Learning with Noisy Triplet Correspondence for Composed Image Retrieval

By XLearning Group

Code

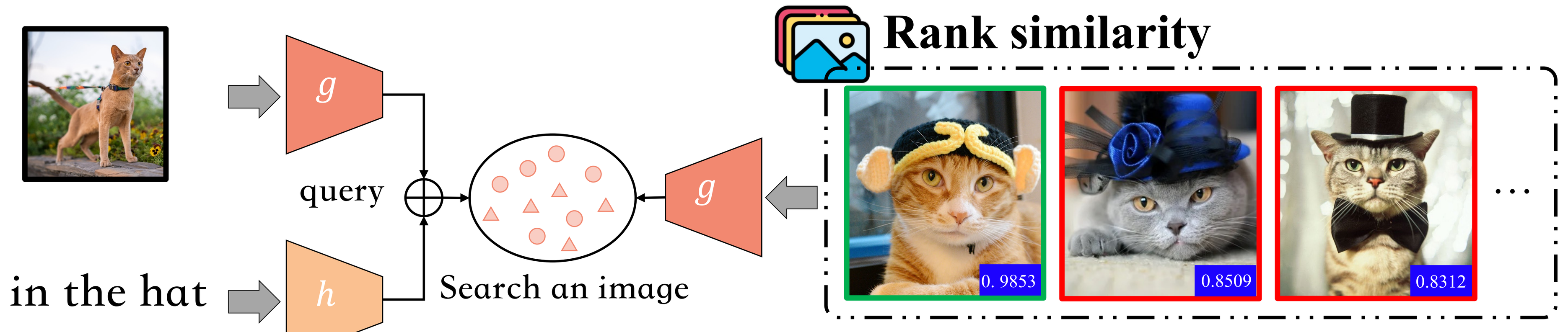
Shuxian Li[1*], Changhao He[1*], Xiting Liu[2], Joey Tianyi Zhou[3], Xi Peng[1,4], Peng Hu[1†]

[1] Sichuan University, China. [2] Georgia Institute of Technology, USA. [3] CFAR, IHPC, A*STAR Singapore.
[4] National Key Laboratory of Fundamental Algorithms and Models for Engineering Simulation, China.

WeChat

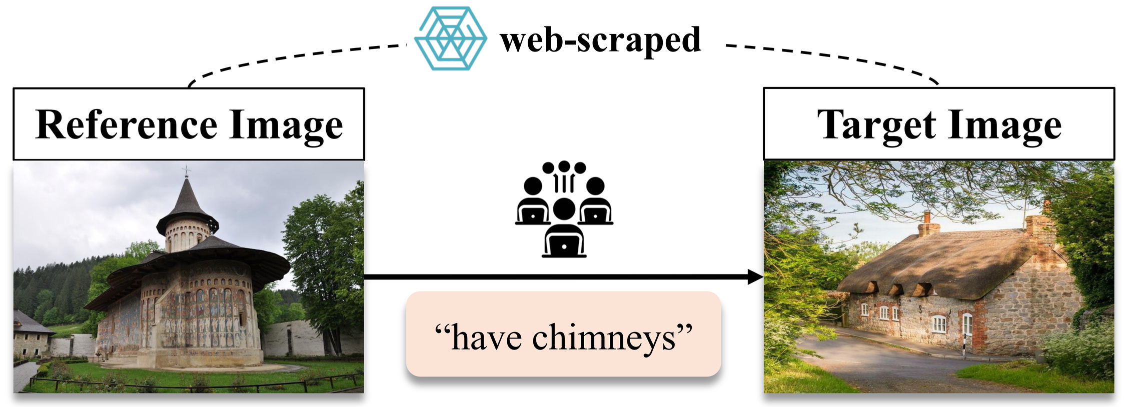
Task

Task: Given a **reference image** and a **modification text**, find the **target image** aligned with the **dual** intention.



Challenge

- Imperfection annotators inevitably introduce noise into the trainset, resulting in Partial Matching and Incorrect Matching triplets—we refer to them as **Noisy Triplet Correspondence (NTC).**
- Existing methods ignore the NTC problem, leading to overfitting and performance degradation.



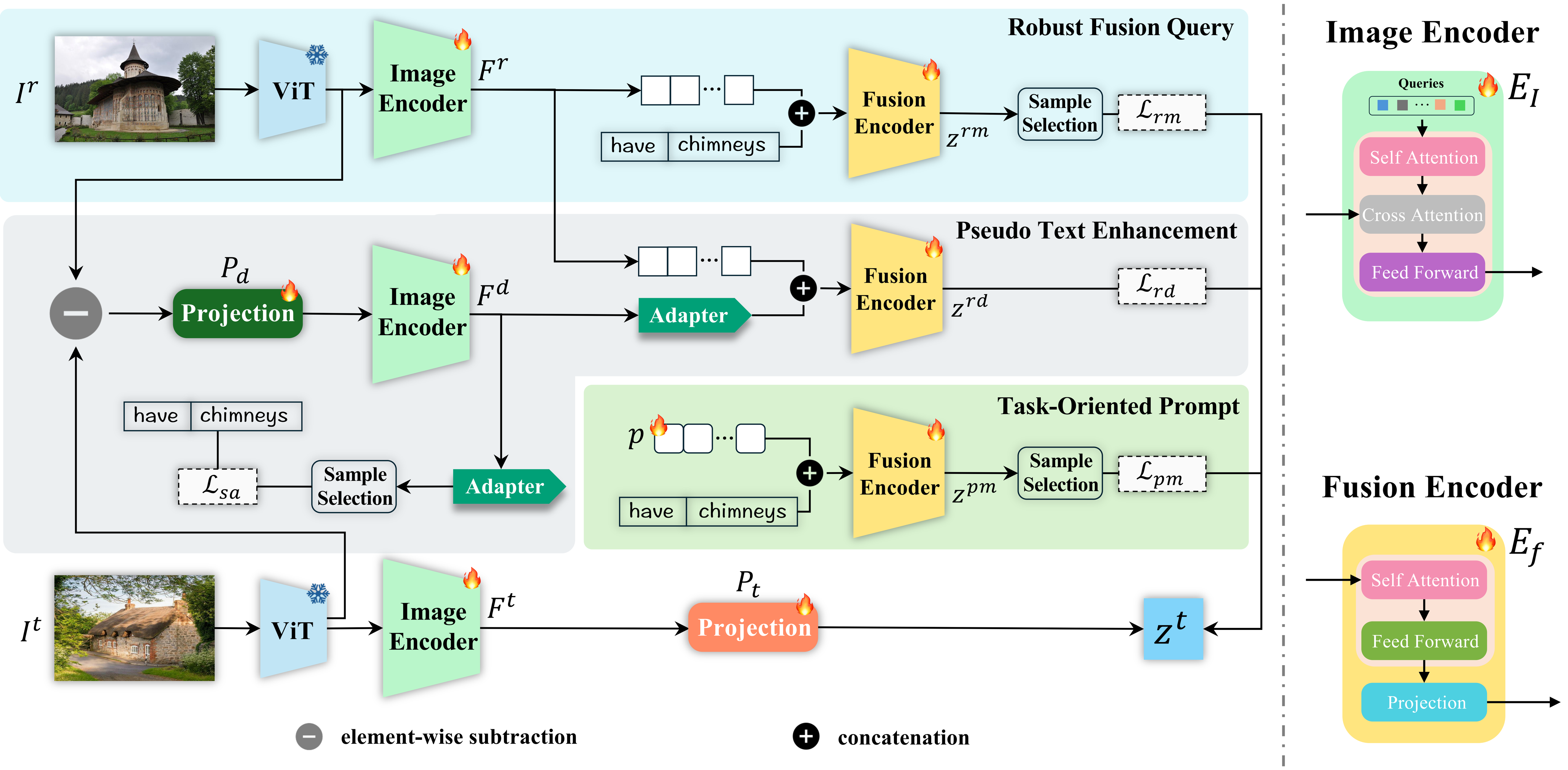
Pratial Matching

Incorret Matching

Contribution

- A **novel setting** in CIR–learning with noisy triplet correspondence--offering a new design perspective for existing supervised methods.
- We proposed a **novel method**, TME, tailored for this setting, enabling intrinsic relationship exploitation and noisy triplet utilization.
- Extensive experiments on two domain-specific datasets confirm the **robustness and effectiveness** of our approach in addressing noisy triplet correspondence.

Task-oriented Modification Enhancement (TME)



$$\mathcal{L} = -\frac{1}{\sum_i^B y_i} \sum_{i,j\neq i}^{B} y_i \log\left(1 - \frac{\exp(\tau(\boldsymbol{z}_i)^T \boldsymbol{z}_j^t)}{\sum_j^B \exp(\tau(\boldsymbol{z}_i)^T \boldsymbol{z}_j^t)}\right)$$

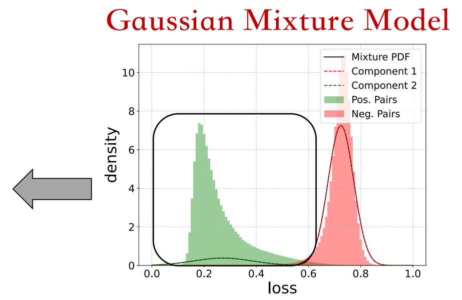
τ: the temperature parameter.

$\boldsymbol{z}_i$: one of the $\boldsymbol{z}_{rm}$, $\boldsymbol{z}_{rd}$, and $\boldsymbol{z}_{pm}$, corresponding to $\mathcal{L}_{rm}$, $\mathcal{L}_{rd}$, and $\mathcal{L}_{pm}$.

$\boldsymbol{L}_t$: text embedding layer of the Q-Former.

$$\mathcal{L}_{sa} = \frac{1}{\sum_i^B y_i} \sum_{i}^{B} y_i ||\boldsymbol{F}_i^d - \boldsymbol{L}_t(\boldsymbol{m}_i)||^2$$

y_i: pseudo-labe, 1 means a clean triplets.



✓More Robust
✓More Stable
✓Less Overfit

- **RFQ** module leverages a **GMM**-based sample selection strategy to filter out noisy triplets.
- **PTE** module generates an **adapter** to reconstruct the relations for noisy triplets, enabling learning from noisy data.
- **TOP** replaces the reference image with a single **learnable prompt**, helps modification-target semantic alignment, thus alleviating the reference-target visually irrelevant noise in partial matching.
- **Losses** $\mathcal{L}_{rm}$, $\mathcal{L}_{rd}$, and $\mathcal{L}_{pm}$ are **complementary contrastive losses** for robust learning, and $\mathcal{L}_{pm}$ is an **MSE loss** to achieve the alignment between the **adapter** with modification from clean triplets.

Experimental Results

CIRR test set:

Noise	Methods	$R@K$ K=1	K=5	K=10	K=50	$R_{subset}@K$ K=1	K=2	K=3	Avg($R@5$, $R_{subset}@1$)
0%	SPRC (ICLR'24)	51.96	82.12	89.74	97.69	80.65	92.31	96.60	81.39
	RCL (TPAMI'23)	53.16	82.41	90.12	**98.34**	79.57	92.02	96.87	80.99
	TME	**53.42**	**82.99**	90.24	98.15	**81.04**	**92.58**	**96.94**	**82.01**
20%	SPRC (ICLR'24)	45.90	75.86	83.52	93.37	78.10	**91.40**	96.05	76.98
	RCL (TPAMI'23)	50.43	**81.11**	**88.82**	96.68	77.52	90.80	95.71	79.31
	TME	**51.35**	81.01	88.53	**97.81**	**78.46**	91.25	**96.39**	**79.74**
50%	SPRC (ICLR'24)	39.93	66.00	73.59	86.48	75.81	89.21	95.37	70.90
	RCL (TPAMI'23)	**48.58**	77.45	85.93	94.70	75.60	89.28	94.80	76.52
	TME	48.48	**78.94**	**87.28**	**96.99**	**76.48**	**90.07**	**95.83**	**77.71**
80%	SPRC (ICLR'24)	29.95	51.25	58.51	73.86	70.22	86.05	93.21	60.74
	RCL (TPAMI'23)	44.94	74.43	82.99	92.31	71.93	86.84	92.96	73.18
	TME	**46.31**	**75.78**	**84.89**	**95.83**	**73.37**	**88.02**	**94.89**	**74.58**

FashionIQ validation set:

Noise	Methods	Dress R@10	Dress R@50	Shirt R@10	Shirt R@50	Toptee R@10	Toptee R@50	Average R@10	Average R@50	AVG.
0%	SPRC (ICLR'24)	49.18	72.43	55.64	73.89	**59.35**	78.58	54.92	74.97	64.85
	RCL(TPAMI'23)	48.79	**72.68**	55.89	73.90	56.91	77.41	53.86	74.66	64.26
	TME	**49.73**	71.69	**56.43**	**74.44**	59.31	**78.94**	**55.15**	**75.02**	**65.09**
20%	SPRC (ICLR'24)	39.81	62.22	48.58	66.29	50.48	70.58	46.29	66.36	56.33
	RCL(TPAMI'23)	47.05	**70.65**	53.14	71.74	55.28	75.62	51.82	72.67	62.25
	TME	**49.03**	70.35	**55.84**	**73.16**	**57.22**	**78.23**	**54.03**	**73.91**	**63.97**
50%	SPRC (ICLR'24)	35.94	57.16	42.25	61.63	44.98	64.76	41.06	61.19	51.12
	RCL(TPAMI'23)	43.68	66.44	50.74	69.19	52.63	73.84	49.01	69.82	59.42
	TME	**46.26**	**68.27**	**53.09**	**71.88**	**55.07**	**76.59**	**51.47**	**72.25**	**61.86**
80%	SPRC (ICLR'24)	28.41	50.77	36.21	54.37	35.90	56.96	33.51	54.03	43.77
	RCL(TPAMI'23)	38.82	60.54	45.44	64.38	47.42	68.38	43.89	64.43	54.16
	TME	**41.45**	**64.35**	**47.30**	**68.20**	**51.25**	**73.23**	**46.67**	**68.60**	**57.63**

Visualization

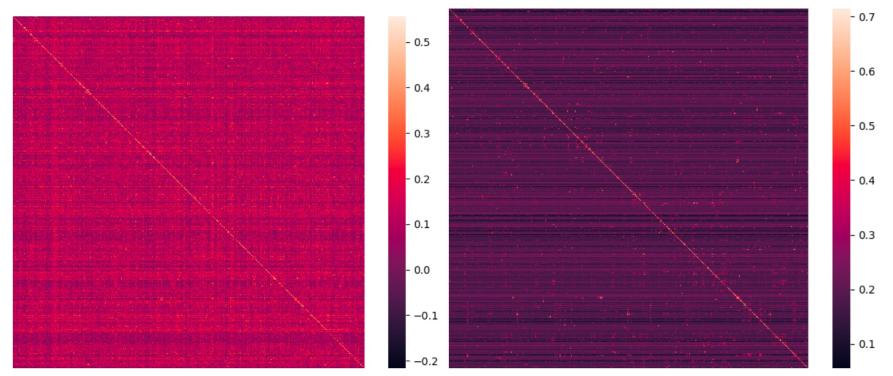


Figure: Visualization of query-target similarity matrices from SPRC (left) and TME (right) on the CIRR validation set with 50% noise.

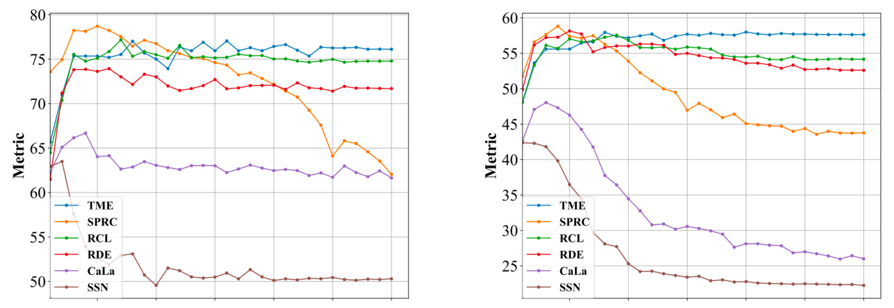


Figure: performance w.r.t epoch on the CIRR (left) and FashionIQ (right) validation set with 80% noise.

Experiment Analysis:

- TME produces a **sharper** diagonal, showing its strength at separating relevant from irrelevant images under heavy noise.
- At 80% noise, TME shows higher **stability and accuracy** than both vanilla CIR methods and 2-tuple NC methods, highlighting its robustness and ability to leverage **intrinsic relations** and noisy triplets.