1 Model Description

In this paper, we use two types of Transformer-based models: non-pre-trained models (Trans, L3) and pre-trained models (T5-Base and T5-Large) in our experiments. The number of parameters for these models are shown in Table 1. Mention Flags in the T5-Base models add approximate 5000 trainable parameters. Mention Flags in the T5-Large models add approximate 9000 trainable parameters. We use a single Tesla P100 GPU with 16 GB memory to train our models. Table 2 shows the approximate training time for our models in all three evaluation tasks. We only apply the T5-Large model to the *CommonGen* task.

Models	# Total P.	# Trainable P.
T5-Base	222904 k	113276 k
T5-Base + **MF**	222909 k	113281 k
T5-Large	737669 k	402739 k
T5-Large + **MF**	737678 k	402739 k
Trans, L3	159189 k	159189 k
Trans, L3 + **MF**	159194 k	159194 k

Table 1: Model size for all used model in this paper.

Models	*CommonGen*	*E2ENLG*	*nocaps*
T5-Base	2 h	1.5 h	30.5 h
+ **MF**	2.5 h	2 h	43.5 h
T5-Large	6 h	-	-
+ **MF**	7.5 h	-	-
Trans, L3	1.75 h	1.3 h	19.3 h
+ **MF**	2 h	1.7 h	27.6 h

Table 2: Training Time for all used model in this paper.

2 Model Hyper-parameters

As discussed above, we use the T5 models and their shallower versions. We do not conduct any model hyper-parameter search. Table 3 shows the training hyper-parameter used in our experiments. In the *CommonGen* task, we use the same batch size as suggested in the Lin et al. (2020). We search the batch size 25, 50, 75, 100 for the *E2ENLG* and *nocaps* task. We search the LR for 1e-4, 5e-4, 1e-3, the LR schedule for constant and with 10% warm-up in the *CommonGen* task. We use the CIDEr score on the *CommonGen* development set.

	CommonGen	*E2ENLG*	*nocaps*
Batch Size	192	50	50
LR	5e-5	5e-5	5e-5
LR Schedule	const.	const.	const.
Optimizer	AdamW	AdamW	AdamW

Table 3: The training hyper-parameters used in our experiments.

3 Datasets

Table 4 shows the number of training data in all three tasks. All of our three datasets follow the same formatting: each training instance has one encoder input sequence and multiple human-annotated ground-truth output sequences. During training, we use all of these output sequences for training. During evaluation, we only feed all encoder input sequences to the model and use the multiple ground-truth output sequences for evaluation. *nocaps* does not release its evaluation and test captions. We submit all data in the *E2ENLG* and *CommonGen* task in the *Data* together with this submission. Due to the large size of the *nocaps* data, we recommend readers to its official website `https://nocaps.org/` for more details. Wang et al. (2021) also share information about the *nocaps* dataset, including captions and visual features. We follow the standard split used in all previous work (Lin et al., 2020; Dušek et al., 2020; Agrawal et al., 2019). As T5 models can handle any words in the input and output sequences, we only lower case all of these sequences before feeding them into the model.

Split	*CommonGen*	*E2ENLG*	*nocaps*
Train	33k / 67k	4862 / 42k	118k / 592k
Val	4018 / 993	547 / 4672	4500 / -
Test	6012 / 1497	630 / 4693	10600 / -

Table 4: Training Data Statistics. Input Sequence / Output Sequence for Train/Val/Test split in all three tasks. *nocaps* does not release its ground-truth captions.

4 Used Evaluation Metrics

In our experiments, we use following metrics to evaluate the quality of generated output text:

1. CIDEr (Vedantam et al., 2015) is the average cosine similarity between the system output and the reference sentences on the level of n-grams (n=1,2,3,4). The importance of the

individual n-grams is given by the TF-IDF. This is proposed for short text in image captioning tasks.

2. SPICE (Anderson et al., 2016) parses human references and system outputs as scene-graphs and calculate F1 score for the graph matching between system outputs and human references. It captures more long-range dependencies and word relationships than n-gram based metrics. This is proposed for short text in image captioning tasks.

3. BLEU (Papineni et al., 2002) is the harmonic mean of n-gram precision of the system output given the ground-truth reference sentences, with n =1, 2, 3, 4. It is lowered by a brevity penalty if the output is shorter than references.

4. NIST (Lin and Hovy, 2003) is a version of BLEU with higher weighting for less frequent (i.e., more informative) n-grams and a different length penalty. It uses n = 1,2,3,4,5.

5. METEOR (Banerjee and Lavie, 2005) measures both precision and recall of uni-grams by aligning the system output with each ground-truth reference sentence. It uses fuzzy matching based on stemming and WordNet synonyms.

5 System Output

Table 5 shows some output examples (including positive and negative ones) from our **MF** model and baseline model. In positive examples, our **MF** model completes all input constraints with meaningful sentences. The baseline outputs, missing some of the constraints, do not make much sense. In the negative examples, our **MF** model produces sentences with all constraints satisfied but makes factual errors.

Positive Cases	
pair, polish, shoe, demonstrate, clean	
T5-B	a woman demonstrates how to clean a pair of shoes
+ **MF**	a woman demonstrates how to clean and polish a pair of shoes
GT	the lady demonstrates how to clean and polish a pair of shoes
guitar, tattoo, stand, front, arm	
T5-B	a man standing in front of a tattooed guitar
+ **MF**	a man with a tattoo on his arm standing in front of a guitar
GT	a man with tattoo 's on his arms is standing in front of a microphone playing guitar
crowd, stage, concert, jump, surfs	
T5-B	a man jumps and surfs on stage during a concert
+ **MF**	a man jumps off stage and surfs the crowd during a concert
GT	a man jumps off the stage and crowd surfs at a concert
couple, take, park, dog, walk	
T5-B	a couple is walking their dog in a park
+ **MF**	a couple takes their dog for a walk in a park
GT	a couple taking their dogs on a walk through a park outside
Negative Cases	
bench, wife, sit, husband	
T5-B	a husband and wife sit on a bench
+ **MF**	a man sitting on a bench with his wife and husband
GT	the husband and wife sit on a bench
ball, catch, dog, jump	
T5-B	a dog jumps to catch a ball
+ **MF**	a dog jumps on a ball and catches it
GT	the dog jumped to catch the ball

Table 5: Representative examples of T5 baseline and **MF** model output in the *CommonGen* task. GT: ground truth text. T5-B: T5 base.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the State-of-the-Art of End-to-End Natural Language Generation: The E2E NLG Challenge. *Computer Speech & Language*, 59:123–156.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Yufei Wang, Ian D. Wood, Stephen Wan, and Mark Johnson. 2021. ECOL-R: Encouraging Copying in Novel Object Captioning with Reinforcement Learning. *arXiv preprint arXiv:2101.09865*.