PV2TEA: Patching Visual Modality to Textual-Established Information Extraction

Hejie Cui[1]*, Rongmei Lin[2], Nasser Zalmout[2], Chenwei Zhang[2], Jingbo Shang[3], Carl Yang[1], Xian Li[2]

[1] Emory University, GA, USA

[2] Amazon.com Inc, WA, USA

[3] University of California, San Diego, CA, USA

{hejie.cui, j.carlyang}@emory.edu, jshang@ucsd.edu

{linrongm, nzalmout, cwzhang, xianlee}@amazon.com

Presented by **Jingbo Shang**

Assistant Professor at UC San Diego

Visiting Academics at Amazon

Outline

1. Introduction and Motivation
2. Proposed Method: PV2TEA
3. Experiment Results

Multimodal Attribute Extraction

- Attribute value extraction: extract structed knowledge triples, i.e., (*sample_id, attribute, value*), from unstructured information, e.g., text descriptions and images

- Existing automatic attribute value extraction methods work well when prediction targets are *inferrable from text*

Visual Information Can Potentially Help in Improving Recall

Good Earth Sensorial Blends Tropical Moringa & Mango Herbal Tea, 15Count
Visit the Good Earth Store
200 ratings | 3 answered questions

Price: $3.85 ($0.26 / Count)

Earn 5% back on this purchase (worth $0.19 when redeemed) with your Prime Store Card.
SNAP EBT eligible

Brand	Good Earth
Item Form	Tea Bags
Flavor	Tropical Mango and Moringa Herbal Tea
Tea Variety	Green
Number of Items	1

About this item

- BORN TO BE BOLD: Not your ordinary English breakfast tea, our blend tantalizes your taste buds for an early morning lift
- ALL NATURAL: No artificial flavors, colors or preservatives
- REFRESHINGLY GOOD: Our flavored teas create a cup of effortless character and depth that is sure to leave you blushing
- ETHICAL TEA: Sustainability is at the core of everything Good Earth does with Rainforest Alliance ingredients on our Sensorial Blends
- Born in the 70s – 1972, to be exact and inspired by sunny Santa Cruz we came up with tantalizing teas to give your days a little lift

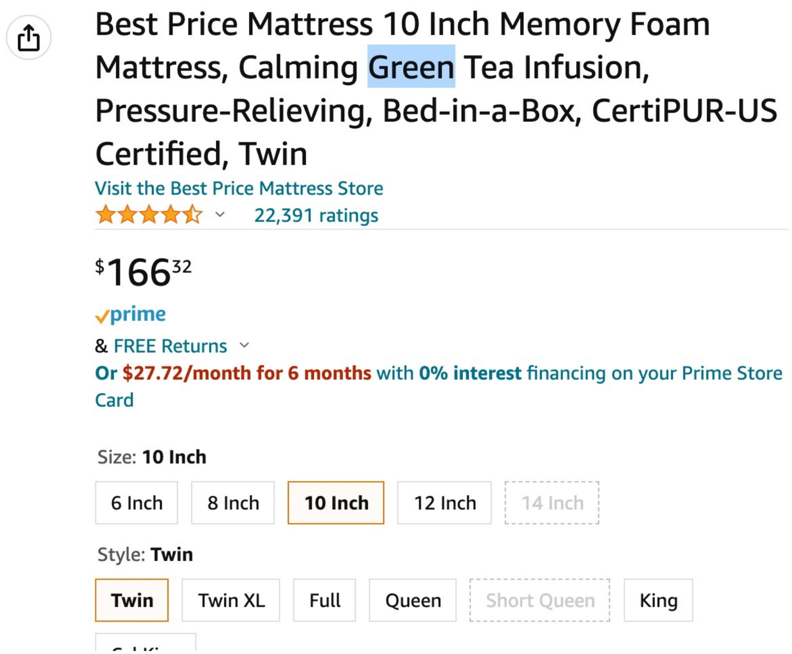
Scenario 1: Attribute value not in text

- The provided images may contain the missing attribute information

➢ Improving Recall

Itemform: tea bag

Visual Information Can Potentially Help in Improving Precision



Scenario 2: Distracting information

- The provided images may potentially help to distinguish noisy labels
  - Improving Precision

Color: white

Task Illustration and Challenges in Cross-Modality Integration

Image

Textual Descriptions: "Best Price Mattress 12 Inch Memory Foam Mattress, Calming ***Green Tea***-Infused Foam, Pressure Relieving, Bed-in-a-Box, Queen"

Question: What is the ***color*** of the mattress?

Weakly Supervised Label: green **True Value:** white

Challenge Explanations:

…

C1 Loosely-aligned product image and textual descriptions:

- intra-sample: weakly related across modalities and difficult to ground;
- inter-samples: images of other products can also pair with the text

C2 Visual bias: noisy contextual backgrounds, e.g., pillow, bed frame, etc.

C3 Textual bias: the training label is misled/biased by 'green tea' in text

Motivating Analysis on the Textual Bias of Attribute Extraction

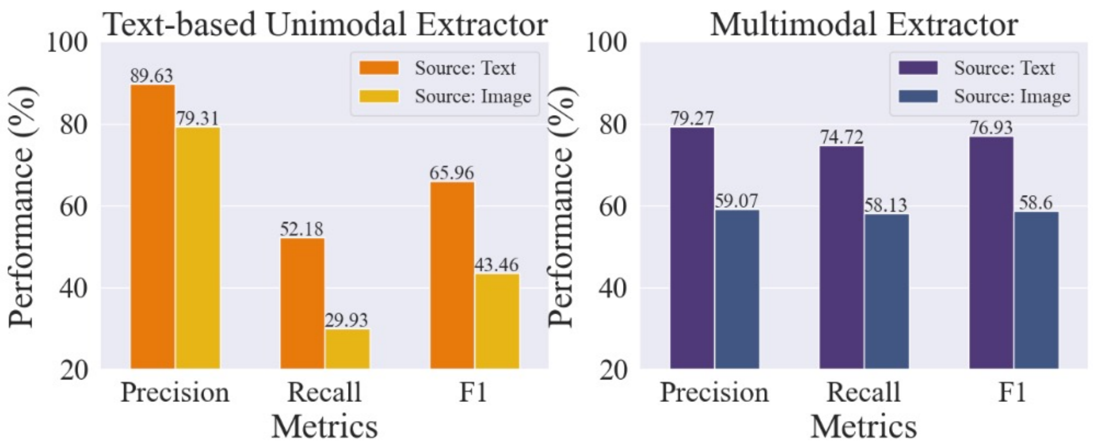


Figure 2: Source-aware evaluation of existing unimodal and multimodal models on the textual-biased issue.

Source: Text indicates the gold value is present in the text; **Source: Image** indicates the gold value is absent from the text and must be inferred from the image

- Two representative unimodal and multimodal methods: **OpenTag** and **PAM**
  - Both achieve impressive results when the gold value is contained in the text
  - When the gold value is not contained in the text and must be derived from visual input, the performance drops dramatically
- Model trained with **textual-shifted labels** will result in a learning ability gap between modalities → strong textual bias and dependence

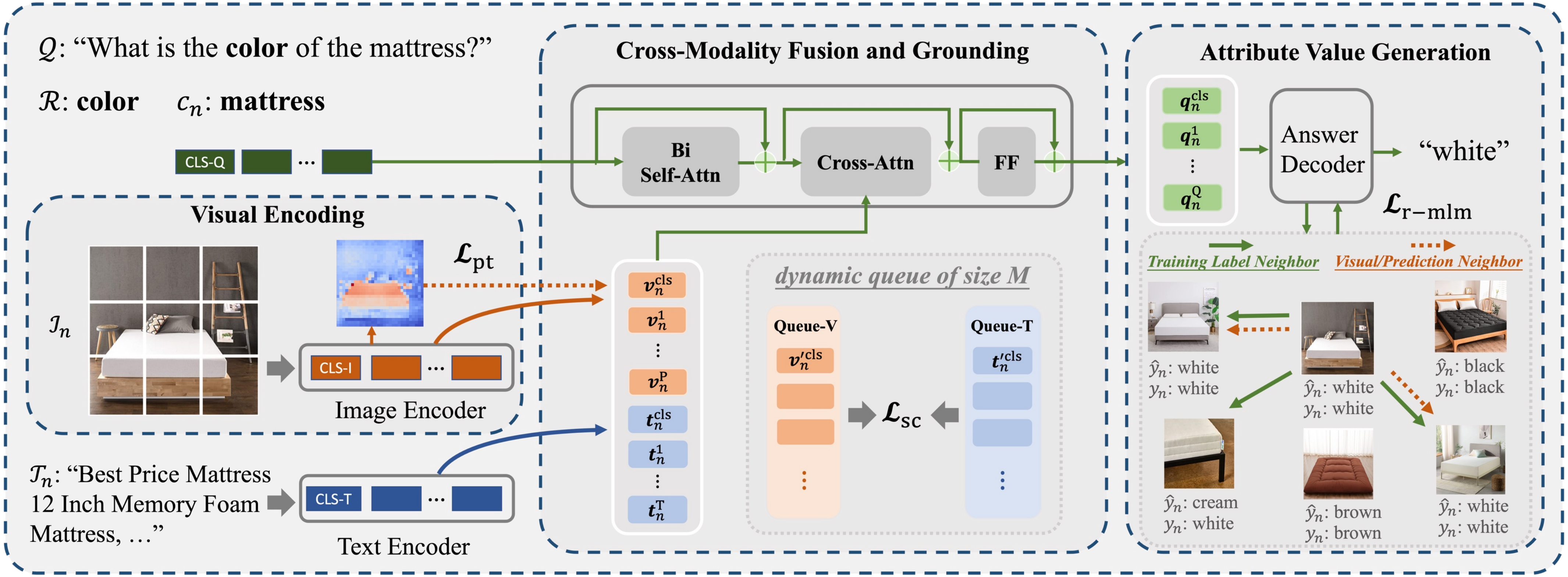
Outline

1. Introduction and Motivation
2. Proposed Method: PV2TEA
3. Experiment Results

Problem Definition

- **Task**: automatic attribute extraction from multimodal input
- **Input**: a query attribute $\mathcal{R}$ and a text-image dataset $\mathcal{D} = \{\mathcal{X}_n\}_{n=1}^{N} = \{(\mathcal{I}_n, \mathcal{T}_n, c_n)\}_{n=1}^{N}$ consisting of N samples (e.g., products)
  - $\mathcal{I}_n$ represents the profile image of $\mathcal{X}_n$
  - $\mathcal{T}_n$ represents the textual description
  - c_n is the sample category (e.g., product type)
- **Output**: infer attribute value y_n of the query attribute $\mathcal{R}$ for sample $\mathcal{X}_n$
- **Setting**: open-vocabulary, the number of candidate values is extensive and y_n can contain either single or multiple values

The Overview of PV2TEA



The PV2TEA model architecture with three modules, each equipped with a bias reduction scheme

Augmented Label-Smoothed Contrast for Multi-modality Loose Alignment (S1)

Augment the contrast to include sample comparison from two queues storing the most recent *M* visual and textual representations：

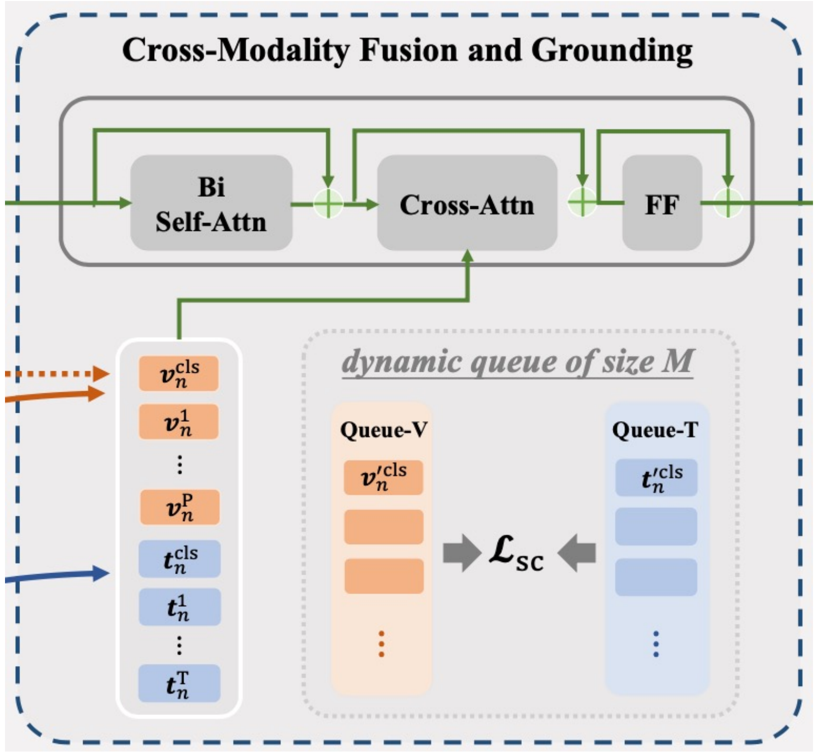
- **Intra-sample weak alignment**
  - Smooth the one-hot pairing label $\mathbf{p}_n^{i2t}$ with the pseudo-similarity $\mathbf{q}_n^{i2t}$

$$\widetilde{\mathbf{p}}_n^{i2t} = (1-\alpha)\mathbf{p}_n^{i2t} + \alpha \mathbf{q}_n^{i2t}$$

$$\mathbf{q}_n^{i2t} = \sigma\left(\mathcal{F}'_v\left(\mathcal{I}_n\right)^\top \mathcal{F}'_t\left(\mathcal{T}_n\right)\right) = \sigma\left(\boldsymbol{v}_n^{\prime\text{cls}\top} \boldsymbol{t}_n^{\prime\text{cls}}\right)$$

- **Potential inter-samples alignment**
  - Compare visual representation $\boldsymbol{v}_n^{\prime\text{cls}}$ with all textual representations $\boldsymbol{T}'$ in the queue to augment contrast

$$\mathbf{d}_n^{i2t} = \frac{\exp\left(\boldsymbol{v}_n^{\prime\text{cls}\top} \boldsymbol{T}'_m / \tau\right)}{\sum_{m=1}^{M} \exp\left(\boldsymbol{v}_n^{\prime\text{cls}\top} \boldsymbol{T}'_m / \tau\right)}$$



$$L_{i2t} = -\frac{1}{N}\left(\sum_{n=1}^{N} \widetilde{\mathbf{p}}_n^{i2t} \cdot \log\left(\mathbf{d}_n^{i2t}\right)\right)$$

$$L_{\text{SC}} = \left(L_{i2t} + L_{t2i}\right)/2$$

Label Noise in Adversarial Training: A Novel Perspective to Study Robust Overfitting. Chengyu Dong, Liyuan Liu and Jingbo Shang. *NeurIPS* 2022 (oral).

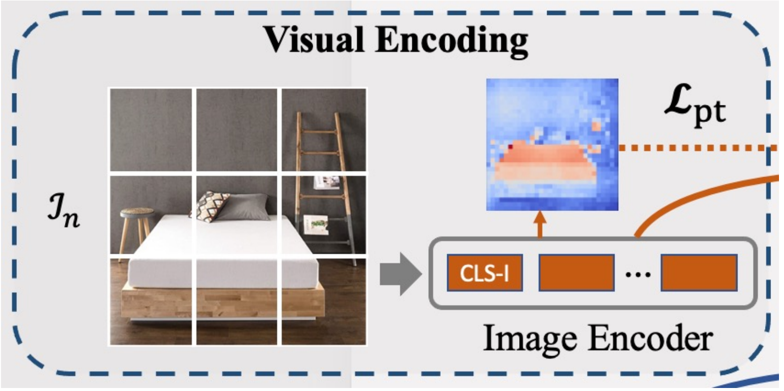
Visual Attention Pruning (S2)

Encourage the ViT encoder $\mathcal{F}$ focus on task-relevant foregrounds given the input image $\mathcal{I}_n$ with a product type aware attention pruning, supervised with product type classification,

$$L_{\text{pt}} = -\frac{1}{N}\left(\sum_{n=1}^{N} c_n \cdot \log\left(\mathcal{F}(\mathcal{I}_n)\right)\right)$$

The learned attention mask $\boldsymbol{M}$ is then applied on the visual representation sequences $\boldsymbol{v}_n$ of the whole image to screen out noisy backgrounds and task-irrelevant patches

$$\boldsymbol{v}_n^{pt} = \boldsymbol{v}_n \odot \sigma(\boldsymbol{M})$$



Two-level Neighborhood-regularized Sample Weight Adjustment (S3)

In each iteration, sample weight $s(\mathcal{X}_n)$ is updated based on its label reliability

$$\mathcal{L}_{\text{r-mlm}} = -\frac{1}{N}\left(\sum_{n=1}^{N} s\left(\mathcal{X}_n\right) \cdot g\left(y_n, \hat{y}_n\right)\right)$$

- **Visual Neighbor Regularization:**

for each sample $\mathcal{X}_n$ with $\boldsymbol{v}_n$, find its KNN neighbors in visual feature spaces: $\mathcal{N}_n = \{\mathcal{X}_n \cup \mathcal{X}_k \in \text{KNN}(\boldsymbol{v}_n, \mathcal{D}, K)\}$

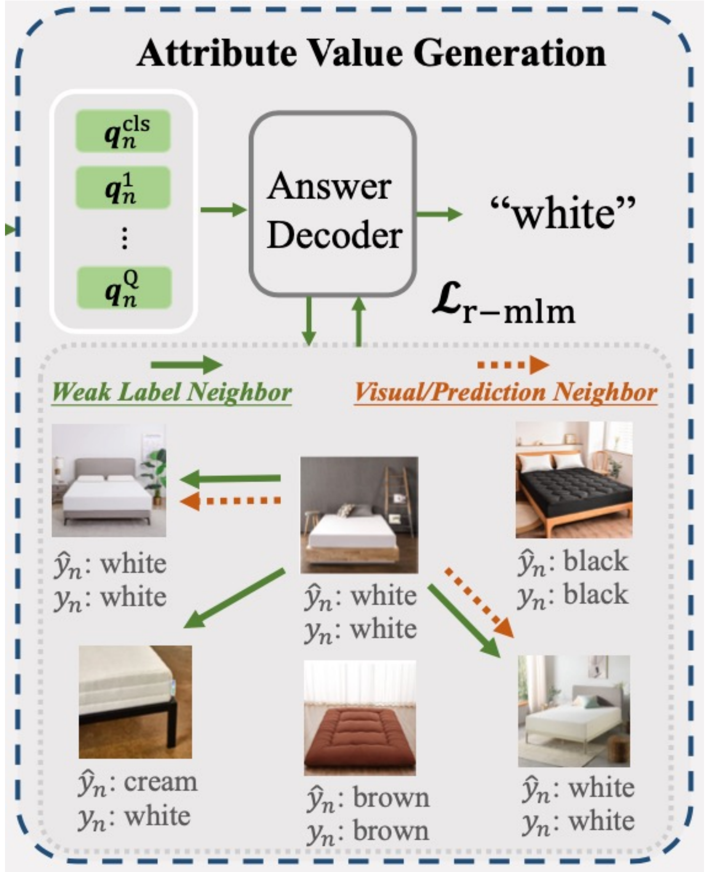
simultaneously, get the set of samples with the same training labels y_i as sample $\mathcal{X}_n$: $\mathcal{Y}_n = \{\mathcal{X}_n \cup \mathcal{X}_j \in \mathcal{D}_{y_j = y_n}\}$

The reliability of sample $\mathcal{X}_n$: $s_v(\mathcal{X}_n) = |\mathcal{N}_n \cap \mathcal{Y}_n| / K$

- **Prediction Neighbor Regularization**

similarly, find the sample set with the same predicted attribute values with $\mathcal{X}_n$ $\hat{\mathcal{Y}}_n = \{\mathcal{X}_n \cup \mathcal{X}_j \in \mathcal{D}_{\hat{y}_j = \hat{y}_n}\}$

The reliability of sample $\mathcal{X}_n$: $s_p\left(\mathcal{X}_n\right) = \left|\hat{\mathcal{Y}}_n \cap \mathcal{Y}_n\right| / \left|\hat{\mathcal{Y}}_n \cup \mathcal{Y}_n\right|$



$$s\left(\mathcal{X}_n\right) = \begin{cases} s_v\left(\mathcal{X}_n\right) & e < E, \\ \text{AVG}\left(s_v\left(\mathcal{X}_n\right), s_p\left(\mathcal{X}_n\right)\right) & e \geq E. \end{cases}$$

Outline

1. Introduction and Motivation
2. Proposed Method: PV2TEA
3. Experiment Results

Overall Performance

Attr	# PT	Value Type	# Valid	# Train & Val	# Test
Item Form	14	Single	142	42,911	4,165
Color	255	Multiple	24	106,176	3,777
Pattern	31	Single	30	119,622	2,093

Table 1: Statistics of the attribute extraction datasets.

Type	Method	Dataset: Item Form			Dataset: Color			Dataset: Pattern		
		Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
Unimodal	OpenTag_{seq}	91.37	44.97	60.27	83.94	24.73	38.20	79.65	19.83	31.75
	OpenTag_{cls}	89.40	51.67	65.49	81.13	28.61	42.30	78.10	24.41	37.19
	TEA	82.71	60.98	70.20	67.58	47.80	55.99	60.87	37.40	46.33
Multimodal	ViLBERT	75.97	65.67	70.45	60.22	51.12	55.30	60.10	40.52	48.40
	LXMERT	75.79	68.72	72.08	60.20	54.26	57.08	60.33	42.28	49.72
	UNITER	76.75	69.10	72.72	61.30	54.69	57.81	62.45	43.38	51.20
	BLIP	78.21	69.25	73.46	62.70	58.23	60.38	58.74	44.01	50.32
	PAM	78.83	74.35	<u>76.52</u>	63.34	60.43	<u>61.85</u>	61.80	44.29	<u>51.60</u>
Ours	PV2TEA w/o S1	80.03	72.49	76.07	71.00	58.41	64.09	60.03	45.59	51.82
	PV2TEA w/o S2	80.48	75.32	77.81	73.77	59.37	65.79	59.01	46.74	52.16
	PV2TEA w/o S3	80.87	72.71	76.57	74.29	59.04	65.79	59.92	44.92	51.35
	PV2TEA	82.46	75.40	**78.77**	77.44	60.19	**67.73**	62.10	46.84	**53.40**

Table 2: Performance comparison with different baselines (%). The performance gains over the baselines have passed the t-test with a p-value<0.05. The best performance is in bold, and the second runner baseline is underlined.

Observations:

- Comparing the unimodal methods with multimodal ones, textual-only models achieve impressive results on precision while greatly suffering from low recall
- Adding visual information can further improve recall, especially for the multi-value attribute, e.g., *Color*
- With the three proposed bias-reduction schemes, PV2TEA improves on all three metrics over multimodal baselines and balances precision and recall compared with unimodal models

Source Aware Evaluation & Case Study

Method	Gold Value Source	Precision	Recall	F_1
OpenTag$_{cls}$	Text ✓	89.78	52.13	65.96
	Text ✗ Image ✓	78.95	31.25	44.78
	GAP ↓	10.83	20.88	21.18
PAM	Text ✓	79.16	74.53	76.78
	Text ✗ Image ✓	66.67	58.33	62.22
	GAP ↓	12.50	16.20	<u>14.56</u>
PV2TEA	Text ✓	82.64	75.71	79.02
	Text ✗ Image ✓	75.00	62.50	68.18
	GAP ↓	7.64	13.21	**10.84**

Table 3: Fine-grained source-aware evaluation of different methods. The *gold value source* indicates whether the gold value is contained in the text, or is not contained in the text and must be inferred from the image.

The performance gap between when the gold value is present or absent in the text is significantly reduced by PV2TEA ⬇ indicates a more balanced and generalized capacity of PV2TEA to learn from different modalities.

Milumia Women Casual 2 Piece Outfits Tie Back Cami Crop Top Belted Pants Sets Navy Medium Material: 100% Polyester. Fabric is Non-stretch. Feature: Cami Crop Top with Pants Sets, Tie Hem, Bow, Spaghetti Strap, Sleeveless, Knot, Belted Pants, Striped Occasion: Perfect for Summer Beach, Vacation, Traveling, Holiday, Party, Weekend Casual, Going Out, Weekend Daily, Shopping and Dating wear. Season: Suitable for Spring, Summer

Q: what is the *pattern* of the *one-piece outfit*? **PV2TEA Prediction: striped**

WSERE 3 Pack Plastic Flip Top Bird Small Poultry Feeder for Pigeon Quails Ducklings Birds, No Mess No Waste Multihole Birds Feeding Dish Dispenser Chick Feeder

Q: what is the *color* of the *wildlife feeder*? **PV2TEA Prediction: red, yellow, green**

URATOT Glittered Christmas Tree Topper Metal Christmas Treetop Hallow Wire Star Topper for Christmas Home Decoration; Product material: this Christmas tree topper is made of quality plastic

Q: what is the *color* of the *decoration*? **PV2TEA Prediction: silver**

Sugar in the raw 500 packets 4 lbs 15 4 ounces cooking raw sugar. A natural unrefined sugar made from sugar cane grown in each packet holds approximately one teaspoon and has five grams of carbohydrates and 20 calories flavor: original; packing type: packets; premeasured: yes; capacity weight : 0 18 oz

Q: what is the *item form* of the *sugar*? **PV2TEA Prediction: crystal**

Figure 6: Qualitatively case studies.

Ablation Studies

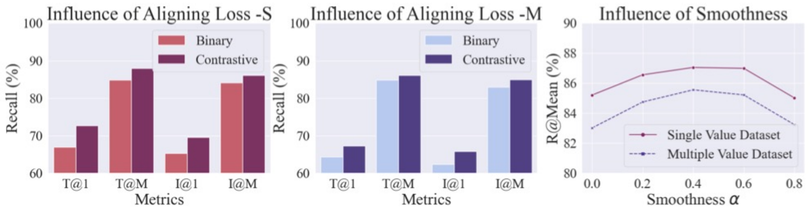


Figure 4: The influence study of alignment objectives, i.e., binary matching v.s. contrastive loss, and the influence of softness α via the task of image-to-text and text-to-image retrieval. The metric T/I@1 is the recall of text/image retrieval at rank 1, T/I@M means the rank average, and R@Mean further averages T@M and I@M.

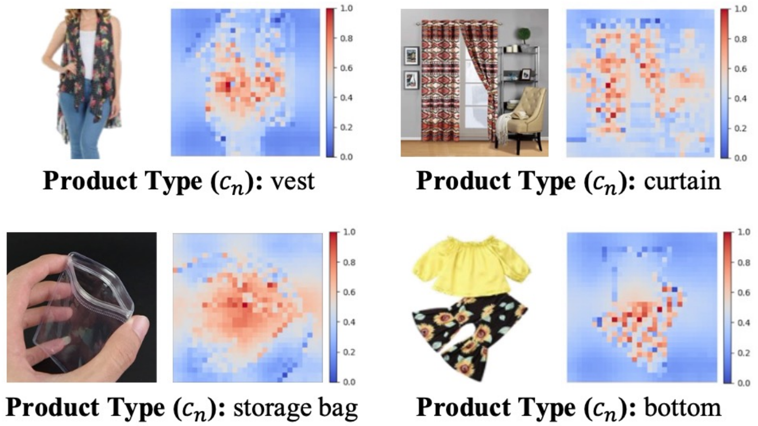


Figure 5: Visualization of learned attention mask with category (e.g., product type) aware ViT classification.

Method	Single Value Dataset			Multiple Value Dataset		
	P	R	F_1	P	R	F_1
w/o L_{sc}	80.03	72.49	76.07	71.00	58.41	64.09
w/o Smooth	81.42	74.41	77.76	75.06	59.99	66.68
PV2TEA	82.46	75.40	78.77	77.44	60.19	67.73

Table 4: Ablation study on the augmented label-smoothed contrast for cross-modality alignment (%).

Method	Single Value Dataset			Multiple Value Dataset		
	P	R	F_1	P	R	F_1
w/o L_{ct}	80.48	75.32	77.81	73.77	59.37	65.79
w/o Attn Prun	80.61	75.49	77.97	74.60	59.42	66.15
PV2TEA	82.46	75.40	78.77	77.44	60.19	67.73

Table 5: Ablation study on the category supervised visual attention pruning (%).

Method	Single Value Dataset			Multiple Value Dataset		
	P	R	F_1	P	R	F_1
w/o NR	80.87	72.71	76.57	74.29	59.04	65.79
w/o Vis-NR	81.87	73.54	77.48	77.07	59.99	67.47
w/o Pred-NR	81.81	73.18	77.25	76.71	59.44	66.98
PV2TEA	82.46	75.40	78.77	77.44	60.19	67.73

Table 6: Ablation study on the two-level neighborhood-regularized sample weight adjustment (%).

Ablation studies for the design modules in S1, S2, and S3 respectively

Summary, Thank You! Q&A

- PV2TEA is a bias-mitigated visual patching-up model for multimodal information extraction
  - Augment label-smoothed contrast promotes accurate & complete cross modal alignment
  - Visual attention pruning improves precision by masking out task-irrelevant regions
  - neighborhood-regularized sample weight adjustment reduces textual bias from noisy samples
- **Generalizable**: we anticipate the investigated challenges and solutions can inspire future scenarios where the task is first established on the text and then expanded to multiple modalities.
- **Limitations**:
  - multimodal alignment and fusion only consider a single image for each sample
  - attention pruning may filter out helpful text information on the images intentionally provided