Supplementary Materials for Enhancing Vision-Language Pre-training with Rich Supervisions

1. Quantitative Results on Joint Training

We provide some additional discussions for the impact of joint training language and location tasks. Directly combining $S4_{NL}$ objectives and $S4_{Loc}$ objectives, which we denote as $S4_{Joint}$, harms the performance on most downstream tasks. For $S4_{Joint}$, we use the following weighted loss as it gives the best average performance across all tasks:

```
loss_weights = {
    'screen2html': 1.0,
    'attribute_prediction': 0.5,
    'title_generation': 0.5,
    'node_relation_prediction':0.1,
    'table_parsing':0.1,
    'ocr':0.1,
    'table_detection':0.1,
    'layout_analysis':0.1,
    'image_grounding':0.1,
    'element_grounding':0.1
}
```

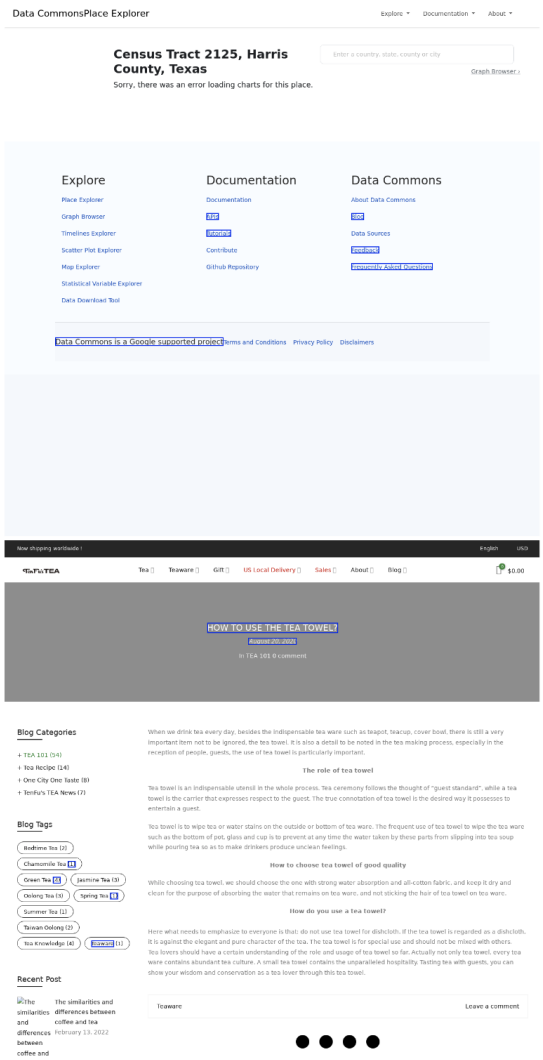
Methods	Pre-training Dataset	Pre-training Objectives	Finetune Batchsize	ChartQA ↑	$\text{RefExp}_{\text{cls}}$↑	Widget ↑ Cap.	Screen ↑ Sum.	WebSRC ↑
$S4^*$ **(Ours)**	S4 Data - 15M	$S4_{NL}$	8	**55.0**	**94.9**	**130.6**	**105.7**	**61.1**
$S4^*$ **(Ours)**	S4 Data - 15M	$S4_{Joint}$	8	52.1	**94.9**	128.4	101.5	60.3

Methods	Pre-training Dataset	Pre-training Objectives	RefExp ↑ $_{\text{cand_free}}$ 30k samples	PublayNet ↑ 1M samples	PubTables1M ↑ (Table Det.) 400k samples	ICDAR 2019 ↑ (Modern Subset) 600 samples
$S4^*$ **(Ours)**	S4 Data - 15M	$S4_{Loc}$	**84.3**	**93.1**	**99.0**	**79.4**
$S4^*$ **(Ours)**	S4 Data - 15M	$S4_{Joint}$	79.2	91.6	97.7	76.7

2. Qualitative Results during Pre-training

We visualize qualitative pre-training results in below figures. For table detection, image grounding, and element grounding, red boxes denotes ground truth boxes and blue boxes denotes predicted boxes. We present four images for each of these three tasks, where the bottom right is a failure case and the others are good cases. For layout analysis, we present a pair of prediction (blue) and ground truth layout (red).

Attribute Prediction



Prediction:
<hyperlink,footer,col md ><hyperlink,footer,col md ><hyperlink,footer,col md ><hyperlink,footer,col md ><hyperlink,footer,col md ><paragraph,footer,col md>

Ground Truth:
<hyperlink,main footer,col col sm col md><hyperlink,main footer,col col sm col md ><hyperlink,main footer,col col sm col md><hyperlink,main footer,col col sm col md ><hyperlink,main footer,col col sm col md><inline grouping,sub footer,col>

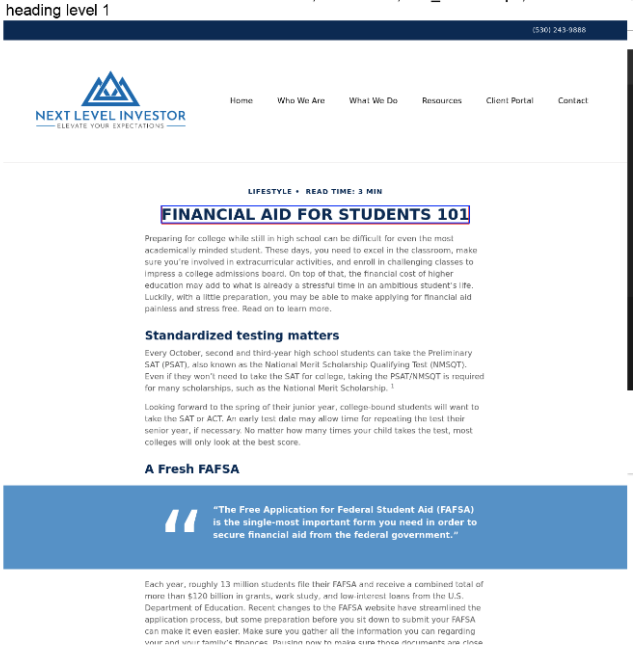
Prediction:
<heading level 1,page,entry title><time or date,page,entry date published><inline grouping,tag cloud,tag link count><inline grouping,tag cloud,tag link count><inline grouping,tag cloud,tag link count><hyperlink,tag cloud,tag cloud link tag link tag link position >

Ground Truth:
<heading level 1,shopify section article heading,tu cw mb ><time or date,shopify section article heading,entry date published updated f libre><inline grouping,shopify section article sidebar,blg count><inline grouping,shopify section article sidebar,blg count><inline grouping,shopify section article sidebar,blg count><hyperlink,shopify section article sidebar,nt filter block nt filter styletag blg count true>

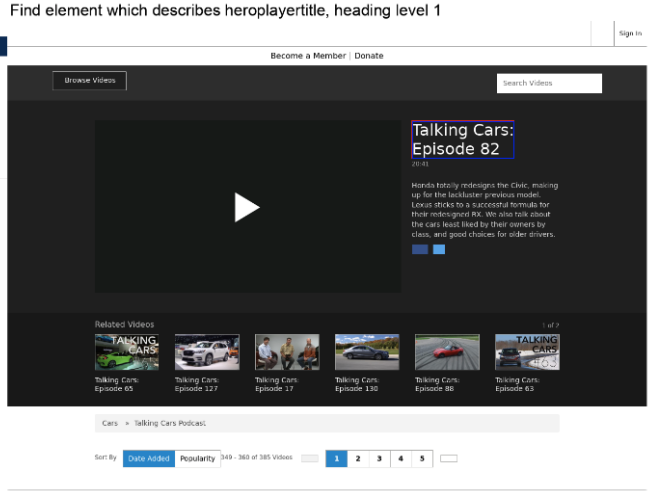
Figure 1. Attribute Prediction

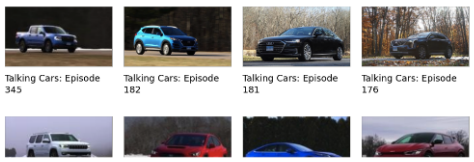
Element Grounding

Find element which describes articleContent, cmattertitle, font_size is 32px, heading level 1

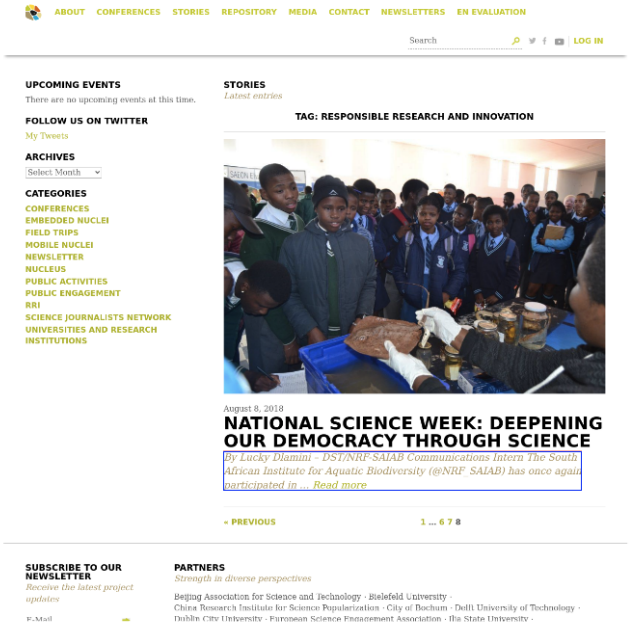


Find element which describes heroplayertitle, heading level 1





Find element which describes entrysummary, font_size is 20px, paragraph



Find element which describes guide header updated, slgguidelabel, inline grouping

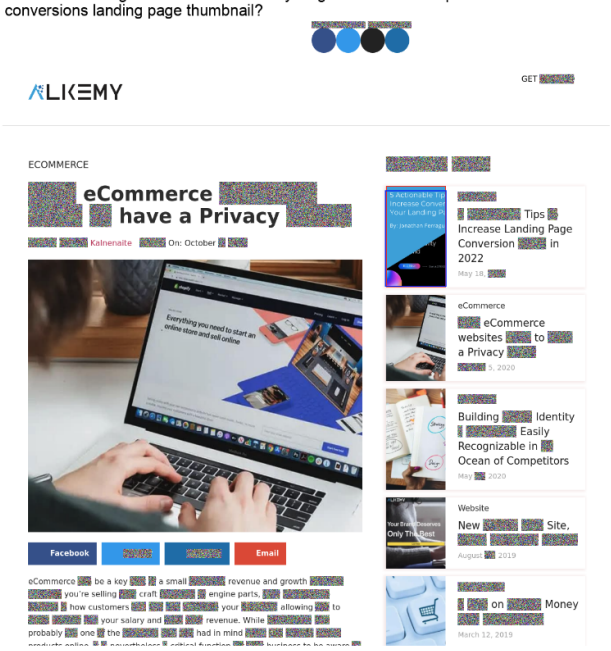
Figure 2. Element Grounding

Image Grounding

where is the image that describes ARC Logo?

where is the image that describes ask a librarian, contact us; phone, email, text?

where is the image that describes Alkemy blog - 5 actionablen tips increase conversions landing page thumbnail?



where is the image that describes (Beautiful Dark Beasts, #1) Published by: City Owl Press Publication date: May 5th 2020 Genres: Adult, Paranormal, Romance?

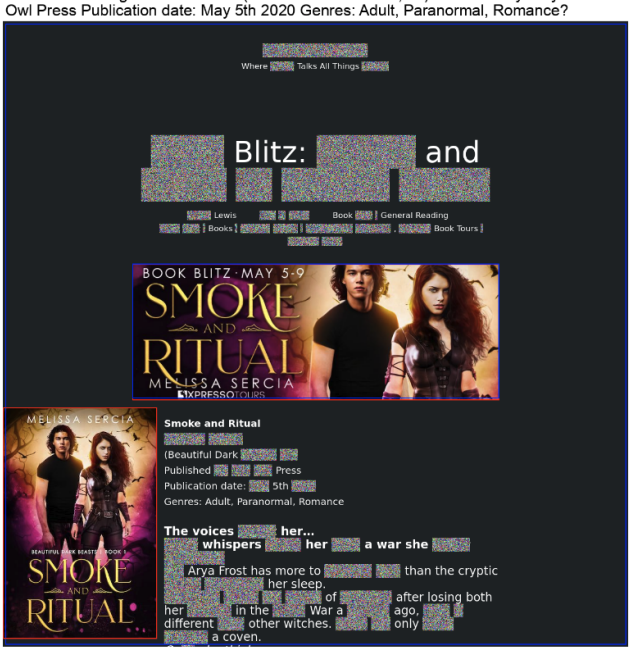


Figure 3. Image Grounding

Layout Analysis

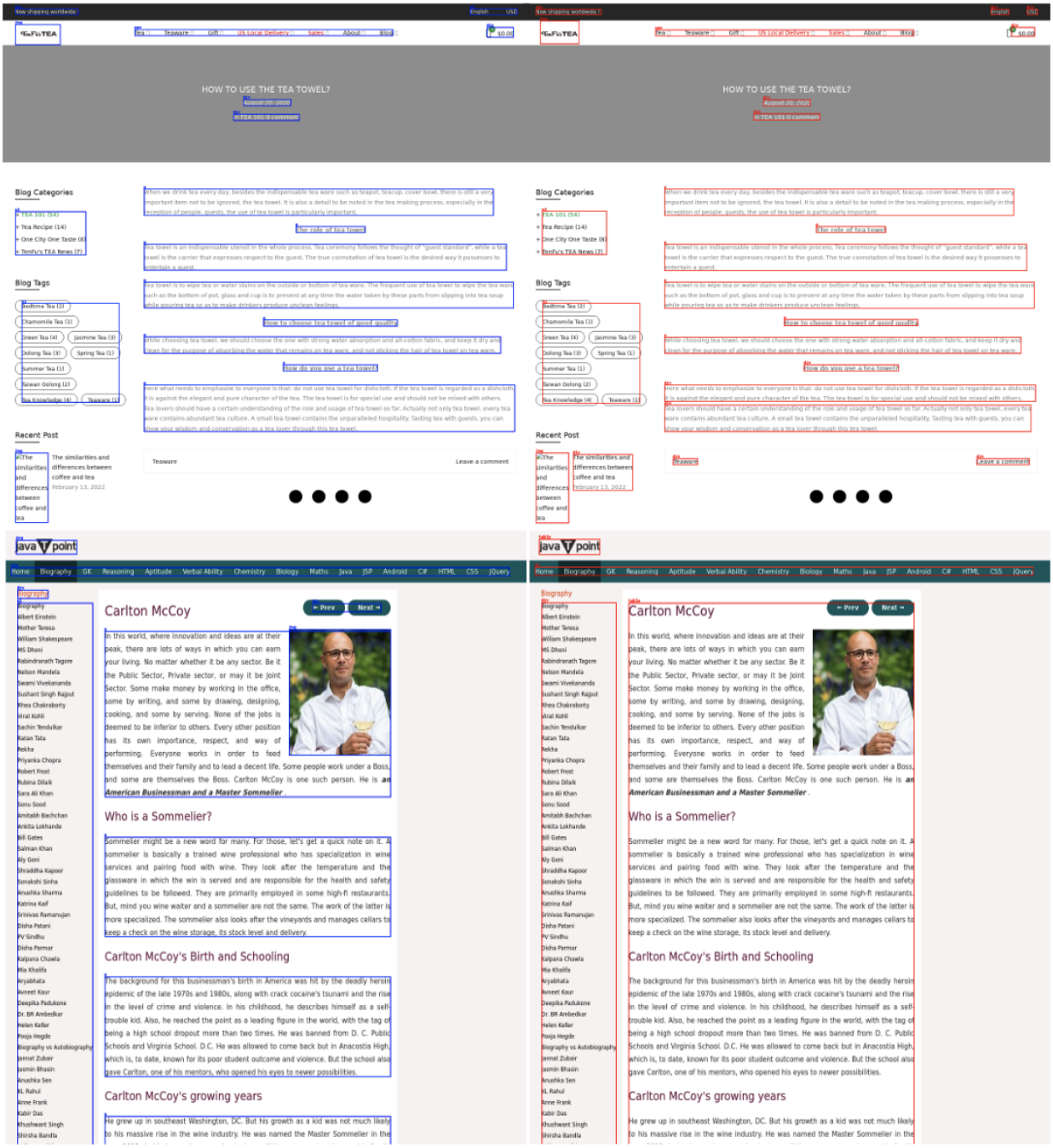
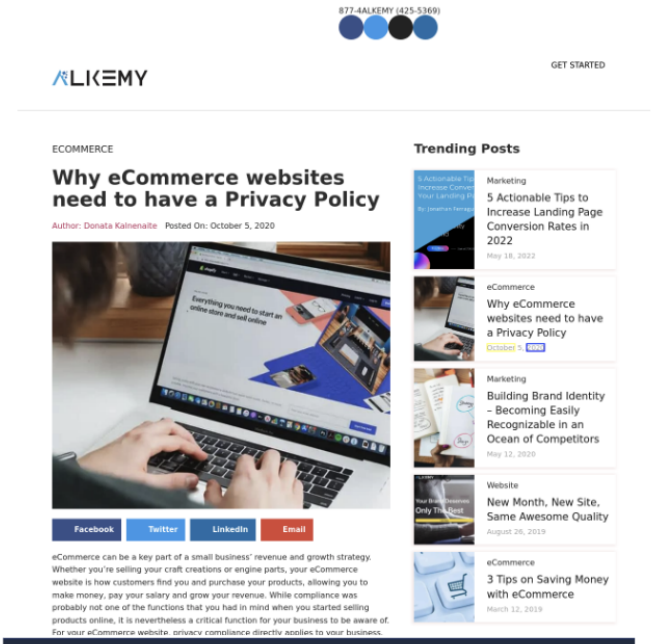


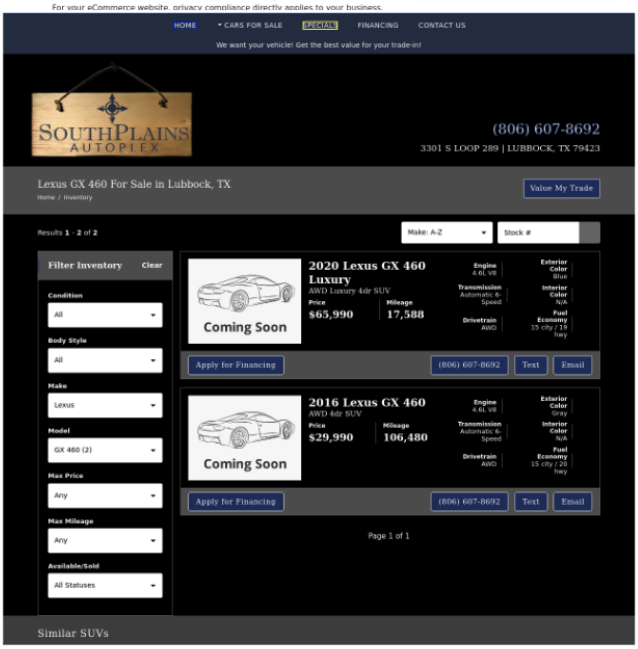
Figure 4. Layout Analysis

Node Relation Prediction



Prediction: Same

Ground Truth: Same

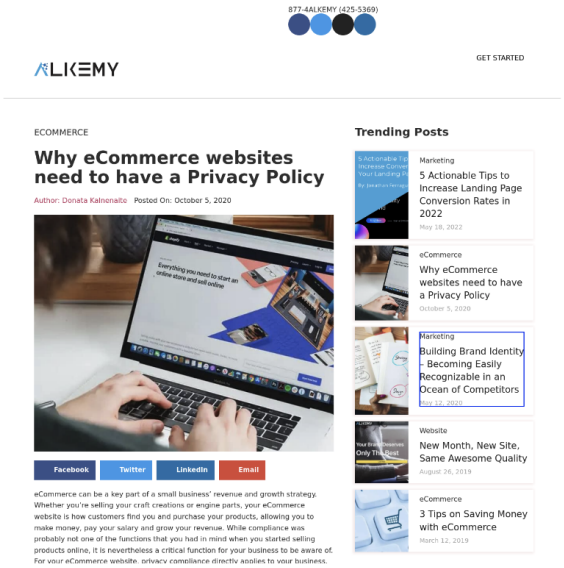


Prediction: Sibling

Ground Truth: Sibling

Figure 5. Node Relation

OCR

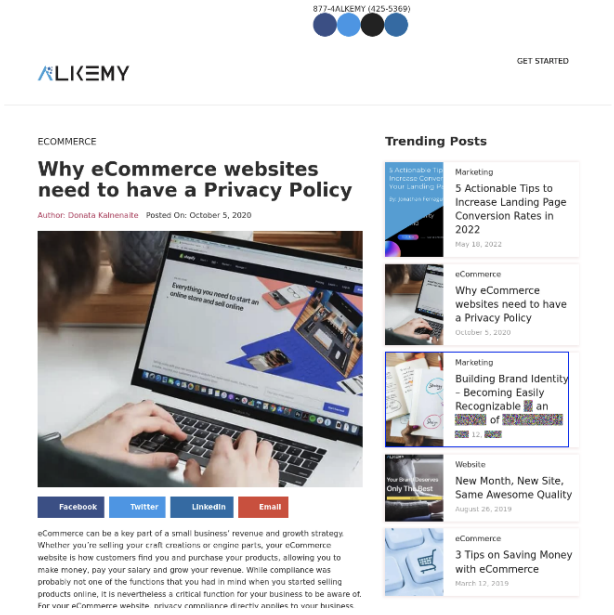


Prediction:
Marketing<949><752><1029><771> Building<949><784><1033><808>
Brand<1040><784><1102><808> Identity<1108><784><1188><808>
–<949><813><959><837> Becoming<966><813><1070><837>
Easily<1076><813><1137><837> Recognizable<949><842><1087><866>
in<1094><842><1113><866> an<1119><842><1145><866>
Ocean<949><871><1015><895> of<1022><871><1042><895>
Competitors<1049><871><1176><895> May<949><906><979><922>
12,<983><906><1006><922> 2020<1011><906><1047><922>

Ground Truth:
Marketing<949><752><1029><771> Building<949><784><1033><808>
Brand<1040><784><1102><808> Identity<1108><784><1188><808>
–<949><813><959><837> Becoming<966><813><1070><837>
Easily<1076><813><1137><837> Recognizable<949><842><1087><866>
in<1094><842><1113><866> an<1119><842><1145><866>
Ocean<949><871><1015><895> of<1022><871><1042><895>
Competitors<1049><871><1176><895> May<949><906><979><922>
12,<983><906><1006><922> 2020<1011><906><1047><922>

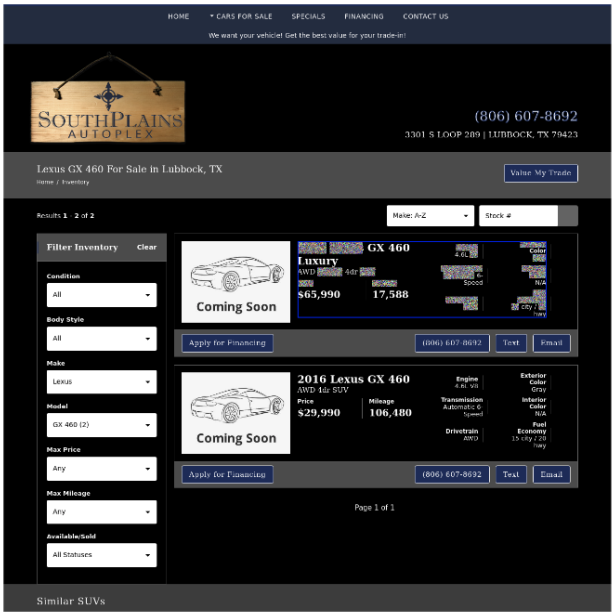
Figure 6. OCR

Screen2html



Prediction:
<<<Marketing><Building Brand Identity – Becoming Easily Recognizable in an Image of eCommerce><May 12, 2020>>>

Ground Truth:
<<<Marketing><Building Brand Identity – Becoming Easily Recognizable in an Ocean of Competitors><May 12, 2020>>>

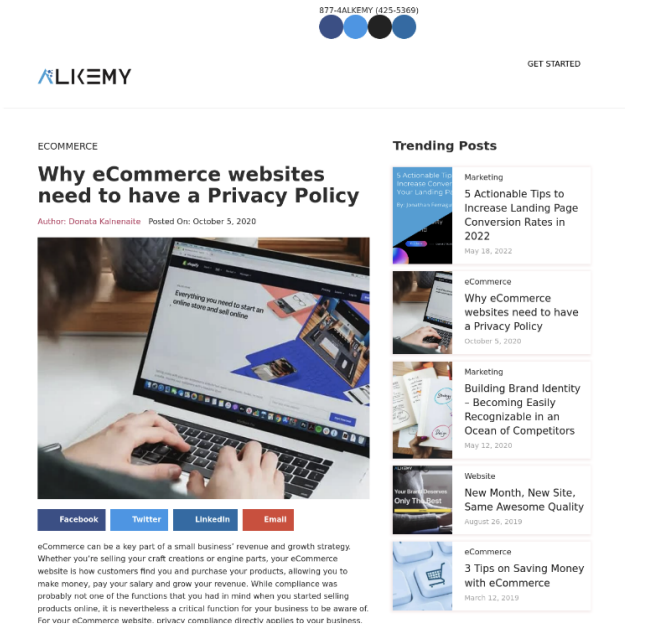


Prediction:
<<<2016 Lexus GX 460 Luxury><AWD 4dr SUV><<<Price><$65,990>><<Mileage><17,588>>>><<<Engine><4.6L V8>><<Color><Gray>><<Transmission><Automatic 4 Speed>><<Eco-Fuel><N/A>><<Eco-Fuel><15 cyl / 20 hwy>>>>

Ground Truth:
<<<2020 Lexus GX 460 Luxury><AWD Luxury 4dr SUV><<<Price><$65,990>><<Mileage><17,588>>>><<<<Engine><4.6L V8>><<Exterior Color><Blue>>><<<Transmission><Automatic 6-Speed>><<Interior Color><N/A>>><<<Drivetrain><AWD>><<Fuel Economy><15 city / 19 hwy>>>>>

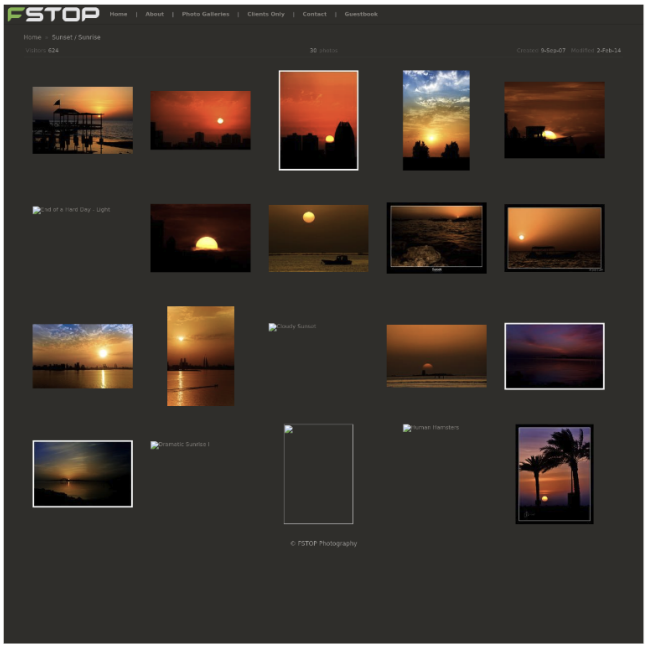
Figure 7. Screen2html

Screen Titling



Prediction:
Why eCommerce websites need to have a Privacy Policy - ALKEMY

Ground Truth:
Why eCommerce websites need to have a Privacy Policy | Alkemy



Prediction:
F5TOP Photography | Sunset / Sunrise | Photo 30

Ground Truth:
Ali Zainal Photography | Sunset / Sunrise

Figure 8. Screen Tiltling

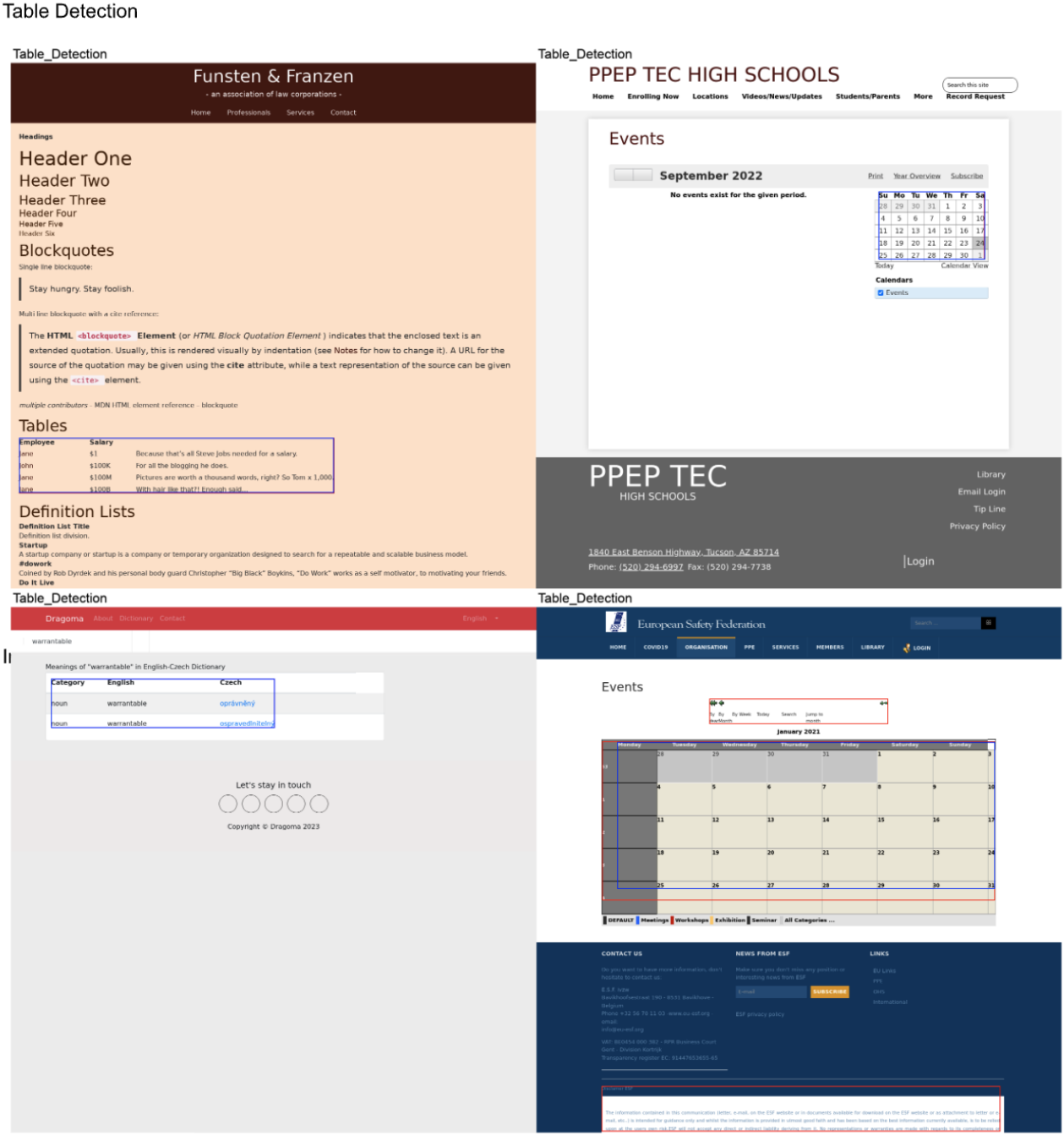


Figure 9. Table Detection

Table Parsing

Prediction:
<table><tbody><tr><td>[MMW Live Shows by Date] </td></tr><tr><td>1990s </td><td>2000s </td><td>2001s </td></tr><tr><td>TSP 1993 </td><td>TSP 2002 </td><td>TSP 2003 </td></tr><tr><td>TSP 1994 </td><td>TSP 2005 </td><td>TSP 2006 </td></tr><tr

Ground Truth:
<table><tbody><tr><td>- MMW Live Shows by Date - </td></tr><tr><td>1990s </td><td>2000s </td><td>2010s </td></tr><tr><td>TSP 1993 </td><td>TSP 2000 </td><td>TSP 2010 </td></tr><tr><td>TSP 1994 </td><td>TSP 2001 </td><td>TSP 2011 </td></tr><tr><td>TSP 1995 </td><td>TSP 2002 </td><td>TSP 2012 </td></tr><tr><td>TSP 1996 </td><td>TSP 2003 </td><td>TSP 2013 </td></tr><tr><td>TSP 1997 </td><td>TSP 2004 </td><td>TSP 2014 </td></tr><tr><td>TSP 1998 </td><td>TSP 2005 </td><td>TSP 2016 </td></tr><tr><td>TSP 1999 </td

Prediction:
<table><thead><tr><th>Category </th><th>License </th><th>Czech </th></tr></thead><tbody><tr><td>2022 </td><td>2022-01-26 </td><td>Unlicensed </td></tr><tr><td>2021 </td><td>2021-01-26 </td><td>Unlicensed-2021 </td></tr></tbody></table>

Ground Truth:
<table><thead><tr><th>Category </th><th>English </th><th>Czech </th></tr></thead><tbody><tr><td>noun </td><td>warrantable </td><td>oprávněný </td></tr><tr><td>noun </td><td>warrantable </td><td>ospravedlnitelný </td></tr></tbody></table>

Figure 10. Table Parsing