Long and Diverse Text Generation with Planning-based Hierarchical Variational Model

Zhihong Shao\textsuperscript{1}, Minlie Huang\textsuperscript{1}, Jiangtao Wen\textsuperscript{1}, Wenfei Xu\textsuperscript{2}, Xiaoyan Zhu\textsuperscript{1}

\textsuperscript{1} Institute for Artificial Intelligence, State Key Lab of Intelligent Technology and Systems
\textsuperscript{1} Beijing National Research Center for Information Science and Technology
\textsuperscript{1} Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
\textsuperscript{2} Baozun, Shanghai, China

szh19@mails.tsinghua.edu.cn, aihuang@tsinghua.edu.cn
jtwen@tsinghua.edu.cn, xuwenfeilittle@gmail.com, zxy-dcs@tsinghua.edu.cn

Abstract

Existing neural methods for data-to-text generation are still struggling to produce long and diverse texts: they are insufficient to model input data dynamically during generation, to capture inter-sentence coherence, or to generate diversified expressions. To address these issues, we propose a Planning-based Hierarchical Variational Model (PHVM). Our model first plans a sequence of groups (each group is a subset of input items to be covered by a sentence) and then realizes each sentence conditioned on the planning result and the previously generated context, thereby decomposing long text generation into dependent sentence generation sub-tasks. To capture expression diversity, we devise a hierarchical latent structure where a global planning latent variable models the diversity of reasonable planning and a sequence of local latent variables controls sentence realization. Experiments show that our model outperforms state-of-the-art baselines in long and diverse text generation.

1 Introduction

Data-to-text generation is to generate natural language texts from structured data (Gatt and Krahmer, 2018), which has a wide range of applications (for weather forecast, game report, product description, advertising document, etc.). Most neural methods focus on devising encoding scheme and attention mechanism, namely, (1) exploiting input structure to learn better representation of input data (Lebret et al., 2016; Liu et al., 2018), and (2) devising attention mechanisms to better employ input data (Mei et al., 2016; Liu et al., 2018; Nema et al., 2018) or to dynamically trace which part of input has been covered in generation (Kiddon et al., 2016). These models are able to produce fluent and coherent short texts in some applications.

However, to generate long and diverse texts such as product descriptions, existing methods are still unable to capture the complex semantic structures and diversified surface forms of long texts. First, existing methods are not good at modeling input data dynamically during generation. Some neural methods (Kiddon et al., 2016; Feng et al., 2018) propose to record the accumulated attention devoted to each input item. However, these records may accumulate errors in representing the state of already generated prefix, thus leading to wrong new attention weights. Second, inter-sentence coherence in long text generation is not well captured (Wiseman et al., 2017) due to the lack of high-level planning. Recent studies propose to model planning but still have much space for improvement. For instance, in (Puduppully et al., 2019) and (Sha et al., 2018), planning is merely designed for ordering input items, which is limited to aligning input data with the text to be generated. Third, most methods fail to generate diversified expressions. Existing data-to-text methods inject variations at the conditional output distribution, which is proved to capture only low-level variations of expressions (Serban et al., 2017).

To address the above issues, we propose a novel Planning-based Hierarchical Variational Model (PHVM). To better model input data and alleviate the inter-sentence incoherence problem, we design a novel planning mechanism and adopt a compatible hierarchical generation process, which mimics the process of human writing. Generally speaking, to write a long text, a human writer first arranges contents and discourse structure (i.e., high-level planning) and then realizes the surface form of each individual part (low-level realization). Motivated by this, our proposed model first performs
Figure 1: Generation process of PHVM. After encoding a list of input attribute-value pairs, PHVM first conducts planning by generating a sequence of groups, each of which is a subset of input items. Each sentence is then realized conditioned on the corresponding group and its previous generated sentences.

planning by segmenting input data into a sequence of groups, and then generates a sentence conditioned on the corresponding group and preceding generated sentences. In this way, we decompose long text generation into a sequence of dependent sentence generation sub-tasks where each sub-task depends specifically on an individual group and the previous context. By this means, the input data can be well modeled and inter-sentence coherence can be captured. Figure 1 depicts the process.

To deal with expression diversity, this model also enables us to inject variations at both high-level planning and low-level realization with a hierarchical latent structure. At high level, we introduce a global planning latent variable to model the diversity of reasonable planning. At low level, we introduce local latent variables for sentence realization. Since our model is based on Conditional Variational Auto-Encoder (CVAE) (Sohn et al., 2015), expression diversity can be captured by the global and local latent variables.

We evaluate the model on a new advertising text generation task from (Kiddon et al., 2016) which requires the system to correctly use the given ingredients and maintain coherence among cooking steps. Experiments on advertising text generation show that our model outperforms state-of-the-art baselines in automatic and manual evaluation. Our model also generalizes well to long recipe text generation and outperforms the baselines. Our contributions are two-fold:

- We design a novel Planning-based Hierarchical Variational Model (PHVM) which integrates planning into a hierarchical latent structure. Experiments show its effectiveness in coverage, coherence, and diversity.

- We propose a novel planning mechanism which segments the input data into a sequence of groups, thereby decomposing long text generation into dependent sentence generation sub-tasks. Thus, input data can be better modeled and inter-sentence coherence can be better captured. To capture expression diversity, we devise a hierarchical latent structure which injects variations at both high-level planning and low-level realization.

2 Related Work

Traditional methods (Reiter and Dale, 1997; Stent et al., 2004) for data-to-text generation consist of three components: content planning, sentence planning, and surface realization. Content planning and sentence planning are responsible for what to say and how to say respectively; they are typically based on hand-crafted (Kukich, 1983; Dalianis and Hovy, 1993; Hovy, 1993) or automatically-learnt rules (Duboue and McKeown, 2003). Surface realization generates natural language by carrying out the plan, which is template-based (McRoy et al., 2003; van Deemter et al., 2005) or grammar-based (Bateman, 1997; Espinosa et al., 2008). As these models are shallow and the two stages (planning and realization) often function separately, traditional methods are unable to capture rich variations of texts.

Recently, neural methods have become the mainstream models for data-to-text generation due to their strong ability of representation learning and scalability. These methods perform well in generating weather forecasts (Mei et al., 2016) or very short biographies (Lebret et al., 2016; Liu
Figure 2: Architecture of PHVM. The model controls planning with a global latent variable $z^p$. The plan decoder conducts planning by generating a sequence of groups $g = g_1g_2...g_T$ where $g_t$ is a subset of input items and specifies the content of sentence $s_t$ to be generated. The sentence decoder controls the realization of $s_t$ with a local latent variable $z^s_t$; dependencies among $z^s_t$ are explicitly modeled to better capture inter-sentence coherence.

et al., 2018; Sha et al., 2018; Nema et al., 2018) using well-designed data encoder and attention mechanisms. However, as demonstrated in Wiseman et al. (2017) (a game report generation task), existing neural methods are still problematic for long text generation: they often generate incoherent texts. In fact, these methods also lack the ability to model diversity of expressions.

As for long text generation, recent studies tackle the incoherence problem from different perspectives. To keep the decoder aware of the crucial information in the already generated prefix, Shao et al. (2017) appended the generated prefix to the encoder, and Guo et al. (2018) leaked the extracted features of the generated prefix from the discriminator to the generator in a Generative Adversarial Nets (Goodfellow et al., 2014). To model dependencies among sentences, Li et al. (2015) utilized a hierarchical recurrent neural network (RNN) decoder. Konstas and Lapata (2013) proposed to plan content organization with grammar rules while Puduppully et al. (2019) planned by reordering input data. Most recently, Moryossef et al. (2019) proposed to select plans from all possible ones, which is infeasible for large inputs.

As for diverse text generation, existing methods can be divided into three categories: enriching conditions (Xing et al., 2017), post-processing with beam search and rerank (Li et al., 2016), and designing effective models (Xu et al., 2018). Some text-to-text generation models (Serban et al., 2017; Zhao et al., 2017) inject high-level variations with latent variables. Variational Hierarchical Conversation RNN (VHCR) (Park et al., 2018) is a most similar model to ours, which also adopts a hierarchical latent structure. Our method differs from VHCR in two aspects: (1) VHCR has no planning mechanism, and the global latent variable is mainly designed to address the KL collapse problem, while our global latent variable captures the diversity of reasonable planning; (2) VHCR injects distinct local latent variables without direct dependencies, while our method explicitly models the dependencies among local latent variables to better capture inter-sentence connections. Shen et al. (2019) proposed ml-VAE-D with multi-level latent variables. However, the latent structure of ml-VAE-D consists of two global latent variables: the top-level latent variable is introduced to learn a more flexible prior of the bottom-level latent variable which is then used to decode a whole paragraph. By contrast, our hierarchical latent structure is tailored to our planning mechanism: the top level latent variable controls planning results and a sequence of local latent variables is introduced to obtain fine-grained control of sentence generation sub-tasks.

We evaluated our model on a new advertising text generation task which is to generate a long and diverse text that covers all given specifications about a product. Different from our task, the advertising text generation task in (Chen et al., 2019) is to generate personalized product description based on product title, product aspect (e.g., “appearance”), and user category.

3 Task Definition

Given input data $x = \{d_1, d_2, ..., d_N\}$ where each $d_i$ can be an attribute-value pair or a keyword, our task is to generate a long and diverse text $y = s_1s_2...s_T$ ($s_t$ is the $t^{th}$ sentence) that covers $x$ as much as possible. For the advertising
text generation task, $x$ consists of specifications about a product where each $d_i$ is an attribute-value pair $<a_i, v_i>$. For the recipe text generation task, $x$ is an ingredient list where each $d_i$ is an ingredient. Since the recipe title $r$ is also used for generation, we abuse the symbol $x$ to represent $<(d_1, d_2, ..., d_N), r>$ for simplification.

4 Approach

4.1 Overview

Figure 2 shows the architecture of PHVM. PHVM first samples a global planning latent variable $z^p$ based on the encoded input data; $z^p$ serves as a condition variable in both planning and hierarchical generation process. The plan decoder takes $z^p$ as initial input. At time step $t$, it decodes a group $g_t$ which is a subset of input items ($d_i$) and specifies the content of the $t^{th}$ sentence $s_t$. When the plan decoder finishes planning, the hierarchical generation process starts, which involves the high-level sentence decoder and the low-level word decoder. The sentence decoder models intersentence coherence in semantic space by computing a sentence representation $h^s_t$ and sampling a local latent variable $z^s_t$ for each group. $h^s_t$ and $z^s_t$, along with $g_t$, guide the word decoder to realize the corresponding sentence $s_t$.

The planning process decomposes the long text generation task into a sequence of dependent sentence generation sub-tasks, thus facilitating the hierarchical generation process. With the hierarchical latent structure, PHVM is able to capture multi-level variations of texts.

4.2 Input Encoding

We first embed each input item $d_i$ into vector $e(d_i)$. The recipe title $r$ is also embedded as $e(r)$. We then encode $x^2$ with a bidirectional Gated Recurrent Unit (GRU) (Cho et al., 2014). For advertising text generation, $x$ is represented as the concatenation of the last hidden states of the forward and backward GRU $\text{enc}(x) = [\overrightarrow{h_N}; \overleftarrow{h_1}]$; for recipe text generation, $\text{enc}(x) = [h_N; h_1; e(r)]$. $h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}]$ is the context-aware representation of $d_i$. Note that input encoder is not necessarily an RNN; other neural encoders or even other encoding schemes are also feasible, such as multi-layer perceptron (MLP) and bag of words.

4.3 Planning Process

The planning process generates a subset of input items to be covered for each sentence, thus decomposing long text generation into easier dependent sentence generation sub-tasks. Due to the flexibility of language, there may exist more than one reasonable text that covers the same input but in different order. To capture such variety, we model the diversity of reasonable planning with a global planning latent variable $z^p$. Different samples of $z^p$ may lead to different planning results which control the order of content. This process can be formulated as follows:

$$g = \text{argmax}_g P(g|x, z^p)$$ (1)

where $g = g_1 g_2 ... g_T$ is a sequence of groups, and each group $g_t$ is a subset of input items which is a main condition when realizing the sentence $s_t$.

The global latent variable $z^p$ is assumed to follow the isotropic Gaussian distribution, and is sampled from its prior distribution $p_0(z^p|x) = N(\mu^p, \sigma^2 I)$ during inference and from its approximate posterior distribution $q_{θ^p}(z^p|x, y) = N(\mu^p_θ, \sigma^p I)$ during training:

$$[\mu^p; \log \sigma^p] = MLP_θ(x)$$ (2)

$$[\mu^p_θ; \log \sigma^p_θ] = MLP_θ(y, x, y)$$ (3)

We solve Eq. 1 greedily by computing $g_t = \text{argmax}_{g_t} P(g_t|g_{<t}, x, z^p)$ with the plan decoder (a GRU). Specifically, at time step $t$, the plan decoder makes a binary prediction for each input item by estimating $P(d_i \in g_t|g_{<t}, x, z^p)$:

$$P(d_i \in g_t) = \sigma(W_p[h_i; h^p_i] + b_p)$$ (4)

where $σ$ denotes the sigmoid function, $h_i$ is the vector of input item $d_i$, and $h^p_i$ is the hidden state of the plan decoder. Each group is therefore formed as $g_t = \{d_i|P(d_i \in g_t) > 0.5\}$ (If this is empty, we set $g_t = \{\text{argmax}_{d_i} P(d_i \in g_t)\}$).

We feed $\text{bow}(g_t)$ (the average pooling of $[h_i|d_i \in g_t]$) to the plan decoder at the next time step, so that $h^p_{t+1}$ is aware of what data has been selected and what has not. The planning process proceeds until the probability of stopping at the next time step is over 0.5:

$$P^\text{stop}_t = \sigma(W_c h^p_t + b_c)$$ (5)

The hidden state is initialized with $\text{enc}(x)$ and $z^p$. The plan decoder is trained with full supervision, which is applicable to those tasks where
reference plans are available or can be approximated. For both tasks we evaluate in this paper, we approximate the reference plans by recognizing the subset of input items covered by each sentence with string match heuristics. The loss function at time step $t$ is given by:

$$- \log P(g_t = \tilde{g}_t|g_{<t}, x, z^p) = - \sum_{d_i \in \tilde{g}_t} \log P(d_i \in g_t) - \sum_{d_i \notin \tilde{g}_t} \log (1 - P(d_i \in g_t))$$

(6)

where $\tilde{g}_t$ is the reference group. As a result, $z^p$ is forced to capture features of reasonable planning.

4.4 Hierarchical Generation Process

The generation process produces a long text $y = s_1 s_2 ... s_T$ in alignment with the planning result $g = g_1 g_2 ... g_T$, which is formulated as follows:

$$c = \{x, z^p\}$$

(7)

$$y = \text{argmax}_y P(y|g, c)$$

(8)

We perform sentence-by-sentence generation and solve Eq. 8 greedily by computing $s_t = \text{argmax}_s_t P(s_t|s_{<t}, g, c)$. $s_t$ focuses more on $g_t$ than on the entire plan $g$. The generation process is conducted hierarchically, which consists of sentence-level generation and word-level generation. Sentence-level generation models inter-sentence dependencies at high level, and interactively controls word-level generation which conducts low-level sentence realization.

Sentence-level Generation

The sentence decoder (a GRU) performs sentence-level generation; for each sentence $s_t$ to be generated, it produces a sentence representation $h^s_t$ and introduces a local latent variable $z^s_t$ to control sentence realization.

The latent variable $z^s_t$ is assumed to follow the isotropic Gaussian distribution. At time step $t$, the sentence decoder samples $z^s_t$ from the prior distribution $p_\phi(z^s_t|s_{<t}, g, c) = \mathcal{N}(\mu^s_t, \sigma^s_t^2 \mathbf{I})$ during inference and from the approximate posterior distribution $q_{\phi'}(z^s_t|s_{<t}, g, c) = \mathcal{N}(\mu'^s_t, \sigma'^s_t^2 \mathbf{I})$ during training. $h^s_t$ and the distribution of $z_t^s$ are given by:

$$h^s_t = \text{GRU}_s([z^s_{t-1}; h^w_{t-1}], h^s_{t-1})$$

(9)

$$\begin{align*}
[\mu^s_t; \log \sigma^s_t^2] &= \text{MLP}_\phi(h^s_t, \text{bow}(g_t)) \\
[\mu'^s_t; \log \sigma'^s_t^2] &= \text{MLP}_{\phi'}(h^s_t, \text{bow}(g_t), s_t)
\end{align*}$$

(10)

(11)

where $h^w_{t-1}$ is the last hidden state of the word decoder after decoding sentence $s_{t-1}$, and $\text{GRU}_s$ denotes the GRU unit of the sentence decoder. By this means, we constrain the distribution of $z^s_t$ in two aspects. First, to strengthen the connection from the planning result $g$, we additionally condition $z^s_t$ on $g_t$ to keep $z^s_t$ focused on $g_t$. Second, to capture the dependencies on $s_{<t}$, we explicitly model the dependencies among local latent variables by inputting $z^s_{t-1}$ to the sentence decoder, so that $z^s_t$ is conditioned on $z^s_{<t}$ and is expected to model smooth transitions in a long text.

We initialize the hidden state $h^s_0$ by encoding the input $x$, the global planning latent variable $z^p$ and the planning result $g$:

$$h^s_0 = \text{GRU}_g(\text{bow}(g_1), h^q_{t-1})$$

(12)

$$h^s_0 = W_s[\text{enc}(x); z^p; h^q_t] + b_s$$

(13)

where $h^q_T$ is the last hidden state of $\text{GRU}_g$ that encodes the planning result $g$.

Word-level Generation

The word decoder (a GRU) conducts word-level generation; it decodes a sentence $s_t = \text{argmax}_s_t P(s_t|s_{<t}, z^s_t, g, c)$ conditioned on $\{h^s_t, z^s_t, g_t\}$. Specifically, we sample word $w^d_t$ of $s_t$ as follows:

$$w^d_t \sim P(w^d_t|w^d_{<t}, s_{<t}, z^s_t, g, c)$$

(14)

4.5 Loss Function

We train our model end-to-end. The loss function has three terms: the negative evidence lower bound (ELBO) of $P(y|x)$ ($L_1$), the loss of predicting the stop signal ($L_2$) and the bag-of-word loss (Zhao et al., 2017) ($L_3$).

We first derive the ELBO:

$$\log P(y|x) \geq E_{q_{\phi'}(z^p|x,y)}[\log P(y|x, z^p)] - D_{KL}(q_{\phi'}(z^p|x, y)||p_{\theta}(z^p|x))$$

(15)

$$\log P(y|x, z^p) = \log P(g, y|x, z^p) = \sum_{t=1}^{T} \log P(g_t|g_{<t}, x, z^p) + \log P(s_t|s_{<t}, g, x, z^p)$$

(16)

$$\log P(s_t|s_{<t}, g, x, z^p) \geq E_{q_{\phi'}(z^s_{t}|s_{<t}, g, x, z^p)}[\log P(s_t|s_{<t}, z^s_t, g, x, z^p)] - D_{KL}(q_{\phi'}(z^s_{t}|s_{<t}, g, x, z^p)||p_{\phi}(z^s_{t}|s_{<t}, g, x, z^p))$$

(17)
We can obtain the ELBO by unfolding the right hand side of Eq. 15 with Eq. 16 and 17. During training, we use linear KL annealing technique to alleviate the KL collapse problem (Bowman et al., 2016).

\[ L_2 = \sum_{t=1}^{T-1} \log P_{t}^{\text{stop}} + \log(1 - P_{T}^{\text{stop}}) \]  

(18)

\[ \mathcal{L}_3 \] is the sum of bag-of-word loss (Zhao et al., 2017) applied to each sentence, which is another technique to tackle the KL collapse problem.

5 Experiments

5.1 Dataset

We evaluated PHVM on two generation tasks. The first task is the new advertising text generation task which is to generate a long advertising text that covers all given attribute-value pairs for a piece of clothing. The second task is the recipe generation task from (Kiddon et al., 2016) which is to generate a correct recipe for the given recipe title and ingredient list.

Advertising Text Generation We constructed our dataset from a Chinese e-commerce platform. The dataset consists of 119K pairs of advertising text and clothing specification table. Each table is a set of attribute-value pairs describing a piece of clothing. We made some modifications to the original specification tables. Specifically, if some attribute-value pairs from a table do not occur in the corresponding text, the pairs are removed from the table. We also recognized attribute values by string matching with a dictionary of attribute values. If a pair occurs in the text but not in the table, the pair is added to the table.

The statistics are shown in Table 1 and Table 2.

| Category | Tops | Dress / Skirt | Pants |
|----------|------|---------------|-------|
| # Type   | 22   | 23            | 9     |
| # Attr.  | 13   | 16            | 11    |
| # Val.   | 264  | 284           | 203   |
| Avg. # Input Pairs | 7.7 | 7.7 | 6.6 |
| Avg. Len. | 110 | 111 | 108 |
| # Instances | 48K | 47K | 24K |

Table 1: Detailed statistics of our dataset. # Attr. / # Val.: the total number of attributes / attribute values.

Recipe Text Generation We used the same train-validation-test split (82,590 / 1,000 / 1,000) and pre-processing from (Kiddon et al., 2016). In the training set, the average recipe length is 102 tokens, and the vocabulary size of recipe title / text is 3,793 / 14,103 respectively. The recipe dataset covers a wide variety of recipe types indicated by the vocabulary size of recipe title.

5.2 Baselines

We compared our model with four strong baselines where the former two do not perform planning and the latter two do:

Checklist: This model utilizes an attention-based checklist mechanism to record what input data has been mentioned, and focuses more on what has not during generation (Kiddon et al., 2016).

CVAE: The CVAE model proposed by Zhao et al. (2017) uses a latent variable to capture the diversity of responses in dialogue generation. We adapted it to our task by replacing the hierarchical encoder with a one-layer bidirectional GRU.

Pointer-S2S: A two-stage method (Puduppully et al., 2019) that decides the order of input data with Pointer Network (Vinyals et al., 2015) before generation with Sequence-to-Sequence (Seq2Seq) (Bahdanau et al., 2015).

3We presented a detailed comparison with other benchmark corpora in Appendix A.2.
| Models               | BLEU (%) | Coverage (%) | Length   | Distinct-4 (%) | Repetition-4 (%) |
|---------------------|----------|--------------|----------|----------------|-----------------|
| Checklist           | 4.17     | 84.52**      | 83.61**  | 21.95**        | 46.40**         |
| CVAE                | 4.02     | 77.65**      | 80.96**  | 41.69**        | 36.58**         |
| Pointer-S2S         | 3.88**   | 85.97**      | 74.88**  | 18.16**        | 36.78**         |
| Link-S2S            | 3.90**   | 70.49**      | 95.65    | 16.64**        | 59.83**         |
| PHVM (ours)         | 2.85**   | 87.05        | 89.20**  | 72.87          | 3.90            |
| w/o $z^p$           | 3.07**   | 84.74**      | 91.97**  | 70.51**        | 4.19            |
| w/o $z^s$           | 3.38**   | 84.89**      | 75.28**  | 42.32**        | 20.88**         |

Table 3: Automatic evaluation for advertising text generation. We applied bootstrap resampling (Koehn, 2004) for significance test. Scores that are significantly worse than the best results (in bold) are marked with ** for p-value < 0.01.

**Link-S2S** Link-S2S (Sha et al., 2018) is a Seq2Seq with link-based attention mechanism where a link matrix parameterizes the probability of describing one type of input item after another.

### 5.3 Implementation Details

For both advertising text generation and recipe text generation, the settings of our model have many in common. The dimension of word embedding is 300. All embeddings were randomly initialized. We utilized GRU for all RNNs. All RNNs, except the input encoder, the plan decoder, and the plan encoder, have a hidden size of 300. The global planning latent variable and local latent variables have 200 dimensions. We set batch size to 32 and trained our model using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001 and gradient clipping threshold at 5. We selected the best model in terms of $L_1 + L_2$ on the validation set.

As we need to train the plan decoder with full supervision, we extracted plans from the texts by recognizing attribute values (or ingredients) in each sentence with string match heuristics. Some sentences do not mention any input items; we associated these sentences with a special tag, which is treated as a special input item for Pointer-S2S, Link-S2S, and our model. Although our extraction method can introduce errors, the extracted plans are sufficient to train a good plan decoder.\(^4\)

**Advertising Text Generation** We embedded an attribute-value pair by concatenating the embedding of attribute and the embedding of attribute value. Embedding dimensions for attribute and attribute value are 30 and 100 respectively. The input encoder, the plan decoder, and the plan encoder all have a hidden size of 100.

\(^4\)Our corpus and code are available at https://github.com/ZhihongShao/Planning-based-Hierarchical-Variational-Model.git.

### 5.4 Automatic Evaluation Metrics

We adopted the following automatic metrics. (1) **Corpus BLEU**: BLEU-4 (Papineni et al., 2002). (2) **Coverage**: This metric measures the average proportion of input items that are covered by a generated text. We recognized attribute values (ingredients) with string match heuristics. For the advertising text generation task, synonyms were also considered. (3) **Length**: The average length of the generated texts. (4) **Distinct-4**: Distinct-n (Li et al., 2016) is a common metric for diversity which measures the ratio of distinct n-grams in generated tokens. We adopted distinct-4. (5) **Repetition-4**: This metric measures redundancy with the percentage of generated texts that repeat at least one 4-gram.

### 5.5 Advertising Text Generation

**5.5.1 Automatic Evaluation**

Table 3 shows the experimental results. As our dataset possesses high expression diversity, there are many potential expressions for the same content, which leads to the low BLEU scores of all models. Our model outperforms the baselines in terms of coverage, indicating that it learns to arrange more input items in a long text. With content ordering, Pointer-S2S outperforms both Checklist and CVAE in coverage. By contrast, our planning mechanism is even more effective in controlling generation: each sentence generation sub-task is specific and focused, and manages to cover...
Table 4: Manual pair-wise evaluation for advertising text generation. We conducted Sign Test for significance test. Scores marked with * mean p-value < 0.05 and ** for p-value < 0.01. \( \kappa \) denotes Fleiss’ kappa, all indicating moderate agreement.

| Models              | Grammaticality | Coherence |
|---------------------|----------------|-----------|
|                     | Win (%)        | Lose (%)  | Tie (%)  | Win (%) | Lose (%) | Tie (%) |
| PHVM vs. Checklist  | 59.0**         | 23.5      | 17.5     | 0.484   | 54.5*    | 42.5    | 3.0     | 0.425   |
| PHVM vs. CVAE       | 69.5**         | 13.5      | 17.0     | 0.534   | 60.0**   | 37.0    | 3.0     | 0.426   |
| PHVM vs. Pointer-S2S| 76.5**         | 17.0      | 6.5      | 0.544   | 56.5**   | 39.0    | 4.5     | 0.414   |
| PHVM vs. Link-S2S   | 66.0**         | 28.5      | 5.5      | 0.462   | 62.5**   | 31.5    | 6.0     | 0.415   |

95.16% of the corresponding group on average. Noticeably, Link-S2S also models planning but has the lowest coverage, possibly because a static link matrix is unable to model flexible content arrangement in long text generation. As for diversity, our model has substantially lower repetition-4 and higher distinct-4, indicating that our generated texts are much less redundant and more diversified. Notably, Link-S2S has the longest texts but with the highest repetition-4, which produces many redundant expressions.

To investigate the influence of each component in the hierarchical latent structure, we conducted ablation tests which removed either global latent variable \( z_p \) or local latent variables \( z_s \). As observed, removing \( z_p \) leads to significantly lower distinct-4, indicating that \( z_p \) contributes to expression diversity. The lower coverage is because the percentage of input items covered by a planning result decreases from 98.4% to 94.4% on average, which indicates that \( z_p \) encodes useful information for planning completeness. When removing \( z_s \), distinct-4 drops substantially, as the model tends to generate shorter and more common sentences. This indicates that \( z_s \) contributes more to capturing variations of texts. The significantly higher repetition-4 is because removing \( z_s \) weakens the dependencies among sentences so that the word decoder is less aware of the preceding generated context. The lower coverage is because each generated sentence covers less planned items (from 95.16% to 93.07% on average), indicating that \( z_s \) keeps sentence \( s_t \) more focused on its group.

### 5.5.2 Manual Evaluation

To better evaluate the quality of the generated texts, we conducted pair-wise comparisons manually. Each model generates texts for 200 randomly sampled inputs from the test set. We hired five annotators to give preference (win, lose or tie) to each pair of texts (ours vs. a baseline, 800 pairs in total).

**Metrics** Two metrics were independently evaluated during annotation: **grammaticality** which measures whether a text is fluent and grammatical, and **coherence** which measures whether a text is closely relevant to input, logically coherent, and well-organized.

**Results** The annotation results in Table 4 show that our model significantly outperforms baselines in both metrics. Our model produces more logically coherent and well-organized texts, which indicates the effectiveness of the planning mechanism. It is also worth noting that our model performs better in terms of grammaticality. The reason is that long text generation is decomposed into sentence generation sub-tasks which are easier to control, and our model captures inter-sentence dependencies through modeling the dependencies among local latent variables.

### 5.5.3 Diversity of Planning

![Figure 3: Average number of distinct planning results (left) / average score of generation quality (right) when the number of input pairs varies.](image)

To evaluate how well our model can capture the diversity of planning, we conducted another manual evaluation. We randomly sampled 100 test inputs and generated 10 texts for each input by re-
Table 5: Automatic evaluation for recipe text generation. Checklist was trained with its own source code. We also re-printed results from (Kiddon et al., 2016) (i.e., Checklist §). We applied bootstrap resampling (Koehn, 2004) for significance test. Scores that are significantly worse than the best results (in bold) are marked with * for p-value < 0.05 or ** for p-value < 0.01.

| Models       | BLEU (%) | Coverage (%) | Length | Distinct-4 (%) | Repetition-4 (%) |
|--------------|----------|--------------|--------|----------------|-----------------|
| Checklist §  | 3.0      | 67.9         | N/A    | N/A            | N/A             |
| Checklist    | 2.6**    | 66.9*        | 67.59  | 30.67**        | 39.1**          |
| CVAE         | 4.6      | 63.0**       | 57.49**| 52.53**        | 38.7**          |
| Pointer-S2S  | 4.3      | 70.4**       | 59.18**| 30.72**        | 36.4**          |
| Link-S2S     | 1.9**    | 53.8**       | 40.34**| 24.93**        | 31.6**          |
| PHVM (ours)  | 4.6      | 73.2         | 70.92  | 67.86          | 17.3**          |

5.6 Recipe Text Generation

Table 5 shows the experimental results. Our model outperforms baselines in terms of coverage and diversity; it manages to use more given ingredients and generates more diversified cooking steps. We also found that Checklist / Link-S2S produces the general phrase “all ingredients” in 14.9% / 24.5% of the generated recipes, while CVAE / Pointer-S2S / PHVM produce the phrase in 7.8% / 6.3% / 5.0% of recipes respectively. These results demonstrate that our model may generalize well to other data-to-text generation tasks.

6 Case Study

We present examples for both tasks in Appendix B.

7 Conclusion and Future Work

We present the Planning-based Hierarchical Variational Model (PHVM) for long and diverse text generation. A novel planning mechanism is proposed to better model input data and address the inter-sentence incoherence problem. PHVM also leverages a hierarchical latent structure to capture the diversity of reasonable planning and sentence realization. Experiments on two data-to-text corpora show that our model is more competitive to generate long and diverse texts than state-of-the-art baselines.

Our planning-based model may be inspiring to other long text generation tasks such as long text machine translation and story generation.

Acknowledgements

This work was supported by the National Science Foundation of China (Grant No. 61936010/61876096) and the National Key R&D Program of China (Grant No. 2018YFC0830200).
We would like to thank THUNUS NExT Joint-Lab for the support.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In International Conference on Learning Representations.

John A. Bateman. 1997. Enabling technology for multilingual natural language generation: the KPML development environment. Natural Language Engineering, 3(1):15–55.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016, pages 10–21.

Qibin Chen, Junyang Lin, Yichang Zhang, Hongxia Yang, Jingren Zhou, and Jie Tang. 2019. Towards knowledge-based personalized product description generation in e-commerce. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019., pages 3040–3050.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014, pages 103–111.

Hercules Dalianis and Eduard H. Hovy. 1993. Aggregation in natural language generation. In Trends in Natural Language Generation, An Artificial Intelligence Perspective, Fourth European Workshop, EWNLG ’93, Pisa, Italy, April 28-30, 1993, Selected Papers, pages 88–105.

Kees van Deemter, Mariët Theune, and Emiel Krahmer. 2005. Real versus template-based natural language generation: A false opposition? Computational Linguistics, 31(1):15–24.

Pablo A Duboue and Kathleen R McKeown. 2003. Statistical acquisition of content selection rules for natural language generation. In Proceedings of the 2003 conference on Empirical methods in natural language processing, pages 121–128. Association for Computational Linguistics.

Dominic Espinosa, Michael White, and Dennis Mehay. 2008. Hypertagging: Supertagging for surface realization with CCG. In ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA, pages 183–191.

Xiaocheng Feng, Ming Liu, Jiahao Liu, Bing Qin, Yibo Sun, and Ting Liu. 2018. Topic-to-essay generation with neural networks. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden., pages 4078–4084.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, pages 179–188.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. J. Artif. Intell. Res., 61:65–170.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pages 2672–2680.

Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. 2018. Long text generation via adversarial training with leaked information. In (McIlraith and Weinberger, 2018), pages 5141–5148.

Eduard H. Hovy. 1993. Automated discourse generation using discourse structure relations. Artif. Intell., 63(1-2):341–385.

Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. 2016. Globally coherent text generation with neural checklist models. In (Su et al., 2016), pages 329–339.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In International Conference on Learning Representations.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In EMNLP, pages 388–395. ACL.

Ioannis Konstas and Mirella Lapata. 2013. Inducing document plans for concept-to-text generation. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL, pages 1503–1514. ACL.

Karen Kukich. 1983. Design of a knowledge-based report generator. In 21st Annual Meeting of the Association for Computational Linguistics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, June 15-17, 1983., pages 145–150.
Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In (Su et al., 2016), pages 1203–1213.

Jiwei Li, Minh Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26–31, 2015, Beijing, China, Volume 1: Long Papers, pages 1106–1115.

Percy Liang, Michael I. Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In ACL/JulCNLP, pages 91–99. The Association for Computer Linguistics.

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2018. Table-to-text generation by structure-aware seq2seq learning. In (McIlraith and Weinberger, 2018), pages 4881–4888.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In ACL, pages 142–150. The Association for Computer Linguistics.

Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based recommendations on styles and substitutes. In SIGIR, pages 43–52. ACM.

Sheila A. McIlraith and Kilian Q. Weinberger, editors. 2018. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. AAAI Press.

Susan W McRoy, Songsak Channarukul, and Syed S Ali. 2003. An augmented template-based approach to text realization. Natural Language Engineering, 9(4):381–420.

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 NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics, San Diego California, USA, June 12-17, 2016, pages 110–119.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Preksha Nema, Shreyas Shetty, Parag Jain, Anirban Laha, Karthik Sankaranarayanan, and Mitesh M. Khapra. 2018. Generating descriptions from structured data using a bifocal attention mechanism and gated orthogonalization. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pages 1539–1550.

Jekaterina Novikova, Ondrej Dusek, 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, Saarbrücken, Germany, August 15-17, 2017, pages 201–206.

Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors. 2017. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017. 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 ACL, pages 311–318. ACL.

Yookoon Park, Jaemin Cho, and Gunhee Kim. 2018. A hierarchical latent structure for variational conversation modeling. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), pages 1792–1801.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. Natural Language Engineering, 3(1):57–87.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9,
Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, pages 654–664.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018, pages 1097–1100.
A Dataset

A.1 Dataset Statistics

| Categories | Attributes | Attribute Values               |
|------------|------------|--------------------------------|
| 上衣(Tops) | 衣样式(Type) | 衬衣(Shirt),卫衣(Sweater),西装(Suit) |
|            | 衣袖型(Sleeve) | 灯笼袖(Lantern Sleeve),喇叭袖(Flare Sleeve),蝙蝠袖(Batwing Sleeve) |
|            | 衣领型(Collar) | 连衣裙(One-piece Dress), 百褶裙(Pleated Skirt),金裙(Full Skirt) |
|            | 裤型(Pants) | 哈伦裤(Harem Pants),铅笔裤(Pencil Pants),背带裤(Overalls) |
|            | 裤腰型(Waist) | 高腰(High-rise),低腰(Low-rise),自然腰(Regular-rise) |
|            | 裤口(Leg) | 开叉(Split),毛边(Cut off),小脚(Ankle-tied) |

Figure 4: Samples of attributes and attribute values for tops, dress / skirt, and pants.

Figure 4 shows some samples of attributes and attribute values from our dataset for advertising text generation.

A.2 Comparisons with other Benchmark Datasets

Compared with other widely used data-to-text datasets (i.e., WEATHERGOV, WIKIBIO, E2E, WebBLG, and ROTOWIRE), our dataset is more suitable for long and diverse text generation. Table 6 presents the comparisons among the datasets. WEATHERGOV (Liang et al., 2009) defines a task to generate a weather forecast from weather statistics; it has a vocabulary size of less than 400, which indicates the simplicity of its language expressions. WIKIBIO (Lebret et al., 2016) defines a task that requires to generate a very short biography (about 26 words) from personal information. E2E (Novikova et al., 2017) and WebNLG (Gar-
Due to diverse yet reasonable planning, the two texts of our model exhibit different discourse structures. The first text adopts a general-to-specific discourse structure where the statement in the beginning (i.e., the elegance of the dress) is supported by the following descriptions of local features. It groups global features (i.e., color, material and length) from the input in the first sentence and realizes each of the remaining sentences with one local feature. The second text adopts a parallel structure which splits global features and arranges some of them in the middle. Despite the difference, the two texts exhibit a global pattern in the data. They both describe the dress from top to bottom (i.e., collar –> sleeve –> shape of the lower part), which verifies the effectiveness of content organization. Noticeably, the two texts show diverse wording, which exemplifies that our model captures the diversity of expressions.

**B.2 Recipe Text Generation**

Figure 6 shows the generated examples. Although the three models fail to cover all given ingredients, our model gives the most complete procedure for making a pumpkin pie which includes five steps: 1. beat eggs –> 2. blend with some other ingredients –> 3. pour into pie shell –> 4. bake –> 5. cool. Our model also gives the most specific and precise instructions for step 4 and step 5. By contrast, all baselines miss step 3 or step 5. Checklist produces the general phrase “combine all of the ingredients”. CVAE suffers from the redundancy and incoherence problems. Pointer-S2S mentions the most ingredients but misses the most important one “pumpkin”. Link-S2S misses “pumpkin” and generates incoherent expressions.
This dress with black lace, the minimalist round collar flatters the curve of your neck which is sexy and attractive. The three-quarter sleeves flatter the curves of your arms, making your arms look slender. A-line shape is figure flattering and fit for all kinds of figures. The design of lace cutting and stitching makes you look very sexy. This black dress, the minimalist design of the round collar and three-quarter sleeves reveal your slender arms. The design of lace cutting and stitching shows a looming presence of your white skin, which is sexy and attractive. The A-line shape is figure flattering and makes you look taller. The black dress with lace, the minimalist round collar flatters your face shape, which makes you look sexier and more attractive. The minimal round collar flatters your face shape, which makes your face look more delicate. This is a black dress with lace cutting and stitching, which has an aura of unique elegance. The round neckline not only shows off your sexy collarbones, but also flatters your face shape. The minimalist three-quarter sleeves betray your slender arms, radiating fashion. The A-line shape heightens your waistline visually, which is figure flattering. Its lightly slimming design flatters your figure, while keeping your body comfortable. The decoration of black lace cutting and stitching is sexy, fashionable and buzzworthy. The minimalist round collar, paired with the three-quarter sleeves, shows your elegant temperament. The dress’s shape is A-line, which accentuates your slender waist and endows you a beautiful look.
Title: Drunken apple/pumpkin pie

Ingredients:
- 2 eggs
- 3/4 c brown sugar; firmly packed
- 1 9-inch pie shell; unbaked
- 1 c cooked pumpkin - mashed and drained
- 1 c applesauce; thick & chunky
- 1 tb flour
- 1/2 ts salt
- 1 ts each cinnamon and ginger
- 1/4 ts nutmeg
- 1/8 ts each allspice and cloves
- 1/4 ts nutmeg
- 1/8 ts each allspice and cloves
- 1/4 ts nutmeg
- 1 c pecan halves
- 1/2 c half-and-half or vaporated milk
- 1 ts vanilla
- 1/2 c half-and-half or vaporated milk
- 1 ts vanilla
- 1 tb rum

Truth
In a bowl, beat together the eggs and sugar until light. Mix in the pumpkin, applesauce, flour, salt, cinnamon, ginger, nutmeg, allspice, cloves, half-and-half and vanilla; blend thoroughly. Pour into pie shell. Arrange pecan halves over top of filling. Bake 30 to 35 minutes longer, or until filling is firm and a knife inserted 1” from the edge comes out clean. Cool on a wire rack. At serving time, warm rum in a small container suitable for pouring. Light the rum with a match and pour immediately while flaming over the pie.

Checklist
Preheat oven to 350 degrees. Grease a 9" pie plate. In large bowl, combine all of the ingredients, mixing well. Pour into a greased 9x13" pan. Bake for 1 hour at 350 degrees for 1 hour. Cool. Store in airtight container.

CVAE
Beat the eggs with the sugar, flour, salt, cinnamon and nutmeg in a bowl. Make a well in the center of the mixture and set aside. Combine the pumpkin, flour, salt, cinnamon, nutmeg, and nutmeg and stir into the eggs. Add the half-and-half, flour, salt, and nutmeg, mixing well. Bake in a 350 f oven for 45 minutes, or until a knife inserted in the center comes out clean. Cool in the pan for 5 minutes, then remove from the pan.

Pointer-S2S
Beat eggs until light and fluffy. Add brown sugar, applesauce, flour, salt, nutmeg, cloves, allspice, and pecans. Pour into pie shell. Bake at 350 degrees for 45 minutes.

Link-S2S
Beat eggs until light and fluffy. Add sugar, eggs, and vanilla. Mix well. Add eggs, one at a time, beating well after each addition. Stir in the flour, baking powder, salt, and vanilla. Pour into a greased and floured 9" x 13" pan. Bake at 350 degrees for 45 minutes.

PHVM
In a large mixing bowl, beat together the eggs and brown sugar until light and fluffy. Stir in the pumpkin. Blend in the flour, salt and nutmeg or half-and-half. Stir in vanilla and pecans. Pour into prepared pie shell. Bake for 30-35 minutes or until toothpick inserted in the center comes out clean. Cool on racks.

Figure 6: Generated recipes from different models.