

## A APPENDIX

### A.1 Additional Dataset

Table 1: Statistics of the SIMMC dataset.

Dataset	SIMMC		
Dataset Stats	Train	Valid	Test
dialogues	7307	1687	1687
Proportion	68%	16%	16%
Avg Rec Turns	4	4	4
Avg Pos Imgs	2	2	2
Avg Neg Imgs	22	22	22

In this paper, we also conduct experiments on SIMMC Dataset. For further insights, detailed statistics are provided in Table 1. Here, “Avg Rec Turns” indicates the average number of recommendations per dialogue; and “Avg Pos Imgs” denotes the number of correct recommendations per turn whereas “Avg Neg Imgs” is the number of distractors for evaluation.

### A.2 Additional Experimental Results

Table 2: The overall results of SeMANTIC and baselines on SIMMC, in which the average and standard deviations of different runs are reported.

SIMMC									
Methods	Precision@5	Recall@5	NDCG@5	Precision@10	Recall@10	NDCG@10	Precision@20	Recall@20	NDCG@20
MHRED	22.93±0.51	67.20±1.41	51.16±1.30	14.46±0.22	85.83±1.12	57.14±1.18	8.27±0.04	94.57±0.45	60.24±1.01
MAGIC	26.95±0.38	78.16±0.98	63.52±1.00	15.62±0.36	90.86±1.08	68.32±1.18	8.56±0.03	97.69±0.32	70.10±0.84
CLIP	29.71±0.49	80.74±1.16	70.46±1.21	17.06±0.15	91.18±0.28	74.33±0.91	9.22±0.07	97.41±0.11	76.18±0.89
LARCH	23.31±0.93	71.15±1.71	57.83±1.84	14.48±0.31	86.85±1.72	63.80±1.48	8.15±0.08	96.10±0.89	66.69±1.23
TREASURE	27.50±0.47	79.43±1.00	64.99±1.31	16.00±0.18	91.66±0.57	69.89±1.24	8.60±0.04	98.10±0.16	71.27±1.07
SeMANTIC	<b>31.99±0.33</b>	<b>87.14±0.71</b>	<b>76.82±0.87</b>	<b>17.85±0.09</b>	<b>95.45±0.41</b>	<b>79.96±0.75</b>	<b>9.35±0.01</b>	<b>98.99±0.14</b>	<b>81.04±0.64</b>

*Additional Main Results on SIMMC.* In Section 4.3, to study the performance of SeMANTIC and other baselines when being trained with small conversational sample sets, we conduct experiments on MMD-v3. Here, we further extend the experiments to SIMMC dataset, and results are provided in Table 2.

*Varying Sizes of Conversational Samples.* In Section 4.4, to study the impacts of sample size, we show the performance of SeMANTIC trained with varying ratio of fully labeled data (with ground-truth dialogue state label) on MMD-v3 in terms of NDCG@5 and Recall@5. Here, we further show the experiments in terms of NDCG@10 and Recall@10, and the results are provided in Figure 1.

*Varying Size of Fully Labeled Data.* In Section 4.4, to study the impacts of sample size, we show the performance of SeMANTIC trained with varying sample sizes on MMD-v2 in terms of NDCG@5 and Recall@5. Here, we further show the experiments in terms of NDCG@10 and Recall@10, and the results are provided in Figure 2.

Furthermore, The results for changing the varying number of samples with dialogue states (ds) on SIMMC dataset are presented in Table 3.

*Ablation Study.* We further extend the ablation study to SIMMC dataset and Table 4 showcases more details of the impact of different loss functions on SeMANTIC.

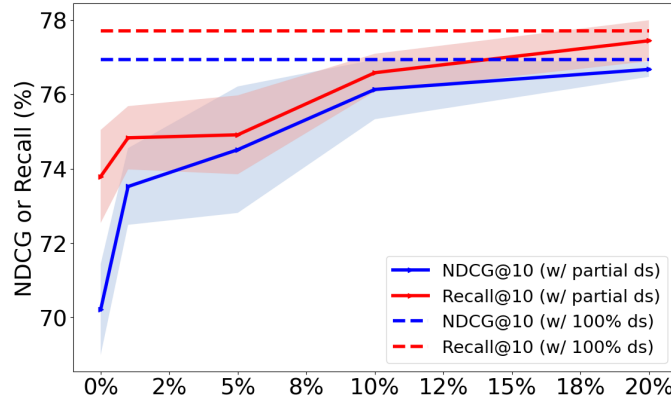


Figure 1: Performance in terms of NDCG@10 and Recall@10 for SeMANTIC trained with varying ratio of fully labeled data on MMD-v3.

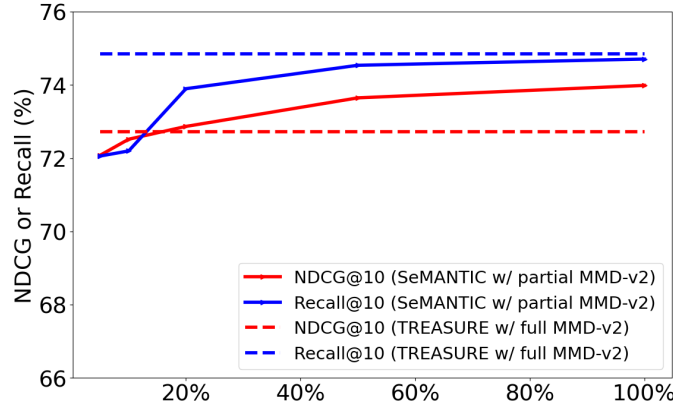


Figure 2: Performance in terms of NDCG@10, Recall@10 of SeMANTIC with different sample sizes on MMD-v2.

Table 3: Performance of SeMANTIC on SIMMC when different size of ground-truth dialogue state labels (labeled ds) is used for training.

	Precision@5	Recall@5	NDCG@5	Precision@10	Recall@10	NDCG@10	Precision@20	Recall@20	NDCG@20
SeMANTIC(0% labeled ds)	59.26±1.14	69.66±1.34	68.46±1.66	31.33±0.52	73.79±1.24	70.21±1.22	16.31±0.27	76.91±1.30	71.30±1.16
SeMANTIC(1% labeled ds)	61.08±0.72	71.87±0.91	72.23±1.06	31.76±0.37	74.83±0.85	73.52±1.03	16.47±0.19	77.69±0.98	74.52±1.04
SeMANTIC(5% labeled ds)	61.47±1.35	72.30±1.49	73.23±1.74	31.95±0.55	74.91±1.06	74.51±1.70	16.45±0.33	77.86±0.97	75.52±1.66
SeMANTIC(10% labeled ds)	62.56±0.56	73.66±0.73	74.89±0.90	32.48±0.19	76.59±0.51	76.13±0.80	16.89±0.07	79.75±0.42	77.20±0.77
SeMANTIC(20% labeled ds)	63.29±0.52	74.67±0.55	75.50±0.20	32.79±0.25	77.44±0.55	76.67±0.19	16.99±0.10	80.30±0.47	77.65±0.16
SeMANTIC(100% labeled ds)	63.80±0.39	75.19±0.54	75.87±0.71	32.96±0.16	77.71±0.53	76.94±0.72	17.06±0.09	80.52±0.47	77.91±0.71

Table 4: Effect of different loss functions on MMD-v3 and SIMMC.

MMD									
Methods	Precision@5	Recall@5	NDCG@5	Precision@10	Recall@10	NDCG@10	Precision@20	Recall@20	NDCG@20
SeMANTIC	63.87±0.39	75.19±0.54	75.87±0.71	32.96±0.16	77.71±0.53	76.94±0.72	17.06±0.09	80.52±0.47	77.91±0.71
w/o co_sim	38.84±1.98	45.02±2.29	43.90±3.51	21.87±0.92	50.84±2.21	46.52±3.21	12.11±0.44	56.47±2.11	48.55±3.04
w/o MSE	59.26±1.14	69.66±1.34	68.46±1.66	31.33±0.52	73.79±1.25	70.21±1.22	16.31±0.27	76.91±1.30	71.30±1.16
w/o JS	63.26±2.09	74.48±2.65	74.85±3.56	32.79±0.85	77.28±2.16	76.05±3.33	16.96±0.37	80.01±1.90	76.99±3.23
SIMMC									
SeMANTIC	31.99±0.33	87.14±0.71	76.82±0.87	17.85±0.09	95.45±0.41	79.96±0.75	9.35±0.01	98.99±0.14	81.04±0.64
w/o co_sim	31.79±0.26	86.31±0.27	75.16±0.13	17.12±0.07	94.64±0.19	78.10±0.18	9.31±0.02	97.28±0.04	80.62±0.41
w/o MSE	31.03±0.19	86.44±0.36	75.23±0.48	17.19±0.02	94.74±0.13	78.00±0.42	9.31±0.01	97.18±0.11	80.73±0.39
w/o JS	31.27±0.37	87.01±0.80	76.74±1.15	17.21±0.10	95.38±0.46	79.34±0.99	9.34±0.01	98.33±0.06	81.09±0.88