A Table-to-Text Framework with Heterogeneous Multidominance Attention and Self-Evaluated Multi-Pass Deliberation

Xi Chen[1]*, **Xinjiang Lu**[2]✉, **Haoran Xin**[3], **Wenjun Peng**[1], **Haoyang Duan**[1], **Feihu Jiang**[1], **Jingbo Zhou**[2], **Hui Xiong**[3]✉

[1] University of Science and Technology of China, [2] Baidu Research,
[3] Hong Kong University of Science and Technology (Guangzhou)
{chenxi0401,pengwj,duanhaoyang,jiangfeihu}@mail.ustc.edu.cn,
{luxinjiang,zhoujingbo}@baidu.com,
hxin883@connect.hkust-gz.edu.cn, xionghui@ust.hk

Abstract

Though big progress in table-to-text works, effectively leveraging table structure signals, *e.g.,* hierarchical structure, remains challenging. Besides, deliberating generated descriptions proves to be effective for table-to-text. However, determining the appropriate outcome when encountering multi-pass candidates is another challenge. To this end, we propose a novel table-to-text approach on top of **S**elf-evaluated multi-pass **G**eneration and **H**eterogenous **M**ultidominance **A**ttention, namely SG-HMA. Specifically, we formulate the table structure into a multidominance (MD) structure and devise a heterogenous multidominance attention (HMA) to comprehensively explore the complex interactions encoded in the hierarchical structure, which can further deliver rich signals for text generation with the help of pre-trained language models (PLMs). Afterward, a contrastive loss is introduced to align the generation objective with evaluation metrics, so the more faithful generated descriptions can be guaranteed. We conduct extensive experiments on three public datasets, demonstrating that SG-HMA outperforms several SOTA methods quantitatively and qualitatively.

1 Introduction

Table-to-text, referring to the task of producing a textual description taking the table as input, has been widely applied in different domains, such as weather forecast (Liang et al., 2009; Mei et al., 2016), logical tabular reasoning (Chen et al., 2020a), and financial report generation (Lin et al., 2022). Automatic description generation may shed light on mitigating the time-consuming procedure in table-to-text tasks with bare hands.

Recent advances in pre-trained language models (PLMs) have demonstrated significant progress in natural language generation (NLG) (Yao et al., 2022, 2023; Fang et al., 2023). To effectively leverage the power of PLM, several table-to-text works (Gong et al., 2020; Suadaa et al., 2021) serialized the table input via manually defined templates. Besides, to preserve the table's structural information, TableGPT (Gong et al., 2020) devises a table structure reconstruction task, and TASD (Chen et al., 2022b) proposes to learn the structure representation explicitly. However, to group the data into categories, the cells in a table (*e.g.,* a pivot table) are often organized in a nested/hierarchical manner using headings and subheadings. The attention map computed by these approaches may fail to harness the perplexing hierarchical table structures.

On the other hand, one can effectively deliberate the generated texts from a global perspective with the multi-pass generation paradigm (Niehues et al., 2016; Chen et al., 2022b). While it is hard to terminate the multi-pass generation procedure and determine the appropriate outcome towards faithful descriptions of tables. Existing works evaluate and finalize the multi-pass generated texts with the help of reinforcement learning (RL) (Geng et al., 2018) or a customized rewrite-evaluator architecture (Li and Yao, 2021). However, extra workload might be brought in due to the intractability of RL or a separate evaluation module.

To this end, in this paper, we propose a PLM-based table-to-text approach with the help of **S**elf-evaluated multi-pass **G**eneration and **H**eterogeneous **M**ultidominance **A**ttention (SG-HMA). Specifically, we first formulate an input table into a multidominance (MD) structure (Gračanin-Yuksek, 2013). In this way, the hierarchical relation of the table content can be preserved since one child node in the MD structure can have more than one parental node. Then, we devise a heterogenous multidominance attention (HMA) mechanism to represent the table content with the awareness of complex hierarchical structure. Afterward, we deliberate the generated texts with a multi-pass paradigm

*This work was done when the first author was an intern at Baidu Research.

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 607–620
December 6-10, 2023

and develop a contrastive loss to equip the model to generate more faithful table descriptions with self-evaluation. The contributions of this work can be summarized as follows:

- We propose to transform the tabular input into a multidominance structure and devise a heterogenous multidominance attention to yield table representation on top of PLMs.
- We innovate a self-evaluated multi-pass generation framework for the table-to-text task with the help of contrastive learning.
- Extensive experiments on benchmark datasets validate the superiority of the proposed SG-HMA framework in generating descriptive texts for tabular inputs.

2 Related Work

2.1 Table-to-Text Generation

With the success of deep neural networks, the seq2seq method has been applied to various natural language generation (NLG) tasks. Based on this framework, researchers (Liu et al., 2018; Puduppully et al., 2019) divide table-to-text generation into content selection, planning, and surface smooth under end-to-end training fed a sequence of table records as input. However, these methods rely on large-scale datasets, showing poor results in few-shot learning.

Since PLMs have shown great potential in transfer learning, fine-tuning PLMs on different downstream tasks becomes a general and effective method in NLG tasks (Kale and Rastogi, 2020; Chen et al., 2020b). Parikh et al. (2020) proposed a novel table-to-text dataset with a controlled generation task applying BERT as a baseline. Gong et al. (2020) transformed the table into natural language text and designed two auxiliary tasks to address the incompatibility between text-to-text PLMs and table-to-text generation. To infer facts from tables, Chen et al. (2020a) introduced reinforcement learning in the training algorithm, and Suadaa et al. (2021) designed a reasoning-based template. Inspired by prompting, Li and Liang (2021) applied prefix-tuning to GPT2 for table-to-text generation and outperformed fine-tuning in low-data settings. Li et al. (2021) introduced table representation learning into fine-tuning of PLMs, showing the potential of table representation in guiding text generation. TASD (Chen et al., 2022b) designed three multi-head attention layers within and among cells, ignoring the inherent hierarchical structure within a table. For the table-to-text task, no work takes the data structure of tables into account. We resolve the table hierarchical structure into an MD structure and devise an HMA to learn the table representation.

2.2 Contrastive Learning for Natural Language Generation

Contrastive learning has been widely applied in NLG tasks such as machine translation (Yang et al., 2019) and summarization (Cao and Wang, 2021). SimCTG (Su et al., 2022), as a contrastive training objective, can calibrate the model's representation space. Xu et al. (2022) propose SeqCo with a contrastive objective trying to map representations of document and summary to the same space.

Most generation models are trained without being exposed to incorrectly generated tokens. To solve the exposure bias problem, Sun and Li (2021) applies margin-based losses in the generated text. Liu and Liu (2021) trains an evaluation model with contrastive learning. Several works (Lee et al., 2021; An et al., 2022; Su et al., 2022) use an N-pairs contrastive loss and Brio (Liu et al., 2022) uses a ranking loss based on their sequence-level scores to learn a better sequence-level distance function between the document and the target. Therefore, contrastive learning can well bridge the gap between training objectives and evaluation metrics (Chen et al., 2022a). We are the first to apply it into text deliberation, implementing a multi-pass generation process that can be self-evaluated.

3 Preliminaries

3.1 Problem Formulation

In a typical table, the caption provides critical information about the entire table and the associated topics, the rows and columns describe the properties of their affiliated cells, while the cells provide additional details within the table framework. A table organizes these by a hierarchical structure, hiding deep semantic information which can be expressed by natural language. The table-to-text task aims at generating an appropriate summary s for each table t.

Specifically, given a structured table t, the model θ is expected to generate a descriptive sentence y in an auto-regressive way:

$$y_i = \arg\max_{y_i} P(y_i|t, y_{<i}; \theta), i = 1, ..., |y|, \quad (1)$$

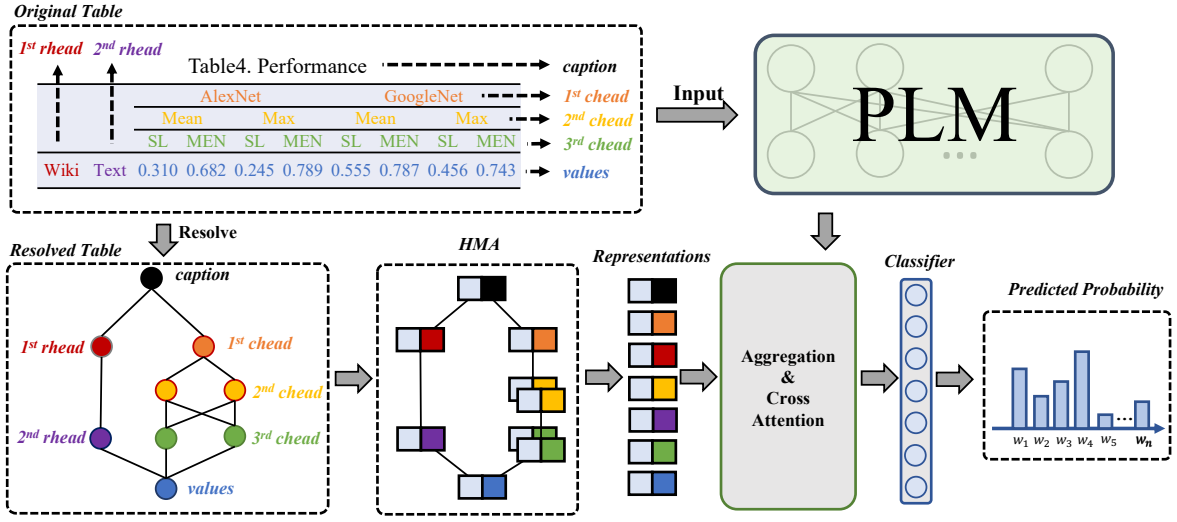


Figure 1: The architecture of the heterogeneous multidominance attention enhanced generation model.

where $|y|$ is the word number of sentence y.

3.2 PLM as Generator

Fine-tuning PLM on different downstream tasks becomes a general and effective method in NLG tasks. We first serialize the table into natural language text N_t that conforms to the standard input format of PLM. Given the serialized table N_t and the reference s, in the training process, the last hidden state H with the input of the decoder s is obtained as follows:

$$H_{t,s} = \text{PLM}([N_t; s]). \tag{2}$$

where H will be used to predict the probability of the next token. Formally, the training objective of text generation that maximizes the likelihood of the reference text is given by:

$$\mathcal{L}_{mle} = -\frac{1}{|S|}\sum_{i=1}^{|S|}\log P(S_i|[N_t; S_{<i}]). \tag{3}$$

4 Methodology

In this section, we introduce the proposed framework in detail. Firstly, as shown in Fig.1, following the table hierarchical structure, we resolve the table into a MD structure and apply a HMA mechanism to propagate and aggregate the information in the nodes as the table representation, guiding the generation of PLM. During the multi-pass generation process, we apply the constrastive learning and table representation to achieve self-evaluation. Finally, we can get a well trained SE-HAN to generate ideal table description with tables as input.

4.1 Tabular Input Representation

Table Structure Formulation. Tables exhibit a natural hierarchical structure, which is comprised of the table components and their complex dependencies. In order to take full advantage of the table structural information, we resolve the hierarchical table structure into the multidominance structure (Gračanin-Yuksek, 2013) where at least one node has more than one parent node. Formally, for each table t, the table hierarchical structure is denoted as the multidominance $\mathcal{D} = \{\mathcal{V}, \mathcal{E}\}$ where $\mathcal{V}$ is the set of table component nodes extracted from t, and $\mathcal{E}$ is the set of directed edges that suggest connectivity among the nodes[1]. Based on the heterogeneity of the component nodes, $\mathcal{V}$ can be further divided into several disjoint subsets following:

$$\mathcal{V} = \mathcal{V}_a \cup \mathcal{V}_r \cup \mathcal{V}_c \cup \mathcal{V}_e, \tag{4}$$

where $\mathcal{V}_a = \{a_i\}_{i=1}^{|\mathcal{V}_a|}$, $\mathcal{V}_r = \{r_i\}_{i=1}^{|\mathcal{V}_r|}$, $\mathcal{V}_c = \{c_i\}_{i=1}^{|\mathcal{V}_c|}$ and $\mathcal{V}_e = \{e_i\}_{i=1}^{|\mathcal{V}_e|}$ are the set of caption, row, column, and content nodes, respectively. Furthermore, we use $h(r_i)$ and $h(c_i)$ to signify the located level of row $r_i \in \mathcal{V}_r$ and column $c_i \in \mathcal{V}_c$, respectively. We also denote the located row and column in the max level of the content $e_i \in \mathcal{V}_e$ as $\rho_R(e_i)$ and $\rho_C(e_i)$, respectively. Based on the heterogeneous node sets, the set of edges $\mathcal{E}$ is defined according

[1]We omit t in the notations in Section 4.1 for brevity.

to the hierarchical structure as:

$$\begin{aligned}\mathcal{E} = &\{(a_i, x_j) | a_i \in \mathcal{V}_a, x_j \in \mathcal{V}_r \cup \mathcal{V}_c, h(x_j) = 0\} \cup \\ &\{(x_i, x_j) | (x_{i,j} \in \mathcal{V}_r \vee x_{i,j} \in \mathcal{V}_c), h(x_j) - h(x_i) = 1\} \cup \\ &\{(r_i, e_j) | r_i \in \mathcal{V}_r, e_j \in \mathcal{V}_e, \rho_R(e_j) = r_i\} \cup \\ &\{(c_i, e_j) | c_i \in \mathcal{V}_c, e_j \in \mathcal{V}_e, \rho_C(e_j) = c_i\} \cup \\ &\{(x_i, x_i) | x_i \in \mathcal{V}_r \cup \mathcal{V}_c\}.\end{aligned} \tag{5}$$

Heterogeneous Multidominance Attention. Multidominance is a tree-like structure organized in a top-down manner with non-uniform abstract meaning among different levels. More importantly, MD distinguishes itself from commonly known tree structure by inheritance non-linearity, i.e., the node may have more than one parent node. To deal with MD that experiences both the directed hierarchical structure and the inheritance non-linearity, we propose a novel heterogeneous multidominance attention (HMA) mechanism, where a flow aggregation layer is employed to adaptively aggregate the heterogeneous flows within the MD table structure.

Specifically, we first obtain the initial representations of the table component nodes by the PLM embedding layer $\mathrm{Emb}(\cdot)$:

$$E_x = \mathrm{Emb}(x), x \in \mathcal{V}, \tag{6}$$

where $E_x \in \mathbb{R}^{l_x \times d}$, l_x is the length of the text in node x and d is the dimensionality. Then, we respectively concatenate the representations of the nodes in terms of their types (i.e., caption, row, column and content) as:

$$U_A = \mathop{\Vert}_{a \in \mathcal{V}_a} E_a, \tag{7}$$

where $U_A \in \mathbb{R}^{n_A \times d}$ is the caption representation, n_a is the length of all caption nodes, and $\Vert$ is the concatenation operation along the first dimension. Analogously, the row, column and content representations $U_R \in \mathbb{R}^{n_R \times d}$, $U_C \in \mathbb{R}^{n_C \times d}$ and $U_E \in \mathbb{R}^{n_E \times d}$ can be also derived.

Afterward, we devise a flow aggregation layer $\mathrm{FA}(\cdot \rightsquigarrow \cdot, \cdot)$ to enable the table information flow from the high-level component (e.g., captions) to the low-level component (e.g., rows and columns) while selectively accumulating beneficial knowledge for each node. Specifically, the row (column) representation can be updated as:

$$\begin{aligned}Z_{R(C)} &= \mathrm{FA}(U_A \rightsquigarrow U_{R(C)}, P^{A \to R(C)}), \\ \widetilde{Z}_{R(C)} &= \mathrm{FA}(Z_{R(C)} \rightsquigarrow U_{R(C)}, P^{R(C) \to R(C)}),\end{aligned} \tag{8}$$

where P is a incidence prior suggesting the hierarchical connectivity based on $\mathcal{D}$ between specific types of nodes, which is illustrated in Fig. 2. For example, consider the connectivity from caption nodes to row nodes, $P^{A \to R} \in \mathbb{R}^{n_A \times n_R}$ is given by:

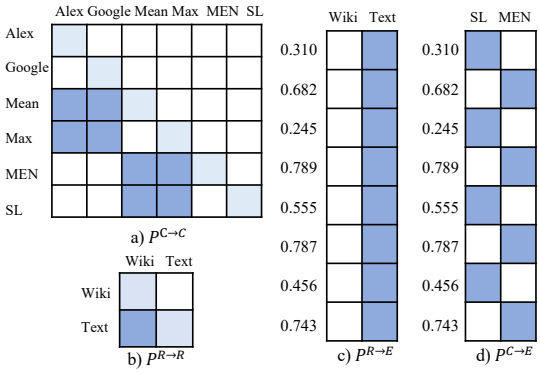


Figure 2: The prior for the table in Fig 1 where the blue cell indicates 1 while the white cell indicates 0.

$$P_{ij}^{A \to R} = \begin{cases} 1, & \text{if } a_i \in \mathcal{V}_a, r_j \in \mathcal{V}_r, (a_i, r_j) \in \mathcal{E}, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

The cell content representation can be further acquired following:

$$Z_E = \mathrm{FA}(\left[\widetilde{Z}_R \parallel \widetilde{Z}_C\right] \rightsquigarrow U_E, \left[P^{R \to E} \parallel P^{C \to E}\right]). \tag{10}$$

Particularly, to fulfill adaptive information aggregation, the flow aggregation layer FA is implemented with a sparse attention (Wang et al., 2019) as:

$$\begin{aligned}&\mathrm{FA}(S \rightsquigarrow T, P) = (P \odot \mathrm{softmax}(\frac{QK^T}{\sqrt{d}}))V, \\ &Q = TW_Q, K = SW_K, V = SW_V,\end{aligned} \tag{11}$$

where S, T are the source and target input of the flow, respectively, $\odot$ is the element-wise multiplication, and W_Q, W_K, W_V are parameters. Finally, the table representation Z_T is obtained by concatenating the learned caption, row, column and content representations as:

$$Z_T = \left[Z_A \parallel \widetilde{Z}_R \parallel \widetilde{Z}_C \parallel Z_E\right]. \tag{12}$$

4.2 Self-Evaluated Multi-pass Generation

For the generation procedure, we first utilize the table representation to enhance the generation ability of PLM by cross attention. So as to achieve self-evaluation that dominates the termination signal in the multi-pass deliberation, we develop a contrastive loss that ranks the similarity between the table and the generation samples conforming to the evaluation metric. Moreover, the deliberation is performed by rewriting not only the output

Algorithm 1: Training Procedure.

Data: Given a training dataset with a table set $\mathcal{T}$, the corresponding reference set $\mathcal{S}$ and a language model LM_Θ with initial parameters Θ

Result: A language model for table-to-text generation LM_{Θ^*} with optimal parameters Θ^*

$p \leftarrow 1$, $\{N_{t_i}\}_{i=1}^{|\mathcal{T}|} \leftarrow$ table serialization ;$M^{(0)} \leftarrow 0$;
do
  $p \leftarrow p + 1$;
  for $t \in \mathcal{T}$ **do**
    $\{u_i^t\}_{i=1}^{|n_u|}, y \leftarrow \mathrm{LM}_\Theta(t, N_t)$;
  end
  compute $\mathcal{L}_{mul}$ by Eq.16;
  $\Theta^{(p)} \leftarrow \Theta^{(p-1)} - \lambda \nabla_\Theta \mathcal{L}_{mul}$;
  $M^{(p)} \leftarrow 0$;
  for $t \in \mathcal{T}$ **do**
    $M^{(p)} \leftarrow \sigma(t, y) + M^{(p)}$;
    $N_t \leftarrow [N_t; y]$;
  end
while $M^{(p)} \geq M^{(p-1)}$;
$\Theta^* \leftarrow \Theta^{(p-1)}$;
return LM_{Θ^*}

Algorithm 2: Inference Procedure.

Data: Given a table t, a well-trained table-to-text generation LM_{Θ^*} with parameters Θ

Result: A textual description of the table

$p \leftarrow 1$; $N_t \leftarrow$ table serialization; $M^{(0)} \leftarrow 0$;
do
  $p \leftarrow p + 1$;
  $y^{(p)} \leftarrow \mathrm{LM}_\Theta(t, N_t)$;
  $M^{(p)} \leftarrow \sigma(t, y^{(p)})$; $N_t \leftarrow [N_t; y^{(p)}]$;
while $M^{(p)} \geq M^{(p-1)}$;
return $y^{(p-1)}$

generation but also the candidates, which are able to incorporate abundant samples into constrastive learning.

Table Structure Enhanced Generation. Despite the capability of the hidden state H learned by the PLM to incorporate the collective information of the table, we propose to fuel the text generation with hierarchical structure of tables, which is crucial to generate a high-quality description. Specifically, we serialize the table based on the MD structure where the detail can be found in Appendix B.5 and leverage the table representation learned in Section 4.1 to make further selection on the table content by multi-head cross attention given by:

$$\widetilde{H}_{t,s} = \mathrm{MultiHeadAttn}(H_{t,s}, Z_T, Z_T) + H_{t,s}, \quad (13)$$

where $\widetilde{H}_{t,s}$ is a structure-enhanced hidden state that determines the probability of text generation. The $\mathcal{L}_{mle}$ defined in Equation 3 is applied as the training objective for text generation.

Contrastive Self-Evaluation. As the table itself and its summary convey different aspects of the same semantic information, it is beneficial to align the table's hidden state for generation with more closely matched summaries. Hence, we propose to empower the model to self-evaluate the quality of candidates based on their compatibility with the table, using a carefully designed ranking-based contrastive loss. Specifically, inspired by sinkborn divergence (SD) (Feydy et al., 2019) that interpolates between MMD (Gretton et al., 2006) and OT (Chen et al., 2019) to attain probability distribution comparison, we first define the similarity score σ between the table t and its generation sample g as:

$$\sigma(t, g) = -\mathrm{SD}(\widetilde{H}_{t,g}, Z_T). \quad (14)$$

Then, we generate the candidates $\{u_i^t\}_{i=1}^{n_u}$ by beam search for table t where n_u is the candidate number. Afterward, we leverage the evaluation metrics (e.g., BLEU) to construct the partial order $\succ$ of the candidates, where higher metrics suggest higher rankings. Finally, We introduce a ranking loss (Zhong et al., 2020) to assign higher score to better candidates as follows:

$$\mathcal{L}_{ctr} = \sum_{u_i^t} \sum_{u_j^t \succ u_i^t} max(0, \sigma(t, u_j^t) - \sigma(t, u_i^t) + \epsilon), \quad (15)$$

where ϵ is the margin value.

Learning Objective and Deliberation. We jointly train the text generation task and candidate self-evaluation task with a composite loss:

$$\mathcal{L}_{mul} = \mathcal{L}_{mle} + \alpha \mathcal{L}_{ctr}, \quad (16)$$

where α is a hyperparameter that is a scale factor. Regarding the multi-pass deliberation, for each pass, we rewrite not only the generation result but the candidates to incorporate abundant samples into

Metrics	GPT2					BART		T5		
	F-T	TableGPT	P-T	TASD	SG-HMA	F-T	SG-HMA	F-T	Cont	SG-HMA
					NumericNLG					
BLEU	4.60	5.08	3.28	5.10	**5.76**	6.41	**8.18**	4.96	5.13	**6.45**
ROUGE-L	17.89	18.39	18.51	20.40	**21.69**	21.36	**22.30**	20.36	21.06	**23.10**
NIST	1.24	2.17	0.65	1.09	**1.36**	2.45	**2.92**	1.21	**1.67**	1.58
METEOR	10.87	11.34	8.89	11.46	**11.97**	12.97	**14.10**	11.36	11.38	**12.30**
CIDEr	0.06	0.04	0.10	0.07	**0.13**	0.10	**0.16**	0.15	0.09	**0.20**
					Totto					
BLEU	18.09	18.10	17.02	18.50	**20.30**	17.03	**17.52**	18.13	18.53	**22.64**
ROUGE-L	36.46	36.75	34.83	36.85	**38.34**	34.93	**38.26**	37.34	36.90	**41.04**
NIST	4.09	3.96	3.60	4.03	**4.23**	3.17	**3.96**	3.96	4.10	**4.43**
METEOR	20.58	18.45	18.61	20.59	**21.04**	19.83	**23.01**	20.21	20.67	**23.80**
CIDEr	1.45	1.39	1.48	1.56	**1.64**	1.32	**1.38**	1.52	1.68	**1.82**
					E2E					
BLEU	67.75	68.80	69.92	68.96	**70.61**	68.70	**69.45**	69.20	69.61	**70.03**
ROUGE-L	70.76	70.56	71.46	70.77	**72.18**	70.77	**70.78**	70.90	71.16	**71.18**
NIST	8.63	8.63	8.81	8.76	**8.92**	8.80	**8.83**	8.80	8.80	**8.82**
METEOR	45.65	45.66	46.27	45.96	**46.75**	45.70	**45.77**	46.03	46.20	**46.30**
CIDEr	2.33	2.34	2.45	2.43	**2.49**	2.44	**2.45**	2.47	2.48	**2.49**

Table 1: Performance comparisons of the automatic evaluation on the fine-tuning, TableGPT, prefix-tuning, TASD and SG-HMA with different backbones. Except fine-tuning, other baselines are implemented according to their source codes. We implement fine-tuning and our method with three backbones to achieve a fair comparison.

the contrastive evaluation. Then, according to the average score, we terminates the deliberation if the score starts to decline. The details of the training pipeline are shown in Algorithm 1.

Table Description Inference. In the inference stage, we generate multiple candidates in an autoregressive manner using the multi-pass deliberation paradigm. To attain a satisfactory deliberation, we further leverage the benefits of contrastive self-evaluation to discern the quality of candidates. Specifically, we terminate the deliberation if the quality of the newly generated candidate, as determined by the similarity score defined in Equation 14, begins to decline. Detailed description of the inference stage is given in Algorithm 2.

5 Experiments

5.1 Experimental Settings

Datasets We conducted experiments on three publicly available table-to-text datasets: numericNLG[2] (Suadaa et al., 2021), Totto[3] (Parikh et al., 2020), and E2E[4] (Novikova et al., 2017). The numericNLG dataset comprises tables with complex hierarchical structures and corresponding descriptions, sourced from scientific papers. The Totto dataset is a diverse English table-to-text dataset. To ensure a similar dataset size to numericNLG, we manually filtered out tables with simple hierarchical structures and constructed a new Totto dataset. The E2E dataset is a straightforward table dataset describing restaurant information. Furthermore, we input the table without knowledge of the highlighted sections to demonstrate the guiding significance of the hierarchical table representation to the table content. The statistics of the datasets are shown in Table 4 of Appendix B.1.

Evaluation Metrics. We included five most widely used automatic metrics: BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), NIST (Doddington, 2002), METEOR (Lavie and Agarwal, 2007), and CIDEr (Vedantam et al., 2015) to evaluate the quality of our generation. These metrics can be evaluated by a public evaluation script[5]. More details are in Appendix B.2.

Backbones. To verify the backbone-agnostic nature of our architecture, we employed three primary PLMs, GPT2 (Radford et al., 2019), BART (Lewis et al., 2020), and T5 (Raffel et al., 2020), as backbones. More details of these backbones can be found in Appendix B.3.

Baselines. We compare our method with several state-of-the-art baselines, such as Fine-tuning, TableGPT (Gong et al., 2020), Prefix-tuning (Li

[2]https://github.com/titech-nlp/numeric-nlg
[3]https://github.com/google-research-datasets/ToTTo
[4]https://github.com/UFAL-DSG/tgen
[5]https://github.com/tuetschek/e2e-metrics

and Liang, 2021), Cont (An et al., 2022) and TASD (Chen et al., 2022b). The details of these baselines can be found in Appendix B.5.

5.2 Overall Performance

As shown in Table 1, we compared the performance of SG-HMA with multiple state-of-the-art methods on three datasets. The results show that SG-HMA outperforms the best baseline on all metrics. Specifically, in terms of BLEU, ROUGE-L, NIST, METEOR, and CIDEr, SG-HMA surpasses the strongest baseline and achieves an improvement of 1.77 (6.41 → 8.18, 27.6%), 0.94 (21.36 → 22.30, 4.40%), 0.47 (2.45 → 2.92, 19.18%), 1.13 (12.97 → 14.1, 8.71%), and 0.06 (0.1 → 0.16, 60%) on the NumericNLG dataset; 4.11 (18.53 → 22.64, 22.2%), 4.14 (36.9 → 41.04, 11.22%), 0.33 (4.1 → 4.43, 8.05%), 3.13 (20.67 → 23.8, 15.14%), and 0.14 (1.68 → 1.82, 8.33%) on the Totto dataset; 0.69 (69.92 → 70.61, 0.97%), 0.72 (71.46 → 72.18, 1.00%), 0.11 (8.81 → 8.92, 1.03%), 0.48 (46.27 → 46.75, 1.04%), and 0.04 (2.45 → 2.49, 1.63%) on the E2E dataset, respectively. Furthermore, it's worth noting that SG-HMA can maintain a leading position when changing its backbone, which demonstrates its backbone-agnostic nature.

Moreover, we observed performance variations of different backbones across datasets. Notably, BART demonstrated the best performance on the numericNLG dataset due to its reputation as a denoising model, allowing for more accurate summaries of complex tables with noise. On the Totto dataset, which features tables of varying types and formats, T5 achieved the best results. This is because T5 is pre-trained with multiple tasks, giving it a strong generalization capability to process various data types with different formats. Lastly, the E2E dataset, which describes restaurant information in a relatively simple format, is more susceptible to overfitting with complex models. Therefore, the GPT2 backbone may be slightly more effective.

5.3 In-depth Analysis

We conducted an in-depth analysis from multiple perspectives on the three public datasets with the best-performing backbones to gain further insights into our proposed method and verify the effectiveness of each component.

Ablation study. To futher explore the effectiveness of our proposed modules, we conduct an ablation study. Specifically, we compare SG-HMA with several variants: **1)** ***w/o*** HMA that removes the heterogeneous multidominance attention, **2)** ***w/o*** ctr that removes the contrastive self-evaluation in the training stage, **3)** FA that replaces our designed HMA with a full attention.

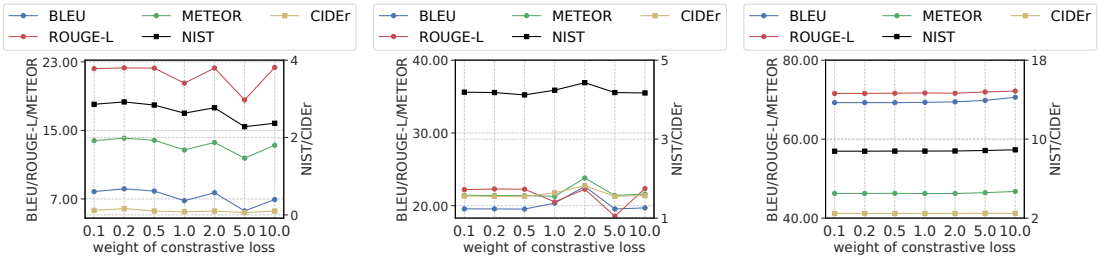
Method	B	R	N	M	C
BART			NumericNLG		
w/o HMA	6.56	21.49	2.45	12.94	0.10
w/o ctr	7.58	22.21	2.74	13.46	0.09
FA	7.05	21.75	2.60	13.18	0.15
SG-HMA	**8.18**	**22.30**	**2.92**	**14.10**	**0.16**
T5			Totto		
w/o HMA	18.24	37.89	4.02	21.07	1.61
w/o ctr	19.34	36.53	4.13	21.30	1.53
FA	19.62	38.46	4.11	21.74	1.65
SG-HMA	**22.64**	**41.04**	**4.43**	**23.80**	**1.82**
GPT2			E2E		
w/o HMA	68.12	70.78	8.69	46.27	2.41
w/o ctr	69.21	71.58	8.78	46.24	2.47
FA	69.07	71.83	8.74	46.37	2.48
SG-HMA	**70.61**	**72.18**	**8.92**	**46.75**	**2.49**

Table 2: Performance comparisons on BLEU(B), ROUGE-L(R), NIST(N), METEOR(M) and CIDEr(C) of SG-HMA and its variants with the best backbone.

As can be seen in Table 2, *w/o* ctr performs worse than SG-HMA under all metrics. This demonstrates the importance of the contrastive loss in guiding the PLM to generate a more reasonable probability distribution. Besides, SG-HMA outperforms *w/o* HMA and FA under all metrics. Since our HMA is a form of flow aggregation that propagates information in a top-down manner, it matches the inherited complex structures of tables, thus performing best on all datasets.

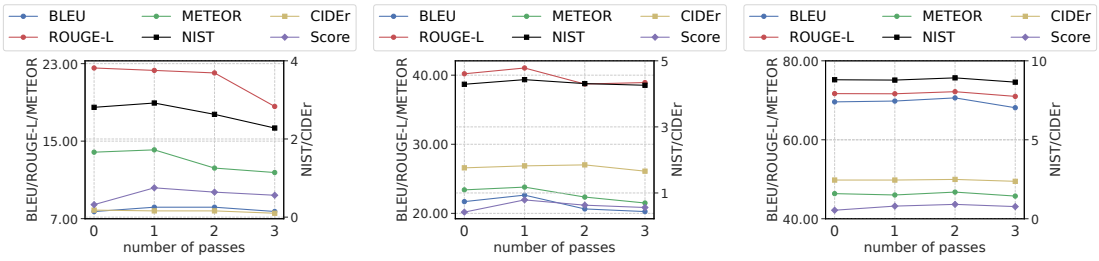
Parameter sensitivity analysis. We study the impact of constrastive loss weight on the model performance. The weight varies from 0.1 to 10. As shown in the first row of Figure 3, on the whole, the model performance initially improves as the weight increase. This is because increasing weight forces the PLM to generate a more reasonable probability distribution and evaluate the deliberation result effectively. However, as the weight continues to increase, the model's performance decreases. This is because when the weight is too large, the model overly focuses on the contrastive loss and degenerates into a discriminative model, losing the ability to generate text.

Effectiveness of self-evaluation on deliberation. As shown in Figure 4, we present the predicted similarity score by our model, along with five met-



(a) Weight NumeircNLG (b) Weight Totto (c) Weight E2E

Figure 3: Parameter sensitivity analysis on contrastive loss weight.



(a) #Pass NumeircNLG (b) #Pass Totto (c) #Pass E2E

Figure 4: Self evaluation effectivess validation.

rics against the golden table description for the description candidates in each pass.We have observed that metrics perform best at the second pass on the numericNLG and Totto datasets, while it achieves the best performance at the third pass on the E2E dataset.Such an observation underscores the significance of identifying an appropriate termination point for the deliberation, as prolonged deliberation may result in a deterioration of the generated description. Furthermore, we observe a positive correlation between the model's predicted similarity score and the BLEU metric owing to the contrastive self-evaluation performed with the BLEU metric during the training stage. As a result, our proposed model can serve as a reliable indicator of the quality of the generated table descriptions during the inference stage, facilitating the assessment of deliberation adequacy.

5.4 Case Study

Fig 5 presents an interesting case we observed when comparing the results on the numericNLG dataset. Specifically, the green highlights indicate text that appears in both the table and the generated result, the red highlights indicate summary statements and their corresponding content, and the blue highlights indicate parts showing the hierarchical semantic relationship. Compared to the fine-tuned BART model, the results generated by

Table 5: Performance on monolingual word similarity computation with seed lexicon size 6000

	Chineses(source)		English(target)	
Method	WS-240	WS-297	WS-353	SL-999
BiLex	60.36	62.17	60.46	27.22
CLSP-WR	61.27	65.25	60.46	27.22
CLSP-SE	60.84	65.25	62.47	28.79

Groundtruth: table 5 shows the results of monolingual word similarity computation on four datasets.from the table, we find that:.(1) our models perform better than bilex on both chinese word similarity datasets.(2) clsp-wr model does not enhance english word similarity results but clspse model does

Finetuned BART: table 5 shows the results of this experiment. we can see that clsp-wr achieves the median accuracy of 60.46% compared to bilex, which indicates that clone-based methods are more effective than conventional word similarity methods.

SG-HMA BART: table 5 shows the results of monolingual word similarity computation with seed lexicon size 6000. we observe that clsp-se outperforms other models on all datasets except for Chinese ws-240 where the accuracy of bilex is slightly worse

Figure 5: An example of the generated table descriptions on the numericNLG dataset.

SG-HMA better restore the original content of the table. In terms of summarization ability, SG-HMA not only correctly identifies "CLSP-SE" as the best performing model, but also acknowledges poor performance on the "ws-240" dataset and captures the hierarchical relationship between "Chinese" and "ws-240". This demonstrates our model's ability to summarize tables by effectively mining hierarchical table structure information.

5.5 Human Evaluation

We randomly selected 40 samples from three datasets and conducted a human evaluation from four aspects: fluency, coverage, relevance, and

Method		Flue	Cove	Rele	Ovqa
			NumericNLG		
GPT2	Finetuning	4.10	3.40	3.72	3.68
GPT2	TASD	4.22	3.52	3.85	3.80
GPT2	SG-HMA	**4.32**	**3.67**	**4.06**	**3.98**
BART	Finetuning	4.33	3.57	3.75	3.69
BART	SG-HMA	**4.44**	**4.00**	**4.16**	**4.11**
			Totto		
T5	Finetuning	4.31	3.47	3.75	3.59
T5	Cont	4.34	3.67	3. 87	3.76
T5	SG-HMA	**4.41**	**4.01**	**4.17**	**4.16**
			E2E		
GPT2	Finetuning	4.40	3.87	4.05	3.98
GPT2	Prefixtuning	4.44	4.00	4.16	4.11
GPT2	SG-HMA	**4.46**	**4.11**	**4.21**	**4.20**

Table 3: Human evaluation results. Flue means fluency, Cove means coverage, Rele means relevance, Ovqa means overall quality.

overall quality. Each criterion was rated on a scale from 1 (worst) to 5 (best). The criterion details are shown in Appendix D. We invited 20 volunteers to evaluate the generated text. According to the results of Table 3, we have the following observations: **1)** All methods perform similarly in fluency. Since PLMs have developed very mature in text generation, their powerful imitation ability allows them to generate text that easily conforms to human reading habits. **2)** Regarding coverage, relevance, and overall quality, our proposed method performs better for considering table structures and self-evaluation generation, correctly summarizing critical information, and generating more reasonable descriptions.

6 Conclusion

In this article, we introduced SG-HMA, a novel approach that can effectively learn table representation and generate accurate summaries. To be specific, SG-HMA firstly resolves the table hierarchical structure into a MD structure and utilizes the HMA to acquire the table representation, which can guide the PLM in generating and evaluating text. Then, the candidate texts and generation results in each pass are rewrited by SG-HMA to create samples for metric ranking with a contrastive objective and obtain more accurate summaries. Self-evaluation enables the model have capability to determine when to terminate training process based on the evaluation results, facilitating the generation of high-quality text. Finally, extensive experiments conducted on three datasets demonstrate the effectiveness of SG-HMA.

7 Limitations

Our proposed method exhibits marginal improvements over the prefix-tuning baseline when the input tabular data, such as the E2E dataset, are relatively simple. To further enhance the model's performance on simple tables, we aim to integrate prompt learning with our hierarchical table representation in future work. Besides, we only take the BLEU as the metric ranking criterion in contrastive learning. In the future, we will consider all metrics to achieve a more balanced model.

8 Ethics Statement

We will abide by the laws, rules, and regulations of our community, school, work, and country. We will conduct ourselves with integrity, fidelity, and honesty. We will openly take responsibility for our actions and only make agreements, which we intend to keep.

Acknowledgements

This research was supported in part by the National Natural Science Foundation of China (Grant No.92370204) and the Science and Technology Planning Project of Guangdong Province (Grant No. 2023A0505050111).

References

Chenxin An, Jiangtao Feng, Kai Lv, Lingpeng Kong, Xipeng Qiu, and Xuanjing Huang. 2022. CoNT: Contrastive neural text generation. In *Advances in Neural Information Processing Systems*.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liqun Chen, Yizhe Zhang, Ruiyi Zhang, Chenyang Tao, Zhe Gan, Haichao Zhang, Bai Li, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019. Improving sequence-to-sequence learning via optimal transport. In *International Conference on Learning Representations*.

Liyi Chen, Zhi Li, Tong Xu, Han Wu, Zhefeng Wang, Nicholas Jing Yuan, and Enhong Chen. 2022a. Multi-modal siamese network for entity alignment. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 118–126.

Miao Chen, Xinjiang Lu, Tong Xu, Yanyan Li, Zhou Jingbo, Dejing Dou, and Hui Xiong. 2022b. Towards table-to-text generation with pretrained language model: A table structure understanding and

text deliberating approach. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8199–8210, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020a. Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7929–7942, Online. Association for Computational Linguistics.

Zhiyu Chen, Harini Eavani, Wenhu Chen, Yinyin Liu, and William Yang Wang. 2020b. Few-shot NLG with pre-trained language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Online. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, page 138–145. Morgan Kaufmann Publishers Inc.

Chuyu Fang, Chuan Qin, Qi Zhang, Kaichun Yao, Jingshuai Zhang, Hengshu Zhu, Fuzhen Zhuang, and Hui Xiong. 2023. Recruitpro: A pretrained language model with skill-aware prompt learning for intelligent recruitment. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3991–4002.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. 2019. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690.

Xinwei Geng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2018. Adaptive multi-pass decoder for neural machine translation. In *EMNLP*, pages 523–532.

Heng Gong, Yawei Sun, Xiaocheng Feng, Bing Qin, Wei Bi, Xiaojiang Liu, and Ting Liu. 2020. TableGPT: Few-shot table-to-text generation with table structure reconstruction and content matching. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1978–1988, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Martina Gračanin-Yuksek. 2013. *Linearizing multidominance structures*, pages 269–294. De Gruyter Mouton, Berlin, Boston.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. 2006. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102, Dublin, Ireland. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Seanie Lee, Dong Bok Lee, and Sung Ju Hwang. 2021. Contrastive learning with adversarial perturbations for conditional text generation. In *International Conference on Learning Representations*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Liang Li, Can Ma, Yinliang Yue, and Dayong Hu. 2021. Improving encoder by auxiliary supervision tasks for table-to-text generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5979–5989, Online. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Yangming Li and Kaisheng Yao. 2021. Rewriter-evaluator architecture for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5701–5710, Online. Association for Computational Linguistics.

Percy Liang, Michael Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 91–99, Suntec, Singapore. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xiexiong Lin, Huaisong Li, Tao Huang, Feng Wang, Linlin Chao, Fuzhen Zhuang, Taifeng Wang, and Tianyi Zhang. 2022. A logic aware neural generation method for explainable data-to-text. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3318–3326.

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2018. Table-to-text generation by structure-aware seq2seq learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive

summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2016. What to talk about and how? selective generation using LSTMs with coarse-to-fine alignment. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 720–730, San Diego, California. Association for Computational Linguistics.

Jan Niehues, Eunah Cho, Thanh-Le Ha, and Alex Waibel. 2016. Pre-translation for neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1828–1836, Osaka, Japan. The COLING 2016 Organizing Committee.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany. Association for Computational Linguistics.

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.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6908–6915.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, page 9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, pages 1–67.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. In *Advances in Neural Information Processing Systems*.

Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura. 2021. Towards table-to-text generation with numerical reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1451–1465, Online. Association for Computational Linguistics.

Shichao Sun and Wenjie Li. 2021. Alleviating Exposure Bias via Contrastive Learning for Abstractive Text Summarization. *arXiv e-prints*, page arXiv:2108.11846.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Yaushian Wang, Hung-Yi Lee, and Yun-Nung Chen. 2019. Tree transformer: Integrating tree structures into self-attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1061–1070, Hong Kong, China. Association for Computational Linguistics.

Shusheng Xu, Xingxing Zhang, Yi Wu, and Furu Wei. 2022. Sequence level contrastive learning for text summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11556–11565.

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. 2019. Reducing word omission errors in neural machine translation: A contrastive learning approach. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6191–6196, Florence, Italy. Association for Computational Linguistics.

Kaichun Yao, Jingshuai Zhang, Chuan Qin, Xin Song, Peng Wang, Hengshu Zhu, and Hui Xiong. 2023. Resuformer: Semantic structure understanding for resumes via multi-modal pre-training. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 3154–3167. IEEE.

Kaichun Yao, Jingshuai Zhang, Chuan Qin, Peng Wang, Hengshu Zhu, and Hui Xiong. 2022. Knowledge enhanced person-job fit for talent recruitment. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 3467–3480. IEEE.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

A Multidominance Structure based Table Serialization

Formally, the serialization of the table according to both the structure and the content is given by: As (the caption is) a, the $r_0^{h_r}$ of r_0^0 ... $r_0^{h_r-1}$ and $c_0^{h_c}$ of c_0^0 ... $c_0^{h_c-1}$ is $e_k^{r_0,c_0}$, ..., the $r_i^{h_r}$ of r_i^0 ... $r_i^{h_r-1}$ and $c_j^{h_c}$ of c_j^0 ... $c_j^{h_c-1}$ is $e_k^{r_i,c_j}$, ... where x_i^j is the i-th x with a j-level, $e_k^{r_i,c_j}$ is the cell e_k with row attribute r_i and column attribute c_j and h_x is the max level of x.

B Extra Experimental Setings

B.1 Dataset Divison

The dataset divison of numericNLG and E2E adapt the official method. For Totto dataset, we filter 1189 samples. Table 4 shows the size of each part of the dataset after division.

Split	NumericNLG	Totto	E2E
Train	1084	960	4862
Val	136	113	547
Test	135	116	630

Table 4: Statistics of the training, validation, and test sets for the NumericNLG, Totto and E2E datasets.

B.2 Evaluation Metrics

BLEU measures the precision of N grams in a sentence against references. ROUGE-L measures the recall of the longest common subsequence between the source and the target. NIST improves the BLEU method by weighing the penalty for incorrectly matching n-grams. METEOR evaluates the generation of word-to-word matching. CIDEr is an automated consistency metric used for evaluating image descriptions.

B.3 Backbones

GPT2 is a pre-trained language model with a decoder-only transformer architecture. It was pre-trained on a large and diverse webtext dataset with the goal of maximizing the probability of generating high-quality text. BART is a denoising autoencoder for pretraining sequence-to-sequence models, with a standard transformer-based architecture. T5 is a pretrained transformer model with an encoder-decoder architecture. providing a unified framework for converting all NLP tasks into text-to-text tasks.

B.4 Implementation Details

Regarding automatic evaluation, all results of deep models were obtained by conducting experiments on a Linux machine with Nvidia P40 GPU. Furthermore, an Adam optimizer was utilized for LM fine-tuning, and training was iterated in 30 epochs for numericNLG, 20 epochs for Totto and 5 epochs for E2E. A beam search algorithm was adopted when generating the text and the beamwidth was set to 4. The learning rate of PLM was searched from 1e-5, 5e-5, 1e-4 and we selected 1e-4 for numericNLG and Totto and 1e-5 for E2E. We use BLEU as the evaluation metric to define the target ordering of the candidate summaries. We fine-tuned on a GPT2 model with 124M parameters, a BART model with 400M parameters and a T5 model with 220M parameters.

B.5 Baselines

We compare SG-HMA with the most relevant baselines as following:

- **Fine-tuning**. To leverage the rich semantic information in PLMs, fine-tuning as a transfer learning method has shown great potential in various downstream tasks, including machine translation, named entity recognition and summarization.

- **TableGPT**. TableGPT is the first attempt to apply table serialization to convert semi-structured data into natural language text and a multi-task learning paradigm to enhance the generation ability, showing the potential in leveraging the table structure information.

- **Prefix-tuning**. Prefix is a sequence of continuous task-specific vectors, the only module that needs to be optimized while keeping basic PLM parameters. Prefix-tuning is a state-of-the-art table-to-text method on the E2E dataset.

- **Cont**. Contrastive learning is an effective solution to solve the exposure bias problem in NLG tasks. This paper proposes a unified framework to break the bottlenecks from tree aspects: contrastive example construction, contrastive loss choice and decoding strategy.

- **TASD**. TASD devises a three-layered multi-head attention network to leverage the table structure information and adapt a multi-pass

Year	Show	Role	Notes
2016	In the Heights	Choreographer	Hangar Theatre
2012	Sweet Charity	Director	New Haarlem Arts Theatre
2008	Whistle Down the Wind	Director / Choreographer	North Carolina Theatre

Reference1	In 2012, Agustin directed Sweet Charity at The New Haarlem Arts Theatre.
Fine-tuning T5	Agustin directed the 2007 show, Out of Line, at the Pennsylvania Center Stage.
Cont T5	In 2008, Julio Agustin played the role of Director Choreographer in Whistle Down the Wind.
SG-HMA T5	Julio Agustin worked as a Director /Choreographer for the North Carolina Theatre , in 2008, and as Director for the 2012 show Sweet Charity at the New Haarlem Arts Theatre .

Table 5: An example of the generated table descriptions on Totto dataset

name	The Phoenix	type	pub
food	French	price	less than £ 20
customer rating	low	area	riverside
family friendly	yes	near	Crowne Plaza Hotel

Reference1: Near Crowne Plaza Hotel by riverside is a pub that is yes family friendly with a low customer rating called The Phoenix and the prices are less than £ 20 .
Fine-tuning GPT2: The Phoenix is a family friendly pub located near the Crowne Plaza Hotel. It is in the low price range.
Prefix-tuning GPT2: The Phoenix is a family friendly pub located near the Crowne Plaza Hotel.
SG-HMA GPT2: The Phoenix is a family friendly French pub in the riverside area near Crowne Plaza Hotel. It has a low customer rating and a price range of less than £ 20 .

Table 6: An example of the generated table descriptions on E2E dataset

decoder framework to polish the generation. It is a most recent baseline on numericNLG and part of Totto dataset.

C Extra Case Study

In addition to the case of numericNLG in the text, we also conducted a case study on the Totto and E2E datasets. Table 5 shows a case from the Totto dataset, which clearly demonstrates the wider coverage of original table content and a more comprehensive summary achieved by SG-HMA. Similarly, on the E2E dataset, SG-HMA outperformed fine-tuning and prefix-tuning by accurately and comprehensively summarizing more table content shown on the table 6. These cases provide compelling evidence of the benefits of SG-HMA in comprehending table representations. They indicate that the HMA method has the potential to extract structural information from tables and generate more extensive summaries through the SG.

D Human Evaluation Settings

Four criteria were used in our human evaluation:

- Fluency: Is this text readable and smooth?
- Coverage: Does this text describe the table content comprehensively?
- Relevance: Is this text related to the table content? Dose it reflect the table content authentically?
- Overall quality: Evaluation of overall quality of generated text.

Each criterion was scored on a scale of 1(worst) to 5(best). In greater detail, we present a comprehensive list of explicit justifications for scoring the generated text according to each criterion, as outlined below.

Fluency

1. Poor Fluency: The text is difficult to understand.
2. Below Average Fluency: The text has some basic elements of fluency but still contains noticeable errors or inconsistencies.
3. Average Fluency: The text demonstrates a moderate level of fluency, with relatively few errors.
4. Above Average Fluency: The text demonstrates a high level of fluency, with minimal errors or disruptions.
5. Excellent Fluency: The text demonstrates exceptional fluency, with virtually no errors or disruptions.

Coverage

1. The generated text does not mention any key elements of the table.
2. The generated text provides limited information about some of the key elements of the table.
3. The generated text provides moderately comprehensive description of the table.
4. The generated text provides a comprehensive description of the table.
5. The generated text offers an exceptional and comprehensive description of the table.

Relevance

1. The generated text does not mention any relevant information about the table, and/or the information provided is entirely fabricated or false.
2. The generated text includes some relevant information about the table, but the description is limited, and/or contains factual inaccuracies.
3. The generated text provides a moderately relevant description of the table, and/or contain minor errors.
4. The generated text offers a highly relevant description of the table, and/or contains mostly accurate and reliable information that can be reasonably verified.
5. The generated text provides an exceptionally relevant and authentic description of the table.

Overall Quality

1. The generated text is of poor quality overall with numerous issues.
2. The generated text is below average in quality, with several issues affecting its overall effectiveness.
3. The generated text is average in quality, with some strengths but also some weaknesses.
4. The generated text is above average in quality, with clear strengths and good overall effectiveness.
5. The generated text is of excellent quality overall, with outstanding strengths and high effectiveness.