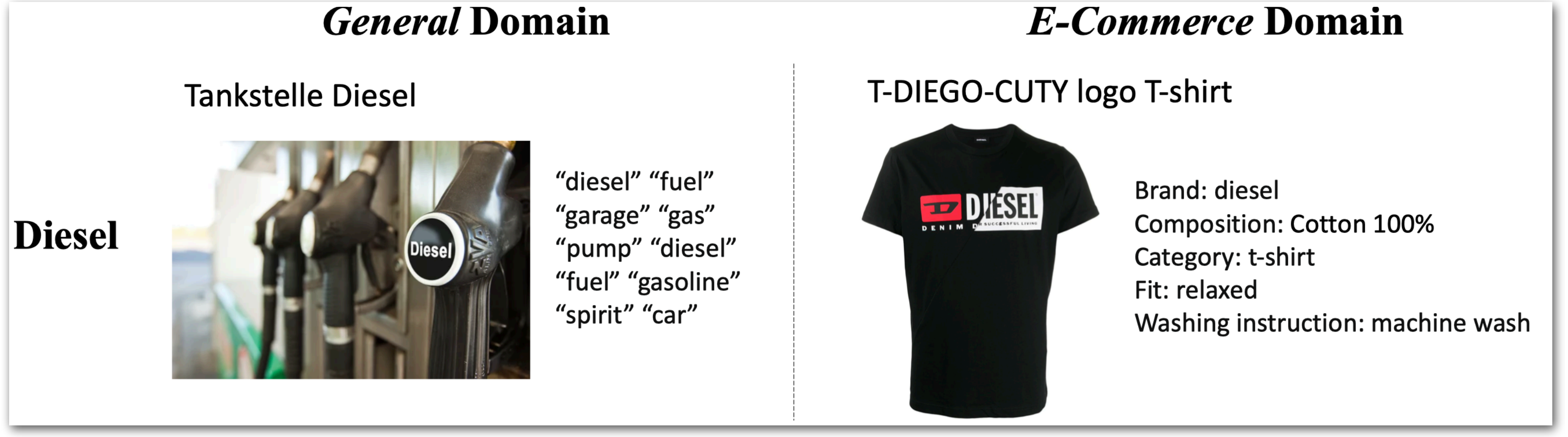
EI-CLIP: Entity-aware Interventional Contrastive Learning for E-commerce Cross-modal Retrieval

Haoyu Ma[1], Handong Zhao[2], Zhe Lin[2], Ajinkya Kale[2], Zhangyang Wang [3], Tong Yu[2], Jiuxiang Gu[2], Sunav Choudhary[2], Xiaohui Xie[1]

[1]University of California, Irvine, [2]Adobe Research [3]University of Texas at Austin

Motivations:

1) Words come up with special meanings in e-commerce.



***General* Domain** | ***E-Commerce* Domain**

Diesel

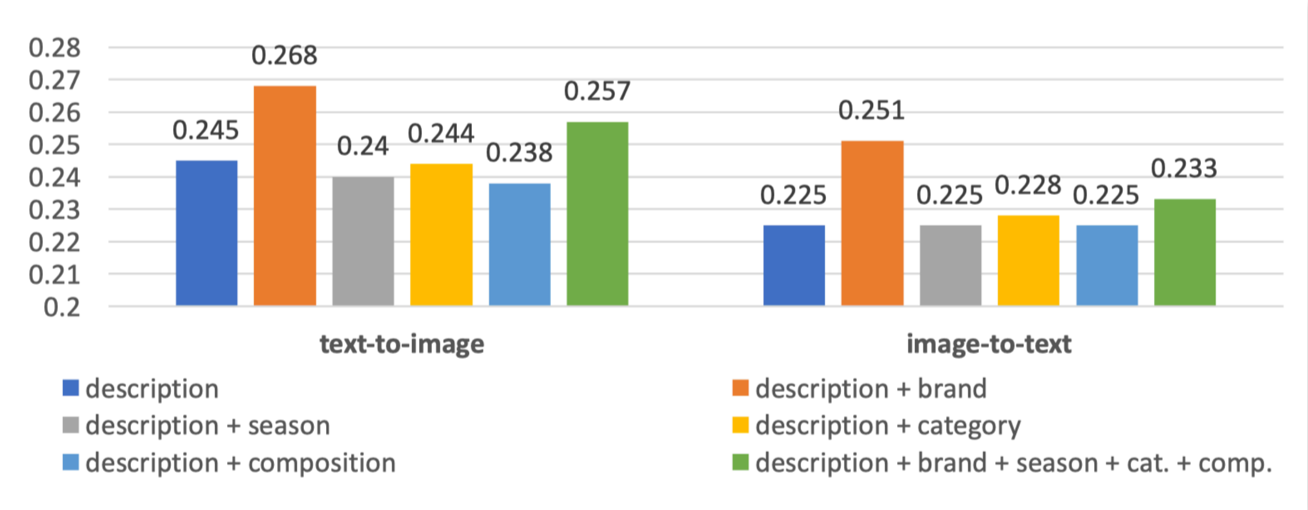
Tankstelle Diesel

"diesel" "fuel"
"garage" "gas"
"pump" "diesel"
"fuel" "gasoline"
"spirit" "car"

T-DIEGO-CUTY logo T-shirt

Brand: diesel
Composition: Cotton 100%
Category: t-shirt
Fit: relaxed
Washing instruction: machine wash

2) Meta data contributes unevenly in cross-modal retrieval.



CLIP in the Causal View:

Confounders $z = g(a, b)$, entity a takes the semantics b.

X: text, Y: image.

Contrastive learning of CLIP:

$$P(Y|X) = \sum_z P(Y, z|X) = \sum_z P(Y|X, z)\underline{P(z|X)}$$

Interve X with do-calculus

mitigate bias towards commonsense in general domain

$$P(Y|do(X)) = \sum_z P(Y|X, z)\underline{P(z)}$$

e.g.

X = "*<u>a T-shirt of golden goose</u>*"

most of the likelihood is assigned to $P(Y|X, z = g(\text{golden goose}, \text{"animal"}))$, as $P(z = g(\text{golden goose}, \text{"animal"})|X)$ is large in the general domain

Framework (EI-CLIP):

EA-Learner: explicity capture each entity information

CE-Selector: select important entities

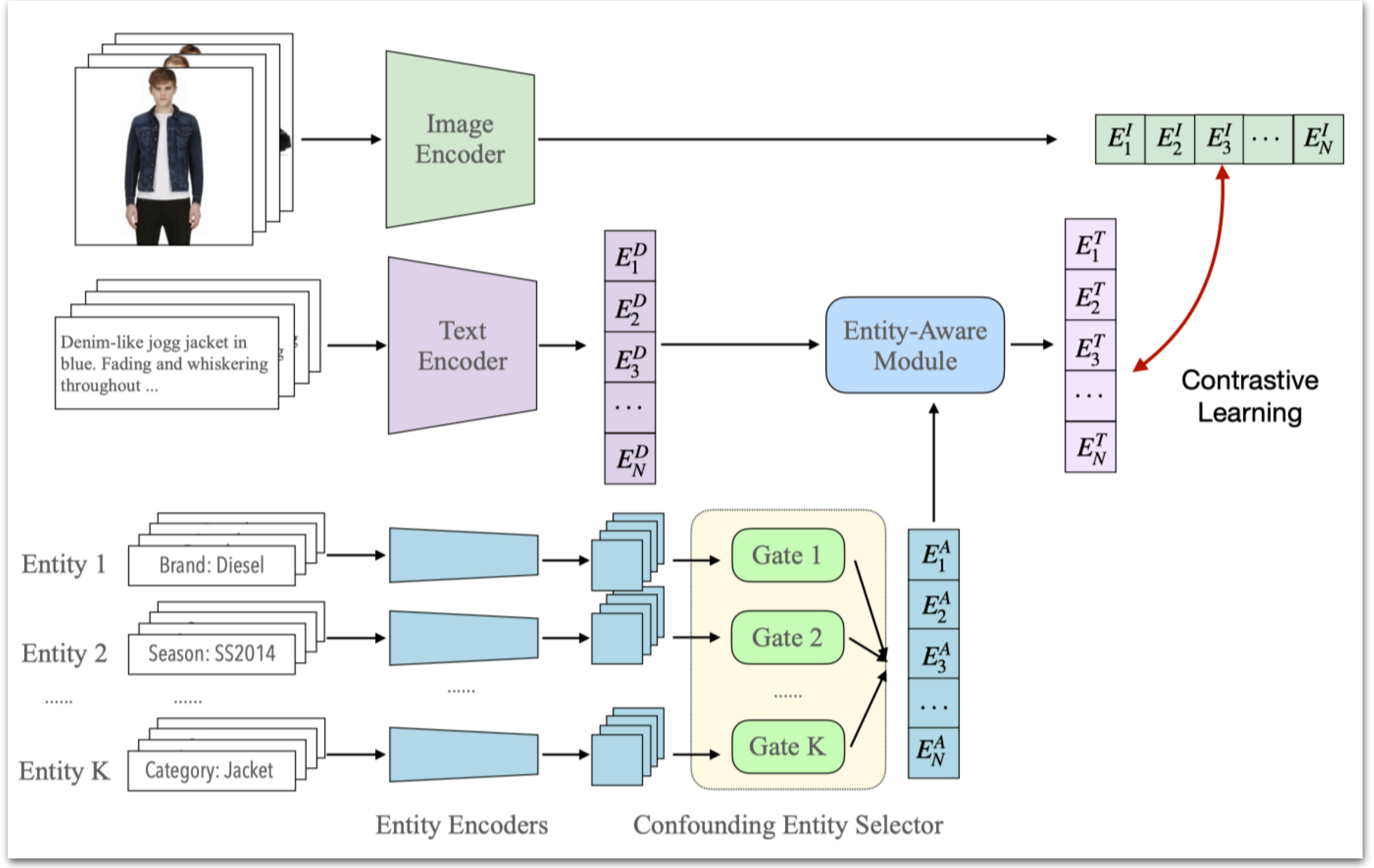


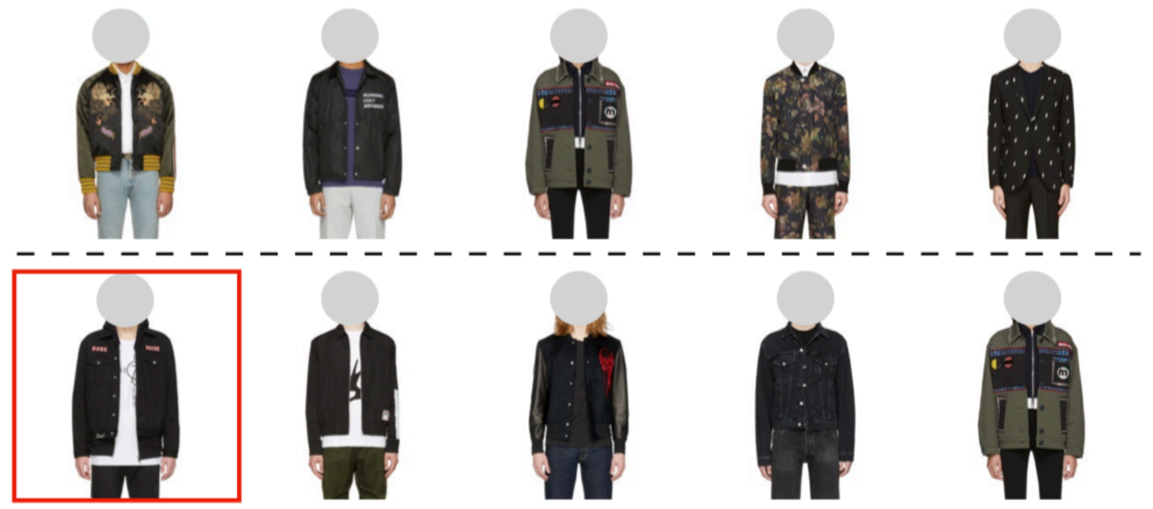
Image Encoder
E_1^I E_2^I E_3^I ⋯ E_N^I
Denim-like jogg jacket in blue. Fading and whiskering throughout ...
Text Encoder
E_1^D E_2^D E_3^D ⋯ E_N^D
Entity-Aware Module
E_1^T E_2^T E_3^T ⋯ E_N^T
Contrastive Learning
Entity 1
Brand: Diesel
Entity 2
Season: SS2014
......
Entity K
Category: Jacket
Gate 1
Gate 2
Gate K
E_1^A E_2^A E_3^A ⋯ E_N^A
Entity Encoders
Confounding Entity Selector

Results:

Fashion-Gen

	Image-to-text			Text-to-image			*SumR*
	R@1	R@5	R@10	R@1	R@5	R@10	
①	9.4	24.5	33.5	10.7	26.8	35.8	141
②	22.5	49.5	62.0	24.5	51.1	63.6	273
③	23.3	51.5	64.6	25.7	53.9	66.5	285
④	25.2	52.6	64.8	28.2	56.6	68.4	296
⑤	**25.7**	**54.5**	**66.8**	**28.4**	**57.1**	**69.4**	**302**
↑	10.3%	5.8%	3.1%	10.5%	5.9%	4.4%	6.0%

Brand: <u>Diesel</u>. Long sleeve denim jacket in black. Fading , distressing, stitched detailing, and multicolor appliqués throughout. Spread collar. Button closure at front. Flap pockets at chest. Seam pockets at waist. White logo embroidered at front hem. Adjustable buttoned tabs at back hem. Silver-tone hardware. Tonal stitching.



Contribution Summaries:

1) The pioneering work to tackle the challenges introduced by e-commerce special entities.

2) The first to formulate the entity-aware retrieval task in causal view.

3) We propose an Entity-aware Intervention-based contrastive learning framework (EI-CLIP), which achieves competitive performance on e-commerce benchmark dataset Fashion-Gen.

References:

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *IICML*, 2021.