Learning Instance-Level Representation for Large-Scale Multi-Modal Pretraining in E-commerce

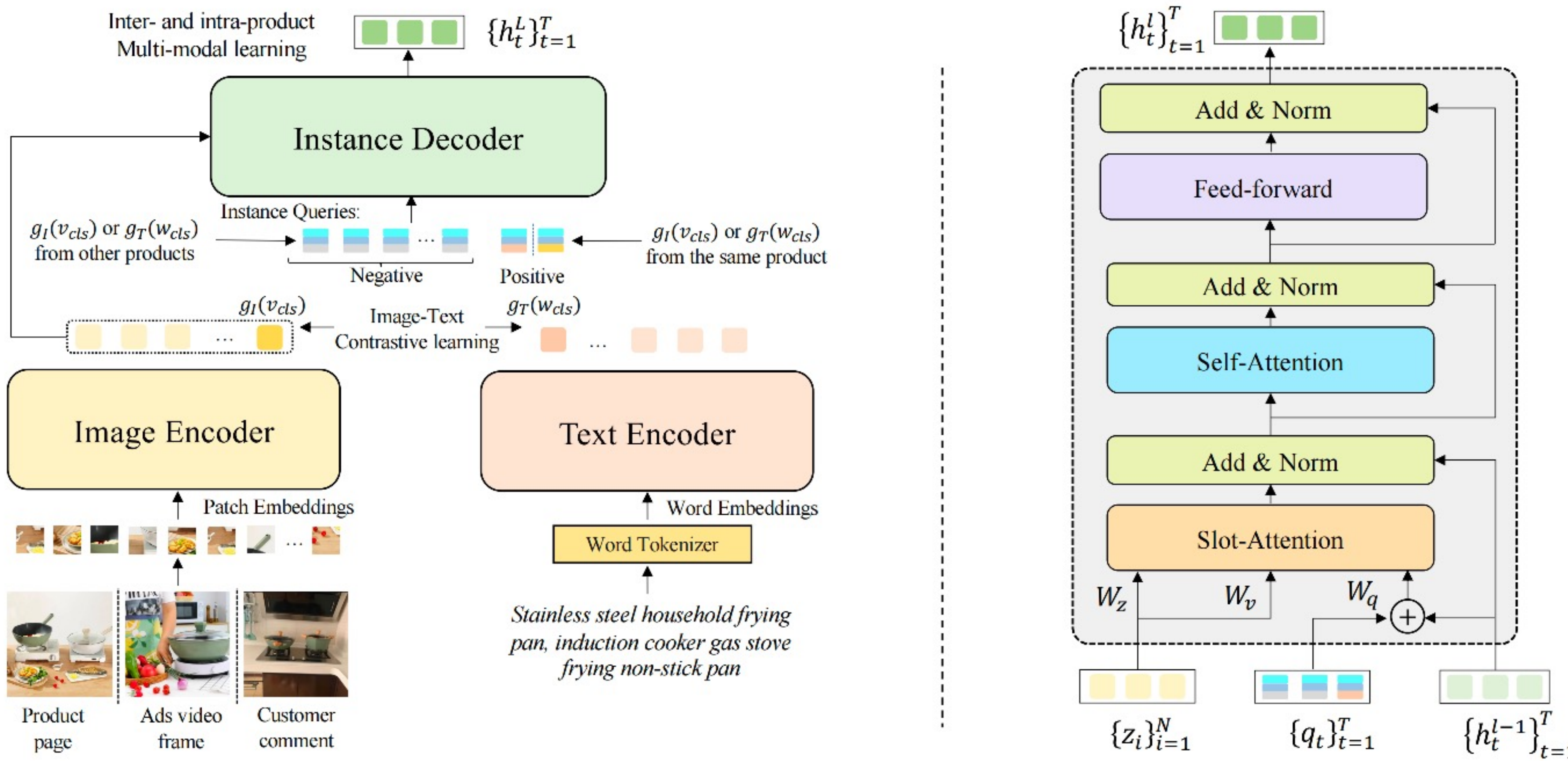
Yang Jin, Yongzhi Li, Zehuan Yuan, Yadong Mu

Peking University, ByteDance Inc.

Poster: Wed-PM-269

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Our Foundation Model : ECLIP

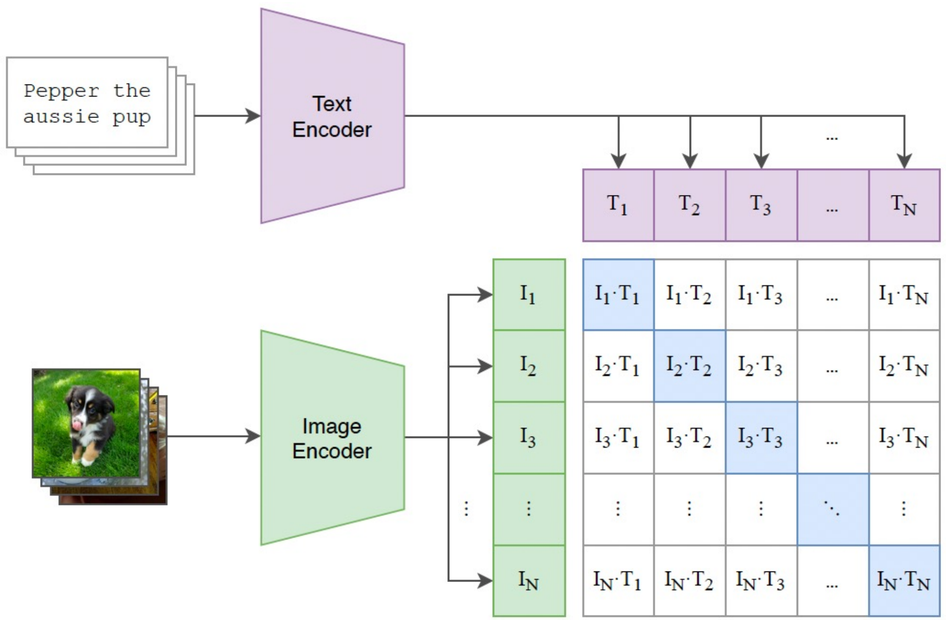


(a) The overall Pipeline

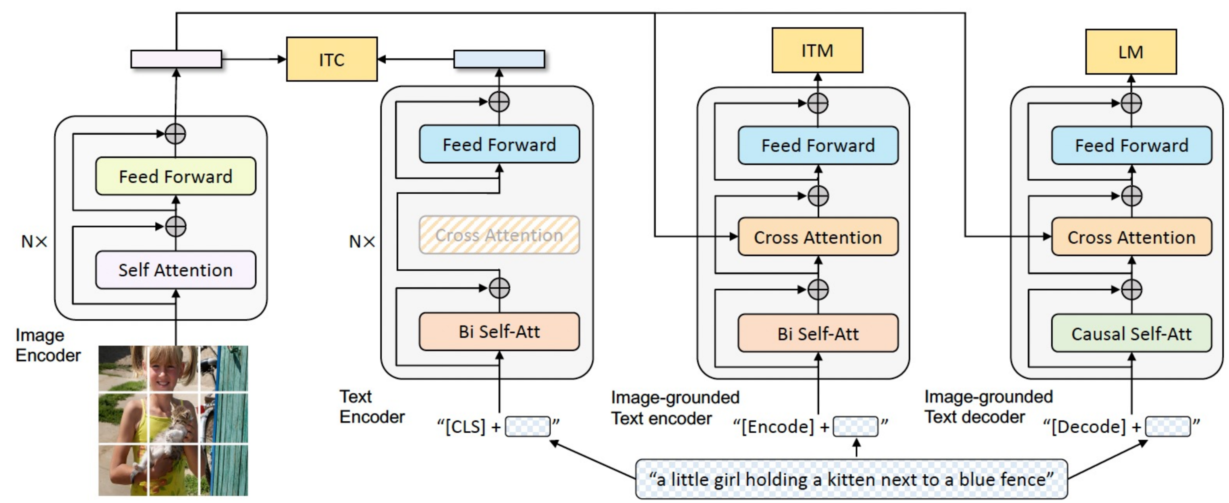
(b) The i-th Decoder Block

Introduction

Given an image-text pair, existing Vision-Language foundation models aims to learn the image-level representations.



CLIP



BLIP

Introduction

We explore the ways to enable vision-language foundation model to obtain instance-level representation in E-commerce.

General Domain

Foreground: horse, people, church

E-commerce Domain

Foreground: frying pan, coffee machine

A group of people on horseback next to a church

Stainless steel frying pan

Italian semi-automatic home coffee maker

The difference of natural image and product image in E-commerce.

Motivation

A product usually has multiple image samples from different sources (e.g., merchant, customer comments, attached advertisement videos, etc.)

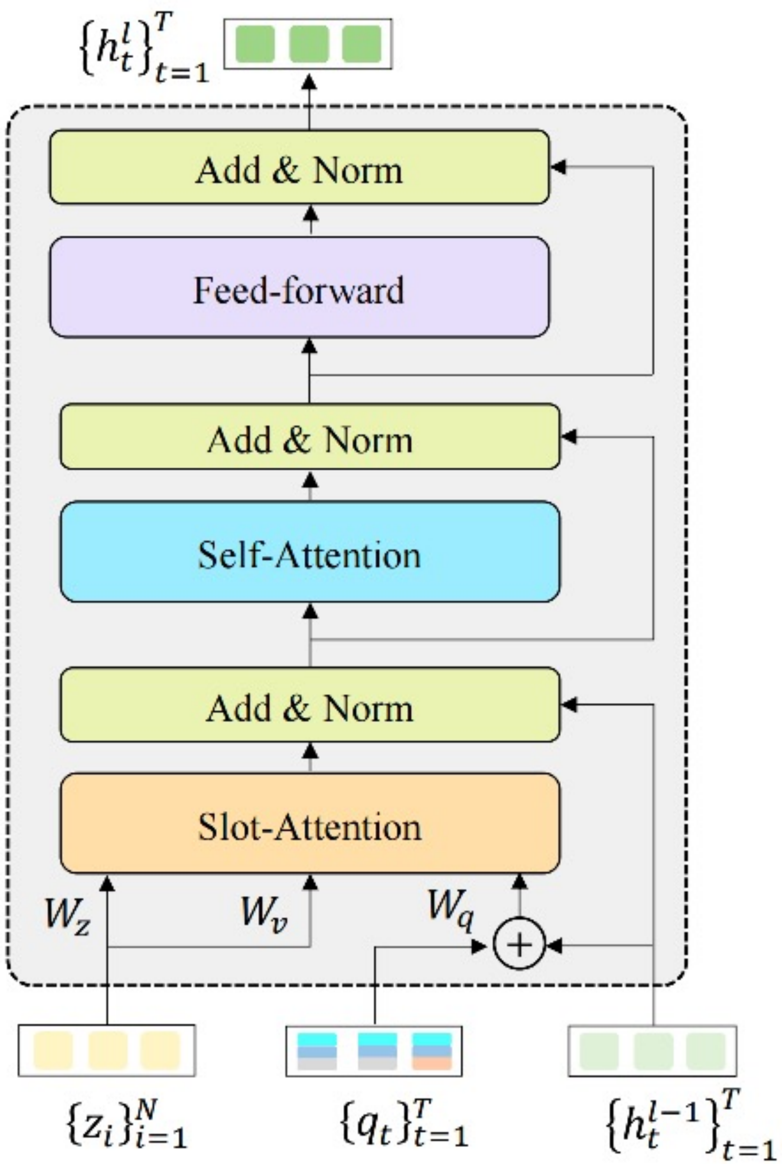
Ruby Face Cream

(a) (b) (c)

PROYA ruby face cream for ladies

The property of product images in E-commerce.

Instance Decoder



Input:

Instance Query

$$Q = \{q_t \in \mathcal{R}^D\}_{t=1}^T, \quad q_t = q_t^{\text{prompt}} + q_t^{\text{pos}} + q_t^{\text{type}}.$$

One positive query, T - 1 negative ones

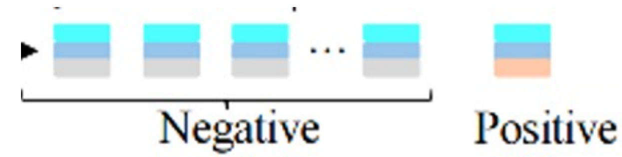


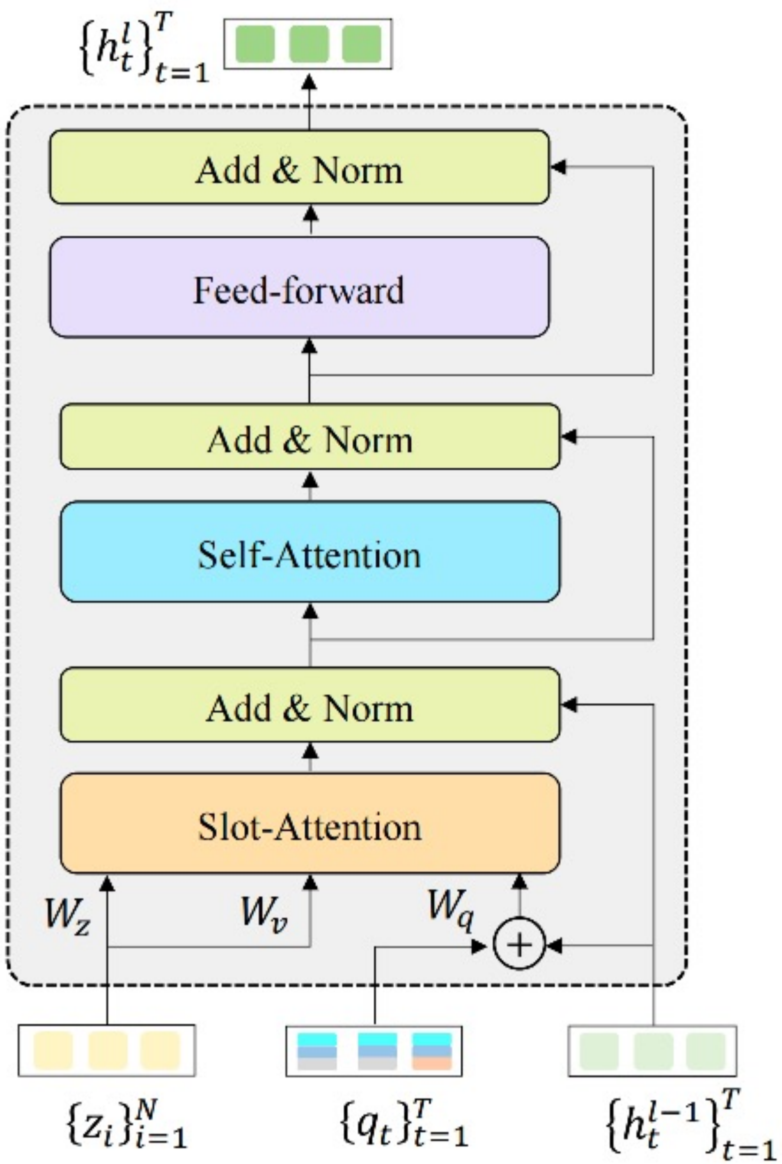
Image Patch Representation

$$Z = \{z_i \in \mathcal{R}^D\}_{i=1}^N.$$

Instance Representations

$$H = \{h_t\}_{t=1}^T \qquad H^0 \text{ are zero-initialized,}$$

Slot-Attention Layer

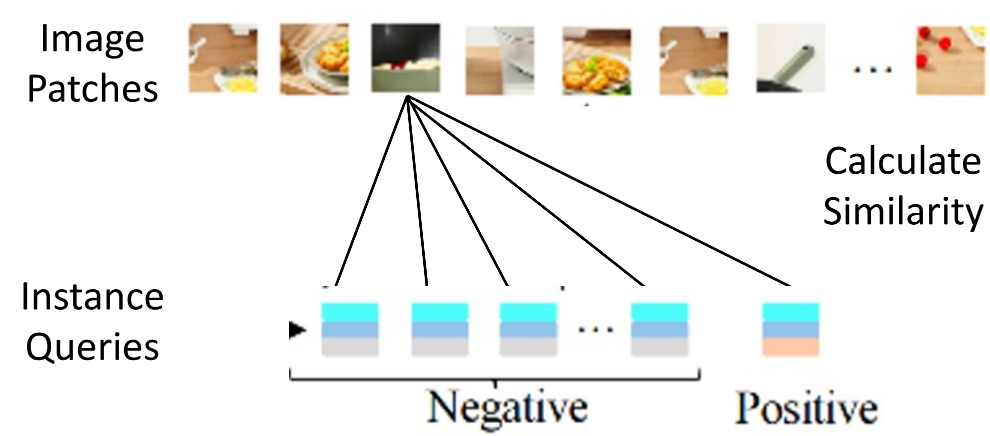


Step 1:

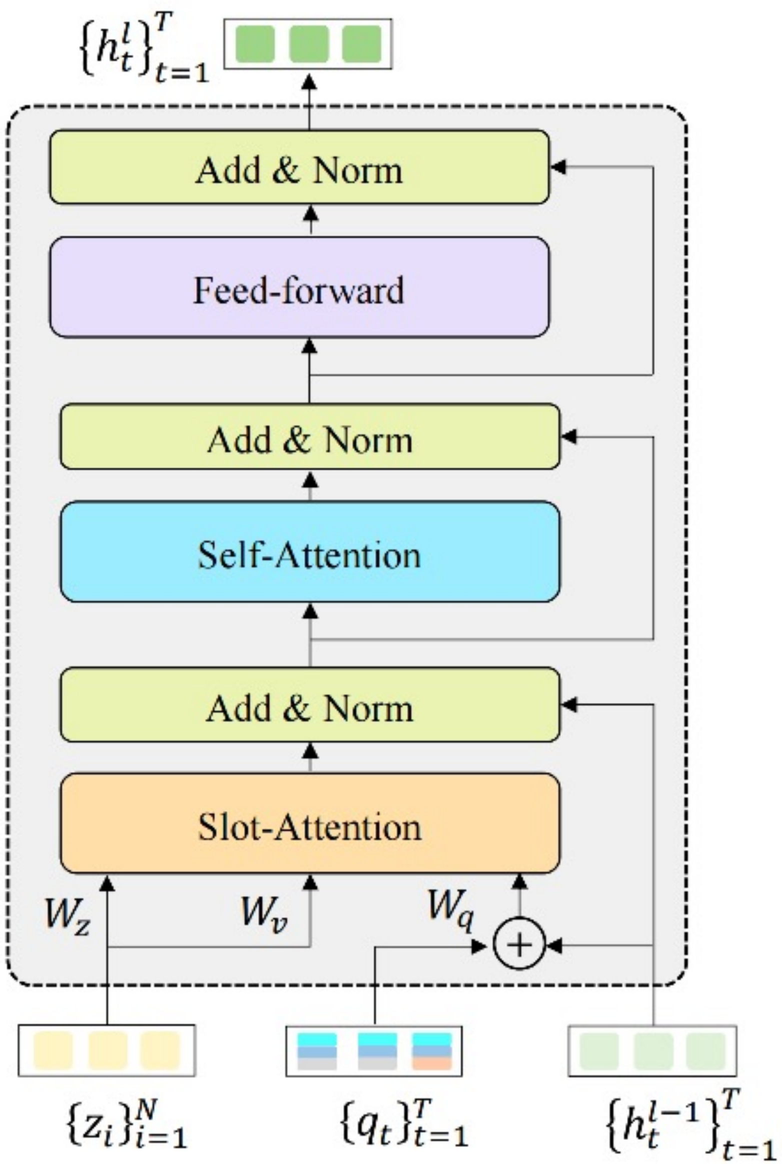
Calculate the similarity matrix

$$M = \frac{1}{\sqrt{D}}(ZW_z) \cdot ((Q + H^{l-1})W_q)^\top,$$

$$M_{ij} = \frac{\exp(M_{ij})}{\sum_{t=1}^{T} \exp(M_{it})} \cdot \ M \in \mathcal{R}^{N \times T},$$



Slot-Attention Layer

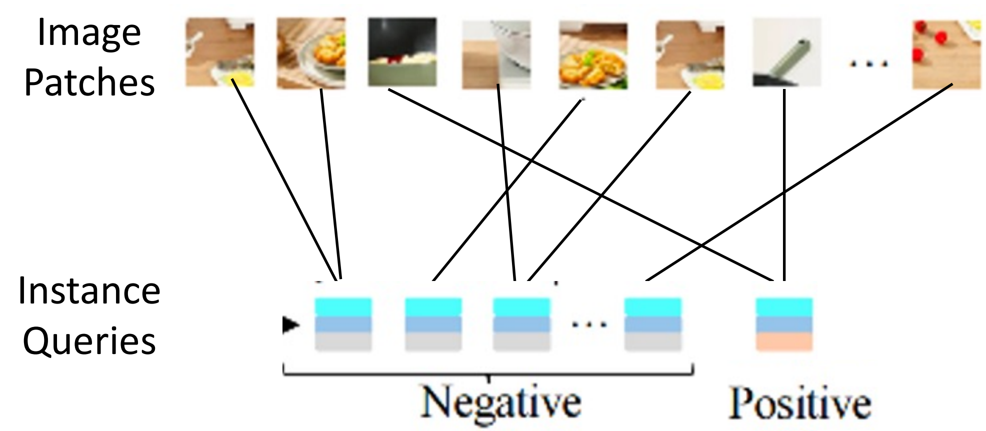


Step 2:

Perform the soft assignment and Update the instance representation

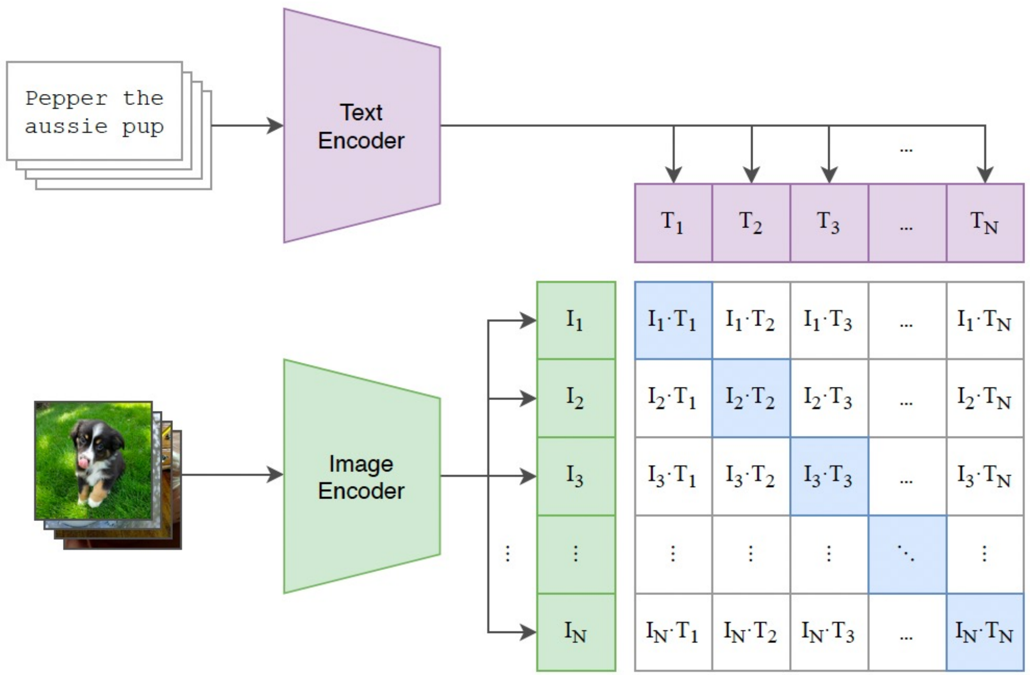
$$\Delta h_t^{l-1} = \frac{1}{\sum_{i=1}^N M_{it}} \sum_{i=1}^N M_{it}(W_v z_i).$$

$$h_t^l = h_t^{l-1} + W_o \Delta h_t^{l-1}.$$



Pretraining Proxy Tasks

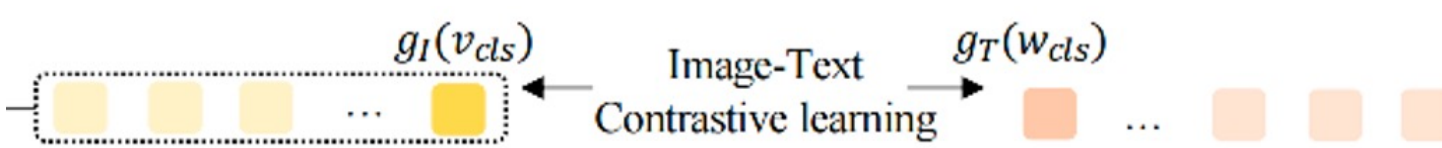
Image-Text Contrastive Learning :



$$s(x^I, x^T) = g_I(v_{cls})^\top g_T(w_{cls}).$$

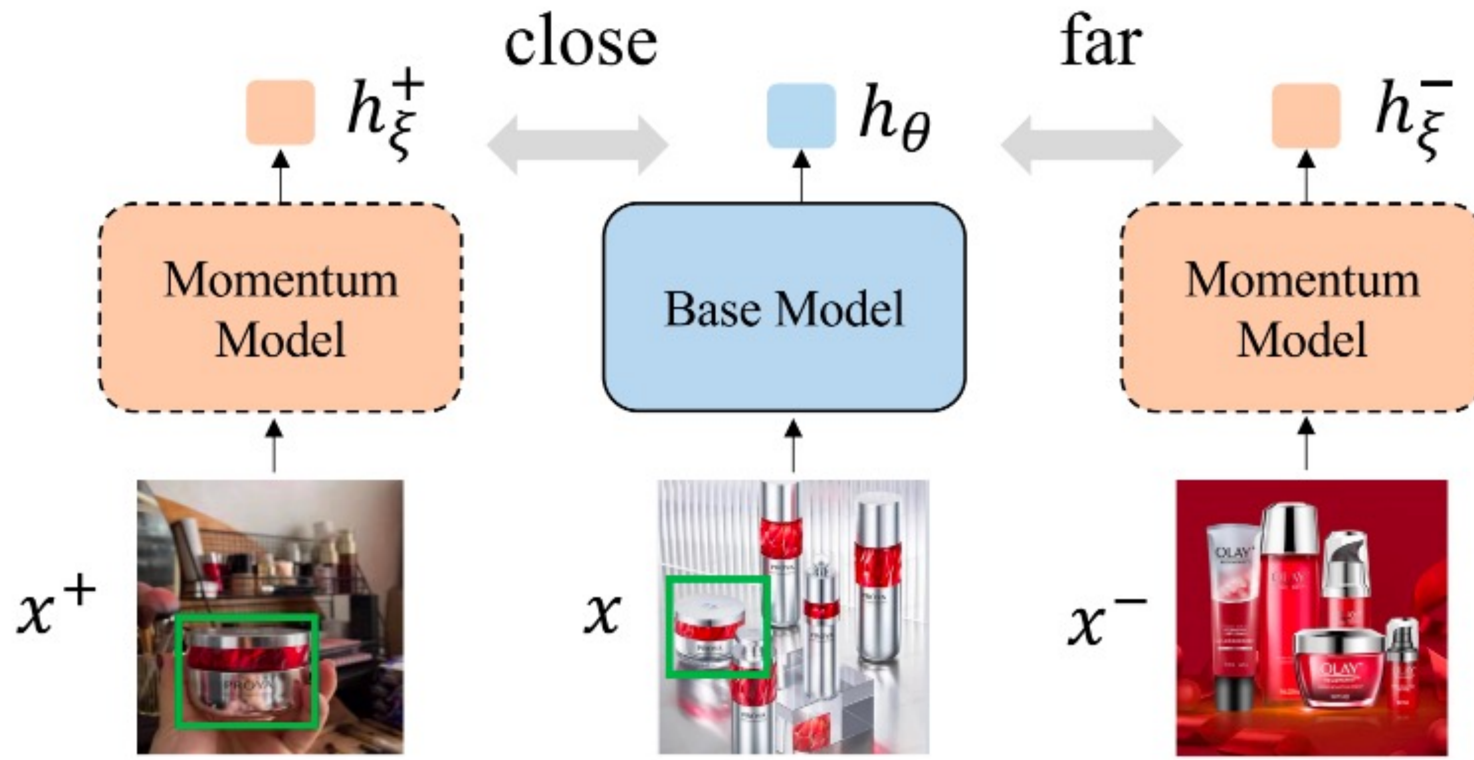
$$\mathcal{L}_{i2t} = -\sum_{i=1}^{B} \log \frac{\exp(s(x_i^I, x_i^T)/\tau)}{\sum_{j=1}^{B} \exp(s(x_i^I, x_j^T)/\tau)},$$

$$\mathcal{L}_{t2i} = -\sum_{i=1}^{B} \log \frac{\exp(s(x_i^T, x_i^I)/\tau)}{\sum_{j=1}^{B} \exp(s(x_i^T, x_j^I)/\tau)},$$



Pretraining Proxy Tasks

Inter-Product Multi-modal Learning

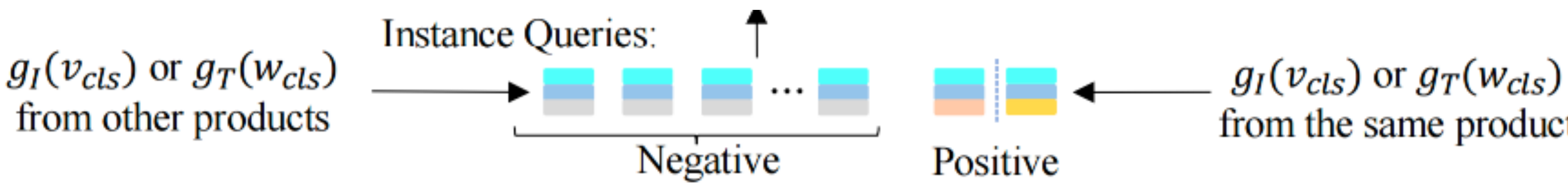


We use the similarity between $g_I(v_{cls})$ and $g_T(w_{cls})$ to sample the hard negative samples.

$$\mathcal{L}_{inter} = -\sum_{i=1}^{B} \log \frac{\exp({h_\theta^i}^\top h_\xi^j/\tau)}{\exp({h_\theta^i}^\top h_\xi^j/\tau) + \sum_{k\in\mathcal{N}^-}\exp({h_\theta^i}^\top h_\xi^k/\tau)},$$

Pretraining Proxy Tasks

Intra-Product Multi-modal Learning



$$\mathcal{L}_{intra} = -\sum_{i=1}^{B} \log \frac{\exp({h_r^i}^\top g_T(w_{\text{cls}}^i)/\tau)}{\sum_{t=1}^{T} \exp({h_t^i}^\top g_T(w_{\text{cls}}^i)/\tau)},$$

r is the positive query index

To regularize the Similarity Matrix M

$$\mathcal{L}_{\mathcal{R}} = \boxed{\sum_{i=1}^{N} M_{i,r} \log(\frac{1}{M_{i,r}}) +} \boxed{\sum_{j=1, j \neq r}^{T} \left(\log N - \sum_{i=1}^{N} M_{i,j} \log(\frac{1}{M_{i,j}}) \right)}$$

For Positive Query

For Negative Queries

Pretraining On 100M E-commerce Data

100M various image-text pairs, from 15M different products

(a) Product Detail Page

(b) Customer Comment

(c) Advertisement Video

Living room oil painting style tissue box

Pretraining Data Format

Method	Classification	Image-to-Text		Text-to-Image	
	Acc@1	R@1	R@5	R@1	R@5
CLIP [21]	37.2	52.6	74.1	58.7	84.0
FILIP [32]	37.1	52.3	73.8	58.0	83.5
DeCLIP [15]	37.8	53.1	75.8	58.8	83.9
ALBEF [13]	38.5	52.9	74.4	58.2	83.3
BLIP [12]	39.3	53.3	75.6	59.1	84.4
Ours$_{\text{ViT-B/16}}$	43.8	53.8	76.0	59.9	84.6
Ours$_{\text{ViT-L/16}}$	**44.8**	**58.2**	**79.6**	**63.8**	**87.4**

Zero-shot transfer to classification and image-text retrieval

Experimental Results

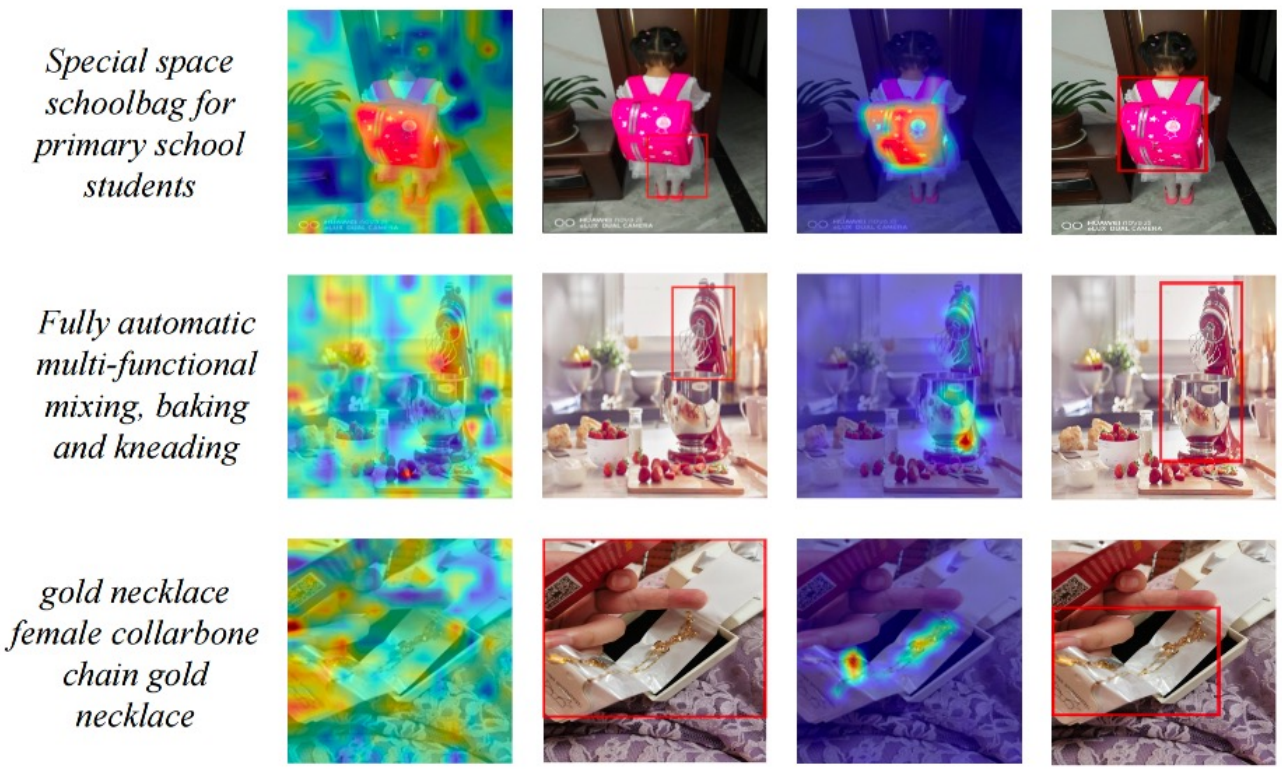
Method	Pretraining Dataset	Coarse Product Retrieval			Fine-grained Product Retrieval					
		mAP@1	mAP@5	mAP@10	R@1	R@5	R@10	mAP@1	mAP@5	mAP@10
ViLBERT [17]	M5Product	58.6	61.7	60.1	-	-	-	-	-	-
UNITER [2]		58.9	62.8	60.9	-	-	-	-	-	-
SCALE [4]		59.8	64.1	62.2	-	-	-	-	-	-
CLIP [20]	ECLIP 100M	68.2	73.2	70.7	34.8	54.2	62.9	34.8	40.2	39.9
FILIP [31]		67.8	73.0	70.3	34.6	53.9	62.2	34.6	40.1	39.7
DeCLIP [14]		68.5	73.4	70.8	35.3	56.4	65.5	35.3	41.2	40.8
ALBEF [12]		68.7	73.6	71.2	35.1	56.1	65.2	35.1	40.7	40.4
BLIP [11]		69.1	74.1	71.6	35.6	56.8	66.0	35.6	41.6	41.3
$\text{Ours}_{\text{ViT-B/16}}$		69.6	74.9	72.5	44.3	63.4	71.1	43.8	48.6	48.2
$\text{Ours}_{\text{ViT-L/16}}$		**70.2**	**75.3**	**72.9**	**45.0**	**64.2**	**72.1**	**45.0**	**50.0**	**49.5**

Zero-shot transfer to Product Retrieval

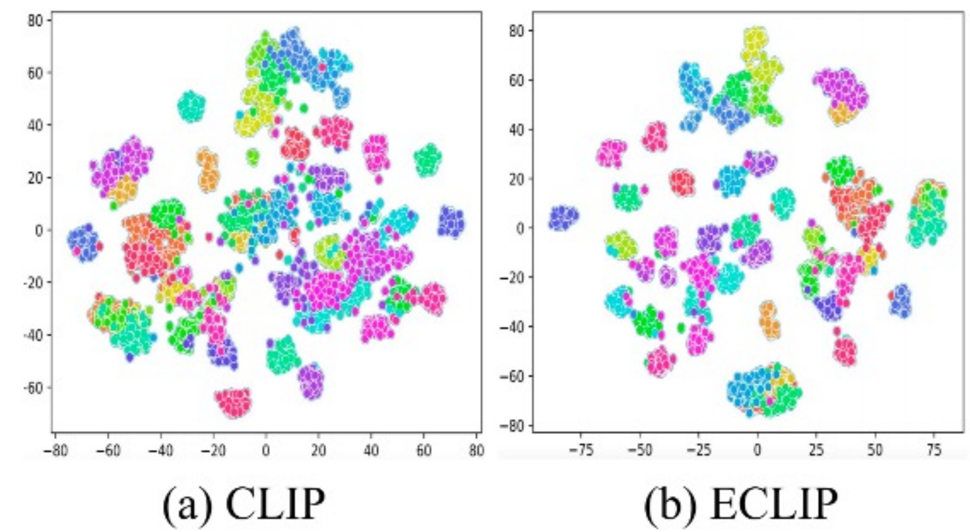
Method	Visual Grounding	
	Acc@0.5	Acc@0.7
CLIP [20]	80.9	75.2
FILIP [31]	81.3	75.6
DeCLIP [14]	81.0	75.3
ALBEF [12]	80.9	74.7
BLIP [11]	81.1	75.1
$\text{Ours}_{\text{ViT-B/16}}$	**91.2**	**89.6**

Zero-shot transfer to Visual Grounding

Visualization Results



ECLIP on Zero-Shot Grounding



The T-SNE visualization of learned representation

Thanks!