PLANET: Dynamic Content Planning in Autoregressive Transformers for Long-form Text Generation

Zhe Hu[1], Hou Pong Chan[2], Jiachen Liu[1], Xinyan Xiao[1], Hua Wu[1], and Lifu Huang[3]

[1]Baidu Inc [2]Faculty of Science and Technology, University of Macau [3]Virginia Tech
[1]{huzhe01,liujiachen,xiaoxinyan,wu_hua}@baidu.com
[2]hpchan@um.edu.mo, [3]lifuh@vt.edu

Abstract

Despite recent progress of pre-trained language models on generating fluent text, existing methods still suffer from incoherence problems in long-form text generation tasks that require proper content control and planning to form a coherent high-level logical flow. In this work, we propose PLANET, a novel generation framework leveraging autoregressive self-attention mechanism to conduct content planning and surface realization dynamically. To guide the generation of output sentences, our framework enriches the Transformer decoder with latent representations to maintain sentence-level semantic plans grounded by bag-of-words. Moreover, we introduce a new coherence-based contrastive learning objective to further improve the coherence of output. Extensive experiments are conducted on two challenging long-form text generation tasks including counter-argument generation and opinion article generation. Both automatic and human evaluations show that our method significantly outperforms strong baselines and generates more coherent texts with richer contents.

1 Introduction

Neural sequence-to-sequence (seq2seq) models are dominant methods for text generation nowadays, which are trained to maximize the log-likelihood over targets in an end-to-end fashion (Cho et al., 2014). Recently, pre-trained methods such as GPT-2 (Radford et al., 2019) and BART (Lewis et al., 2020) have achieved promising results by leveraging large-scale data. While these models can generate fluent results, they still fall short of producing coherent long-form texts with multiple sentences (Dou et al., 2021).

Long text generation, especially opinion generation, usually requires the model to (1) conduct proper content selection and ordering (i.e., "*what to say and when to say it*") to form a coherent high-level logical flow, and (2) appropriately reflect the
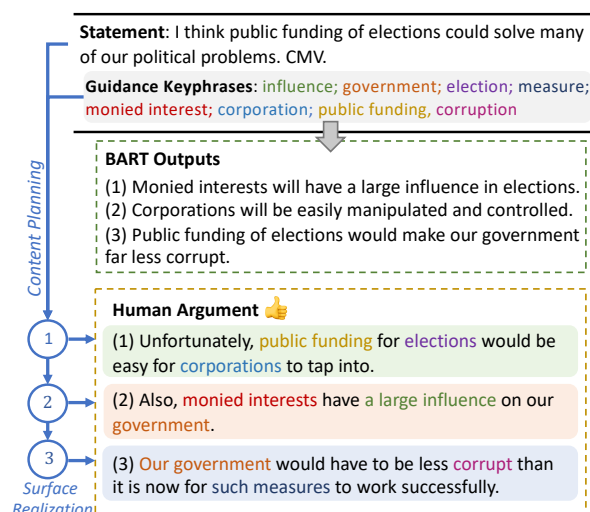

Figure 1: Sample counter-arguments on Reddit Change-MyView. Given a statement and a set of unordered keyphrases as guidance talking points, BART generates an incoherent output. In contrast, human writer conducts content planning and keyphrase selection for each sentence to form a coherent counter-argument.

text plans into final outputs (i.e.,"*how to say it*"). We present an example of counter-argument generation in Figure 1: given a statement on a controversial topic and a set of keyphrases as guidance talking points, the task aims to produce an argument with a different stance to refute the statement (Hua et al., 2019). Human writer assigns keyphrases for each sentence to form a coherent logical flow (e.g., "*corporations easily tap into public funding*" → "*they also have large influence on government*" → "*the current government is still corrupt*") and produces the final counter-argument that "*public funding won't solve the election problems*". In contrast, although BART learns to include keyphrases and generate an argument relevant to the statement, it suffers from incoherence issues such as incorrect usage of keyphrases (not "*corporations*" but '*election*" that "*be manipulated and controlled*") and wrong stance ("*public funding would make government less corrupt*"), and fails to maintain smooth transitions between sentences (e.g., sentence 2 and

3 are unrelated) and form a coherent text.

To solve the above defects, various text planning methods were proposed to improve the coherence of the generated text. The first type of methods (Kang and Hovy, 2020; Fu et al., 2020; Kong et al., 2021) leverage a latent variable as a global plan to guide the generation process, as illustrated in Figure 2 (a). However, these methods do not consider fine-grained sentence-level planning. The second line of methods (Hua and Wang, 2020; Goldfarb-Tarrant et al., 2020) first produce sentence-level content plans, and then pass content plans to a surface realization module to generate the output words, as shown in Figure 2 (b). Nevertheless, the planning and surface realization components are disjointed and may lead to cascading errors (Hua et al., 2021).

In this work, we propose **PLANET**, a novel text generation framework that dynamically performs content planning and surface realization in autoregressive Transformers. As shown in Figure 2 (c), for each target sentence, an autoregressive decoder first performs dynamic content planning by producing a latent representation (SN_j) as a semantic guidance, and then generates the sentence words. Both the content planning and surface realization are achieved dynamically by the autoregressive self-attention in a unified way: to generate a sentence (e.g., sentence 3), the latent representation (SN_3) attends the previous latent representations ($SN_{1,2}$, solid blue arrows) and previous context (sentence 1 and 2, dashed blue arrows) to plan its overall semantic content; Then, each output position in the sentence attends the corresponding latent representation (SN_3, solid green arrow) and the previous words (dashed green arrows), and optionally select keyphrases (gray arrow) to decide the exact wording. To supervise the latent representations, we further introduce a sentence-level bag-of-words prediction auxiliary task to provide supervision signals of the lexical semantics of the corresponding sentence. In this way, our framework can be trained end-to-end and easily applied to pre-trained autoregressive Transformers.

Furthermore, to empower our model to distinguish coherent and incoherent targets and generate more coherent outputs, we propose a novel coherence-based contrastive learning objective with different strategies to construct negative samples. We evaluate our model on two long-form opinion generation tasks: (1) counter-argument
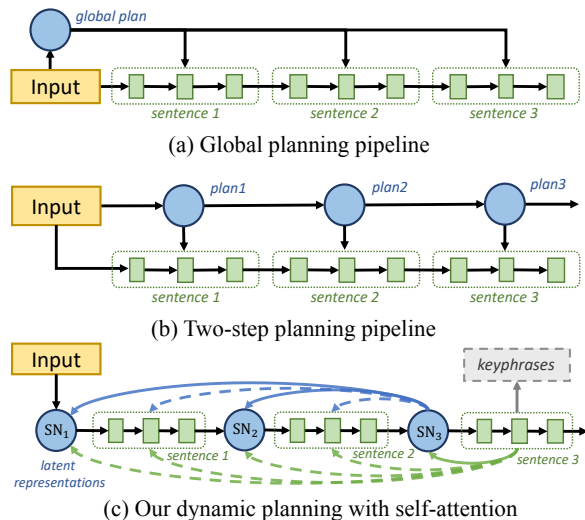

(a) Global planning pipeline

(b) Two-step planning pipeline

(c) Our dynamic planning with self-attention

Figure 2: Comparison of different content planning. For (c), the blue arrow denotes the attention flows for latent representations and the green one for target words. The attention of tokens within the same sentence is omitted. We highlight the attention flows related to the content planning with solid lines for sentence 3. Best viewed in color.

generation with Reddit/ChangeMyView dataset, and (2) opinion article generation from the New York Times Opinion corpus. Automatic evaluations show that our proposed method significantly outperforms strong baselines and generates more coherent texts with richer contents. Human evaluations further indicate that our model can properly leverage guidance keyphrases and generate better results on both datasets.

The overall contributions of our work are:

• A unified framework that dynamically conducts content planning and surface realization by leveraging the autoregressive self-attention, with a novel sentence-level bag-of-words auxiliary task to guide the semantic content of each sentence;

• A new coherence-based contrastive learning method with different negative sample construction strategies to improve the coherence of outputs;

• Our approach outperforms strong baselines for both automatic and human evaluations on two challenging long-form text generation tasks.

2 Related Work

Text Planning for Neural Generation. Traditional text generation pipeline leverages text planning component to decide on the high-level structures (McKeown, 1985; Reiter and Dale, 1997; Hovy, 1990; Carenini and Moore, 2006). Earlier work incorporates text planning into neural seq2seq structures by introducing hierarchical decoders (Yao et al., 2019; Moryossef et al., 2019;

Shen et al., 2019). However, these methods are hard to be applied to pre-trained models because of the modifications of model architecture. Several studies design separate modules for text planning and surface realization (Hua and Wang, 2020; Tan et al., 2021; Goldfarb-Tarrant et al., 2020), which lead to a disconnection of the two components and often produce undesired outputs (Castro Ferreira et al., 2019). Recently, Rashkin et al. (2020) present a memory-based model to keep track of the content usage and generate paragraphs recurrently. Nevertheless, they do not consider sentence-level text planning which is critical to maintain high-level logical flow for opinion text generation. Hua et al. (2021) propose a mixed language model to perform content selection and ordering. However, they encode multiple content items separately and do not fully consider the interactions among content items. In contrast to these prior studies, our model conducts sentence-level text planning and surface realization dynamically by introducing high-level latent representations for target sentences, and can be incorporated into pre-trained autoregressive Transformers.

Coherent Long-form Text Generation. Recent work tackles this problem on the tasks including story generation (Fan et al., 2019; Xu et al., 2020), paragraph completion (Kang and Hovy, 2020), text infilling (Huang et al., 2020), long-form conversation (Xu et al., 2021) and news article generation (Rashkin et al., 2020; Tan et al., 2021). To solve the incoherence issue, one type of work adopts the plan-then-generate strategy as discussed above. Some work also incorporates discourse and structured information into generation process to improve output coherence (Jiang et al., 2021; Ji and Huang, 2021; Bosselut et al., 2018). Recently, Guan et al. (2021) propose two auxiliary objectives of similarity prediction and order discrimination to improve coherence. In this work, we focus on long-form opinion text generation which requires an appropriate combination of credible talking points with rigorous reasoning (Hua et al., 2019), and apply dynamic content planning with a coherence-based contrastive objective to improve output coherence.

Controllable Text Generation. Our work is closely related to controllable generation (Prabhumoye et al., 2020). In this regard, typical studies manipulate sentiments (Hu et al., 2017), style (Gao et al., 2019; Du and Ji, 2021; Hu et al., 2021), syn-

tax (Chen et al., 2019), and keywords (Keskar et al., 2019; He et al., 2020; Wu et al., 2020) to steer the generation process. We use topical keyphrases as guidance talking points and require the model to properly organize and reflect keyphrases for long-form opinion text generation.

3 Our PLANET Framework

3.1 Framework Overview

Task Description. We follow the previous work (Hua and Wang, 2020) and model the long-form opinion generation task by considering the input of (1) a statement x which can be a proposition for argument generation or a title for opinion-article generation, and (2) a set of unordered keyphrases $m = \{m_i\}$ related to the statement, serving as topical guidance signal. The output y is an opinion text consisting of multiple sentences and properly reflects the keyphrases in a coherent way.

Our framework is based on the seq2seq structure, and we adopt BART (Lewis et al., 2020) as the base model. [1] The overall framework is shown in Figure 3. The bi-directional encoder first encodes the statement and keyphrases, and the decoder then generates the output in an autoregressive manner:

$$\hat{y} = \operatorname{argmax} \prod_{t=1}^{n} P(y_t|y_{1:t-1}, x, m), \quad (1)$$

where n is the number of target words. The statement and keyphrases are concatenated, with a segmenter inserted between adjacent keyphrases to indicate the keyphrase boundary.

We conduct content planning and surface realization dynamically by leveraging the autoregressive self-attention mechanism. For each target sentence, we introduce a latent representation SN to represent its global semantic information and guide surface realization (§ 3.2), then the sentence words attend the latent representation and dynamically select keyphrases (§ 3.3). After that, a sentence-level bag-of-words planning is introduced to enhance the latent representations (§ 3.4). Finally, we devise a contrastive learning (CL) objective to further improve the coherence of the output text (§ 3.5).

3.2 Latent Representation Learning

We introduce a latent representation for each target sentence to represent the overall semantic information and guide the generation of the sentence words.

---

[1] Our method can be also applied to other autoregressive pre-trained language models.
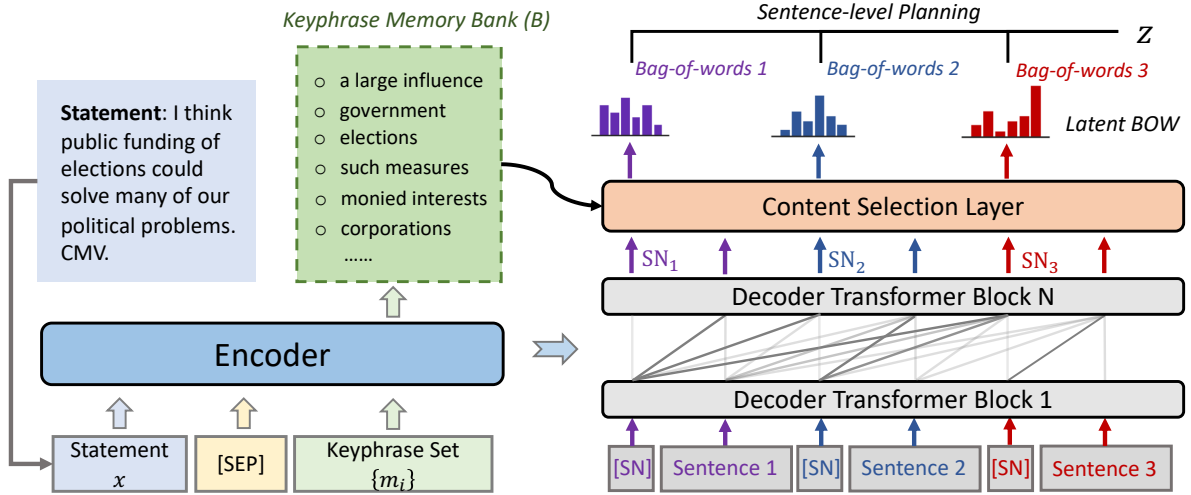
Figure 3: Overview of our framework. The encoder takes as input a statement and a set of keyphrases, and generates a keyphrase memory bank $\mathcal{B}$. The decoder conducts content planning and surface realization dynamically by the autoregressive self-attention to produce a coherent output. Meanwhile, the latent representations (SN) predict bag-of-words as global semantic plans and guide the surface realization of each target sentence. We highlight attention flows related to the content planning.
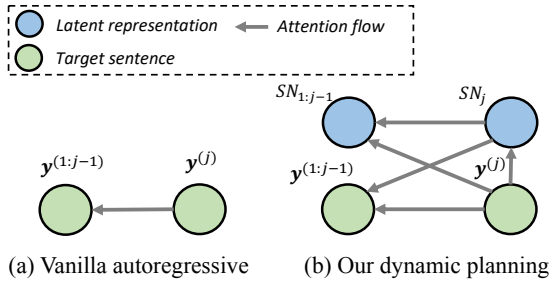

Figure 4: Attention flow of our dynamic planning and surface realization. $\boldsymbol{y}^{(j)}$ represents the words of the j-th sentence.

In particular, we insert a special token [SN] before every target sentence, and regard the hidden states of the decoder at the positions corresponding to [SN] as the latent representations of the target sentences. This has been shown effective by previous work (Guan et al., 2021; Li et al., 2021).

The workflow of our dynamic planning and realization is shown in Figure 4. For the vanilla autoregressive decoder, the generation of each token only depends on the previously generated tokens. In our framework, when producing the j-th output sentence $\boldsymbol{y}^{(j)}$, the latent representation SN_j is first obtained by attending the previous latent representations $\text{SN}_{1:j-1}$ and words in previous sentences $\boldsymbol{y}^{(1:j-1)}$. Then for sentence-level surface realization, each token in the current sentence $\boldsymbol{y}^{(j)}$ attends the previously generated words and latent representations $\text{SN}_{1:j-1}$, as well as the current latent representation SN_j as the guidance. A unique advantage of such modeling is that *the content planning and surface realization can be performed simul-*

taneously and incorporated into any pre-trained autoregressive language models, further optimized in an end-to-end fashion.

3.3 Content Selection

Based on the guidance of latent representations, each sentence word conducts content selection by incorporating keyphrases into decoder hidden states to decide which keyphrases to be reflected during generation. We first feed the keyphrases to the encoder to obtain hidden representations. We then construct a keyphrase memory bank $\mathcal{B}$ by gathering the top layer representations of the segment tokens (each keyphrase is represented by the segment token before it). After that, a content selection layer retrieves keyphrase information from the keyphrase bank and integrates the selected information into the decoding process.

Content Selection Layer. At each decoding step t, the top layer representation of the Transformer decoder $\boldsymbol{h}_t$ attends the keyphrase memory bank via multi-head attention:

$$\boldsymbol{c_t} = \text{MH-ATTENTION}(\boldsymbol{h}_t, \mathcal{B}, \mathcal{B}), \qquad (2)$$

where $\boldsymbol{c_t}$ is a context vector that embeds the selected keyphrase information, $\boldsymbol{h_t}$ is the query, and $\mathcal{B}$ acts as the key and value for multi-head attention. Then we incorporate the keyphrase context $\boldsymbol{c_t}$ into the decoder hidden state via a feed-forward layer followed by a residual connection (RC):

$$\boldsymbol{h}_t^d = \text{RC}(\boldsymbol{W}_s\text{tanh}(\boldsymbol{W}_h\boldsymbol{h}_t + \boldsymbol{W}_c\boldsymbol{c}_t + \boldsymbol{b}_s), \boldsymbol{h}_t). \quad (3)$$

Finally, the enhanced hidden state $\boldsymbol{h}_t^d$ will be passed to another feed-forward layer with softmax to estimate the probability of each output word:

$$P(y_t|\boldsymbol{y}_{1:t-1}) = \mathrm{softmax}(\boldsymbol{W}_o\boldsymbol{h}_t^d + \boldsymbol{b}_o), \quad (4)$$

where $\boldsymbol{W}_*$ and $\boldsymbol{b}_*$ are trainable parameters.

3.4 Sentence-level Bag-of-words Planning

We propose an auxiliary task of sentence-level bag-of-words (BOW) planning to supervise the latent representations. The goal is to ground the meaning of the latent representations with the bag-of-words (Fu et al., 2020) of target sentences to reflect the global semantic plans. Formally, we define the BOW of the j-th target sentence z_j as a categorical distribution over the entire vocabulary:

$$p(z_j|\mathrm{SN}_j) = \mathrm{softmax}(\mathrm{MLP}(\mathrm{SN}_j)), \quad (5)$$

where $\mathrm{MLP}(*)$ is parameterized as a multi-layer feed-forward network. We expect this distribution to capture the overall semantic plan of the corresponding sentence, and enhance SN to guide the surface realization of sentence words by conditioning the probability of each word on the latent representations: $p(y_t|\boldsymbol{y}_{1:t-1}, \mathrm{SN}_{1:s_{j_t}})$, where s_{j_t} denotes the sentence index of the token y_t. This conditional probability can be naturally satisfied by the autoregressive decoding process.

The loss of the task is to maximize the likelihood of predicting the BOW of each target sentence:

$$\mathcal{L}_{\mathrm{BOW}} = -\frac{1}{J}\sum_j\sum_l \log p(z_{jl}|\mathrm{SN}_j), \quad (6)$$

where J is the number of target sentence, and $p(z_{jl}|\mathrm{SN}_j)$ denotes the estimated probability of the l-th element in the bag of words for the j-th target sentence.

3.5 Coherence-based Contrastive Learning

We further design a contrastive learning (CL)-based training objective to enhance the content planning and drive our model to learn a preference of coherent outputs over incoherent ones.

Negative Sample Construction. One challenge for contrastive learning is how to construct negative samples to effectively train the model towards the desired goals. We consider the original target as a positive sample representing a logically coherent output with gold planning, and construct negative samples as incoherent ones. In particular,

for a positive target, we create 4 negative samples based on the following strategies: (1) *SHUFFLE*, where we randomly shuffle the target sentences to encourage the model to learn the correct sentence order; (2) *REPLACE*, where we randomly replace 50% of the original target sentences with random sentences from the corpus to facilitate the model to learn better content organization; (3) *DIFFERENT*, where we completely replace the original target sentences with a new set that are annotated as the target of a different input from the corpus; (4) *MASK*, where we randomly mask 20% of the non-stop target words that are related to any keyphrases from the keyphrase set, and adopt BART to fill the masked tokens since BART is naturally a denoising model. We enforce the filled negative target to be different from the original one.

Coherence-based Contrastive Loss. Since we aim to encourage the model to distinguish between coherent and incoherent targets and generate outputs with coherent logical flows, we design a novel coherence-based contrastive learning objective. Given a source-target pair, the model projects the output feature from the content selection layer to a coherence score between 0 and 1. Formally, for the i-th source-target pair, we enforce the score of the original target (r_i^+) to be larger than all corresponding negatives ($\{r_{ik}^-\}$) by a fixed margin ϕ:

$$\mathcal{L}_{\mathrm{CL}}(r_i^+, \{r_{ik}^-\}) = \sum_k \max(0, \phi + r_{ik}^- - r_i^+), \quad (7)$$

$$r_i^+ = \mathrm{F}(\mathrm{AvgPool}(\boldsymbol{W}_{cl}\boldsymbol{H}_i^{d+} + \boldsymbol{b}_{cl})), \quad (8)$$

$$r_{ik}^- = \mathrm{F}(\mathrm{AvgPool}(\boldsymbol{W}_{cl}\boldsymbol{H}_{ik}^{d-} + \boldsymbol{b}_{cl})), \quad (9)$$

where $\mathrm{F}(*)$ is a nonlinear transformation with sigmoid, $\boldsymbol{H}_i^{d+}$ and $\boldsymbol{H}_{ik}^{d-}$ are output features from the content selection layer for the positive and the k-th negative sample, and $\mathrm{AvgPool}(*)$ is the average pooling to compute a fixed-size vector. In this way, we expect the model to assign higher probability to the coherent target than incoherent ones.

3.6 Training Objective

We jointly optimize our model for content planning and surface realization by combining the objectives for the sentence-level BOW planning ($\mathcal{L}_{\mathrm{BOW}}$), the word-level generation by cross-entropy loss over the target tokens ($\mathcal{L}_{\mathrm{GEN}}$), and the contrastive learning loss ($\mathcal{L}_{\mathrm{CL}}$): $\mathcal{L} = \mathcal{L}_{\mathrm{GEN}} + \alpha\mathcal{L}_{\mathrm{BOW}} + \beta\mathcal{L}_{\mathrm{CL}}$, where α and β are tuned as hyper-parameters.

Dataset	Train	Val.	Test	|State|	|Target|	# KP
ArgGen	42.5k	6.5k	7.5k	19.4	116.6	20.6
OpinionGen	47.6k	5.0k	5.0k	9.0	198.2	16.2

Table 1: Statistics of the datasets. |State| and |Target| represent number of words of input statement and target, and #KP denotes the average number of guidance keyphrases.

4 Experimental Setups

4.1 Tasks and Datasets

We conduct experiments on two long-form opinion generation datasets of distinct domains: (1) Argument Generation (**ArgGen**) (Hua et al., 2019), where the model is required to generate a counter-argument to refute a given proposition; (2) Opinion Article Generation (**OpinionGen**) (Hua and Wang, 2020), to produce an opinion article given a title. The data statistics are shown in Table 1.

Argument Generation. We first apply data from Reddit *r/ChangeMyView* (CMV) for argument generation. We consider the original poster (OP) title as the statement, and the high-quality argument replies (with community endorsement) as the targets. Note that we consider the full argument replies as targets. The noun phrases and verb phrases that contain at least one topic signature word (Lin and Hovy, 2000) are extracted to form the guidance keyphrases.

Opinion Article Generation. For generating opinion articles, we consider samples from the New York Times (NYT) corpus (Sandhaus, 2008), with articles whose taxonomy labels include *Top/Opinion*. The articles with less than three sentences or more than 10 sentences are discarded. We further exclude articles containing more than 250 tokens considering the limited computing resources. 57,600 articles are randomly selected as the final dataset. We apply the same method as in argument generation to extract topical guidance keyphrases. The article title is regarded as the input statement.

4.2 Baselines and Comparisons

We compare our model against the following baselines : (1) **RETRIEVAL** (Stab et al., 2018) which retrieves targets based on TF-IDF weights of words from the training set. We keep the top-ranked results as outputs; (2) **HIERPLAN** (Hua et al., 2019) which is an end-to-end trained generation model with a hierarchical decoder to perform sentence-

level content planning and surface generation; (3) **FULLSEQ2SEQ** (Schiller et al., 2021) where we fine-tune BART with keyphrases concatenated to the input statements; (4) **SSPLANER** (Kang and Hovy, 2020) is a global planning method which first conducts content prediction and then guides the surface generation with the predicted contents; (5) **SEPPLAN** is a two-stage planning model similar to Hua and Wang (2020), where we first fine-tune a BART as the planner to generate the ordered keyphrase plans for each target sentence, and then fine-tune another BART as the generator to produce final outputs based on the statement and keyphrase plans. The details of SEPPLAN are in the Appendix A.2.

4.3 Training and Decoding Details

We use the BART-base version in all experiments for both our method and baselines. We truncate both input statement and output target to at most 256 tokens during training. For the BOW planning loss ($\mathcal{L}_{BOW}$), we consider the salient content words as the ground-truth bag of words for each target sentence. For the training objective, we set α as 0.2 for ArgGen and 0.3 for OpinionGen, and β as 0.2 based on the validation performance. The margin for contrastive loss is set as 0.5 for ArgGen and OpinionGen according to the validation performance. We optimize our model with AdamW (Loshchilov and Hutter, 2017). During the decoding time, we apply nucleus sampling (Holtzman et al., 2019) with a cumulative probability threshold of 0.9, and the maximum of generation steps are 150 for ArgGen and 200 OpinionGen. More training and decoding details are in the Appendix A.2.

5 Results and Analysis

5.1 Automatic Results

We first evaluate our model with BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Denkowski and Lavie, 2014). The results are shown in Table 2.

Our PLANET$_{w/o CL}$ model (without contrastive loss) consistently outperforms all baseline methods. In particular, compared with FULLSEQ2SEQ and SSPLANER which are also fine-tuned based on BART with the same inputs, the substantial improvements underscore the effectiveness of our dynamic content planning to generate better outputs. Meanwhile, the significant lead over HIER-PLAN indicates the importance of incorporating

System	ArgGen				OpinionGen			
	BLEU-2	**ROUGE-2**	**METEOR**	**Len.**	**BLEU-2**	**ROUGE-2**	**METEOR**	**Len.**
RETRIEVAL	10.95	4.02	20.70	113	18.16	6.98	24.87	153
HIERPLAN	14.29	8.38	19.03	115	10.66	5.84	17.50	107
FULLSEQ2SEQ	36.69	26.73	42.54	97	34.71	22.75	39.48	146
SEPPLAN	32.38	24.84	39.79	85	31.20	19.36	33.29	151
SSPLANER	36.92	26.82	42.72	105	35.04	22.55	39.50	140
PLANET$_{\text{w/o CL}}$	38.39	28.24*	44.22*	99	36.41	**23.82***	40.84*	145
− SEL.	37.66	27.71	43.76	96	35.91	23.38	40.33	142
− BOW	37.90	27.80	43.83	95	35.68	23.42	40.39	143
PLANET (ours)	**38.55***	**28.38***	**44.36***	100	**36.79***	23.65*	**40.91***	146

Table 2: Experimental results on argument generation (ArgGen) and opinion article generation (OpinionGen). PLANET$_{\text{w/o CL}}$ is our model variant without contrastive loss. We report BLEU-2, ROUGE-2 recall, METEOR and average output lengths (Len.). *: significantly better than all other methods without asterisks (Welch's t-test, $p < 0.05$).
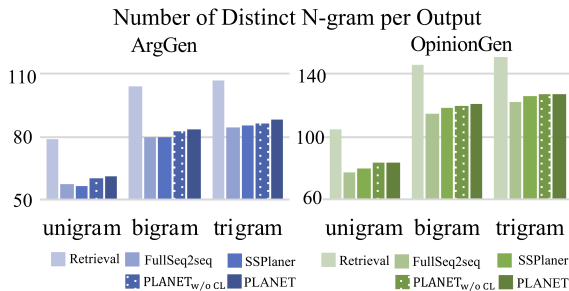

Figure 5: Average number of distinct n-grams per output.

Figure 6: Automatic evaluation on output coherence.

content planning into pre-trained language models. Furthermore, PLANET$_{\text{w/o CL}}$ significantly outperforms SEPPLAN, which confirms that the end-to-end training in our approach can mitigate the disconnection issue of the two-stage generation pipeline and produce superior results.

Among our model variants, removing content selection (w/o SEL.) and BOW planning (w/o BOW) both lead to performance decrease. This demonstrates the importance of the components that help the model conduct effective content planning. In addition, we observe that incorporating the contrastive loss (PLANET) brings performance gains on automatic results, especially with significant improvements on BLEU scores. This suggests that *our contrastive loss can guide the model to more precisely use keyphrases and reflect the keyphrase information in the outputs*. We provide further analysis on the keyphrase usage in Section 5.2.

Content Richness. To evaluate content richness, we employ Distinct n-gram (Li et al., 2016) that calculates the number of distinct n-grams per output in Figure 5. RETRIEVAL achieves the highest distinct results on both datasets since it returns top-ranked human-written texts with the most distinct words. Among generative methods, our dynamic
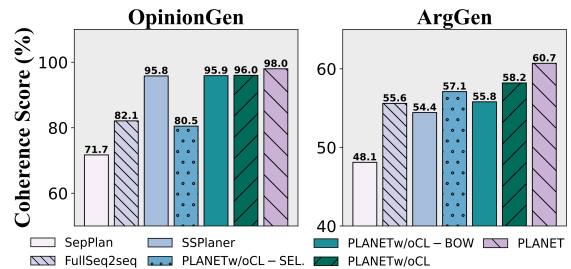
planning model PLANET$_{\text{w/o CL}}$ outperforms all baselines on both datasets. In addition, after applying contrastive loss, our PLANET model generates even more unique n-grams. The results imply our dynamic content planning and contrastive loss can enable the model to generate richer contents.

Automatic Evaluation on Coherence. We fine-tune BERT (Devlin et al., 2019) on each dataset to automatically evaluate the output coherence, which predicts a score between 0 and 1 for each output. The higher score indicates a more coherent output. The coherence model details are in Appendix A.3.

The results are shown in Figure 6. Among all methods, PLANET achieves the highest coherence scores on both datasets, suggesting that our dynamic planning and contrastive loss are effective to improve the coherence of outputs. In contrast, SEPPLAN has the lowest scores, indicating that decoupling planning and decoding stages may lead to cascading errors. Compared to FULLSEQ2SEQ and SSPLANER, our PLANET$_{\text{w/o CL}}$ model without contrastive loss also maintains better coherence, which confirms that incorporating dynamic content planning essentially promotes coherence for long text generation. Moreover, we observe that the results on OpinionGen are consistently better than

System	OpinionGen (%)	ArgGen (%)
PLANET	98.03	60.71
w/o SHUFFLE	96.20	59.30
w/o REPLACE	96.02	58.41
w/o DIFFERENT	96.11	59.95
w/o MASK	96.16	59.58

Table 3: Coherence scores for different negative strategies.

Task	Model	Rel.	Coh.	Rich.	Top-1
ArgGen	FULLSEQ2SEQ	2.25	2.47	2.57	20.7%
	PLANET$_{w/o\ CL}$	2.79	2.83	3.10	30.0%
	PLANET	**2.83**	**2.89**	**3.21**	**33.3%**
OpinionGen	FULLSEQ2SEQ	3.65	3.19	3.44	16.0%
	PLANET$_{w/o\ CL}$	3.81	3.27	3.64	28.7%
	PLANET	**3.89**	**3.47**	**3.81**	**37.3%**

Table 4: Human evaluation on relatedness (Rel.), coherence (Coh.), content richness (Rich.) and % of evaluations a model being ranked in top 1 based on the overall quality. All Krippendorff's $\alpha \geq 0.34$, with specific values in the Appendix B.
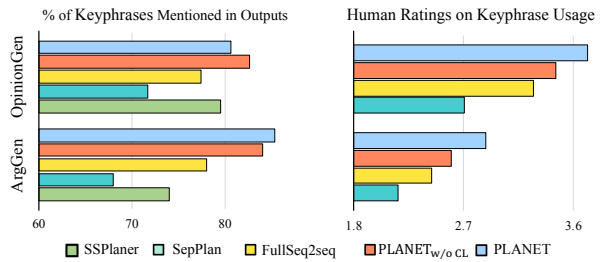

Figure 7: Left: percentage of keyphrases that are mentioned in outputs. Right: human evaluation on keyphrase usage.

those on the ArgGen dataset. A possible reason is that arguments in ArgGen are collected from social networks and contain more colloquial and informal expressions, making it harder to learn the implicit logical coherence. We leave this for future work.

Ablation on Contrastive Sample Construction. We study the contribution of each negative sample construction strategy for improving the coherence of the outputs. As in Table 3, removing each strategy leads to a performance degradation, indicating the effectiveness of all types of negative samples to enhance the contrastive learning. Among all negatives, removing *REPLACE* shows the most effects on both datasets. We hypothesize that replacing target sentences breaks the original logical flow and thus is more likely to encourage the model to focus on the global coherence. In contrast, *DIFFERENT* shows the least effects. One possible explanation is that this strategy focuses more on topical relatedness between the input and output, instead of the logical flow within the output as the negative sample itself is inherently coherent.

5.2 Human Evaluation

We hire three proficient English speakers as human judges to evaluate model outputs on a scale of 1 (worst) to 5 (best) for: (1) **topic relatedness** which measures whether the output is relevant and consistent to the input; (2) **coherence** which measures the high-level logical flow and transition among sentences; and (3) **content richness**, measuring the amount of informative talking points and specific details. We also ask judges to select top-ranked results based on the overall quality, and ties are allowed. 50 random samples are selected from each task. The detailed guidelines of human evaluations are provided in the Appendix B.

The results are shown in Table 4. Both our model variants achieve better results than FULLSEQ2SEQ on all aspects, underscoring the effectiveness of our dynamic planning to promote output coherence. Moreover, introducing contrastive objective further improves output quality on the above aspects, and

the outputs are more likely to be top-ranked. Overall, the human results verify the capability of our dynamic planning and contrastive objective to generate high-quality long-form texts.

Appropriateness of Keyphrase Usage. We further study how keyphrases are utilized in outputs. We first compute the percentage of keyphrases mentioned in outputs, as in the left of Figure 7. Among all models, SEPPLAN uses the least keyphrases in final outputs. However, its intermediate planning results cover more than 95% of keyphrases. This confirms that the two-stage method results in a disconnection problem between the planning module and the surface realization module, and the outputs are not guaranteed to reflect the plans. Compared to FULLSEQ2SEQ and SSPLANER, our methods cover more keyphrases, suggesting that our dynamic planning and keyphrase selection are useful to help the model better incorporate the guidance signal into outputs.

We further select 50 random samples for both tasks and ask the same human judges to score the outputs from 1 (worst) to 5 (best) on the correctness of keyphrase usage: whether the model uses keyphrases adequately as main talking points when generating outputs. Results in Figure 7 (right) indicate that our models tend to use more keyphrases and properly organize them in the out-

Statement: Absolute free speech doesn't exist - real debate is about where to draw the line
PLANET: ① Absolute free speech exists in a **conceptual**, **legal**, and **practical** way. ② The **simple fact** that it exists is <u>obvious</u>. ③ If you're going to say something in **a legal way**, at any time, the **consequences** should be the consequences. ④ It's not that absolute <u>free</u> speech doesn't exist, but the **conceptual way** you could interpret it is much **more practical**. ⑤ **The** <u>internet</u> has made it much more obvious that you could use it any way you wanted.
BOW: ① free, speech, concept ② simple, obvious ③ consequences, legal, illegal ④ freedom, case, practical ⑤ internet, easy
FULLSEQ2SEQ: Absolute free speech exists in a conceptual (and probably legal) way. There is the simple fact that absolute free speech exists in a legal way. It's obvious what the consequences are for someone, but it can be done any time. In many cases, it's very practical to do something like this in a legal way because the internet makes it very obvious.
Statement: Millions of mothers can't choose not to work
PLANET: ① **Single** <u>mothers</u> have to decide how to raise their <u>children</u>. ② But the <u>choice</u> <u>mothers</u> have has often been made to <u>work</u>, **juggling financial responsibilities** that make it all the more difficult. ③ While it is true that many **families** have **no such choice**, the **reality is the same**: **single mothers** have **little options**. ④ And while it is true that **a family of two or more lowers pay** and is likely to take many times the **cost of similar** <u>work</u>, the reality is that it may not be that simple (...)
BOW: ① child, parent, children ② work, choice, mother ③ choice, family, mother ④ work, pay, children, family
FULLSEQ2SEQ: Crittenden is right about single mothers' choice to choose not to work, in her book "the choice mothers make" But the sad reality of working families is that it is the reality that Ms. Crittenden and many others, in juggling financial responsibilities, are forced to choose not to work. If they are lucky enough to be able to keep their jobs, they can be at similar work as nannies. But the sad reality is that the choice mothers make is no longer one wage earner (...)

Figure 8: Sample outputs on ArgGen (Upper) and Opinion-Gen (Lower). For our model results, the phrases relevant to the guidance keyphrases are highlighted in colors, and the words related to the corresponding BOW are underlined. Best viewed in color.

puts compared to all baseline methods. Although on OpinionGen our contrastive model mentions fewer keyphrases, human judges rate it with higher scores for keyphrase usage. We speculate that this can be attribute to the *MASK* strategy for negative sample construction in contrastive learning, which helps to improve the model ability on the appropriate usage of keyphrases. The above results confirm that PLANET can properly utilize the keyphrases and reflect the contents in the outputs.

5.3 Sample Outputs and Discussions

We show two sample outputs on both tasks and highlight the phrases relevant to the guidance

keyphrases in Figure 8. We can see that on both tasks, our model effectively leverages guidance keyphrases as main talking points, and properly organizes and reuses the keyphrases to form a coherent output. In contrast, FULLSEQ2SEQ suffers from incoherence issues such as repetition (e.g., the first and second argument sentences) and inconsistent stance (e.g., "*choose not to work*" in generated opinion article). This indicates that our dynamic planning is effective to guide the model to better leverage keyphrases in the outputs.

We also present the predicted BOW of our model for each generated sentence. As can be seen, our model predicts most of the salient content words of the target sentences and effectively reflects the semantic plans in the generated sentences, suggesting that our latent representations are useful to capture the global semantic information of each sentence and conduct content planning during the generation process. However, there is still a large gap compared with human written texts, inspiring the future work on long-form text generation. More sample outputs are provided in Appendix D.

6 Conclusion

We present a novel generation framework to dynamically conduct content planning and surface realization in large autoregressive Transformers by leveraging self-attention and high-level latent representations. The latent representations are grounded by bag-of-words that measures the overall semantic plan of each target sentence. We further introduce a novel coherence-based contrastive objective with different negative sample construction strategies to improve output coherence. Experiment results on two opinion text generation tasks demonstrate that our model can generate high-quality outputs with better coherence and content richness.

Acknowledgements

Ethics Statement

We recognize that our method may generate fabricated and potentially harmful contents due to the systematic biases of pre-training using heterogeneous web corpora and the open-ended generation characteristics of the opinion generation tasks. Therefore, we urge the users to carefully examine the ethical influence of the generated outputs and cautiously apply the system in real-world applications.

References

Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. Discourse-aware neural rewards for coherent text generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 173–184, New Orleans, Louisiana. Association for Computational Linguistics.

Giuseppe Carenini and Johanna D. Moore. 2006. Generating and evaluating evaluative arguments. *Artificial Intelligence*, 170(11):925–952.

Thiago Castro Ferreira, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. 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 552–562, Hong Kong, China. Association for Computational Linguistics.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. Controllable paraphrase generation with a syntactic exemplar. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5972–5984, Florence, Italy. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A Smith, and Yejin Choi. 2021. Scarecrow: A framework for scrutinizing machine text. *arXiv preprint arXiv:2107.01294*.

Wanyu Du and Yangfeng Ji. 2021. SideControl: Controlled open-domain dialogue generation via additive side networks. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2175–2194, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660, Florence, Italy. Association for Computational Linguistics.

Yao Fu, Yansong Feng, and John P Cunningham. 2020. Paraphrase generation with latent bag of words. *arXiv preprint arXiv:2001.01941*.

Xiang Gao, Yizhe Zhang, Sungjin Lee, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2019. Structuring latent spaces for stylized response generation. 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 1814–1823, Hong Kong, China. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics.

Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. 2021. Long text generation by modeling sentence-level and discourse-level coherence. 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 6379–6393, Online. Association for Computational Linguistics.

Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization. *arXiv preprint arXiv:2012.04281*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Eduard H Hovy. 1990. Pragmatics and natural language generation. *Artificial Intelligence*, 43(2):153–197.

Zhe Hu, Zhiwei Cao, Hou Pong Chan, Jiachen Liu, Xinyan Xiao, Jinsong Su, and Hua Wu. 2021. Controllable dialogue generation with disentangled multi-grained style specification and attribute consistency reward. *arXiv preprint arXiv:2109.06717*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596, International Convention Centre, Sydney, Australia. PMLR.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy. Association for Computational Linguistics.

Xinyu Hua, Ashwin Sreevatsa, and Lu Wang. 2021. DYPLOC: Dynamic planning of content using mixed language models for 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 6408–6423, Online. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2020. PAIR: Planning and iterative refinement in pre-trained transformers for long text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 781–793, Online. Association for Computational Linguistics.

Yichen Huang, Yizhe Zhang, Oussama Elachqar, and Yu Cheng. 2020. INSET: Sentence infilling with INter-SEntential transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2502–2515, Online. Association for Computational Linguistics.

Haozhe Ji and Minlie Huang. 2021. DiscoDVT: Generating long text with discourse-aware discrete variational transformer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4224, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yichen Jiang, Asli Celikyilmaz, Paul Smolensky, Paul Soulos, Sudha Rao, Hamid Palangi, Roland Fernandez, Caitlin Smith, Mohit Bansal, and Jianfeng Gao. 2021. Enriching transformers with structured tensor-product representations for abstractive summarization. In *Proceedings of the 2021 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4780–4793, Online. Association for Computational Linguistics.

Dongyeop Kang and Eduard Hovy. 2020. Plan ahead: Self-supervised text planning for paragraph completion task. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6533–6543, Online. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Xiangzhe Kong, Jialiang Huang, Ziquan Tung, Jian Guan, and Minlie Huang. 2021. Stylized story generation with style-guided planning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2430–2436, Online. Association for Computational Linguistics.

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.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, pages 110–119.

Jiwei Li and Dan Jurafsky. 2017. Neural net models of open-domain discourse coherence. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 198–209, Copenhagen, Denmark. Association for Computational Linguistics.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. 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 128–138, Online. 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.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Kathleen R McKeown. 1985. Discourse strategies for generating natural-language text. *Artificial intelligence*, 27(1):1–41.

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, Volume 1 (Long and Short Papers)*, pages 2267–2277, Minneapolis, Minnesota. 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.

Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. 2020. Exploring controllable text generation techniques. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. PlotMachines: Outline-conditioned generation with dynamic plot state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4274–4295, Online. Association for Computational Linguistics.

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

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-controlled neural argument generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online. Association for Computational Linguistics.

Eva Sharma, Luyang Huang, Zhe Hu, and Lu Wang. 2019. An entity-driven framework for abstractive summarization. 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 3280–3291, Hong Kong, China. Association for Computational Linguistics.

Dinghan Shen, Asli Celikyilmaz, Yizhe Zhang, Liqun Chen, Xin Wang, Jianfeng Gao, and Lawrence Carin. 2019. Towards generating long and coherent text with multi-level latent variable models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2079–2089, Florence, Italy. Association for Computational Linguistics.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. ArgumenText: Searching for arguments in heterogeneous sources. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 21–25, New Orleans, Louisiana. Association for Computational Linguistics.

Bowen Tan, Zichao Yang, Maruan Al-Shedivat, Eric Xing, and Zhiting Hu. 2021. Progressive generation of long text with pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4313–4324, Online. Association for Computational Linguistics.

Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, et al. 2020. A controllable model of grounded response generation. *arXiv preprint arXiv:2005.00613*.

Linzi Xing and Giuseppe Carenini. 2021. Improving unsupervised dialogue topic segmentation with utterance-pair coherence scoring. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 167–177, Singapore and Online. Association for Computational Linguistics.

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2831–2845, Online. Association for Computational Linguistics.

Peng Xu, Hamidreza Saghir, Jin Sung Kang, Teng Long, Avishek Joey Bose, Yanshuai Cao, and Jackie Chi Kit

Cheung. 2019. A cross-domain transferable neural coherence model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 678–687, Florence, Italy. Association for Computational Linguistics.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

A Experiment Details

A.1 Additional Experimental Results

In table 2 we report automatic results on both tasks. Here we present additional automatic results of BLEU-3 and ROUGLE-L (recall) in Table 5 and Table 6.

A.2 Training and Decoding Details

Model Training. Our model is built based on BART, and we use BART-base version for all experiments. Our model contains 185M parameters in total. The batch size is set to be 8, and the maximum training epoch is set as 15 for non-contrastive training and 18 for contrastive training. We truncate both the input statement and output target to be at most 256 tokens during training. We resize the BART embedding matrix with a new token [SN] and insert a [SN] token before each target sentence. This is also done for baselines for a fair comparison. For computing resources, we use NVIDIA Tesla V100 GPUs with 32 GB memory for all experiments, and utilize the mixed-precision (FP16) to improve the computational efficiency. For contrastive learning, for each positive target, we construct 4 negatives using the strategies described in Section 3.5 respectively. The best model checkpoint is chosen based on the validation loss. Our model takes around 4-5 hours for training, and 30 minutes for decoding on V100 GPUs.

Decoding. During decoding time, we apply the nucleus sampling (Holtzman et al., 2019), and set $k = 10$ and $p = 0.9$. Considering the computational cost, we limit the maximum of generation steps to 150 for argument generation on ArgGen and 200 for opinion article generation on Opinion-Gen. To reduce variance introduced by sampling-based decoding method, we decode three times and average the results for automatic evaluations. For our model, we enforce each target sentence to start with a [SN] token during inference: we pre-define a list of sentence end markers, and when the model

System	BLEU-3 (%)	ROUGE-L (%)
RETRIEVAL	4.52	16.13
HIERPLAN	9.28	19.11
FULLSEQ2SEQ	25.83	26.88
SEPPLAN	22.17	23.24
SSPLANER	25.85	26.99
PLANET	27.11	27.42*
− SEL.	26.58	27.01
− BOW	26.78	26.97
PLANET (ours)	**27.21***	**27.54***

Table 5: Additional experimental results of BLEU-3 and ROUGE-L (recall) on ArgGen.

System	BLEU-3 (%)	ROUGE-L (%)
RETRIEVAL	10.98	17.99
HIERPLAN	5.81	15.98
FULLSEQ2SEQ	25.71	26.29
SEPPLAN	21.23	21.68
SSPLANER	25.67	26.49
PLANET	26.91	27.08*
− SEL.	26.49	26.79
− BOW	26.40	26.77
PLANET (ours)	**27.01***	**27.18***

Table 6: Additional experimental results of BLEU-3 and ROUGE-L (recall) on OpinionGen.

finishes generating a sentence, we enforce the next generated token to be [SN], although we find in most cases the model can automatically generate [SN]. The generation process stops when the model generates the <EOS> token. In this way, the model can automatically decide on how many sentences to be generated, and conduct content planning and surface realization in a dynamic way.

Evaluation Scripts. We use NLTK [2] to implement BLEU and METEOR, and the ROUGE_SCORE package [3] to implement ROUGE.

Details for SEPPLAN. We design a two-stage generation method, SEPPLAN, as a baseline model by fine-tuning two independent BART models for content planning and surface realization respectively, similar to Hua and Wang (2020). In particular, the planner BART takes a statement and unordered keyphrase as inputs, and autoregressively generates content plans as a sequence of tokens for every target sentence, where each content plan is represented by the ordered keyphrases with the same order as they appear in the corresponding sentence. Segmenter is added between sentence plans to indicate the sentence boundary. Then the generator

---

[2] https://www.nltk.org/

[3] https://pypi.org/project/rouge-score/

BART consumes the concatenation of the statement and content plans to produce the final results. During training, the ground-truth content plans are used to train the generator, and during inference the predicted plans are used. For decoding, we apply beam search for the planner and nucleus sampling for the generator. Note that Hua and Wang (2020) applies BERT as planner in their original paper, and we replace BERT with BART as BART gives better performance in our experiments.

A.3 Training Details for Coherence Model

We propose a neural coherence model to evaluate output coherence. Concretely, we fine-tune BERT (Devlin et al., 2019) on each dataset to compute the coherence scores. Instead of computing the overall coherence scores by measuring and aggregating the coherence of its adjacent sentence pairs (Xu et al., 2019), we fine-tune BERT on the whole text to better learn the global coherence (Xing and Carenini, 2021).

For training, we follow Sharma et al. (2019) and adopt hinge loss to teach the model to assign higher scores to coherent targets than incoherent ones. The score is normalized into [0, 1] with sigmoid function, and the margin is set to be 0.8. Since each target usually contains multiple sentences, we insert a separator token [SEP] between each adjacent sentence pair. For data construction, we consider the original text as a positive sample, and randomly shuffle sentences to construct negative ones. The test accuracy is 94.3% on OpinionGen and 73.0% on ArgGen, respectively. This implies that our coherence model can be used as a reliable metric to evaluate the output coherence.

B Details for Human Evaluation

We present 55 random samples on each task for human evaluation, and the first 5 samples are used only for calibration [4]. We anonymize the models and shuffle the outputs to the annotators. We evaluate model outputs on the following aspects, and the detailed guidelines are in Table 8:

- **Relatedness**: whether the output is relevant and consistent to the input;
- **Coherence:** whether the overall logical flow is appropriate and the transitions among sentences are natural and smooth;
- **Content Richness:** whether outputs contain substantial talking points and convey specific de-

---

[4]The payment for each human judge is 20 dollars per hour.

Task	Rel.	Coh.	Rich.	KP-Use.
ArgGen	0.49	0.34	0.40	0.44
OpinionGen	0.41	0.46	0.37	0.36

Table 7: Krippendorff's α for human evaluation on relatedness (Rel.), coherence (Coh.), content richness (Rich.) and keyphrase usage (KP-Use.).

tails;

- **Overall Ranking:** this is a general assessment that whether you think the output ranks top among all candidates. Ties are allowed, which means you can choose multiple outputs as top-ranking for a sample.

To measure agreement among human judges, we compute Krippendorff's α for each aspects. The values for all aspects on both datasets are presented in Table 7. As can be seen, all values are equal or larger than 0.34, indicating a general consensus among the judges.

C Discussions on Limitations and Future Directions

Here we discuss the limitations of our work and the potential directions for future studies. Long-form text generation is a challenging task which requires the model to properly select and organize contents, and faithfully reflect the plans in surface realization, in order to form a coherent output. The results suggest that our dynamic content planning can effectively leverage keyphrases and generate more coherent and richer texts than strong baseline methods. Nevertheless, there is still a gap compared with human written outputs. Also, in this paper we follow previous work to study the keyphrases guided generation (Hua and Wang, 2020; Rashkin et al., 2020), where we assume the availability of keyphrases as guidance signals. For the scenarios where guided keyphrases are not available in test time, one can use either retrieval-based methods (Hua et al., 2019; Wu et al., 2020) or a separate knowledge-enhanced generative module to obtain guided keyphrases. However, this is out of the scope of this work.

We believe there are several promising directions to explore in the future. First direction can be applying our dynamic planning method into pre-trainning or post-pretrainning stage. One advantage of our model is that it does not require additional annotated data (the keyphrases and BOW labels can be automatically constructed with off-the-shelf tools as described in data processing). Leveraging

massive pretraining data would be very helpful to further improve the model performance on long-text generation in various domains.

Second, one can study different supervision signals to train the latent representations. In this work we apply bag-of-words to ground the latent representations, which aims to capture the overall semantic information. Other supervision signals such as discourse structures and entity usage are also very important for modeling coherence. Considering these aspects into planning can further improve the output coherence. Meanwhile, coherence is a broad definition including topical relatedness, causal relationship, temporal ordering and discourse structures (Li and Jurafsky, 2017). Designing different supervision signals to tackle specific aspects for coherence would also be a promising direction.

Third, in this work we consider keyphrases as guidance signals to control the generation. Future work can incorporate different guidance signals from heterogeneous sources such as structured knowledge and commonsense information to further improve the output quality.

D Additional Sample Outputs

We present additional examples on argument generation in ArgGen and opinion article generation in OpinionGen from Figure 9 to Figure 12.

Relatedness

1	The output is very generic and irrelevant to the statement
3	The output is tangential to the statement and mentions some relevant concepts or entities, but in general is not precisely on topic
5	The output is highly relevant to the statement and fully on topic

Coherence

1	The sentences are completely disconnected and unrelated, without any overall logical structure
3	The output sentences are somehow connected or on the same topic, but the transition is not smooth, and the overall logical flow is not strong to convey a progression of idea
5	The output is completely coherent with natural and smooth transition among sentences, and the high-level logical flow is strong to form a good progression

Content Richness

1	The output is very generic with no specific information or talking points
3	The output contains several talking points or key information to form an opinion text
5	The output contains sufficient key points and informative contents to from a strong counter-argument or opinion article.

Table 8: Explanations on human evaluation for aspect scales.

Statement: CMV: journalism has become so biased in this country that it has lost its relevance.

PLANET: ① True journalism is **biased by definition**. ② You have to have **sources** in order to read the news, especially when **the facts and propaganda** are coming out. ③ You need to know **a piece** or two about their policies. ④ This includes reading from **a wide range of opinions**. ⑤ You can be **swayed by the emotion** of rage. ⑥ A lot of **great journalists** and pundits are coming out with their favor. ⑦ The AP is no actor. ⑧ BBC has been relentless in **their editorial and their propaganda**. ⑨ NPR has tried to **push conservative candidates** and push their favor in a public and this is reflected in their **editorial and news**.

> **BOW**: ① journal, biased, media ② news, fact, propaganda, media ③ news, media, ④ people, political, lot ⑤ emotion, rage, news ⑥ good, public, news, ⑦ media, main ⑧ BBC, news, ⑨ public, candidates, political, popular

- -

FULLSEQ2SEQ: The AP is biased by definition. They are biased because they are biased by their bias. A lot of true journalism is achieved by the emotion of outrage and propaganda, and no actor can objectively prove this. They are biased by their biases, because they want to push conservative candidates' policies and favorites to an audience that is more informed than they are themselves. So if you want a piece of true journalism from a public that doesn't want to read the news, then you need sources and sources to get the facts. You need great journalists to listen to your sources. BBC, NPR, the AP, and the like.

Figure 9: Sample outputs on ArgGen. For our results, we present the bag-of-words predicted for each sentence, and the phrases relevant to the guidance keyphrases are highlighted in bold.

Statement: CMV: the European Parliament should be elected on transnational lists

---

PLANET: ① A big issue with the EU is that **the formation of government** would **face a similar problem** as the **federalization** of the United States. ② **The US constitution** included **major guarantees** for autonomy, a **separate national voice** for all and a broader level of **governance**. ③ A lot of **Euro-skeptic countries** have been concerned about secession at the start of the EU, but the EU has been able to **reduce these concern**. ④ Euro-skeptic countries are more likely to **balance state autonomy**, and thus need a **slow-motion version** of government. ⑤ if the EU were to step down, it would be violating by the

> **BOW**: ① government, issue, EU, federalization ② democracy, constitution, governance, independence ③ EU, countries, states ④ government, federal, state ⑤ violated, EU, constitution, government

- -

FULLSEQ2SEQ: The United States was a major hot point in US politics from the start, and was a slow-motion version of the US constitution that included major guarantees of independence, and a separate national voice for all. Many countries now want to balance state autonomy with autonomy. This would mean that many would face a similar problem with the federalization of the US. But it would also reduce these concern to a broader level. I don't think we should be at war with the EU for the federalization of the US, but that is a big issue. It would be very different from creeping federalization of the US as a whole. I don't think that the United States would do well without a big issue like the

Figure 10: Sample outputs on ArgGen. For our results, we present the bag-of-words predicted for each sentence, and the phrases relevant to the guidance keyphrases are highlighted in bold.

Statement: Endangered species will gain from reform

PLANET: ① To the editor: I agree with your April 19 **column attack legislation** to reform **the endangered species act**. ② But it is important to focus on how it **improves the recovery** of endangered species. ③ In an April 28 editorial, **representative Richard Pombo** of California and 125 other democrats called on the federal government to **provide habitat** for the endangered species. ④ This would have been good for the species as well as the economy. ⑤ It would have been good for the species, as it would help the economy. ⑥ The endangered species act allows homeowners to **protect habitat** without **forcing private landowners** to pay taxes to protect it. ⑦ If landowners were allowed to sell off their **critical habitat**, they would pay for the maintenance of homes and homes. ⑧ And if owners paid taxes to protect habitat, it would help **attract these species** to their area. ⑨ David h. Lindman Washington, April 30, 1997 the writer is an environmental advocate.

    BOW: ① endangered, act, April, species ② species, endangered, protect, ③ protect, endangered, species, land, support ④ life, species, economy, ⑤ protect, species, park ⑥ owners, species, protection, bill ⑦ landowners, habitat, park, housing, ⑧ homes, build, land, ⑨ writer, address

- -

FULLSEQ2SEQ: To the editor: while i applaud your call to reform the endangered species act, your April 19 column attack legislation that would allow the states to force private landowners to provide habitat for endangered species. In an April 28 editorial, representative Richard Pombo of Texas and 125 other democrats wrote that "the species cannot be exploited to attract these species to this program." However, there are other ways to exploit these species: the endangered species act is a law requiring the state to provide habitat for endangered species and requiring the states to provide a plan to protect habitat for the species. If the endangered species act is enacted, it will be in effect, and will be a significant step toward conservation. The bill is a response to the plight of the endangered species act and will help improve its financing. Daniel s. Bennett New York, April 30, 1999 the writer is chairman of the house appropriations committee.

Figure 11: Sample outputs on OpinionGen. For our results, we present the bag-of-words predicted for each sentence, and the phrases relevant to the guidance keyphrases are highlighted in bold.

Statement: Protect inmates' rights

PLANET: ① To the editor: **Elisabeth Sara** (op-ed, Jan. 1) argues that **stripping inmates** of **civil legal assistance** would not result in a lasting peace or in freeing them from incarceration. ② Yet the Bush administration has not spent the last 10 years **exorcising the demons** of discrimination against federally **financed prisoners**. ③ Prisoners' **legitimate civil rights complaints** should be heard in court. ④ it is time for the administration to begin **stripping inmates** of their rights and **redeeming the debts** owed them by the government. ⑤ The government must make sure that they receive **adequate funds** for housing, education and legal services. ⑥ It must ensure that their rights are protected and that they receive the same level of support for the **legal services** they deserve. ⑦ David Jacobs New York, Jan. 2, 2002 the writer is a fellow at the Brennan center for justice, Columbia university.

    BOW: ① editor, prison, rights, civil ② prisoners, legal, treatment, fair ③ legitimate, rights, prison ④ system, stripping, access ⑤ housing, cost, services, affordable ⑥ support, money, services, deserve ⑦ writer, director, center, new, york

- -

FULLSEQ2SEQ: To the editor: Elisabeth Sara's article "exorcising the demons of jail" (April 19) is a powerful reminder that the system of state-financed prisoners' rights is stripping inmates of their right to equal opportunity, access to legal services and a chance to make real progress in redeeming the debts they have been stripped to the bone. Without adequate financing and affordable legal counsel, thousands of prison-financed prisoners will never be able to get their civil legal assistance they deserve. Joseph Jacobs new York, April 20, 2006 the writer is a staff attorney at the Brennan center for justice.

Figure 12: Sample outputs on OpinionGen. For our results, we present the bag-of-words predicted for each sentence, and the phrases relevant to the guidance keyphrases are highlighted in bold.