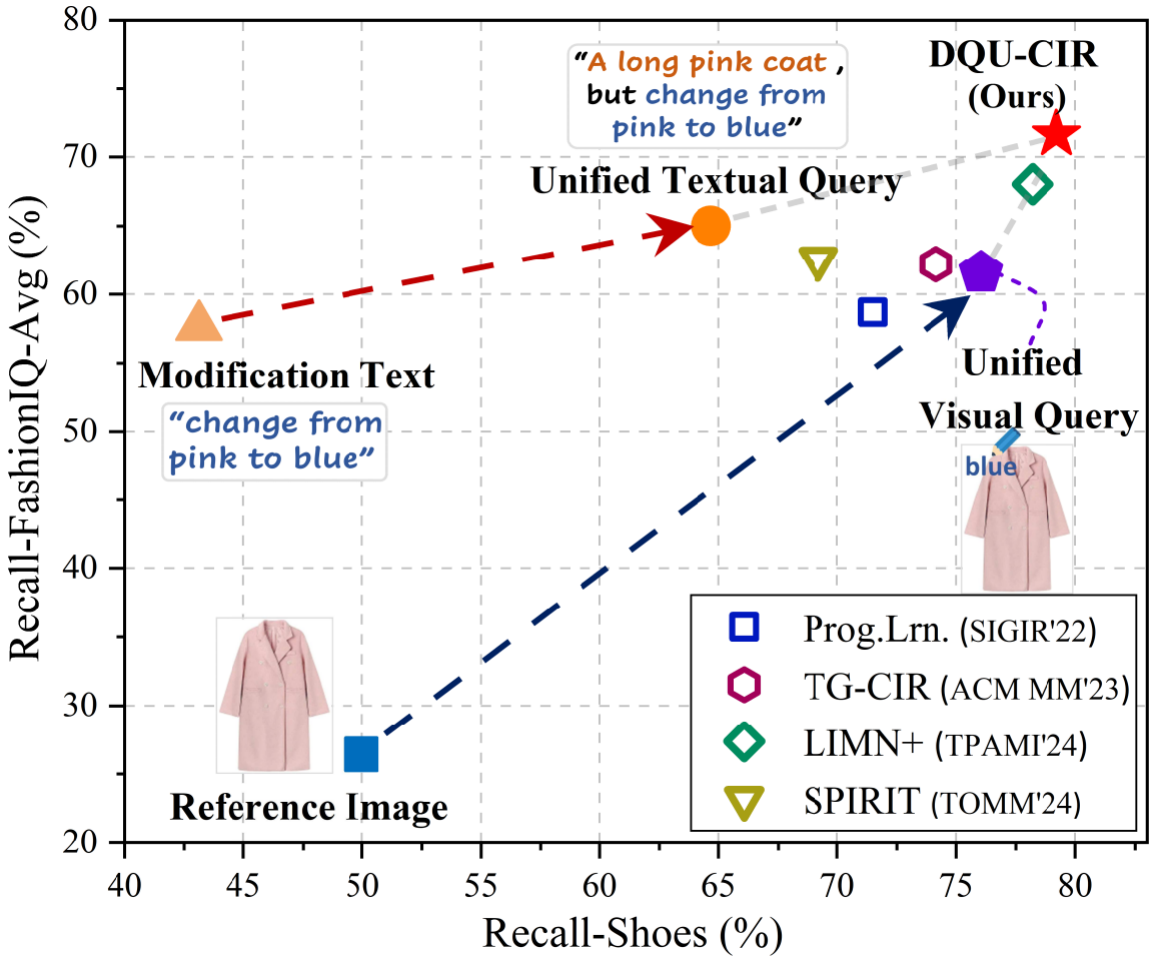
❑ **Performance Comparison on fashion domain: FashionIQ, Shoes, and Fashion200K**

Method	R@1	R@10	R@50	Avg.
Traditional Model-Based Methods				
TIRG [34] (CVPR'19)	12.60	45.45	69.39	42.48
VAL [5] (CVPR'20)	16.49	49.12	73.53	46.38
CLVC-Net [37] (SIGIR'21)	17.64	54.39	79.47	50.50
ARTEMIS [7] (ICLR'22)	18.72	53.11	79.31	50.38
EER [49] (TIP'22)	20.05	56.02	79.94	52.00
CRR [48] (MM'22)	18.41	56.38	79.92	51.57
AMC [53] (TOMM'23)	19.99	56.89	79.27	52.05
CRN [44] (TIP'23)	18.92	54.55	80.04	51.17
CMAP [21] (TOMM'24)	21.48	56.18	81.14	52.93
VLP Model-Based Methods				
Prog. Lrn. [51] (SIGIR'22)	22.88	58.83	84.16	55.29
TG-CIR [39] (MM'23)	25.89	63.20	85.07	58.05
LIMN+ [38] (TPAMI'24)	–	68.37	88.07	–
SPIRIT [6] (TOMM'24)	–	56.90	81.49	–
DQU-CIR	$31.47^{\pm 1.31}$	$69.19^{\pm 0.99}$	$88.52^{\pm 0.31}$	$63.06^{\pm 0.69}$

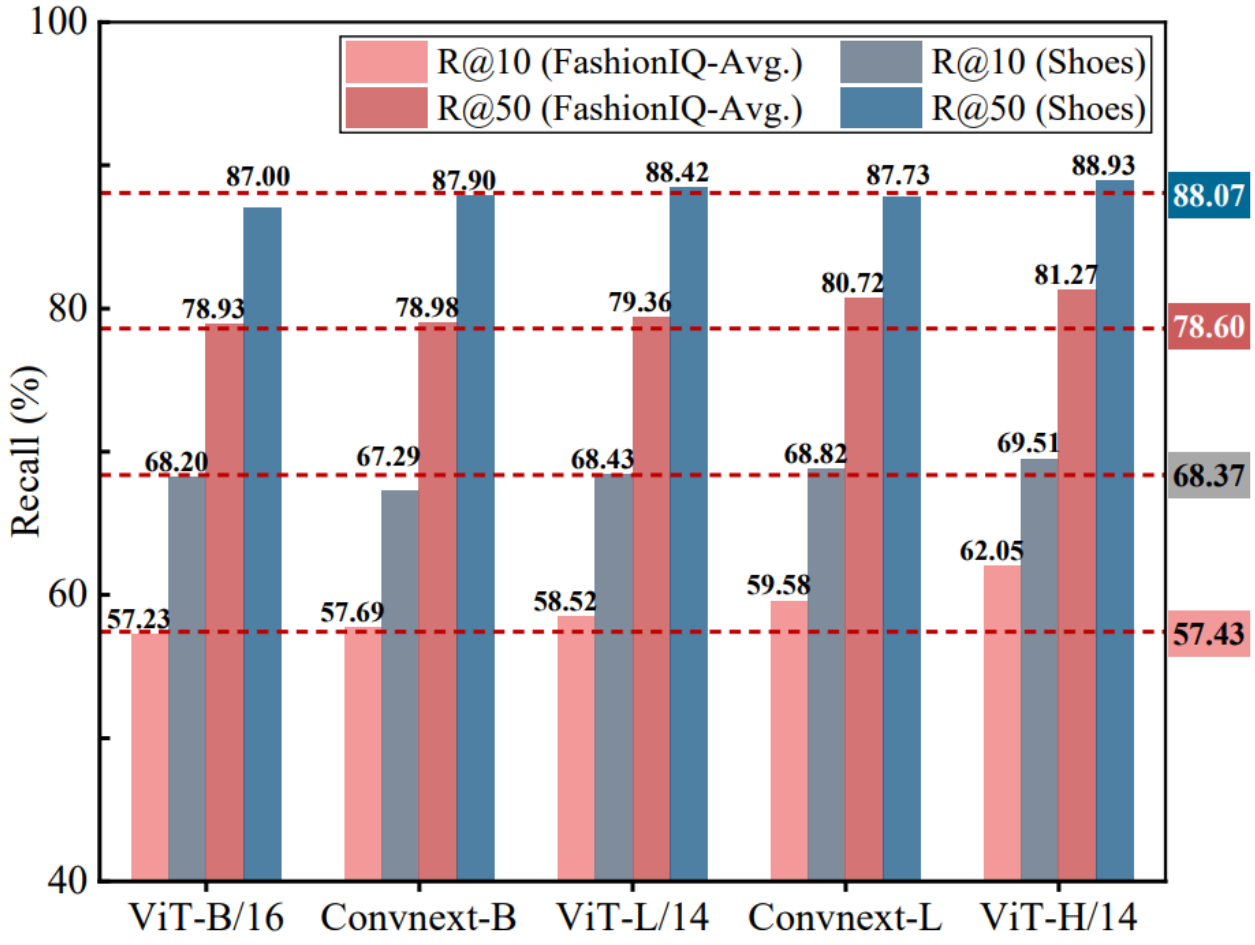
Method	R@1	R@10	R@50	Avg.
Traditional Model-Based Methods				
TIRG [34] (CVPR'19)	14.1	42.5	63.8	40.1
VAL [5] (CVPR'20)	22.9	50.8	72.7	48.8
CLVC-Net [37] (SIGIR'21)	22.6	53.0	72.2	49.3
ARTEMIS [7] (ICLR'22)	21.5	51.1	70.5	47.7
EER [49] (TIP'22)	–	55.3	73.4	–
CRR [48] (MM'22)	24.9	56.4	73.6	51.6
CRN [44] (TIP'23)	–	53.5	74.5	–
CMAP [21] (TOMM'24)	24.2	56.9	75.3	52.1
VLP Model-Based Methods				
LIMN [38] (TPAMI'24)	–	57.2	76.6	–
SPIRIT [6] (TOMM'24)	–	55.2	73.6	–
DQU-CIR	$36.8^{\pm 3.8}$	$67.9^{\pm 2.1}$	$87.8^{\pm 0.3}$	$64.1^{\pm 1.7}$

Experiment

❑ Intuitive performance comparison



❑ Performance with different backbones



Experiment

❑ Case Study

	Multimodal Query	Unified Textual Query	Unified Visual Query	Retrieved Images (Top 5)
(a) FashionIQ	 + "has a red belt and less revealing and has more black and white"	"a woman in a black and white dress, but has a red belt and less revealing and has more black and white" λ=0.66	 black white less revealing red belt 1-λ=0.34	
(b) Shoes	 + "is darker in brown with plainer snake like pattern"	"a women's clogger with colorful flowers on it, but is darker in brown with plainer snake like pattern" λ=0.41	 darker brown plainer snake-like pattern 1-λ=0.59	
(c) CIRR	 + "change focus onto a singular wild wolf, must be in full profile view and looking to the left"	"a lion and hyenas are fighting in the wild, but change focus onto a singular wild wolf, must be in full profile view and looking to the left" λ=0.54	 singular wild wolf full profile view looking to the left 1-λ=0.46	DQU-CIR  TQU-CIR 

Conclusion

- We designed **two training-free multimodal fusion methods at the raw-data level** in the context of CIR, which can fully leverage the VLP model's multimodal encoding and cross-modal retrieval capabilities.
- We surprisingly found that **directly writing descriptive words onto the image** can achieve promising multimodal fusion results, which indicates the superior **OCR potential of the image encoder** of the VLP model. We believe this would inspire the multimodal learning community to approach multimodal fusion from a new perspective.
- Extensive experiments on four real-world datasets demonstrate the superiority of our method over the SOTA baselines.

Thanks for your listening!

Codes are available!