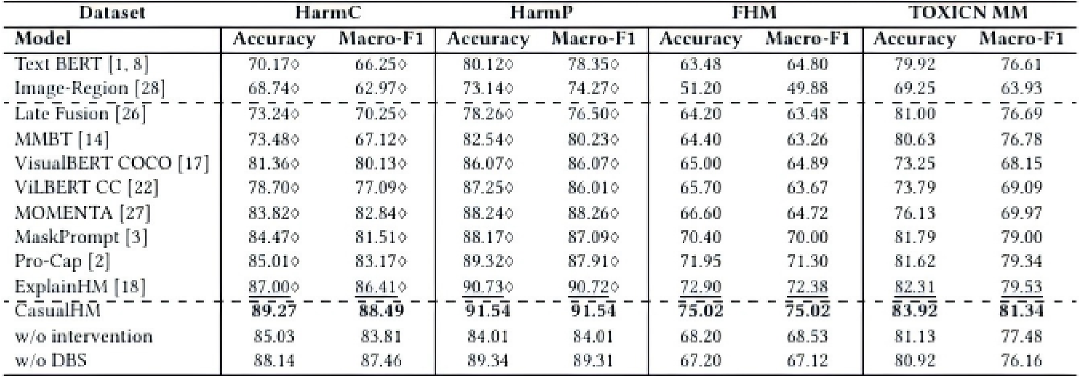
Table 1: Harmful meme detection results on four datasets. The numbers marked with ◇ are taken from [18]. We mark the best and second-best results in bold and underlined, respectively.



Dataset	HarmC		HarmP		FHM		TOXICN MM	
Model	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
Text BERT [1, 8]	70.17◇	66.25◇	80.12◇	78.35◇	63.48	64.80	79.92	76.61
Image-Region [28]	68.74◇	62.97◇	73.14◇	74.27◇	51.20	49.88	69.25	63.93
Late Fusion [26]	73.24◇	70.25◇	78.26◇	76.50◇	64.20	63.48	81.00	76.69
MMBT [14]	73.48◇	67.12◇	82.54◇	80.23◇	64.40	63.26	80.63	76.78
VisualBERT COCO [17]	81.36◇	80.13◇	86.07◇	86.07◇	65.00	64.89	73.25	68.15
ViLBERT CC [22]	78.70◇	77.09◇	87.25◇	86.01◇	65.70	63.67	73.79	69.09
MOMENTA [27]	83.82◇	82.84◇	88.24◇	88.26◇	66.60	64.72	76.13	69.97
MaskPrompt [3]	84.47◇	81.51◇	88.17◇	87.09◇	70.40	70.00	81.79	79.00
Pro-Cap [2]	85.01◇	83.17◇	89.32◇	87.91◇	71.95	71.30	81.62	79.34
ExplainHM [18]	87.00◇	86.41◇	90.73◇	90.72◇	72.90	72.38	82.31	79.53
CasualHM	**89.27**	**88.49**	**91.54**	**91.54**	**75.02**	**75.02**	**83.92**	**81.34**
w/o intervention	85.03	83.81	84.01	84.01	68.20	68.53	81.13	77.48
w/o DBS	88.14	87.46	89.34	89.31	67.20	67.12	80.92	76.16