ASAP: Advancing Semantic Alignment Promotes Multi-Modal Manipulation Detecting and Grounding

Supplementary Material

6. DGM^4 Loss

Given an image-text pair (I, T), we define four sub-task losses following HAMMER as follows:

6.1. Manipulated Image Bounding Box Grounding

For the manipulated image grounding task, we input the multimodal feature i^v_{pat} into a BBox Detector D_v and calculate the Image Manipulation Grounding Loss as:

$$\mathcal{L}_{\text{IMG}} = \mathbb{E}_{(I,T)} \left[\|\text{Sigmoid}(D_v(i^v_{pat})) - y_{box}\| + \mathcal{L}_{IOU}(\text{Sigmoid}(D_v(i^v_{pat})), y_{box}) \right]$$

6.2. Binary Classification

For the binary classification task, we input multimodal feature M_{it} into Binary Classifier C_b and calculate Binary Classifier Loss as follows:

$$\begin{cases} \mathcal{L}_{\text{IMG}} = \mathbb{E}_{(I,T)} \left[\boldsymbol{H}(C_b(M_{it}), y_{bin}) \right] \\ M_{it} = \delta \mathcal{E}^v_{mm}(\mathcal{E}_v(I), \mathcal{E}_t(T)) + \mathcal{E}^t_{mm}(\mathcal{E}_t(T), \mathcal{E}_v(I)) \end{cases}$$

where $\boldsymbol{H}(\cdot)$ is the cross-entropy function.

6.3. Manipulation Type Detection

For the binary classification task, we input the multimodal feature M_{it} into the Binary Classifier C_b and compute the Binary Classifier Loss as:

$$\mathcal{L}_{\text{MLC}} = \mathbb{E}_{(I,T)} \left[\boldsymbol{H}(C_m(M_{it}), y_{mul}) \right]$$

6.4. Manipulated Text Token Grounding

For the manipulated text token grounding task, we use a Token Detector D_t to predict the label of each token in t^t_{tok} and calculate the cross-entropy loss as follows:

$$\begin{cases} \mathcal{L}_{\text{TMG}} = (1-\alpha)\mathcal{L}_{\text{tok}} + \alpha \mathcal{L}^{mom}_{tok} \\ \mathcal{L}_{\text{tok}} = \mathbb{E}_{(I,T)} \left[\boldsymbol{H}(D_t(t^t_{tok}), y_{tok}) \right] \\ \mathcal{L}^{mom}_{tok} = \mathbb{E}_{(I,T)} KL \left[D_t(t^t_{tok}) || \hat{D}_t(\hat{t}^t_{tok}) \right] \\ \{t^t_{cls}, t^t_{tok}\} = \mathcal{E}^t_{mm}(\mathcal{E}_t(T), \mathcal{E}_v(I)) \end{cases}$$

where $\hat{D}_t(\hat{t}^t_{tok})$ represents the pseudo-labels generated by the momentum Token Detector, used to modulate the original token predictions, and KL denotes the Kullback-Leibler divergence between the original token predictions and the momentum-based pseudo-labels.

Tasks	Binary Cls			Image Grounding		
Methods	AUC↑	EER↓	ACC↑	mAP↑	CF1↑	OF1↑
Text	92.89	13.26	86.43	78.90	72.98	75.01
MultiModal	**94.58**	**12.79**	**87.64**	**79.79**	**73.40**	**76.46**

Table 5. Ablation study of text modality.

Tasks	Binary Cls			Image Grounding		
Methods	AUC↑	EER↓	ACC↑	mAP↑	CF1↑	OF1↑
Image	93.13	13.48	86.55	76.16	83.46	75.13
MultiModal	**94.50**	**12.62**	**87.27**	**77.30**	**84.22**	**77.61**

Table 6. Ablation study of image modality.

7. Discussion

7.1. Effectiveness of Cross-modality learning

We evaluated the multimodal fusion mechanism by comparing single-modal and multimodal learning. Table 5 and Table 6 show that "Text" and "Image" represent single-modal learning, while "MultiModal" indicates multimodal learning. The results confirm that our ASAP model improves detection and grounding through multimodal fusion. The difference between the two "MultiModal" results stems from the use of different loss functions.

7.2. Discussion of different Large Models

To assess the effectiveness of our approach and the validity of large model selection, we employed Qwen and LLaMA 2b in the LMA mechanism, conducting ablation studies against our ASAP method. Table 7 demonstrate that incorporating preliminary texts significantly improves performance across all tasks except image grounding, confirming our approach as the optimal solution.

7.3. Discussion of each Hyperparameter

We fine-tuned multiple hyperparameters for the ASAP model, selecting final values of $\delta = 0.5$, $\alpha = 0.1$, and $\lambda = 0.01$ based on model performance. As shown in Tables 8,9, and10, these values provided a balanced optimization of various performance metrics, enabling ASAP to achieve peak results while preserving efficient detection capabilities.

Tasks	Binary Cls			Multi-Label			Image Grounding			Text Grounding		
Different models	AUC↑	EER↓	ACC↑	mAP↑	CF1↑	OF1↑	IoUmean↑	IoU50↑	IoU75↑	Precision↑	Recall↑	F1↑
Baseline	93.16	14.13	86.23	86.23	79.59	80.54	**76.49**	**83.82**	**75.97**	75.25	68.21	71.83
Qwen-VL & LLaMA	94.23	**12.83**	87.49	86.90	80.15	82.01	75.78	83.21	74.86	78.44	73.60	75.04
VisCPM & Mistral	**94.28**	12.86	**87.53**	**88.10**	**81.71**	**82.61**	75.90	83.27	74.98	**78.59**	**74.10**	**76.28**

Table 7. Performance Comparison Across Different Large Models of **LMA**. This table compares two approaches of large model assistance, where Mistral is used as the LLM and Viscpm as the MLLM, and LLaMA is used as the LLM and Qwen VL as the MLLM, to generate auxiliary labels in the LMA module.

Tasks	Binary Cls			Multi-Label			Image Grounding			Text Grounding		
Different δ	AUC↑	EER↓	ACC↑	mAP↑	CF1↑	OF1↑	IoUmean↑	IoU50↑	IoU75↑	Precision↑	Recall↑	F1↑
$\delta = 0.1$	94.33	12.89	87.75	88.32	81.57	82.66	77.01	84.23	76.19	79.02	73.72	**76.66**
$\delta = 0.5$	**94.38**	**12.73**	87.71	**88.53**	**81.72**	**82.89**	**77.35**	**84.75**	**76.54**	**79.38**	**73.86**	76.52
$\delta = 1.0$	94.32	12.79	**87.77**	88.35	81.71	82.88	77.03	84.45	76.10	78.66	73.71	76.11

Table 8. Performance Comparison Across Different Initial Values of the **Hyperparameter** δ in equation 4.

Tasks	Binary Cls			Multi-Label			Image Grounding			Text Grounding		
Different α	AUC↑	EER↓	ACC↑	mAP↑	CF1↑	OF1↑	IoUmean↑	IoU50↑	IoU75↑	Precision↑	Recall↑	F1↑
$\alpha = 0.1$	**94.38**	**12.73**	**87.71**	**88.53**	**81.72**	**82.89**	77.35	84.75	76.54	**79.38**	**73.86**	**76.52**
$\alpha = 0.5$	94.30	12.60	87.65	88.05	81.71	82.81	**77.51**	**84.83**	**77.12**	79.31	72.79	75.92
$\alpha = 1.0$	94.26	12.99	87.32	87.98	81.55	82.34	77.23	84.48	76.57	78.90	72.45	75.68

Table 9. Performance Comparison Across Different Values of the **Hyperparameter** α in equation 13.

Tasks	Binary Cls			Multi-Label			Image Grounding			Text Grounding		
Different λ	AUC↑	EER↓	ACC↑	mAP↑	CF1↑	OF1↑	IoUmean↑	IoU50↑	IoU75↑	Precision↑	Recall↑	F1↑
$\lambda = 0.01$	94.38	**12.73**	**87.71**	88.53	**81.72**	**82.89**	**77.35**	**84.75**	**76.54**	79.38	**73.86**	**76.52**
$\lambda = 0.05$	**94.42**	12.80	87.63	**88.59**	81.68	82.79	77.19	84.49	76.61	**79.44**	73.85	76.49
$\lambda = 0.10$	94.35	12.86	87.55	88.55	81.71	82.80	77.10	84.38	76.52	79.21	73.84	75.98

Table 10. Performance Comparison Across Different Values of the **Hyperparameter** λ in equation 13.