Table R1: Comparison with MaskDINO on the COCO validation set when the backbone is Swin-L. The experiments are conducted on 4 A6000 GPUs with the batch size of 4.

Methods	Epochs	AP^{box}	AP^{box}_S	AP^{box}_M	AP^{box}_L	AP^{mask}	AP^{mask}_S	AP^{mask}_M	AP^{mask}_L	FPS
MaskDINO	50	56.8	40.2	60.2	72.3	51.0	31.3	54.1	71.2	3.4
DI-MaskDINO (Ours)	50	**57.8 (+1.0)**	41.5	61.2	73.9	**51.8 (+0.8)**	31.8	55.1	72.2	3.0

Table R2: Comparison with MaskDINO on the BDD100K validation set with Swin-L backbone.

Methods	Epochs	AP^{box}	AP^{box}_S	AP^{box}_M	AP^{box}_L	AP^{mask}	AP^{mask}_S	AP^{mask}_M	AP^{mask}_L	FPS
MaskDINO	68	30.2	19.0	37.5	48.6	27.0	15.4	32.6	50.5	3.2
DI-MaskDINO (Ours)	68	**31.4 (+1.2)**	19.4	40.4	48.7	**27.9 (+0.9)**	16.6	34.1	51.2	2.8

Table R3: Comparison between DI-MaskDINO configured with different numbers of decoder layers and MaskDINO with 9 decoder layers.

Methods	AP^{box}	AP^{mask}	Params
MaskDINO	45.7	42.4	52.1
DI-MaskDINO (3 decoder layers)	45.8	41.3	47.6
DI-MaskDINO (6 decoder layers)	46.9	42.3	52.3
DI-MaskDINO (9 decoder layers)	46.9	42.5	56.9

Table R4: The results of diagnostic experiments on Token Interaction.

Configurations	BDD100K		COCO	
	AP^{box}	AP^{mask}	AP^{box}	AP^{mask}
w/o Token Interaction	28.3	25.4	46.1	41.7
w/ Token Interaction	**29.5**	**25.7**	**46.9**	**42.3**

Table R5: Comparison with MaskDINO on the COCO test-dev with Swin-L backbone.

Methods	Epochs	AP^{box}	AP^{box}_{75}	AP^{box}_{50}	AP^{box}_S	AP^{box}_M	AP^{box}_L
MaskDINO	50	57.2	77.2	62.7	37.9	59.9	72.4
DI-MaskDINO (Ours)	50	**57.9 (+0.7)**	77.6	63.5	38.8	60.7	73.0