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

Abstract

Information extraction, e.g., attribute value extraction, has been extensively studied and formulated based only on text. However, many attributes can benefit from image-based extraction, like color, shape, pattern, among others. The visual modality has long been underutilized, mainly due to multimodal annotation difficulty. In this paper, we aim to patch the visual modality to the textual-established attribute information extractor. The cross-modality integration faces several unique challenges: (C1) images and textual descriptions are loosely paired intra-sample and inter-samples; (C2) images usually contain rich backgrounds that can mislead the prediction; (C3) weakly supervised labels from textual-established extractors are biased for multimodal training. We present PV2TEA, an encoder-decoder architecture equipped with three bias reduction schemes: (S1) Augmented label-smoothed contrast to improve the cross-modality alignment for loosely-paired image and text; (S2) Attention-pruning that adaptively distinguishes the visual foreground; (S3) Two-level neighborhood regularization that mitigates the label textual bias via reliability estimation. Empirical results on real-world e-Commerce datasets[1] demonstrate up to 11.74% absolute (20.97% relatively) F_1 increase over unimodal baselines.

1 Introduction

Information extraction, e.g., attribute value extraction, aims to extract structured knowledge triples, i.e., (*sample_id, attribute, value*), from the unstructured information. As shown in Figure 1, the inputs include text descriptions and images (optional) along with the queried attribute, and the output is the extracted value. In practice, textual description has played as the main or only input in mainstream approaches for automatic attribute value extraction (Zheng et al., 2018; Xu et al., 2019; Wang et al., 2020; Karamanolakis et al., 2020; Yan et al., 2021; Ding et al., 2022). Such models perform well when the prediction targets are inferrable from the text.

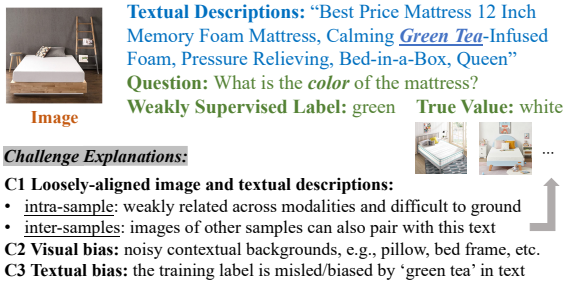

Figure 1: Illustration of multimodal attribute extraction and the challenges in cross-modality integration.

As the datasets evolve, interest in incorporating visual modality naturally arises, especially for image-driven attributes, e.g., *Color*, *Pattern*, *Item Shape*. Such extraction tasks rely heavily on visual information to obtain the correct attribute values. The complementary information contained in the images can improve recall in cases where the target values are not mentioned in the texts. In the meantime, the cross-modality information can help with ambiguous cases and improve precision.

However, extending a single-modality task to multi-modality can be very challenging, especially due to the lack of annotations in the new modality. Performing accurate labeling based on multiple modalities requires the annotator to refer to multiple information resources, leading to a high cost of human labor. Although there are some initial explorations on multimodal attribute value extraction (Zhu et al., 2020; Lin et al., 2021; De la Comble et al., 2022), all of them are fully supervised and overlook the resource-constrained setting of building a multimodal attribute extraction framework based on the previous textual-established models. In this paper, we aim to patch the visual modality to attribute value extraction by leveraging

*Work was done when Hejie was an intern at Amazon.
[1]The code and the human-annotated datasets with fine-grained source modality labels of gold values are available at https://github.com/HennyJie/PV2TEA.

textual-based models for weak supervision, thus reducing the manual labeling effort.

Challenges. Several unique challenges exist in visual modality patching: **C1.** Images and their textual descriptions are usually *loosely aligned* in two aspects: From the intra-sample aspect, they are usually weakly related considering the rich characteristics, making it difficult to ground the language fragments to the corresponding image regions; From the inter-samples aspect, it is commonly observed that the text description of one sample may also partially match the image of another. As illustrated in Figure 1, the textual description of the *mattress* product is fragmented and can also correspond to other images in the training data. Therefore, traditional training objectives for multimodal learning such as binary matching (Kim et al., 2021) or contrastive loss (Radford et al., 2021) that only treat the text and image of the same sample as positive pairs may not be appropriate. **C2.** Bias can be brought by the *visual input* from the *noisy contextual background*. The images usually not only contain the interested object itself but also demonstrate a complex background scene. Although the backgrounds are helpful for scene understanding, they may also introduce spurious correlation in a fine-grained task such as attribute value extraction, which leads to imprecise prediction (Xiao et al., 2021; Kan et al., 2021). **C3.** Bias also exists in *language perspective* regarding the *biased weak labels* from textual-based models. As illustrated in Figure 1, the color label of *mattress* is misled by '*green tea infused*' in the text. These noisy labels can be more catastrophic for a multimodal model due to their incorrect grounding in images. Directly training the model with these biased labels can lead to gaps between the stronger language modality and the weaker vision modality (Yu et al., 2021).

Solutions. We propose PV2TEA, a sequence-to-sequence backbone composed of three modules: visual encoding, cross-modality fusion and grounding, and attribute value generation, each with a bias-reduction scheme dedicated to the above challenges: **S1.** To better integrate the *loosely-aligned texts and images*, we design an augmented label-smoothed contrast schema for cross-modality fusion and grounding, which considers both the intra-sample weak correlation and the inter-sample potential alignment, encouraging knowledge transfer from the strong textual modality to the weak visual one. **S2.** During the visual encoding, we equip PV2TEA with an attention-pruning mechanism that adaptively distinguishes the distracting background and *attends to the most relevant regions* given the entire input image, aiming to improve precision in the fine-grained task of attribute extraction. **S3.** To mitigate the bias from *textual-biased weak labels*, a two-level neighborhood regularization based on visual features and previous predictions, is designed to emphasize trustworthy training samples while mitigating the influence of textual-biased labels. In this way, the model learns to generate more balanced results rather than being dominated by one modality of information. In summary, the main contributions of PV2TEA are three-fold:

- We propose PV2TEA, an encoder-decoder framework effectively patching up visual modality to textual-established attribute value extraction.
- We identify three unique challenges in patching visual modality for information extraction, with solutions for *intra-sample and inter-samples loose alignment* and bias from *complex visual background* and *textual-biased labels*.
- We release three human-annotated datasets with modality source labels of the gold values to facilitate fine-grained evaluation. Extensive results validate the effectiveness of PV2TEA.

2 Preliminaries

2.1 Problem Definition

We consider the task of automatic attribute extraction from multimodal input, i.e., textual descriptions and images. Formally, the input is a query attribute $\mathcal{R}$ and a text-image pairs 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), where $\mathcal{I}_n$ represents the profile image of $\mathcal{X}_n$, $\mathcal{T}_n$ represents the textual description and c_n is the sample category (e.g., product type). The model is expected to infer attribute value y_n of the query attribute $\mathcal{R}$ for sample $\mathcal{X}_n$. We consider the challenging setting with open-vocabulary attributes, where the number of candidate values is extensive and y_n can contain either single or multiple values.

2.2 Motivating Analysis on the Textual Bias of Attribute Information Extraction

Existing textual-based models or multimodal models directly trained with weak labels suffer from a strong bias toward the texts. As illustrated in Figure 1, the training label for the *color* attribute of the *mattress* is misled by '*green tea infused*'

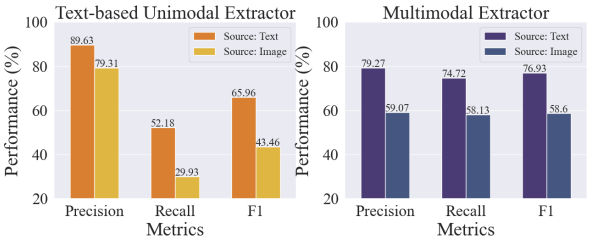

Figure 2: Source-aware evaluation of existing unimodal and multimodal models on the textual-biased issue.

from the textual profile. Models trained with such textual-shifted labels will result in a learning ability gap between modalities, where the model learns better from the textual than the visual modality. To quantitatively study the learning bias, we conduct fine-grained source-aware evaluations on a real-world e-Commerce dataset with representative unimodal and multimodal methods, namely OpenTag (Zheng et al., 2018) with the classification setup and PAM (Lin et al., 2021). Specifically, for each sample in the test set, we collect the source of the gold value (i.e., text or image). Experiment results are shown in Figure 2, where label *Source: Text* indicates the gold value is present in the text, while label *Source: Image* indicates the gold value is absent from the text and must be inferred from the image. It is shown that both the text-based unimodal extractor and multimodal extractor achieve impressive results when the gold value is contained in the text. However, when the gold value is not contained in the text and must be derived from visual input, the performance of all three metrics drops dramatically, indicating a strong textual bias and dependence of existing models.

3 PV2TEA

We present the backbone architecture and three bias reduction designs of PV2TEA, shown in Figure 3. The backbone is formulated based on visual question answering (VQA) composed of three modules:
(1) **Visual Encoding.** We adopt the Vision Transformer (ViT) (Dosovitskiy et al., 2021) as the visual encoder. The given image $\mathcal{I}_n$ is divided into patches and featured as a sequence of tokens, with a special token `[CLS-I]` appended at the head of the sequence, whose representation $\boldsymbol{v}_n^{\text{cls}}$ stands for the whole input image $\mathcal{I}_n$.
(2) **Cross-Modality Fusion and Grounding.** Following the VQA paradigm, we define the question prompt as "*What is the* $\mathcal{R}$ *of the* c_n?", with a special token `[CLS-Q]` appended at the beginning. A unimodal BERT (Devlin et al., 2019) encoder is adopted to produce token-wise textual representation from sample profiles (title, bullets, and descriptions). The visual representations of P image patches $\boldsymbol{v}_n = [\boldsymbol{v}_n^{\text{cls}}, \boldsymbol{v}_n^1, \ldots, \boldsymbol{v}_n^{\text{P}}]$ are concatenated with the textual representation of T tokens $\boldsymbol{t}_n = [\boldsymbol{t}_n^{\text{cls}}, \boldsymbol{t}_n^1, \ldots, \boldsymbol{t}_n^{\text{T}}]$, which is further used to perform cross-modality fusion and grounding with the question prompt through cross-attention. The output $\boldsymbol{q}_n = [\boldsymbol{q}_n^{\text{cls}}, \boldsymbol{q}_n^1, \ldots, \boldsymbol{q}_n^{\text{Q}}]$ is then used as the grounded representation for the answer decoder.
(3) **Attribute Value Generation.** We follow the design from (Li et al., 2022a), where each block of the decoder is composed of a causal self-attention layer, a cross-attention layer, and a feed-forward network. The decoder takes the grounded multimodal representation as input and predicts the attribute value $\hat{y}_n$ in a generative manner[2].
Training Objectives. The overall training objective of PV2TEA is formulated as

$$\mathcal{L} = \mathcal{L}_{\text{sc}} + \mathcal{L}_{\text{ct}} + \mathcal{L}_{\text{r-mlm}}, \quad (1)$$

where the three loss terms, namely augmented label-smoothed contrastive loss $\mathcal{L}_{\text{sc}}$ (Section 3.1), category aware ViT loss $\mathcal{L}_{\text{ct}}$ (Section 3.2), and neighborhood-regularized mask language modeling loss $\mathcal{L}_{\text{r-mlm}}$ (Section 3.3) correspond to each of the three prementioned modules respectively.

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

Contrastive objectives have been proven effective in multimodal pre-training (Radford et al., 2021) by minimizing the representation distance between different modalities of the same data point while keeping those of different samples away (Yu et al., 2022). However, for attribute value extraction, the image and textual descriptions are typically *loosely aligned* from two perspectives: (1) *Intra-sample weak alignment*: The text description may not necessarily form a coherent and complete sentence, but a set of semantic fragments describing multiple facets. Thus, grounding the language to corresponding visual regions is difficult. (2) *Potential inter-samples alignment*: Due to the commonality of samples, the textual description of one sample may also correspond to the image of another. Thus, traditional binary matching and contrastive objectives become suboptimal for these loosely-aligned texts and images.

To handle the looseness of images and texts, we

[2]We compared the settings of generation and classification for the attribute value extractor. See results in Section 5.2.

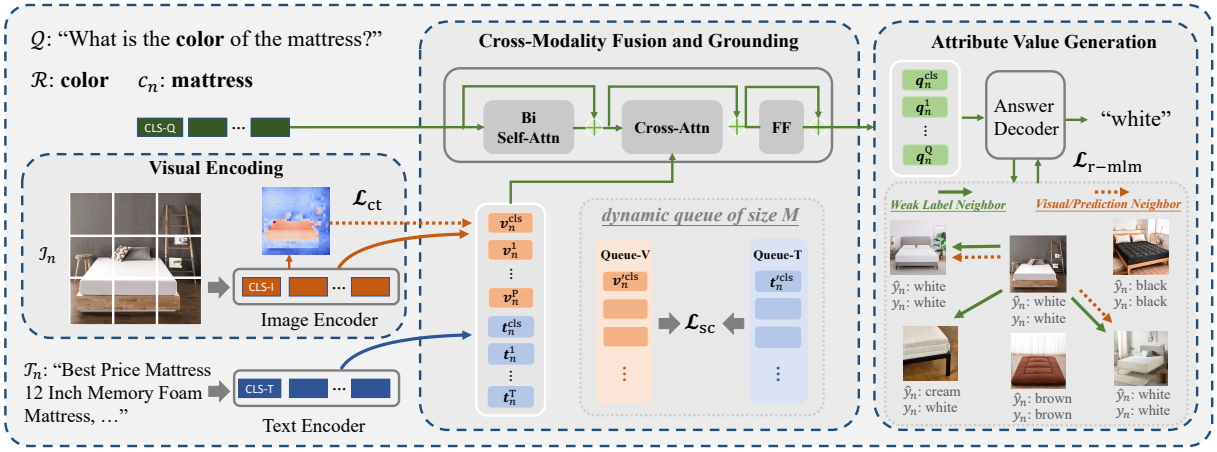

Figure 3: The overview of PV2TEA model architecture with three modules, where each of them is equipped with a bias reduction scheme corresponding to the discussed challenges in Figure 1.

augment the contrast to include sample comparison outside the batch with two queues storing the most recent M ($M \gg$ batch size B) visual and textual representations, inspired by the momentum contrast in MoCo (He et al., 2020) and ALBEF (Li et al., 2021). For the *intra-sample weak alignment* of each given sample $\mathcal{X}_n$, instead of using the one-hot pairing label $\mathbf{p}_n^{i2t}$, we smooth the pairing target with the pseudo-similarity $\mathbf{q}_n^{i2t}$,

$$\widetilde{\mathbf{p}}_n^{i2t} = (1-\alpha)\mathbf{p}_n^{i2t} + \alpha \mathbf{q}_n^{i2t}, \tag{2}$$

where α is a hyper-parameter and $\mathbf{q}_n^{i2t}$ is calculated by softmax over the representation multiplication of the `[CLS]` tokens, $\boldsymbol{v}_n^{'\text{cls}}$ and $\boldsymbol{t}_n^{'\text{cls}}$, from momentum unimodal encoders $\mathcal{F}'_v$ and $\mathcal{F}'_t$,

$$\mathbf{q}_n^{i2t} = \sigma\left(\mathcal{F}'_v\left(\mathcal{I}_n\right)^\top \mathcal{F}'_t\left(\mathcal{T}_n\right)\right) = \sigma\left(\boldsymbol{v}_n^{'\text{cls}\top} \boldsymbol{t}_n^{'\text{cls}}\right). \tag{3}$$

For *potential inter-samples pairing relations*, the visual representation $\boldsymbol{v}_n^{'\text{cls}}$ is compared with all textual representations $\boldsymbol{T}'$ in the queue to augment contrastive loss. Formally, the predicted image-to-text matching probability of $\mathcal{X}_n$ is

$$\mathbf{d}_n^{i2t} = \frac{\exp\left(\boldsymbol{v}_n^{'\text{cls}\top} \boldsymbol{T}'_m / \tau\right)}{\sum_{m=1}^{M} \exp\left(\boldsymbol{v}_n^{'\text{cls}\top} \boldsymbol{T}'_m / \tau\right)}. \tag{4}$$

With the smoothed targets from Equation (2), the *image-to-text* contrastive loss L_{i2t} is calculated as the cross-entropy between the smoothed targets $\widetilde{\mathbf{p}}_n^{i2t}$ and contrast-augmented predictions $\mathbf{d}_n^{i2t}$,

$$L_{i2t} = -\frac{1}{N}\left(\sum_{n=1}^{N} \widetilde{\mathbf{p}}_n^{i2t} \cdot \log\left(\mathbf{d}_n^{i2t}\right)\right), \tag{5}$$

and vise versa for the *text-to-image* contrastive loss L_{t2i}. Finally, the augmented label-smoothed contrastive loss L_{sc} is the average of these two terms,

$$L_{\text{sc}} = \left(L_{i2t} + L_{t2i}\right)/2. \tag{6}$$

3.2 Visual Attention Pruning (S2)

Images usually contain not only the visual foreground of the concerned category but also rich background contexts. Although previous studies indicate context can serve as an effective cue for visual understanding (Doersch et al., 2015; Zhang et al., 2020; Xiao et al., 2021), it has been found that the output of ViT is often based on supportive signals in the background rather than the actual object (Chefer et al., 2022). Especially in a fine-grained task such as attribute value extraction, the associated backgrounds could distract the visual model and harm the prediction precision. For example, when predicting the color of *birthday balloons*, commonly co-occurring contexts such as *flowers* could mislead the model and result in wrongly predicted values.

To encourage the ViT encoder $\mathcal{F}$ focus on task-relevant foregrounds given the input image $\mathcal{I}_n$, we add a category-aware attention pruning schema, supervised with category classification,

$$L_{\text{ct}} = -\frac{1}{N}\left(\sum_{n=1}^{N} c_n \cdot \log\left(\mathcal{F}(\mathcal{I}_n)\right)\right). \tag{7}$$

In real-world information extraction tasks, 'category' denote classification schemas for organizing and structuring diverse data, exemplified by the broad range of product types in e-commerce, such as electronics, clothing, or books. These categories not only display vast diversity but also have distinct data distributions and properties, adding layers of complexity to the information extraction scenarios.

The learned attention mask $\boldsymbol{M}$ in ViT can gradually resemble the object boundary of the interested category and distinguishes the most important task-

related regions from backgrounds by assigning different attention weights to the image patches (Selvaraju et al., 2017). The learned $\boldsymbol{M}$ is then applied on the visual representation sequences $\boldsymbol{v}_n$ of the whole image,

$$\boldsymbol{v}_n^{pt} = \boldsymbol{v}_n \odot \sigma(\boldsymbol{M}), \tag{8}$$

to screen out noisy background and task-irrelevant patches before concatenating with the textual representation $\boldsymbol{t}_n$ for further cross-modal grounding.

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

Weak labels from established models can be noisy and biased toward the textual input. Directly training the models with these labels leads to a learning gap across modalities. Prior work on self-training shows that embedding similarity can help to mitigate the label errors issue (Xu et al., 2023; Lang et al., 2022). Inspired by this line of work, we design a two-level neighborhood-regularized sample weight adjustment. In each iteration, sample weight $s(\mathcal{X}_n)$ is updated based on its label reliability, which is then applied to the training objective of attribute value generation in the next iteration,

$$\mathcal{L}_{\text{r-mlm}} = -\frac{1}{N}\left(\sum_{n=1}^{N} s(\mathcal{X}_n) \cdot g(y_n, \hat{y}_n)\right), \tag{9}$$

where g measures the element-wise cross entropy between the training label y_n and the prediction $\hat{y}_n$. As illustrated by the right example in Figure 3[3], where green arrows point to samples with the same training label as y_n, and red arrows point to either visual or prediction neighbors, a higher consistency between the two sets indicates a higher reliability of y_n, formally explained as below:

(1) Visual Neighbor Regularization. The first level of regularization is based on the consistency between the sample set with the same training label y_n and visual feature neighbors of $\mathcal{X}_n$. For each sample $\mathcal{X}_n$ with visual representation $\boldsymbol{v}_n$, we adopt the K-nearest neighbors (KNN) algorithm to find its neighbor samples in the visual feature space:

$$\mathcal{N}_n = \{\mathcal{X}_n \cup \mathcal{X}_k \in \text{KNN}(\boldsymbol{v}_n, \mathcal{D}, K)\}, \tag{10}$$

where $\text{KNN}(\boldsymbol{v}_n, \mathcal{D}, K)$ demotes K samples in $\mathcal{D}$ with visual representation nearest to $\boldsymbol{v}_n$. Simultaneously, we obtain the set of samples in $\mathcal{D}$ with the same training label y_j as that of the sample $\mathcal{X}_n$,

$$\mathcal{Y}_n = \{\mathcal{X}_n \cup \mathcal{X}_j \in \mathcal{D}_{y_j = y_n}\}. \tag{11}$$

The reliability of sample $\mathcal{X}_n$ based on the visual neighborhood regularization is

$$s_v(\mathcal{X}_n) = |\mathcal{N}_n \cap \mathcal{Y}_n| / K. \tag{12}$$

(2) Prediction Neighbor Regularization. The second level of regularization is based on the consistency between the sample set with the same training label and the prediction neighbors from the previous iteration, which represents the learned multimodal representation. Prediction regularization is further added after E epochs when the model can give relatively confident predictions, ensuring the predicted values are qualified for correcting potential noise. Formally, we obtain the set of samples in $\mathcal{D}$ whose predicted attribute value p_j from the last iteration is the same as that of the sample $\mathcal{X}_n$,

$$\hat{\mathcal{Y}}_n = \{\mathcal{X}_n \cup \mathcal{X}_j \in \mathcal{D}_{\hat{y}_j = \hat{y}_n}\}. \tag{13}$$

With the truth-value consensus set $\mathcal{Y}_n$ from Equation (11), the reliability based on previous prediction neighbor regularization of the sample $\mathcal{X}_n$ is

$$s_p(\mathcal{X}_n) = \left|\hat{\mathcal{Y}}_n \cap \mathcal{Y}_n\right| / \left|\hat{\mathcal{Y}}_n \cup \mathcal{Y}_n\right|. \tag{14}$$

Overall, $s(\mathcal{X}_n)$ is initially regularized with visual neighbors and jointly with prediction neighbors after E epochs when the model predicts credibly,

$$s(\mathcal{X}_n) = \begin{cases} s_v(\mathcal{X}_n) & e < E, \\ \text{AVG}(s_v(\mathcal{X}_n), s_p(\mathcal{X}_n)) & e \geq E. \end{cases} \tag{15}$$

[3]See Appendix G for additional demo examples.

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.

4 Experimental Setup

4.1 Dataset and Implementation Details

We build three multimodal attribute value extraction datasets by collecting profiles (title, bullets, and descriptions) and images from the public `amazon.com` web pages, where each dataset corresponds to one attribute $\mathcal{R}$. The dataset information is summarized in Table 1, where **Attr** is the attribute $\mathcal{R}$, **# PT** represents the number of unique categories (i.e., product types), **Value Type** indicates whether y_n contain single or multiple values, and **# Valid** represents the number of valid values. To better reflect real-world scenarios, we use the attribute-value pairs from the product information section on web pages as weak training labels instead of highly processed data. We follow the same filtering strategy from prior text established work (Zalmout and Li, 2022) to denoise training data. For the testing, we manually annotate gold

Type	Method	Dataset: Item Form			Dataset: Color			Dataset: Pattern		
		Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
Unimodal	$\text{OpenTag}_{\text{seq}}$	91.37	44.97	60.27	83.94	24.73	38.20	79.65	19.83	31.75
	$\text{OpenTag}_{\text{cls}}$	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.

labels on the benchmark dataset to ensure preciseness. Besides, the label sources are marked down, indicating whether the attribute value is present or absent in the text, to facilitate fine-grained source-aware evaluation. The human-annotated benchmark datasets will be released to encourage the future development of modality-balanced multimodal extraction models. See Appendix A for the implementation and computation details of PV2TEA.

4.2 Evaluation Protocol

We use Precision, Recall, and F1 score based on synonym normalized exact string matching. For single value type, an extracted value $\hat{y}_n$ is considered correct when it exactly matches the gold value string y_n. For multiple value type where the gold values for the query attribute $\mathcal{R}$ can contain multiple answers $y_n \in \{y_n^1, \ldots, y_n^m\}$, the extraction is considered correct when all the gold values are matched in the prediction. Macro-aggregation is performed across attribute values to avoid the influence of class imbalance. All reported results are the average of three runs under the best settings.

4.3 Baselines

We compare our proposed model with a series of baselines, spanning unimodal-based methods and multimodal-based ones. For unimodal baselines, OpenTag (Zheng et al., 2018) is considered a strong text-based model for attribute extraction. $\text{OpenTag}_{\text{seq}}$ formulates the task as sequence tagging and uses the BiLSTM-CRF architecture with self-attention. $\text{OpenTag}_{\text{cls}}$ replaces the BiLSTM encoder with a transformer encoder and tackles the task as classification. TEA is another text-only uni-

Method	Gold Value Source	Precision	Recall	F_1
$\text{OpenTag}_{\text{cls}}$	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.

modal generative model with the same architecture as PV2TEA but without the image patching, which is included to demonstrate the influence of the generation setting. For multimodal baselines, we consider discriminative encoder models, including ViLBERT (Lu et al., 2019), LXMERT (Tan and Bansal, 2019) with dual encoders, and UNITER (Chen et al., 2020) with a joint encoder. We also add generative encoder-decoder models for comparisons. BLIP (Li et al., 2022a) adopts dual encoders and an image-grounded text decoder. PAM (Lin et al., 2021) uses a shared encoder and decoder separated by a prefix causal mask.

5 Experimental Results

5.1 Overall Comparison

Table 2 shows the performance comparison of different types of extraction methods. It is shown that PV2TEA achieves the best F_1 performance, especially compared to unimodal baselines, demon-

strating the advantages of patching visual modality to this text-established task. Comparing the unimodal methods with multimodal ones, textual-only models achieve impressive results on precision while greatly suffering from low recall, which indicates potential information loss when the gold value is not contained in the input text. With the generative setting, TEA sort of mitigates the information loss and improves recall over OpenTag under the tagging and classification settings. Besides, adding visual information can further improve recall, especially for the multi-value attribute *Color*, where multimodal models can even double that of text-only ones. However, the lower precision performance of the multimodal models implies the challenges beneath cross-modality integration. With the three proposed bias-reduction schemes, PV2TEA improves on all three metrics over multimodal baselines and balances precision and recall to a great extent compared with unimodal models. Besides the full PV2TEA, we also include three variants that remove one proposed schema at a time. It shows that the visual attention pruning module mainly helps with precision while the other two benefit both precision and recall, leading to the best F_1 performance when all three schemes are equipped. We include several case studies in Section 5.3 for qualitative observation.

Source-Aware Evaluation. To investigate how the modality learning bias is addressed, we conduct fine-grained source-aware evaluation similarly to Section 2.2, as shown in Table 3[4]. The performance gap between when the gold value is present or absent in the text is significantly reduced by PV2TEA when compared to both unimodal and multimodal representative methods, which suggests a more balanced and generalized capacity of PV2TEA to learn from different modalities. When the gold value is absent in the text, our method outperforms $\text{OpenTag}_{\text{cls}}$ by more than twice as much on recall, and also outperforms on precision under various scenarios compared to the multimodal PAM.

5.2 Ablation Studies

Augmented Label-Smoothed Contrast. We look into the impact of label-smoothed contrast on both single- and multiple-value type datasets [5]. Table 4 shows that removing the contrastive objective leads to a drop in both precision and recall. For the multiple-value dataset, adding the contrastive objective significantly benefits precision, suggesting it encourages cross-modal validation when there are multiple valid answers in the visual input. With label smoothing, the recall can be further improved. This indicates that the augmented and smoothed contrast can effectively leverage the cross-modality alignment inter-samples, hence improving the coverage rate when making predictions.

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 (%).

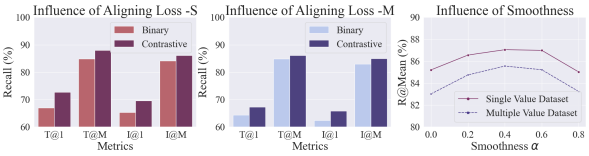

Figure 4: The influence study of alignment objectives, i.e., binary matching v.s. contrastive loss, and the influence of softness α via the task of image-to-text and text-to-image retrieval. The metric T/I@1 is the recall of text/image retrieval at rank 1, T/I@M means the rank average, and R@Mean further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

In addition, we conduct cross-modality retrieval to study the efficacy of aligning objectives, i.e., binary matching and contrastive loss, for cross-modality alignment and the influence of the softness α, as shown in Figure 4. Across different datasets and metrics, the contrastive loss consistently outperforms the binary matching loss. This consolidates our choice of contrasting objectives and highlights the potential benefits of label-smoothing and contrast augmentation, given that both are neglected in a binary matching objective. Retrieval performance under different smoothness values shows a trend of first rising and then falling. We simply take 0.4 for α in our experiments.

Category Aware Attention Pruning. We study

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 (%).

[4]We demonstrate results on the Item Form dataset due to limited space. For more results, please refer to Appendix B.

[5]For ablation analysis, we select Item Form as the representative for single-value and Color for multiple-value type dataset. More ablation results can be referred in Appendix C.

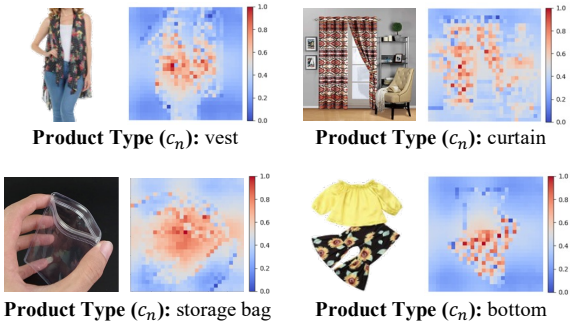

Figure 5: Visualization of learned attention mask with category (e.g., product type) aware ViT classification.

Method	Single Value Dataset			Multiple Value Dataset		
	P	R	F_1	P	R	F_1
w/o 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 (%).

the influence of the category aware attention pruning, as shown in Table 5. The results imply that adding the category classification helps to improve precision performance without harming recall, and the learned attention mask can effectively highlight the foreground regions of the queried sample. Figure 5 presents several visualizations of the learned attention mask.

Neighborhood Regularization. We consider the influence of the two-level neighborhood regularization by removing the visual neighborhood regularization (Vis-NR), prediction neighborhood regularization (Pred-NR), or both (NR) from the full model. Results in Table 6 show all the metrics decrease when both regularizations are removed, indicating the validity of the proposed neighborhood regularized sample weight adjustment in mitigating the influence of hard, noisy samples. Besides, since the second-level prediction-based neighbor regularization is independent of the multimodal extraction framework, it can be incorporated flexibly into other frameworks as well for future usage.

Classification vs. Generation To determine which architecture is better for multimodal attribute value extraction, we compare the generation and classification settings for the module of the attribute information extractor. The results are demonstrated in Table 7. It is shown that the setting of generation achieves significant advantages over classification. Especially on the recall performance for multi-value type attribute Color, where the gold value can be multiple, the improvement of recall can be up to 20% relatively. This indicates that the generation setting can extract more complete results from the multimodal input, leading to a higher coverage rate. Therefore, we choose the generation setting in the attribute value extraction module in the final architecture design of PV2TEA.

Setting	$\mathcal{D}$: Item Form			$\mathcal{D}$: Color			$\mathcal{D}$: Pattern		
	P	R	F_1	P	R	F_1	P	R	F_1
Classification	79.93	70.47	74.90	72.21	50.18	59.21	59.08	42.16	49.21
Generation	82.46	75.40	78.77	77.44	60.19	67.73	62.10	46.84	53.40

Table 7: Attribute extraction performance comparison between the settings of classification and generation.

5.3 Case Study

Figure 6: Qualitatively case studies.

To qualitatively observe the extraction performance, we attach several case studies in Figure 6. It shows that even when the attribute value is not contained in the text, PV2TEA can still perform the extraction reliably from images. In multiple value datasets such as Color, PV2TEA can effectively differentiate related regions and extract multiple values with comprehensive coverage.

6 Related Work

Attribute Information Extraction. Attribute extraction has been extensively studied in the literature primarily based on textual input. OpenTag (Zheng et al., 2018) formalizes it as a sequence tagging task and proposes a combined model leveraging bi-LSTM-CRF, and attention to perform end-to-end tagging. Xu et al. (2019) scales the sequence-tagging-based model with a global set of BIO tags. AVEQA (Wang et al., 2020) develops a question-answering model by treating each

attribute as a question and extracting the best answer span from the text. TXtract (Karamanolakis et al., 2020) uses a hierarchical taxonomy of categories and improves value extraction through multi-task learning. AdaTag (Yan et al., 2021) exploits an adaptive CRF-based decoder to handle multi-attribute value extractions. Additionally, there have been a few attempts at multimodal attribute value extraction. M-JAVE (Zhu et al., 2020) introduces a gated attention layer to combine information from the image and text. PAM (Lin et al., 2021) proposes a transformer-based sequence-to-sequence generation model for multimodal attribute value extraction. Although the latter two use both visual and textual input, they fail to account for possible modality bias and are fully supervised.

Multi-modality Alignment and Fusion. The goal of multimodal learning is to process and relate information from diverse modalities. CLIP (Radford et al., 2021) makes a gigantic leap forward in bridging embedding spaces of image and text with contrastive language-image pretraining. ALBEF (Li et al., 2021) applies a contrastive loss to align the image and text representation before merging with cross-modal attention, which fits loosely-aligned sample image and text. Using noisy picture alt-text data, ALIGN (Jia et al., 2021) jointly learns representations applicable to either vision-only or vision-language tasks. The novel Vision-Language Pre-training (VLP) framework established by BLIP (Li et al., 2022a) is flexibly applied to both vision-language understanding and generation tasks. GLIP (Li et al., 2022b) offers a grounded language-image paradigm for learning semantically rich visual representations. FLAVA (Singh et al., 2022) creates a foundational alignment that simultaneously addresses vision, language, and their interconnected multimodality. Flamingo (Alayrac et al., 2022) equips the model with in-context few-shot learning capabilities. SimVLM (Wang et al., 2022b) is trained end-to-end with a single prefix language modeling and investigates large-scale weak supervision. Multi-way Transformers are introduced in BEIT-3 (Wang et al., 2022a) for generic modeling and modality-specific encoding.

7 Conclusion

In this work, we propose PV2TEA, a bias-mitigated visual modality patching-up model for multimodal information extraction. Specifically, we take attribution value extraction as an example for illustration. Results on our released source-aware benchmarks demonstrate remarkable improvements: the augmented label-smoothed contrast promotes a more accurate and complete alignment for loosely related images and texts; the visual attention pruning improves precision by masking out task-irrelevant regions; and the neighborhood-regularized sample weight adjustment reduces textual bias by lowering the influence of noisy samples. We anticipate the investigated challenges and proposed solutions will inspire future scenarios where the task is first established on the text and then expanded to multiple modalities.

Limitations

There are several limitations that can be considered for future improvements: (1) In multimodal alignment and fusion, we only consider a single image for each sample, whereas multiple images can be available. A more flexible visual encoding architecture that can digest an indefinite number of input images can improve the visual information coverage; 2) The empirical results in this work focus on three attribute extraction datasets (i.e., item form, color, and pattern) that can clearly benefit from visual perspectives, while there are also various attribute types that rely more on the textual input. Different traits of attributes may influence the preferred modalities during the modeling, which is out of scope for this work but serves as a natural extension of this study; 3) Currently there is no specific design to improve the efficiency based on the visual question answering architecture. It can be not scalable as the number of attributes increases.

There could be a dual-use regarding the attention-pruning mechanism, which can be a potential risk of this work that could arise and harm the result. The attention-pruning mechanism encourages the model to focus on the task-relevant foreground on the given image selected with category supervision, which can improve the prediction precision given the input image is visually rich and contains noisy context background. While for some types of images, such as infographics, there may be helpful text information on the images or intentionally attached by providers. These additional texts may be overlooked by the attention-pruning mechanism, resulting in potential information losses. A possible mitigation strategy is to add an OCR component along with the visual encoder to extract potential text information from given images.

Ethics Statement

We believe this work has a broader impact outside the task and datasets in the discussion. The studied textual bias problem in our motivating analysis and the potential of training a multimodal model with weakly-supervised labels from text-established models are not restricted to a specific task. Also, it becomes common in the NLP domain that some tasks first established based on pure text input are expected to further include the consideration multimodal input. The discussion in this work can be generalized to a lot of other application scenarios. The proposed solutions for multimodal integration and modality bias mitigation are independent of model architecture, which we expect can be applied to other downstream tasks or inspire designs with similar needs.

Regarding the human annotation involved in this work, we create three benchmark datasets that are manually labeled by human laborers to facilitate the source-aware evaluation. The annotation includes both gold attribute value as well as label sources, i.e., image or text. The profiles and images are all collected based on the publicly accessible Amazon shopping website. We depend on internal quality-assured annotators with balanced demographic and geographic characteristics, who consent and are paid adequately based in the US. The data collection protocol is approved by the ethics review board. We attach detailed human annotation instructions and usage explanations provided to the annotators in Appendix F for reference.

Acknowledgements

We would like to thank Binxuan Huang and Yan Liang for their insightful advice and thank anonymous reviewers for their feedback. This work was partially supported by Amazon.com Services LLC, internal funds by the Computer Science Department of Emory University, and the University Research Committee of Emory University.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning.

Hila Chefer, Idan Schwartz, and Lior Wolf. 2022. Optimizing relevance maps of vision transformers improves robustness. In *NeurIPS*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Aloïs De la Comble, Anuvabh Dutt, Pablo Montalvo, and Aghiles Salah. 2022. Multi-modal attribute extraction for e-commerce. *arXiv preprint arXiv:2203.03441*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Yifan Ding, Yan Liang, Nasser Zalmout, Xian Li, Christan Grant, and Tim Weninger. 2022. Ask-and-verify: Span candidate generation and verification for attribute value extraction. In *EMNLP*.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. 2015. Unsupervised visual representation learning by context prediction. In *ICCV*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

Xuan Kan, Hejie Cui, and Carl Yang. 2021. Zero-shot scene graph relation prediction through commonsense knowledge integration. In *ECML PKDD*.

Giannis Karamanolakis, Jun Ma, and Xin Luna Dong. 2020. Txtract: Taxonomy-aware knowledge extraction for thousands of product categories. In *ACL*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Hunter Lang, Aravindan Vijayaraghavan, and David Sontag. 2022. Training subset selection for weak supervision. In *NeurIPS*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022a. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022b. Grounded language-image pre-training. In *CVPR*.

Rongmei Lin, Xiang He, Jie Feng, Nasser Zalmout, Yan Liang, Li Xiong, and Xin Luna Dong. 2021. Pam: understanding product images in cross product category attribute extraction. In *SIGKDD*.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *CVPR*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Qifan Wang, Li Yang, Bhargav Kanagal, Sumit Sanghai, D Sivakumar, Bin Shu, Zac Yu, and Jon Elsas. 2020. Learning to extract attribute value from product via question answering: A multi-task approach. In *KDD*.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022a. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022b. Simvlm: Simple visual language model pretraining with weak supervision.

Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. 2021. Noise or signal: The role of image backgrounds in object recognition. In *ICLR*.

Huimin Xu, Wenting Wang, Xinnian Mao, Xinyu Jiang, and Man Lan. 2019. Scaling up open tagging from tens to thousands: Comprehension empowered attribute value extraction from product title. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5214–5223.

Ran Xu, Yue Yu, Hejie Cui, Xuan Kan, Yanqiao Zhu, Joyce Ho, Chao Zhang, and Carl Yang. 2023. Neighborhood-regularized self-training for learning with few labels. *AAAI*.

Jun Yan, Nasser Zalmout, Yan Liang, Christan Grant, Xiang Ren, and Xin Luna Dong. 2021. Adatag: Multi-attribute value extraction from product profiles with adaptive decoding. In *ACL*.

Yue Yu, Chenyan Xiong, Si Sun, Chao Zhang, and Arnold Overwijk. 2022. Coco-dr: Combating distribution shifts in zero-shot dense retrieval with contrastive and distributionally robust learning. *arXiv preprint arXiv:2210.15212*.

Yue Yu, Simiao Zuo, Haoming Jiang, Wendi Ren, Tuo Zhao, and Chao Zhang. 2021. Fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach. In *NAACL*.

Nasser Zalmout and Xian Li. 2022. Prototype-representations for training data filtering in weakly-supervised information extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. 2020. Putting visual object recognition in context. In *CVPR*.

Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. 2018. Opentag: Open attribute value extraction from product profiles. In *KDD*.

Tiangang Zhu, Yue Wang, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. Multimodal joint attribute prediction and value extraction for e-commerce product. In *EMNLP*.

A Implementation Details

Our models are implemented with PyTorch (Paszke et al., 2019) and Huggingface Transformer library and trained on an 8 Tesla V100 GPU node. The model is trained for 10 epochs, where the Item Form dataset takes around 12 hours, the Color dataset takes about 32 hours, and the Pattern dataset needs around 35 hours to run on a single GPU. The overall architecture of PV2TEA consists of 361M trainable parameters, where a ViT_{base} (Dosovitskiy et al., 2021) is used as the image encoder and initialized with the pre-trained model on ImageNet of 85M parameters, and the text encoder is initialized from $\text{BERT}_{\text{base}}$ (Devlin et al., 2019) of 123M parameters. We use AdamW (Loshchilov and Hutter, 2019) as the optimizer with a weight decay of 0.05. The learning rate of each parameter group is set using a cosine annealing schedule (Loshchilov and Hutter, 2016) with the initial value of $1e$-5. The model is trained for 10 epochs, with both training and testing batch sizes of 8. The memory queue size M is set as 57600 and the temperature τ of in Equation 4 is set as 0.07. We performed a grid search for the softness α from [0, 0.2, 0.4, 0.6, 0.8] and used the best-performed 0.4 for reporting the final results. The K for two-level neighborhood regularization is set at 10. The input textual description is cropped to a maximum of 100 words. The input image is divided into 30 by 30 patches. The hidden dimension of both the visual and textual encoders is set to 768 to produce the representations of patches, tokens, or the whole image/sequence. The epoch E for adding the second-level prediction neighbor regularization to reliability score $s\left(\mathcal{X}_n\right)$ is set as 2.

B More Source-Aware Evaluation

Method	Gold Value Source	$\mathcal{D}$: Color			$\mathcal{D}$: Pattern		
		P	R	F_1	P	R	F_1
$\text{OpenTag}_{\text{cls}}$	Text ✓	85.06	43.28	57.37	85.00	42.96	57.07
	Text ✗ Image ✓	66.28	10.24	17.74	66.23	12.02	20.35
	GAP ↓	18.78	33.04	39.63	18.77	30.94	36.72
PAM	Text ✓	73.20	71.88	72.53	75.00	57.04	64.80
	Text ✗ Image ✓	50.30	45.45	47.75	51.82	36.23	42.64
	GAP ↓	22.90	26.43	24.78	23.18	20.81	22.16
PV2TEA	Text ✓	81.74	74.25	77.82	71.19	61.25	65.85
	Text ✗ Image ✓	71.89	47.19	56.98	54.48	37.26	44.25
	GAP ↓	9.85	27.06	**20.84**	16.71	23.99	**21.59**

Table 8: Fine-grained source-aware evaluation for the Color and Pattern datasets.

The source-aware evaluation of the Color and Pattern datasets is shown in Table 8. We can observe that similarly to the discussions in Section 5.1, compared with the baselines, the proposed PV2TEA effectively mitigates the performance gap of F_1 when the gold value is not contained in the text. More specifically, we observed that compared with the unimodal method, PV2TEA mainly reduces the recall performance gap across modalities, while compared with the multimodal method, the reduction happens mainly in precision, which all corresponds to the weaker metrics for each type of method. This indicates the stronger generalizability and more balanced learning ability of PV2TEA.

C Ablation Studies on Pattern Dataset

We further include the ablation results on the single-value type dataset Pattern for each proposed mechanism in Table 9, Table 10, and Table 11, respectively. The observations are mostly consistent with the discussion in section 5.2, where all three proposed mechanisms support improvements in the overall performance of F_1. It is noted that the recall performance with attention-pruning drops a bit compared with that without. This may indicate potential information losses on the challenging dataset such as Pattern with only the selected foreground. We discuss this potential risk in detail in the Limitation section.

Method	**Single Value Dataset: Pattern**		
	Precision	Recall	F_1
PV2TEA w/o L_{sc}	60.03	45.59	51.82
PV2TEA w/o smooth	61.87	45.72	52.58
PV2TEA	62.10	46.84	53.40

Table 9: Ablations on the augmented label-smoothed contrast for cross-modality alignment (%).

Method	**Single Value Dataset: Pattern**		
	Precision	Recall	F_1
PV2TEA w/o L_{ct} & Attn Prun	59.01	46.74	52.16
PV2TEA w/o Attn Prun	60.14	46.98	52.75
PV2TEA	62.10	46.84	53.40

Table 10: Ablation study on the category supervised visual attention pruning (%).

Method	**Single Value Dataset: Pattern**		
	Precision	Recall	F_1
PV2TEA w/o NR	59.92	44.92	51.35
PV2TEA w/o Vis-NR	61.59	46.24	52.82
PV2TEA w/o Pred-NR	60.77	45.11	51.78
PV2TEA	62.10	46.84	53.40

Table 11: Ablations on the two-level neighborhood-regularized sample weight adjustment (%).

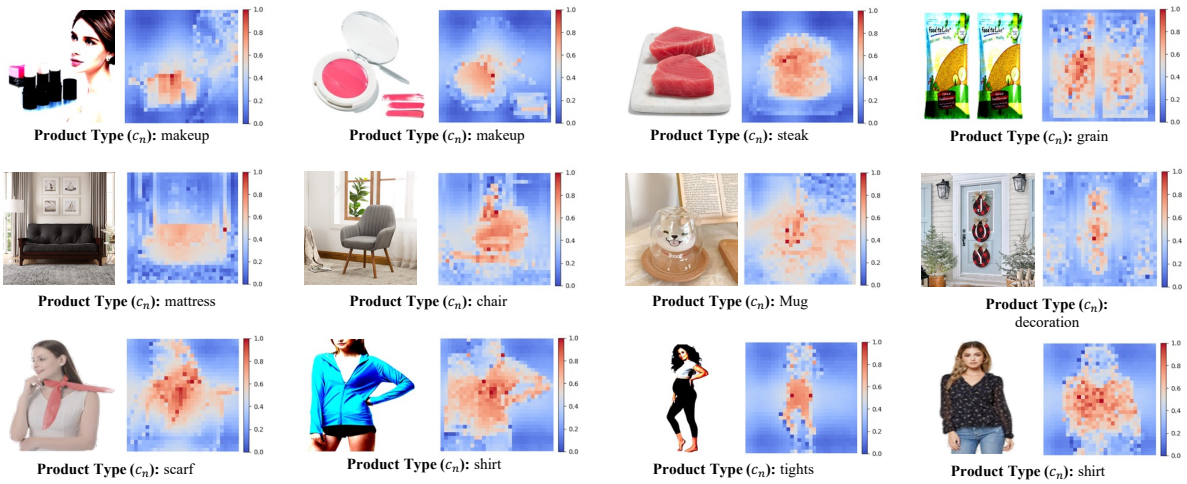

Figure 7: Visualization examples of the learned category aware attention pruning mask.

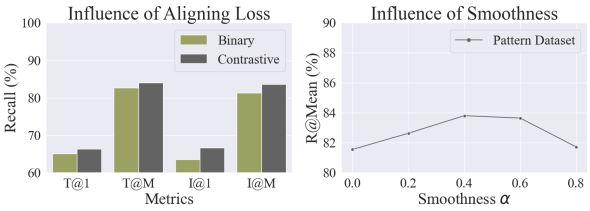

Figure 8: The influence study of alignment objectives, i.e., binary matching v.s. contrastive, and softness α study via cross-modality retrieval on the Pattern dataset.

D Retrieval Ablation on Pattern Dataset

Similar to Figure 4, we also demonstrate the cross-modality retrieval results on the pattern dataset in Figure 8. The conclusion is consistent with our observations mentioned in Section 5.2, where the contrastive objective demonstrates advantages in cross-modal alignment and fusion, and the best smoothness choice peaks at 0.4.

E Visualizations of Attention Pruning

Examples of visualization on the learned attention mask are demonstrated in Figure 7. It is observed that the visual foreground is highlighted under the supervision of category classification, which potentially encourages a higher prediction precision for fine-grained tasks like attribute extraction, as proved by the experimental results.

F Human Annotation Instruction

We create source-aware fine-grained datasets with internal human annotators. Below are the instruction texts provided to annotators:

The annotated attribute values are used for research model development of multimodal attribute information extraction and fine-grained error analysis. The datasets are named source-aware multimodal attribute extraction evaluation benchmarks and will be released to facilitate public testing and future studies in bias-reduced multimodal attribute value extraction model designs. All the given sample profiles (title, bullets, and descriptions) and images are collected from the public `amazon.com` web pages, so there is no potential legal or ethical risk for annotators. Specifically, the annotation requirements compose two tasks in order: (1) Firstly, for each given sample_id in the given ASINs set, first determine the category of the sample by referring to ID2Category.csv mapping file, then label the gold value for the queried attribute by selecting from the candidates given the category. The annotation answer candidates for the Item Form dataset can be referred to in Table 12. Note that this gold value annotation step requires reference to both sample textual titles, descriptions, and images; (2) For each annotated ASIN, mark down which modality implies the gold value with an additional source label, with different meanings as below:

- **0**: *the gold attribute value can be found in text.*
- **1**: *the gold attribute value cannot be inferred from the text but can be found in the image.*

The annotated attribute values and source labels are assembled in fine-grained source-aware evaluation.

G Neighborhood Regularization Demos

We provide two more demo examples for illustrating the two-level neighborhood-regularized sample weight adjustment in Figure 9. The example on the left demonstrates a higher consistency between the green arrows (which point to samples with the same training label as y_n) and red arrows (which point

Category (Product Type)	Candidate Attribute Values Given the Category
cereal	grain, flake, seed, liquid, powder, ground
dishwasher detergent	gel, capsule, pac, liquid, tablet, pod, powder
face shaping makeup	powder, pencil, cream, liquid, stick, oil, spray, gel, cushion, blush, drop, balm, gloss
fish	fillet, chunk, steak, solid, stick, whole, slice, ground
herb	powder, root, leaf, thread, flake, seed, tea bag, stick, oil, slice, pod, ground, bean, paste
honey	jelly, capsule, lozenge, candy, cream, powder, granule, flake, liquid, stick, oil, crystal, butter, drop, syrup, comb
insect repellent	wipe, spray, band, granular, liquid, stick, candle, coil, oil, lotion, gel, capsule, tablet, powder, balm, patch, roll on
jerky	strip, slab, shredded, bite, bar, slice, stick, ground
sauce	puree, jelly, paste, seed, liquid, gravy, ground, oil, powder, cream
skin cleaning agent	powder, capsule, toothpaste, wipe, cream, spray, mousse, bar, flake, liquid, lotion, gel, serum, mask, ground, balm, paste, foam
skin foundation concealer	powder, pencil, cream, mousse, liquid, stick, oil, lotion, spray, cushion, gel, drop, serum, balm, airbrush
sugar	granule, crystal, pearl, liquid, powder, cube, ground
sunscreen	wipe, cream, spray, mousse, liquid, ointment, stick, fluid, oil, lotion, milk, compact, gel, drop, serum, powder, balm, foam, mist
tea	leaf, powder, granule, tea bag, liquid, pod, ground, brick

Table 12: The annotation candidates provided to annotators given each sample type on the Item Form dataset.

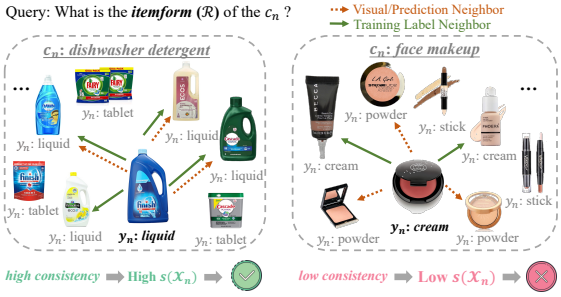

Figure 9: Demo examples for illustrating S3: two-level neighborhood-regularized sample weight adjustment.

to k-nearest neighbor samples in visual feature and previous prediction space), indicating a higher reliability of y_n. Thus the sample weight of $\mathcal{X}_n$ will be increased in the next training epoch. In contrast, the training label neighbors and visual/prediction neighbors of the right example show a large inconsistency, which implies a relatively lower reliability of y_n. Therefore, the sample weight $s\left(\mathcal{X}_n\right)$ of the right $\mathcal{X}_n$ will be degraded in the next epoch. This regularization process adjusts the sample weights of all the training samples in each epoch.