養天地正氣 法古今完人

How to Understand "Support"? An Implicit-enhanced Causal Inference Approach for Weakly-supervised Phrase Grounding

Jiamin Luo, Jianing Zhao, Jingjing Wang, and Guodong Zhou

School of Computer Science and Technology, Soochow University, Suzhou, China

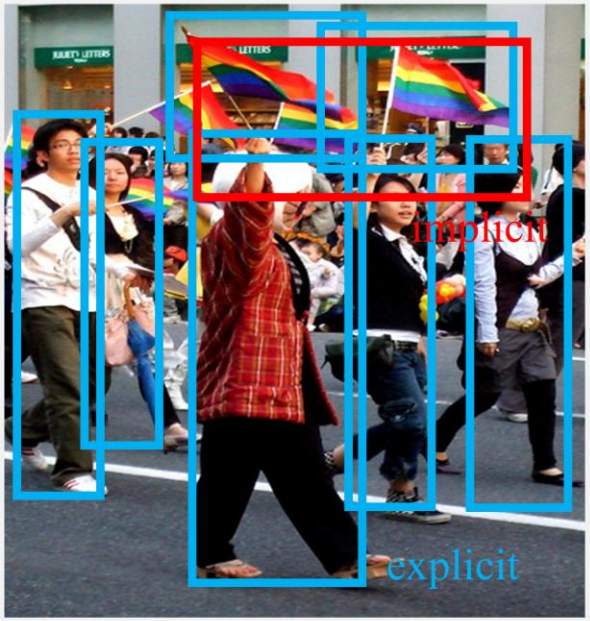
Motivations

- For information, existing WPG studies ignore the implicit phrase-region matching relations, which are crucial for evaluating the capability of models in understanding the deep multimodal semantics.
- For approach, we consider introducing both the intervention and counterfactual techniques to model the implicit relations and highlight them beyond explicit.
- For dataset, we annotate a high-quality implicit-enhanced dataset to evaluate the effectiveness of the proposed IECI approach.

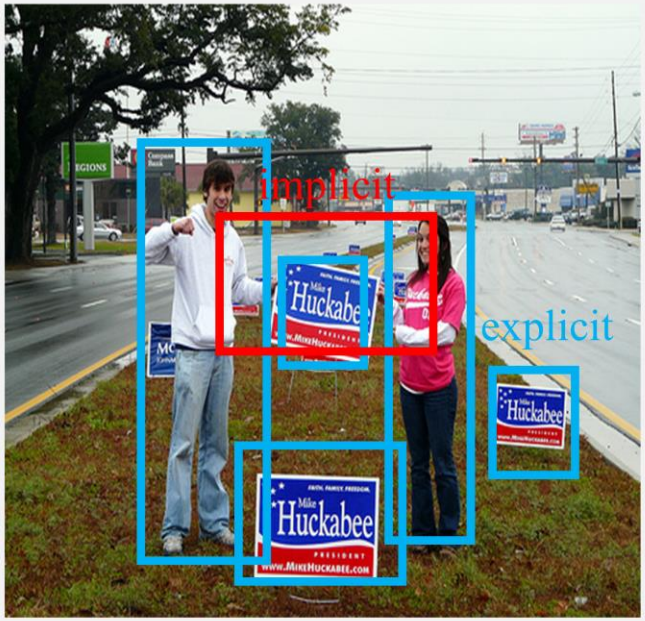
WPG Task

Examples of WPG Task:

During a gay pride parade in an Asian city , some people hold up rainbow flags to show support

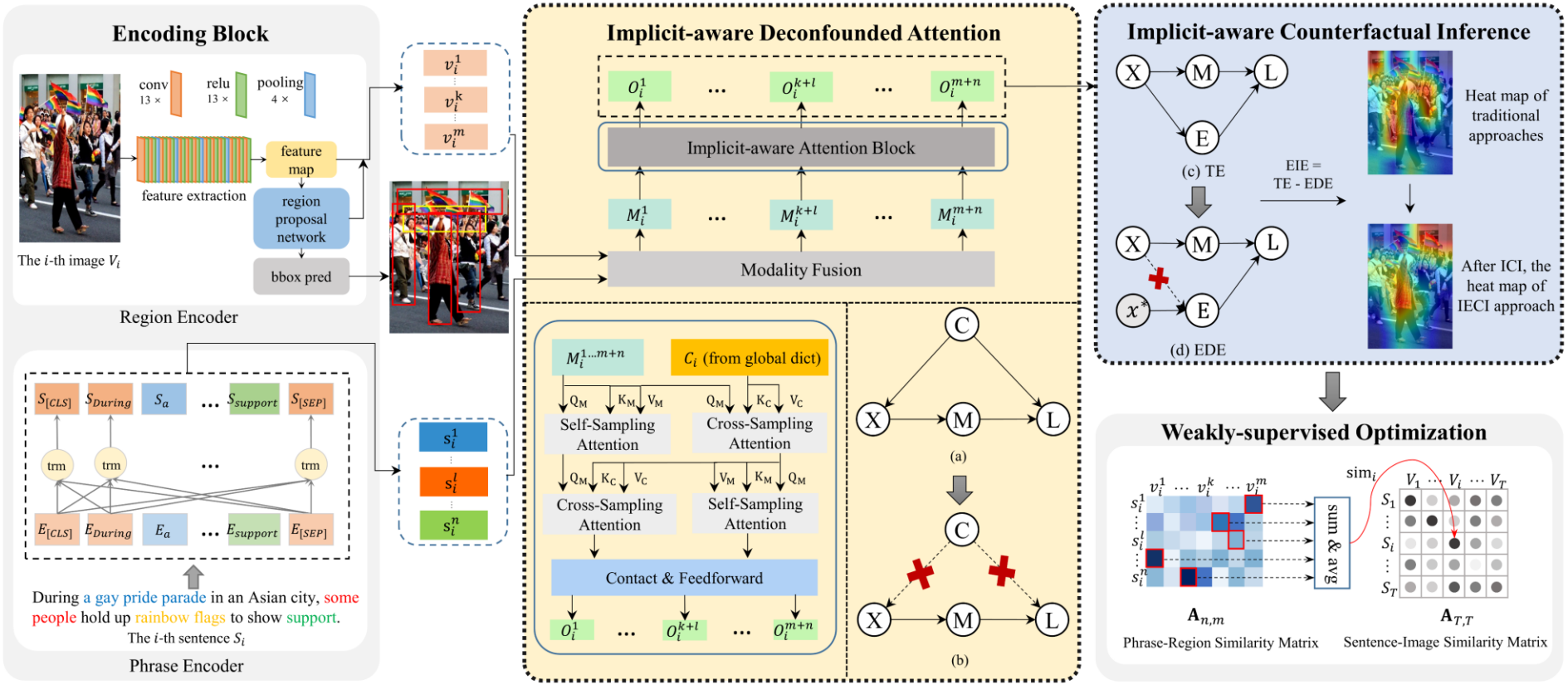


A man and a woman show support for the campaign of Mike Huckabee as they hold up a sign

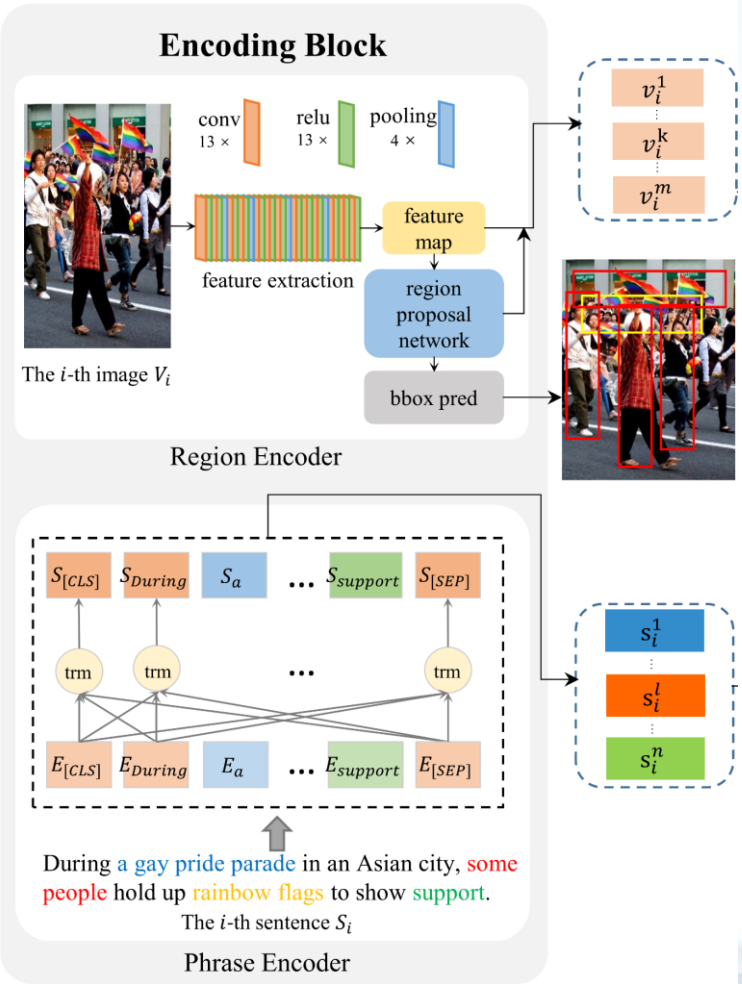


IECI Approach

The overall framework of our Implicit-Enhanced Causal Inference (IECI) approach to WPG :



IECI Approach

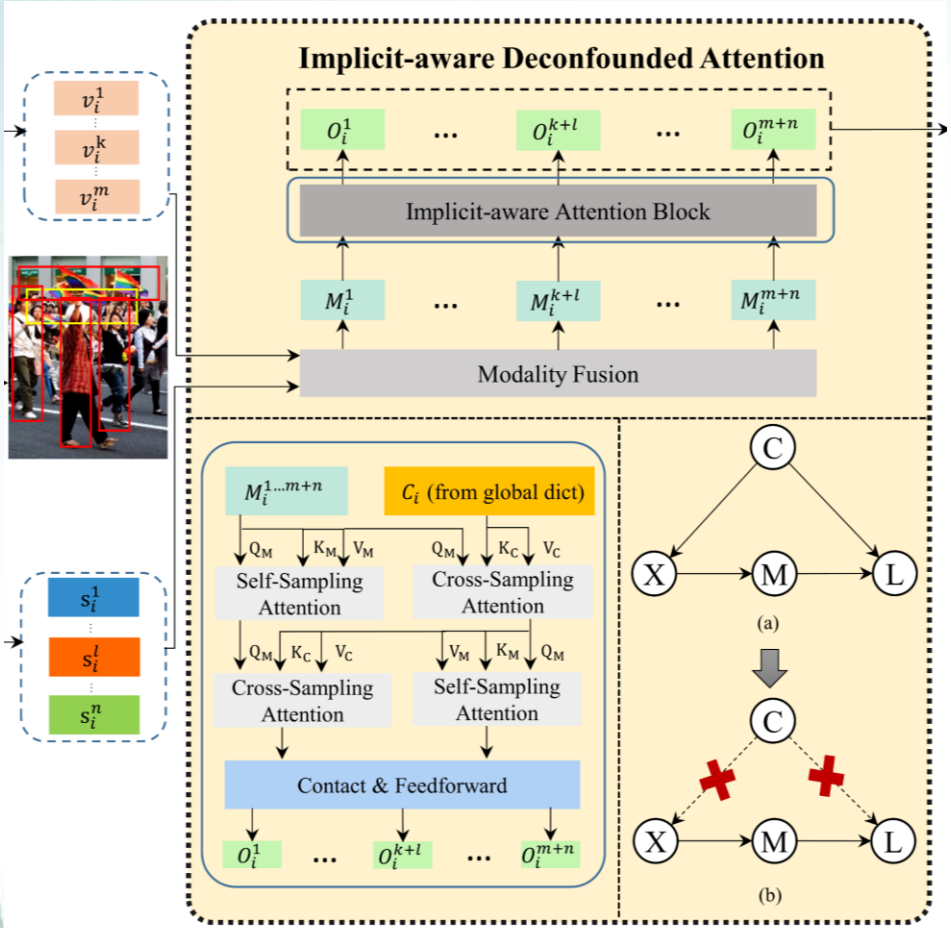


Encoding Block

1 Phrase Encoder: BERT-based model is adopted as the phrase encoder, which is a light-weighting language encoding model.

2 Region Encoder: Faster R-CNN is adopted as the region encoder, which generates the encoding of the region along with the corresponding bounding box.

IECI Approach

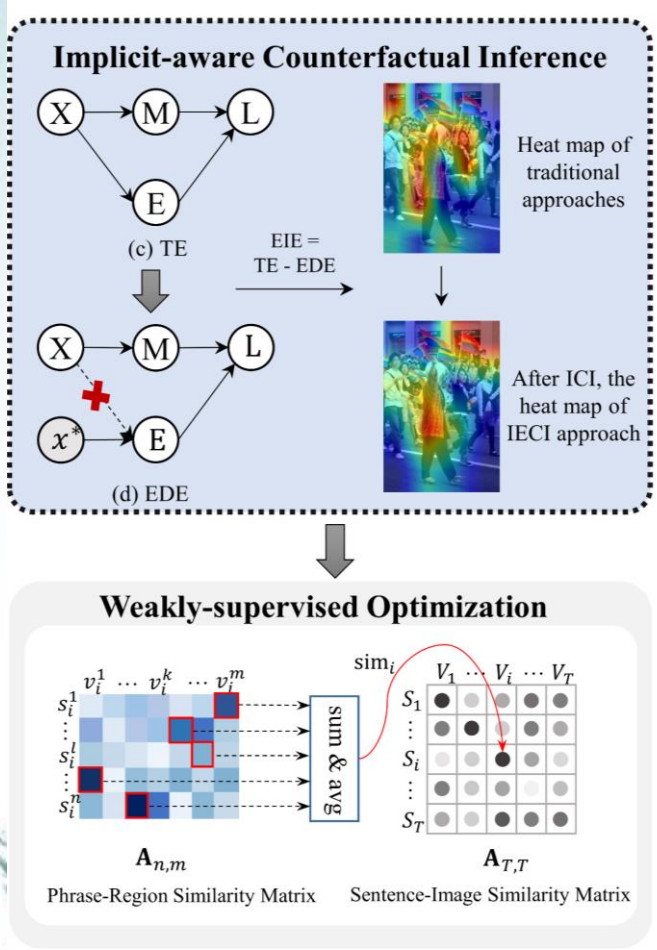


Implicit-aware Deconfounded Attention

1 Deconfounded Causal Graph: we construct a causal graph (a) to mitigate the confounding bias through the front-door adjustment strategy (b).

2 Implicit-aware Attention: we implement the front-door adjustment strategy through the utilization of attention mechanisms.

IECI Approach



Implicit-aware Counterfactual Inference

1 Counterfacual Causal Graph: we construct a causal graph (c) to analyze the direct effect of the explicit relations (d).

2 Implicit-aware Inference: we reduce the direct effect of explicit relations to improve the alignment of implicit phrase region pairs.

Weakly-supervised Optimization

We convert the phrase-region similarity matrix to sentence-image similarity matrix for weakly-supervised optimization.

Experiments

Annotation for our implicit-enhanced dataset:

1 We summarize four main types of implicit relations through preliminary annotation.

2 Assign two annotators to tag each phrase-region pair.

3 We annotate 2K sentence-image samples and obtain1.4K implicit phrase-region pairs

Implicit Relations	Ratios (%)	Sentence-Image Examples with Implicit Relations
Commonsense Understanding (CU)	34.5	During a gay pride parade in an Asian city , some people hold up rainbow flags to show support. 
Complex Context Understanding (CCU)	26.9	A man aboard a red train helps a small child onto it while another person tries to get on. 
Spatial Relations Understanding (SRU)	23.5	A man is playing a guitar next to another man who is sitting behind a green cart wearing a mask. 
Numerical Understanding (NU)	15.1	Six ladies at the dining table and three of them are knitting. 

Experiments

Experimental results as follows:

Approach	Flickr30K						COCO					
	Implicit		Explicit		Full		Implicit		Explicit		Full	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
KAC-Net	31.99	51.14	49.11	63.50	38.71$^{\sharp}$	-	39.51	56.85	54.81	73.72	45.88	66.27
ARN	33.81	49.87	46.13	64.50	34.87	50.42	40.35	57.68	57.07	74.98	41.93	58.27
KPRN	30.95	44.50	47.31	62.93	33.41	47.33	35.31	55.22	55.77	76.51	38.30	57.00
InfoGround	44.72	74.79	55.07	80.87	47.88†	76.63†	45.66	74.08	61.17	84.67	51.67†	77.69†
ALBEF	56.40	78.27	69.37	85.03	57.64	77.56	52.00	76.23	66.19	84.06	54.22	76.34
CL&KD	50.33	73.75	62.50	82.00	53.10$^{\natural}$	-	50.37	75.02	64.42	83.78	51.36	74.98
ReIR	57.98	**78.72**	69.65	85.33	59.27§	-	54.01	77.60	66.77	84.45	55.26	76.72
BLIP	20.31	41.51	26.62	62.62	23.30	57.07	26.99	63.19	34.14	70.02	31.15	67.13
IECI	**61.32**	78.36	**72.37**	**86.27**	**62.29**	**79.28**	**56.32**	**78.01**	**68.62**	**85.25**	**56.92**	**78.31**
w/o IDA	57.72	77.16	71.10	84.96	59.07	77.87	53.96	76.88	67.19	84.13	54.92	77.25
w/o ICI	58.09	77.79	71.87	85.33	59.50	78.51	54.44	77.12	67.61	84.71	55.05	77.17
w/o Both	56.05	76.48	69.32	83.89	57.87	77.78	52.20	76.07	66.05	83.57	54.17	76.54

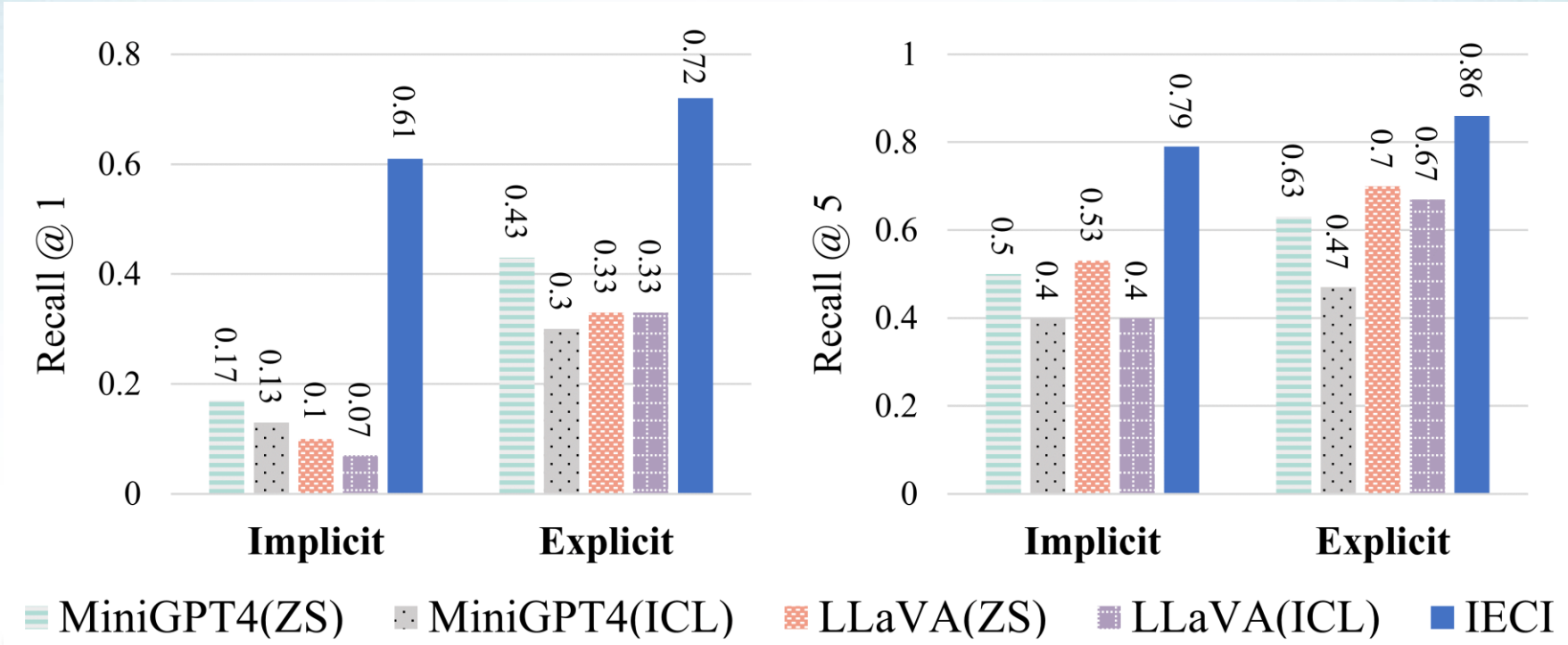
Experiments

Ablation study for the contributions of causal inference in our IECI approach:

Approach	Flickr30K						COCO					
	Implicit		Explicit		Full		Implicit		Explicit		Full	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
KAC-Net	31.99	51.14	49.11	63.50	38.71$^{\sharp}$	-	39.51	56.85	54.81	73.72	45.88	66.27
ARN	33.81	49.87	46.13	64.50	34.87	50.42	40.35	57.68	57.07	74.98	41.93	58.27
KPRN	30.95	44.50	47.31	62.93	33.41	47.33	35.31	55.22	55.77	76.51	38.30	57.00
InfoGround	44.72	74.79	55.07	80.87	47.88†	76.63†	45.66	74.08	61.17	84.67	51.67†	77.69†
ALBEF	56.40	78.27	69.37	85.03	57.64	77.56	52.00	76.23	66.19	84.06	54.22	76.34
CL&KD	50.33	73.75	62.50	82.00	53.10$^{\natural}$	-	50.37	75.02	64.42	83.78	51.36	74.98
ReIR	57.98	**78.72**	69.65	85.33	59.27§	-	54.01	77.60	66.77	84.45	55.26	76.72
BLIP	20.31	41.51	26.62	62.62	23.30	57.07	26.99	63.19	34.14	70.02	31.15	67.13
IECI	**61.32**	78.36	**72.37**	**86.27**	**62.29**	**79.28**	**56.32**	**78.01**	**68.62**	**85.25**	**56.92**	**78.31**
w/o IDA	57.72	77.16	71.10	84.96	59.07	77.87	53.96	76.88	67.19	84.13	54.92	77.25
w/o ICI	58.09	77.79	71.87	85.33	59.50	78.51	54.44	77.12	67.61	84.71	55.05	77.17
w/o Both	56.05	76.48	69.32	83.89	57.87	77.78	52.20	76.07	66.05	83.57	54.17	76.54

Experiments

Comparison with our IECI approach with Multimodal LLMs:



Further Study

- For **multimodal representation**, we would like to incorporate the multimodal LLMs, such as MiniGPT4 and LLaVA, to enhance the multimodal representation abilities of our approach to the WPG task.
- For **knowledge injection**, we would like to consider integrating the external knowledge like multimodal knowledge graph to help capture implicit relations.
- For **evaluation of each implicit relation**, we would like to leverage the multimodal LLMs to automatically annotate different types of implicit relations.

Contributions

- We first address the implicit relations problem in the WPG task.
- We propose a new implicit-enhanced causal inference (IECI) approach, which integrates both the intervention and counterfactual techniques for modeling the implicit relations and highlighting the implicit beyond the explicit.
- We meticulously annotate a high-quality implicit-enhanced dataset to evaluate the ability of models in understanding deep multimodal semantics.
- We compare the results of our IECI approach with the advanced multimodal LLMs on the annotated implicit-enhanced dataset, which may further facilitate the evaluation of multimodal LLMs in this direction.

Thank you.

Q&A