	R	H	L	T	E
|training set|	3114	3223	7944	4221	42061
|validation set|	1039	1075	2649	1407	4672
|test set|	1039	1075	2649	1407	4693
total	5192	5373	13242	7035	47366
DA types	8	8	14	14	1
slot types	12	12	20	16	8

Table 2: The details of different datasets.

We adopt the same configurations for all five datasets. The dimensionalities for all embeddings are 256. The hidden units of all layers are set as 512. We adopt 3 layers of self-attention and each of them has 4 heads. L2 regularization is set as 1×10^{-6} and the dropout ratio is assigned 0.4 for reducing overfit. Above setting is obtained by using grid search. We use Adam (**?**) to optimize model parameters. All the studies are conducted at GeForce RTX 2080T. For each DA, we over-generate 10 utterances through beam search and select the top 5 candidates. In experiments, we select the model that works the best on the dev set, and then evaluate it on the test set. The improvements of our models over the baselines are statistically significant with $p < 0.05$ under t-test.

4.2 Main Results

The baselines for comparison are as follows:

- HLSTM (**?**) designs a heuristic gate to control the states of slot values, guaranteeing that all of them are accurately captured;
- SCLSTM (**?**) introduces an extra "reading gate to improve standard LSTM;
- TGen (**?**) uses encoder-decoder architecture augmented with attention mechanism to generate utterances from input DA;
- RALSTM (**?**) presents a RNN-based decoder to select and aggregate the semantic elements produced by attention mechanism;
- NLG-LM **?** incorporates language modeling into response generation. In particular, it provides current state-of-the-art performances on BLEU score for all the datasets.

We follow the baseline results as reported in **??**.

Table 1 presents main results. First of all, VQ-VAE is more effective than gumbel-softmax for modeling mode switcher. Except for Laptop dataset, using VQ-VAE obtains better performances than using gumbel-softmax. In terms of BLEU scores, it improves the performances by 1.68% on Restaurant dataset, 1.99% on Hotel dataset, 2.84% on Television dataset, and 2.06% on E2E-NLG dataset. For correctness, it also reduces the ERR by 14.29% on Restaurant dataset, 28.57% on Hotel dataset, and 11.39% on Television dataset. Secondly, our model is not only interpretable but also competitive with current state-of-the-art method. For instance, our BLEU scores are lower than those of NLG-LM by only 0.75% on Restaurant dataset and 0.44% on E2E-NLG dataset.

Method	Hotel	E2E-NLG
Our Model	**0.921**	**0.681**
w/o Self-attention	0.908	0.651
w/o Self-attention, w/ LSTM	0.914	0.672
w/o Pointer Net	0.862	0.665
w/o Pointer Net, w/ Delex.	0.910	0.669

Table 3: Ablation experiments on two datasets.

Lastly, we discover that there is a trade-off between Interpretability and effectiveness. The results shown in the last row are obtained by directly using the categorical distribution $[z_j^p, z_j^c, z_j^l]$ predicted in Equation 12 as the output $\mathbf{o}_j$ of mode switcher. Although it degrades the interpretability, the performances are notably improved. Compared with NLG-LM, it achieves the increasements on BLEU scores of 2.14% on Restaurant dataset, 0.85% on Laptop dataset, and 1.90% on E2E-NLG dataset. For ERR, the scores are also reduced by 26.19% on Laptop dataset and 19.05% on Television dataset. We guess that using discrete representations limits the flexibility of neural networks.

4.3 Case Study

In this section, we show that HRM is capable of interpreting the rendering process well.

Using the procedure described in Section 2.3, we label a generated item as $P(i)$ if it's copied from the i-th slot value, $C(i)$ if it's reworded from the i-th slot value, or L otherwise. Several examples are depicted in Figure 4, where blue lines and red lines denote paraphrasing and copying, respectively. In most of the cases, HRM is of great interpretability. For example, in the third case, phrase "has low customer ratings" is aligned with value "1 out of 5". In the second case, phrase "an adult restaurant" is aligned with slot value ("familyFriendly", "no"). There also exists a few mistakes. For instance, in the fourth case, phrase "outside Cambridge City Centre" is not tied with value "riverside". We attribute this to the high fault tolerance of end-to-end neural networks. In addition, here we omit act type for brevity.

4.4 Ablation Study

As shown in Table 3, we conduct ablation study to explore the impacts of components.

Effect of DA-level Encoder. By removing self-attention, the BLEU scores sharply drop by 1.41% and 4.41%. Hence it's useful to capture the semantic correlations among slot values. Additionaly, replacing self-attention with LSTM decreases the BLEU scores by 0.76% and 1.32%. We credit this to the order-insensitivity of self-attention.

Effect of Copying Mechanism. By removing pointer network, the BLEU scores decrease by 6.41% and 2.35%. Hence directly copying words from input DA is useful. Replacing pointer network with delexicalization reduces the

Method	E2E-NLG
TGen	0.562
NLG-LM	0.495
Our Model w/ Gumbel-softmax	0.815
Our Model w/ VQ-VAE	0.872
Our Model w/ Softmax	0.621

Table 4: Human evaluation on E2E-NLG dataset.

BLEU scores by 1.19% and 1.76%. Besides, delexicalization can only apply to delexicalizable slots. These indicate that pointer network is more effective.

4.5 Human Evaluation

We conduct human evaluation to quantitively make comparisons on interpretability among different models. The task consists of three stages. Firstly, we introduce an alignment score to assess the interpretability of a model. It's computed as p/N. p is the number of slot values that are correctly aligned with the utterance by a model (see Figure 4). N is the total amount of slot values. Then, we define how various generators interpret their rendering processes. For our models, we do interpretation as described in Section 2.3. For the baselines (e.g., TGen), we use the computed attention weights to align the input DA with the generated sentence. Eventually, we randomly sample 200 cases from the test set as the benchmark set and recruit 15 people to judge whether a slot value is aligned well with the generated utterance (i.e., getting p). The designed task is very simple and highly consistent among workers. Table 4 demonstrates the experiment results. From it, we can draw the following conclusions. Firstly, better performances may lead to worse interpretability. For example, the state-of-the-art model, NLG-LM, underperforms a simple baseline, TGen, by 13.5%. From Table 1, we can see that, for our model, using softmax generally obtains higher F1 scores than adopting discrete latent variable models. However, here we find the interpretability score of using softmax is lower than using VQ-VAE or Gumbel-softmax. Secondly, our models consistently and significantly outperform prior baselines. For example, the F1 score of Our Mode w/ VQ-VAE outnumbers TGen by 35.6% and NLG-LM by 43.2%. Thirdly, VQ-VAE is better than Gumbel-softmax in terms of both BLEU score and interpretability score. For example, in Table 1, using VQ-VAE outperforms using Gumbel-softmax by 1.65% on Restaurant, 2.84% on Television, and 2.06% on E2E-NLG. In Table 4, the increase percentage of interpretability score is 6.54%.

5 Related Work

In task-oriented dialogue systems, NLG is the final module to produce user-facing system utterances, which is directly related to the perceived quality and usability. Traditional approaches generally divide the NLG task into a pipeline of sentence planning and surface realization (**????**). Sentence planning first converts an input DA into a tree-like structure, and then surface realization maps the intermediate structure into the final surface form. For example, **?** use a class-based n-gram language model and a template-based reranker. **?** address the limitations of n-gram language models by using more complex syntactic trees. **?** employ a phrase-based generator that learns from a semantically aligned corpus. Although these methods are adequate and of great interpretability, they are heavily dependent on handcraft rules and expert knowledge. Moreover, the sentences generated from rule-based systems are often rigid, without the diversity and naturalness of human language. Lately, there is a surge of interest in utilizing neural networks to build corpus-based NLG models (**????**). The main superiority is facilitating end-to-end training on the unaligned corpus. For example, **?** present a heuristic gate to guarantee that all slot value pairs are accurately captured during generation. **?** introduce a novel SC-LSTM with an additional reading cell to learn gated mechanism and language model jointly. **?** use encoder-decoder architecture augmented with attention mechanism to generate utterances from input DA. **?** use a RNN-based decoder to select and aggregate the semantic elements produced by attention mechanism. Most recently, **?** incorporate a language model task into the response generation process to boost the naturalness of generated utterances. **?** study the slot consistency issue and propose a novel iterative rectification network to address it. While plenty of state-of-the-art performances have been obtained, they are all treated as black boxes, and thus lack interpretability. Delexicalization (**???**) to some extent raises the interpretability as it directly locates the position of slot values in the utterance. Nevertheless, it is applicable for delexicalizable slots only. In E2E-NLG dataset, most of the slots are reworded or indicative. **?** also observe that using delexicalization results in mistakes.

6 Conclusion

In this paper, we present heterogeneous rendering machines (HRM) to improve the interpretability of NLG models. It consists of a renderer set and a mode switcher. The renderer set contains multiple decoders that vary in structure and functionality. The mode switcher is a discrete latent variable that chooses an appropriate decoder from the renderer set in every generation step. Extensive experiments have been conducted on five datasets, demonstrating that our model is competitive with the current state-of-the-art method. Qualitative studies show that our model can interpret the rendering process well. Human evaluation further confirms its effectiveness in interpretability. Currently, a severe problem in interpretable NLG is lacking a proper evaluation metric. Mainstream metrics such as BLEU are not applicable. Using our alignment score demands massive annotation efforts. We will work hard on this issue in future research.

Acknowledgments

This work was supported by Ant Group through Ant Research Program. We thank anonymous reviewers for their valuable and constructive comments.

Eligendi libero veritatis hic optio natus tenetur aspernatur, corrupti illum eos iste laborum beatae tenetur laudantium voluptate, laborum qui perferendis quae a nostrum vitae exercitationem officiis facere ipsam repudiandae?Dolorum corporis repudiandae repellat ratione