A Details on Hyper-Parameter Tuning

Here are the final parameters used for the **BiLSTM** and **HierLSTM** baselines. 32 for batch size, 128 for maximum paragraph length, 300 for word embedding size initialized by GloVe (Pennington et al., 2014) for baseline models, 1 LSTM layer (Hochreiter and Schmidhuber, 1997) with 512 size, clipping by 0.25, 0.2 learning rate and 0.5 decay rate with Adagrad (Duchi et al., 2011) optimizer, and $50,000$ for the vocabulary size. For **FlowNet** variants, wefollow the setup used in the original paper (Kang et al., 2019).

For BERT and GPT2 models, we use 32 for batch size, 2e-4 leraning rate with 1.0 maximum gradient norm and 0.02 weight decay using Adam (Kingma and Ba, 2014) optimizer.

Since we don't have industry-level of large computing resources, we restrict the maximum number of target sentences to 3 even though it could be up to half of the paragraph size in the full permutation. Making our training procedure on ParCom at larger scale and even larger dataset will be an interesting future direction.

B Details on Ablation test

The full table for our ablation test is in Table 10.

	`Romance`		`WikiText`		`CNNDM`	
Models	**M**	**VE**	**M**	**VE**	**M**	**VE**
SSPlanner	8.1	73.9	7.7	66.8	7.8	59.9
w/o Sent. Position	-1.4	-8.5	-1.7	-9.7	-1.1	-6.2
w/o Plan Predict	-2.1	-12.7	-2.1	-13.9	-2.0	-13.0
w/o Next Predict.	-0.2	-2.9	-0.6	-3.2	-0.6	-4.7
SSPlanner trained by:						
w Random	6.4	65.2	6.2	53.9	5.9	42.9
w Syntac(Verb)	7.8	73.7	7.5	62.8	7.6	54.6
w Syntac(Noun)	7.6	71.3	7.5	61.5	7.5	53.8
w Syntac(N+V)	**8.3**	**74.5**	**7.8**	65.9	**7.9**	58.6
w Off-the-shelf	8.1	73.9	7.7	**66.8**	7.8	**59.9**
w Attention	7.9	72.4	7.4	63.0	7.6	55.4
SSPlanner tested with different types of plan keywords						
w Random	5.7	52.9	5.8	55.0	5.9	58.9
w predicted	8.1	73.9	7.7	66.8	7.8	59.9
w Off-the-shelf ($\hat{p}$)	13.3	84.4	13.0	87.8	12.9	85.9
SSPlanner tested with different ratio of plan keywords						
w 10% keywords	6.1	55.2	5.9	57.7	6.1	56.8
w 50% keywords	7.5	67.0	6.4	59.5	6.7	56.9
w 100% keywords	8.1	73.9	7.7	66.8	7.8	59.9

Table 10: Ablation on SSPlanner's self-supervision modules (1st group), plan types for training (2nd group), and plan types (3rd group) and ratios (4th group) for testing.