Target-Guided Composed Image Retrieval

Haokun Wen[1], Xian Zhang[1], Xuemeng Song[2*], Yinwei Wei[3], and Liqiang Nie[1*]

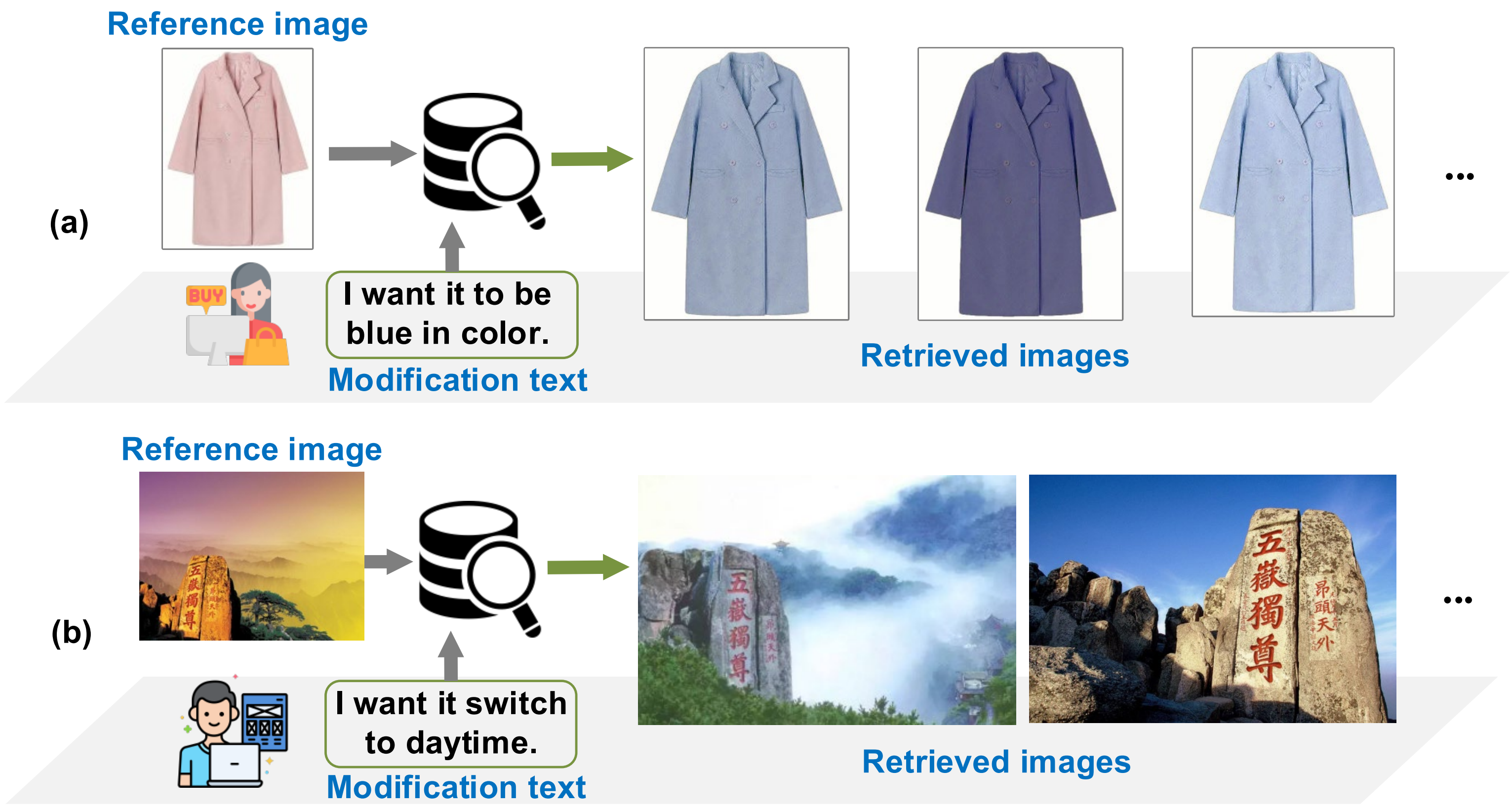
1 Harbin Institute of Technology (Shenzhen), Shenzhen, China
2 Shandong University, Qingdao, China
3 Monash University, Melbourne, Australia

whenhaokun@gmail.com

Outline

1. Background

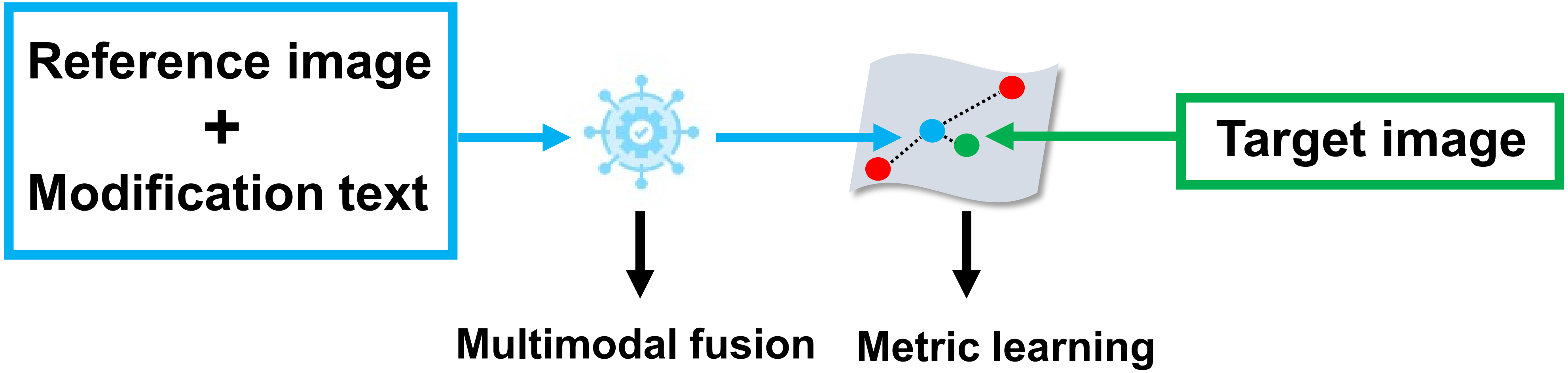
Traditional single-model query-based image retrieval system cannot well deliver the user's sophisticated search intention. Composed image retrieval (CIR) allows users using the multimodal query to express the search intentions more flexibly.



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

1. Background

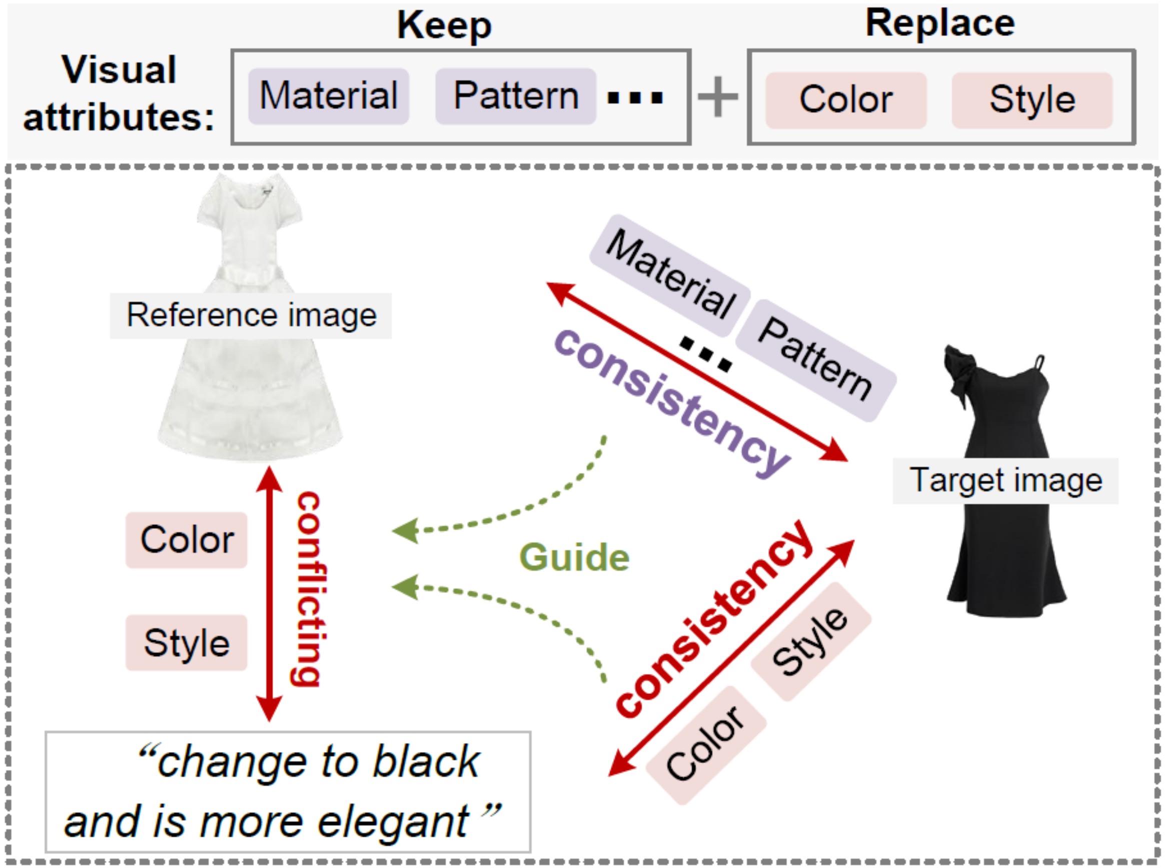
- Composed Image Retrieval (CIR)



The key to CIR lies in two key points:
(1) Multimodal fusion for accurately capturing the user’s search intention;
(2) Metric learning for accurately ranking the candidate images.

2. Motivation

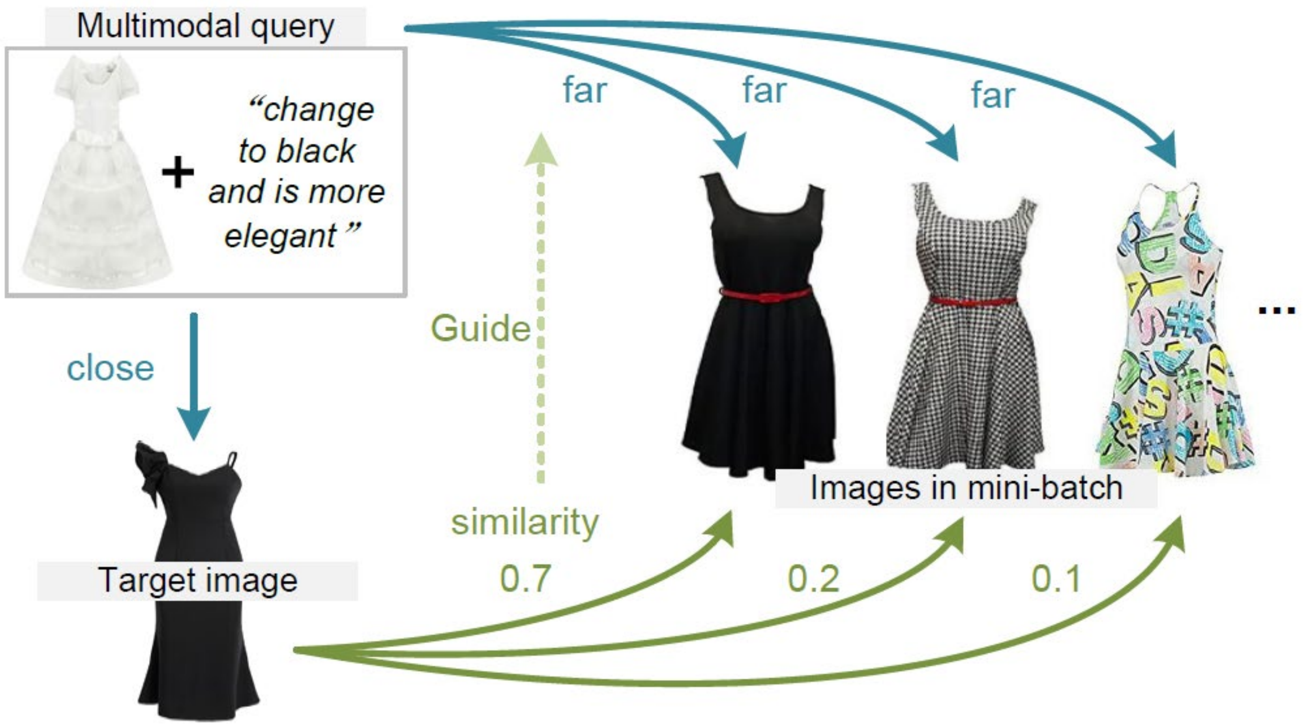
- **On multimodal fusion:** Existing methods ignore the intrinsic conflicting relationship between the multimodal query.



Leverage the target-query relationship to model the conflicting relationship

2. Motivation

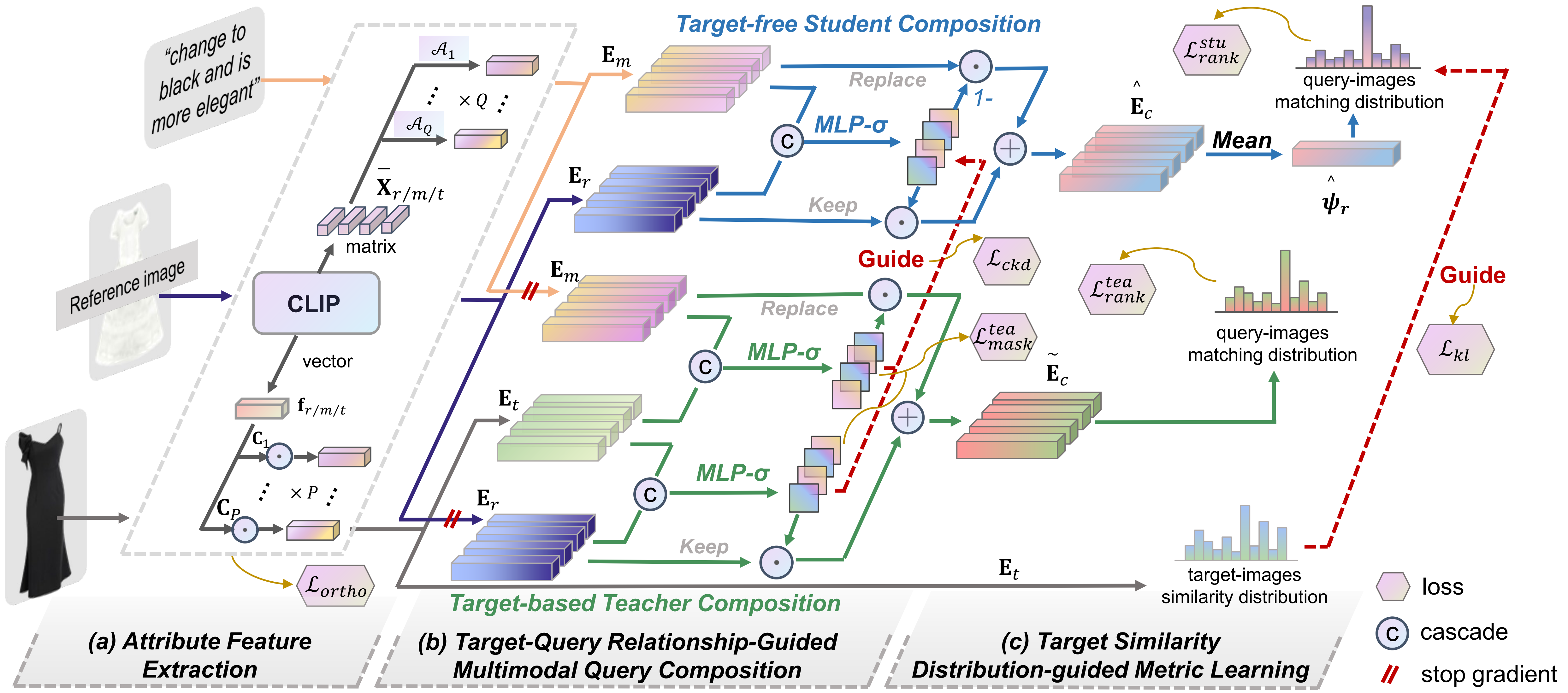
- **On metric learning:** The widely-used batch-based classification loss can affect the metric learning process.



Leverage the target visual similarity to promote the metric learning

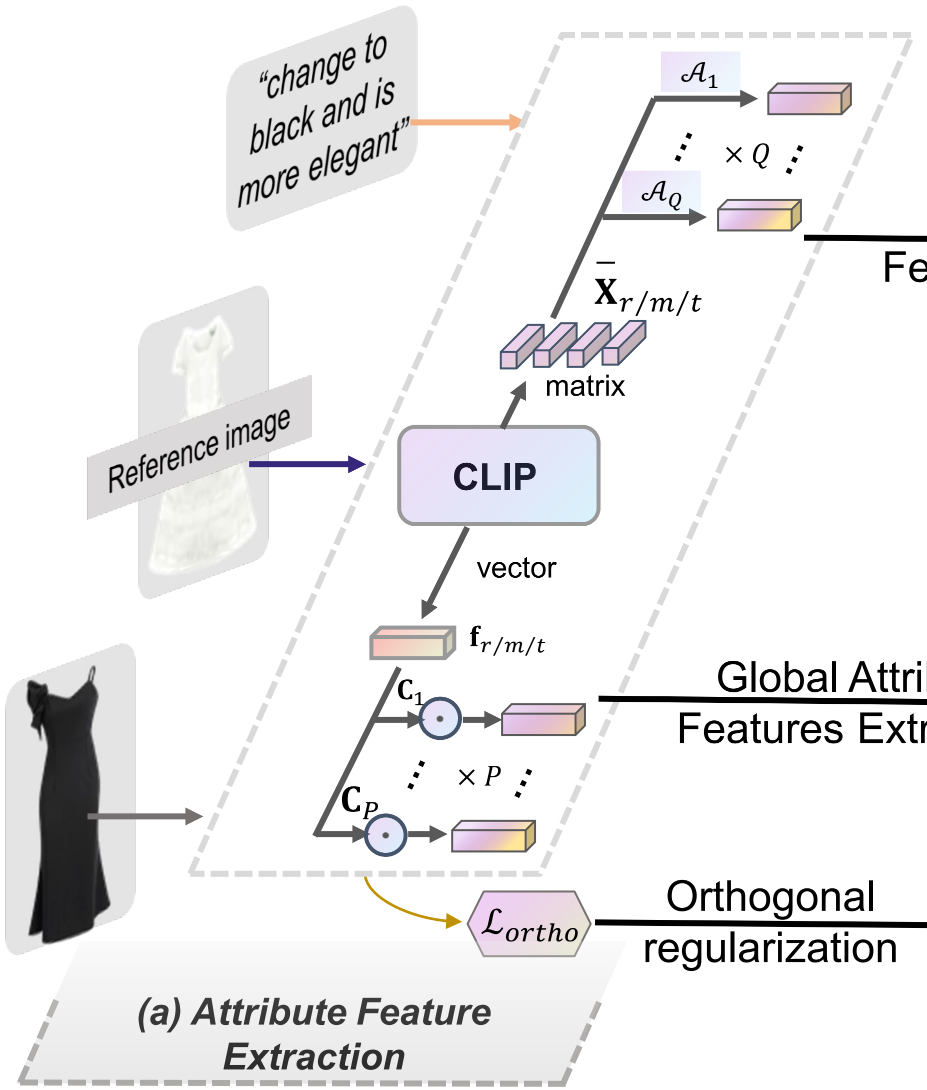
3. Framework

➢ Target-Guided Composed Image Retrieval network (TG-CIR)



3. Framework

- **Target-Guided Composed Image Retrieval network (TG-CIR)**



Local Attribute Features Extraction

$$\begin{cases} \mathbf{s}_r^j = \sigma\left(\mathrm{Conv}^j\left(\overline{\mathbf{X}}_r\right)\right), \\ \mathbf{v}_r^j = \mathbf{s}_r^{j\top} \otimes \overline{\mathbf{X}}_r, \end{cases}$$

Global Attribute Features Extraction

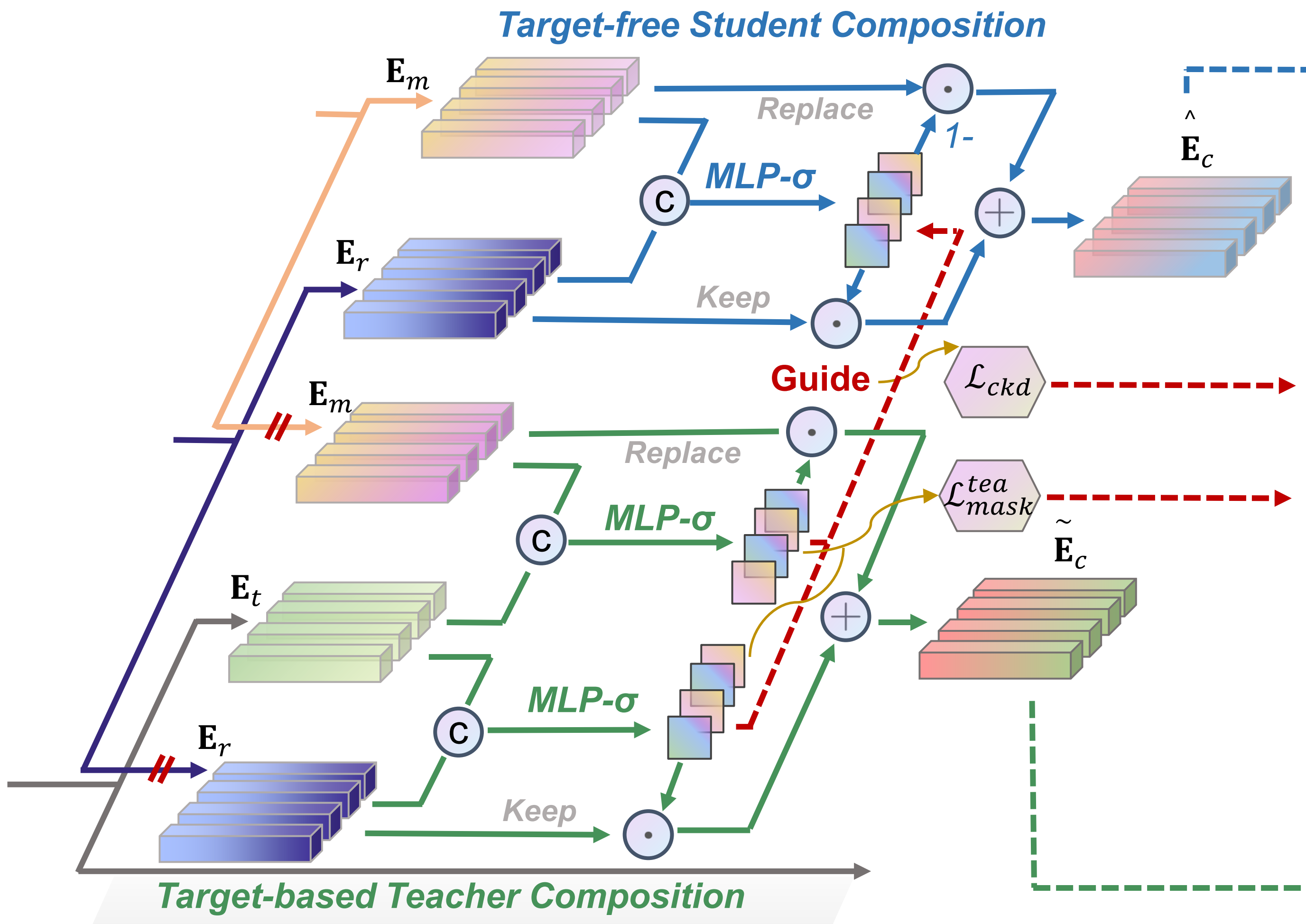
$$\mathbf{u}_r^i = \mathbf{f}_r \odot \mathbf{C}_i$$

Orthogonal regularization

$$\mathcal{L}_{ortho} = \left\|\mathbf{E}_r\mathbf{E}_r^\top - \mathbf{I}\right\|_F^2 + \left\|\mathbf{E}_m\mathbf{E}_m^\top - \mathbf{I}\right\|_F^2 + \left\|\mathbf{E}_t\mathbf{E}_t^\top - \mathbf{I}\right\|_F^2,$$

3. Framework

➢ **Target-Guided Composed Image Retrieval network (TG-CIR)**



(b) Target-Query Relationship-Guided Multimodal Query Composition

$$\begin{cases} \hat{\mathbf{m}}_k = \sigma\left(\mathrm{MLP}_s\left(\boxed{[\mathbf{E}_r, \mathbf{E}_m]}\right)\right), \\ \hat{\mathbf{m}}_r = 1 - \hat{\mathbf{m}}_k, \end{cases}$$

$$\hat{\mathbf{E}}_c = \hat{\mathbf{m}}_k \odot \mathbf{E}_r + \hat{\mathbf{m}}_r \odot \mathbf{E}_m.$$

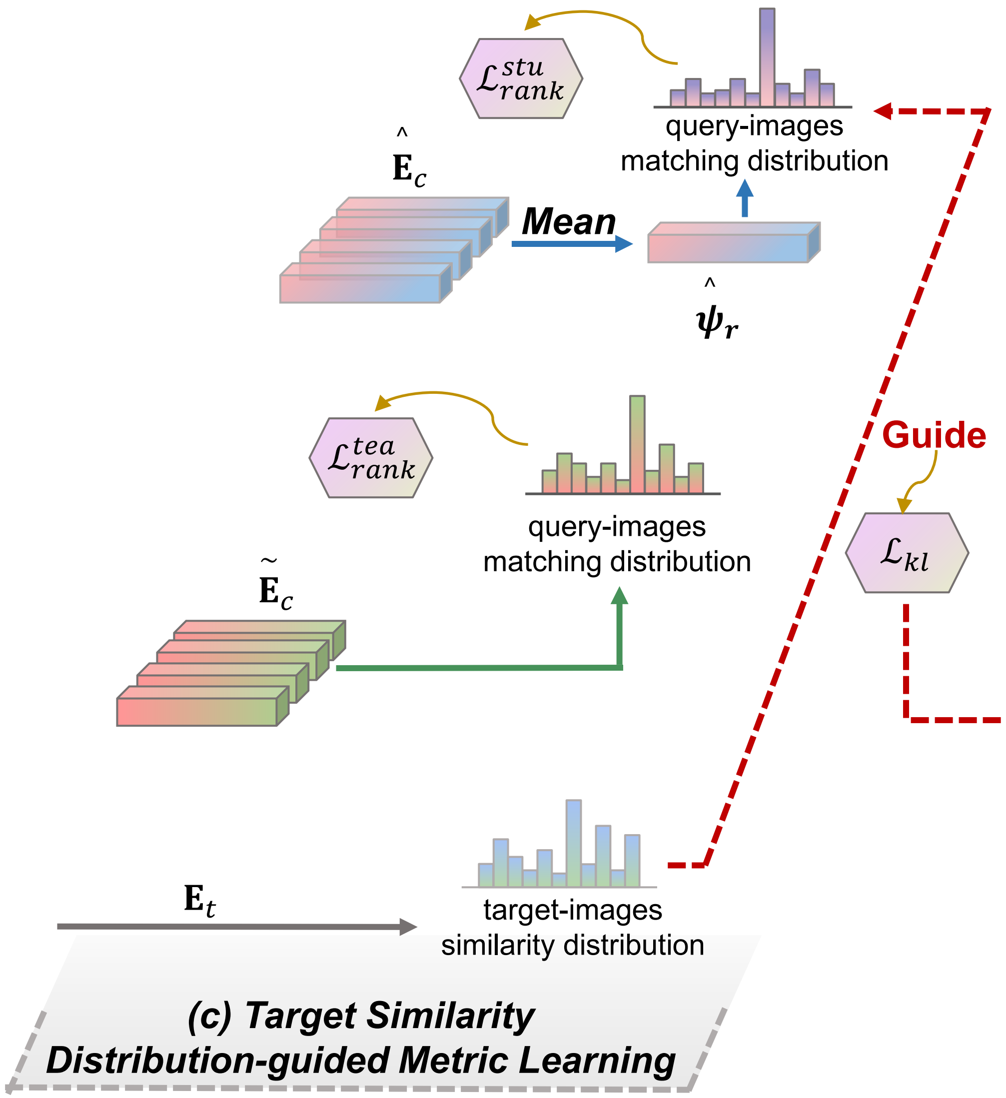
$$\mathcal{L}_{ckd} = \|\tilde{\mathbf{m}}_k - \hat{\mathbf{m}}_k\|^2 + \|\tilde{\mathbf{m}}_r - \hat{\mathbf{m}}_r\|^2.$$

$$\mathcal{L}^{tea}_{mask} = \|\tilde{\mathbf{m}}_r, 1 - \tilde{\mathbf{m}}_k\|^2$$

$$\begin{cases} \tilde{\mathbf{m}}_k = \sigma\left(\mathrm{MLP}_{t1}\left([\boxed{\mathbf{E}_t}, \mathbf{E}_r]\right)\right), \\ \tilde{\mathbf{m}}_r = \sigma\left(\mathrm{MLP}_{t2}\left([\boxed{\mathbf{E}_t}, \mathbf{E}_m]\right)\right), \\ \tilde{\mathbf{E}}_c = \tilde{\mathbf{m}}_k \odot \mathbf{E}_r + \tilde{\mathbf{m}}_r \odot \mathbf{E}_m, \end{cases}$$

3. Framework

➢ Target-Guided Composed Image Retrieval network (TG-CIR)



$$
\begin{cases}
\mathcal{L}_{rank}^{tea} = \frac{1}{B}\sum_{i=1}^{B} -\log\left\{\frac{\exp\left\{\left\{\sum_{k=1}^{K} \mathrm{s}\left(\tilde{\mathbf{E}}_{ci}[k], \mathbf{E}_{ti}[k]\right)\right\}/\tau\right\}}{\sum_{j=1}^{B}\exp\left\{\left\{\sum_{k=1}^{K} \mathrm{s}\left(\tilde{\mathbf{E}}_{ci}[k], \mathbf{E}_{tj}[k]\right)\right\}/\tau\right\}}\right\} \\
\mathcal{L}_{rank}^{stu} = \frac{1}{B}\sum_{i=1}^{B} -\log\left\{\frac{\exp\left\{\mathrm{s}\left(\hat{\psi}_{ci}, \psi_{ti}\right)/\tau\right\}}{\sum_{j=1}^{B}\exp\left\{\mathrm{s}\left(\hat{\psi}_{ci}, \psi_{tj}\right)/\tau\right\}}\right\},
\end{cases}
$$

$$
\mathcal{L}_{kl} = \frac{1}{B}\sum_{i=1}^{B} D_{KL}\left(\mathbf{p}_i^t \| \mathbf{p}_i^c\right) = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B} p_{ij}^t \log\frac{p_{ij}^t}{p_{ij}^c}
$$

4. Experiment

➢ Performance comparison on FashionIQ and Shoes

Method	FashionIQ								Shoes			
	Dresses		Shirts		Tops&Tees		Avg		R@1	R@10	R@50	Avg
	R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50				
TIRG [32] (CVPR'19)	14.87	34.66	18.26	37.89	19.08	39.62	17.40	37.39	12.60	45.45	69.39	42.48
VAL [5] (CVPR'20)	21.12	42.19	21.03	43.44	25.64	49.49	22.60	45.04	16.49	49.12	73.53	46.38
CIRPLANT [24] (ICCV'21)	17.45	40.41	17.53	38.81	21.64	45.38	18.87	41.53	–	–	–	–
CosMo [21] (CVPR'21)	25.64	50.30	24.90	49.18	29.21	57.46	26.58	52.31	16.72	48.36	75.64	46.91
DATIR [11] (ACM MM'21)	21.90	43.80	21.90	43.70	27.20	51.60	23.70	46.40	17.20	51.10	75.60	47.97
MCR [38] (ACM MM'21)	26.20	51.20	22.40	46.00	29.70	56.40	26.10	51.20	17.85	50.95	77.24	48.68
CLVC-Net [35] (SIGIR'21)	29.85	56.47	28.75	54.76	33.50	64.00	30.70	58.41	17.64	54.39	79.47	50.50
ARTEMIS [7] (ICLR'22)	27.16	52.40	21.78	43.64	29.20	54.83	26.05	50.29	18.72	53.11	79.31	50.38
EER [37] (TIP'22)	30.02	55.44	25.32	49.87	33.20	60.34	29.51	55.22	<u>20.05</u>	56.02	<u>79.94</u>	52.00
FashionVLP [9] (CVPR'22)	32.42	60.29	31.89	58.44	38.51	68.79	34.27	62.51	–	49.08	77.32	–
CRR [36] (ACM MM'22)	30.41	57.11	30.73	58.02	33.67	64.48	31.60	59.87	18.41	56.38	79.92	51.57
AMC [41] (TOMM'23)	31.73	59.25	30.67	59.08	36.21	66.60	32.87	61.64	19.99	<u>56.89</u>	79.27	<u>52.05</u>
Clip4cir [1] (CVPRW'22)	33.81	59.40	39.99	60.45	41.41	65.37	38.32	61.74	–	–	–	–
FAME-ViL[17] (CVPR'23)	<u>42.19</u>	<u>67.38</u>	<u>47.64</u>	<u>68.79</u>	<u>50.69</u>	<u>73.07</u>	<u>46.84</u>	<u>69.75</u>	–	–	–	–
TG-CIR	**45.22**	**69.66**	**52.60**	**72.52**	**56.14**	**77.10**	**51.32**	**73.09**	**25.89**	**63.20**	**85.07**	**58.05**
Improvement(%)	↑ 7.18	↑ 3.38	↑ 10.41	↑ 5.42	↑ 10.75	↑ 5.52	↑ 9.56	↑ 4.79	↑ 29.13	↑ 11.09	↑ 6.42	↑ 11.53

4. Experiment

➢ Performance comparison on CIRR

Method	R@k				$R_{subset}@k$			Avg
	$k=1$	$k=5$	$k=10$	$k=50$	$k=1$	$k=2$	$k=3$	
TIRG [32] (CVPR'19)	14.61	48.37	64.08	90.03	22.67	44.97	65.14	35.52
ARTEMIS [7] (ICLR'22)	16.96	46.10	61.31	87.73	39.99	62.20	75.67	43.05
CIRPLANT [24] (ICCV'21)	15.18	43.36	60.48	87.64	33.81	56.99	75.40	38.59
Clip4cir [1] (CVPRW'22)	<u>38.53</u>	<u>69.98</u>	<u>81.86</u>	<u>95.93</u>	<u>68.19</u>	<u>85.64</u>	<u>94.17</u>	<u>69.09</u>
TG-CIR	**45.25**	**78.29**	**87.16**	**97.30**	**72.84**	**89.25**	**95.13**	**75.57**
Improvement(%)	↑ 17.44	↑ 11.87	↑ 6.47	↑ 1.43	↑ 6.82	↑ 4.22	↑ 1.02	↑ 9.38

4. Experiment

➢ Ablation study

Method	FashionIQ-Avg		Shoes	CIRR
	R@10	R@50	Avg	Avg
Local-AttriFea_Only	41.92	67.37	52.35	55.68
Global-AttriFea_Only	49.50	72.89	56.31	74.50
w/o_ortho	50.77	72.23	57.50	75.04
w/o_target_guide	48.84	72.28	55.80	72.62
w/o_target_guide_c	50.00	72.82	55.84	74.17
w/o_target_guide_m	48.95	72.31	56.33	74.58
TG-CIR	**51.32**	**73.09**	**58.05**	**75.57**

5. Conclusion

1. We propose a **target-query relationship-guided multimodal query composition module** with the “keep-and-replace” paradigm.

2. We propose a batch-based **target similarity-guided matching degree regularization** that can improve the performance of metric learning for CIR.

3. We propose an attribute feature extraction module, which can extract **unified attribute features** of the three elements of the CIR task from both **local** and **global** perspectives, to facilitate the conflicting relationship modeling

Q&A

Thanks for your listening!

Codes are available!