General then Personal: Decoupling and Pre-training for Personalized Headline Generation

Yun-Zhu Song[1], **Yi-Syuan Chen**[1], **Lu Wang**[2], and **Hong-Han Shuai**[1]
[1]National Yang Ming Chiao Tung University, Taiwan
[2]University of Michigan, Ann Arbor, MI, USA
{yzsong.ee07,yschen.ee09,hhshuai}@nycu.edu.tw
wangluxy@umich.edu

Abstract

Personalized Headline Generation aims to generate unique headlines tailored to users' browsing history. In this task, understanding user preferences from click history and incorporating them into headline generation pose challenges. Existing approaches typically rely on predefined styles as control codes, but personal style lacks explicit definition or enumeration, making it difficult to leverage traditional techniques. To tackle these challenges, we propose **G**eneral **T**hen **P**ersonal (GTP), a novel framework comprising user modeling, headline generation, and customization. We train the framework using tailored designs that emphasize two central ideas: (a) task decoupling and (b) model pre-training. With the decoupling mechanism separating the task into generation and customization, two mechanisms, i.e., information self-boosting and mask user modeling, are further introduced to facilitate the training and text control. Additionally, we introduce a new evaluation metric to address existing limitations. Extensive experiments conducted on the PENS dataset, considering both zero-shot and few-shot scenarios, demonstrate that GTP outperforms state-of-the-art methods. Furthermore, ablation studies and analysis emphasize the significance of decoupling and pre-training. Finally, the human evaluation validates the effectiveness of our approaches.[1]

1 Introduction

The task of headline generation aims to produce a concise sentence for expressing the salient information of the document (Liu et al., 2020). In addition to preserving the content information, recent studies have further proposed generating appealing headlines. For instance, some researchers propose to inject specific styles, such as humor and romance, into news headlines or generate interrogative ones to attract readers' attention (Shu et al., 2018; Jin et al., 2020; Zhan et al., 2022; Zhang et al., 2018b). However, these approaches only consider one type of style to catch the attention, neglecting that individuals may possess different preferences. Therefore, Personalized Headline Generation (PHG) (Ao et al., 2021) has emerged as a new research direction, which customizes news headlines by considering user information. Specifically, given body-headline pairs along with users' click history, the goal is to infer the implicit user preference and further incorporate it into generating personalized headlines. Besides, the task is usually formulated as zero-shot learning due to the high cost of collecting large-scale personalized headlines (Zhang et al., 2022).

Nevertheless, the task formulation brings the first critical challenge, i.e., generating personalized headlines without ground-truth annotation. In such a scenario, models are required to incorporate the implicit user preference without explicit supervision for personal text styles. The previous work (Ao et al., 2021) tackles the task by reinforcement learning, taking the users' click history as a learning signal to construct the style supervision. However, their learning framework does not leverage the news headlines that can contribute to improving headline quality. Therefore, we make the first attempt to facilitate personalization learning without or with limited ground-truth annotations while enhancing the generation quality by leveraging the news headlines.

In this paper, we propose a framework named **G**eneral **T**hen **P**ersonal (GTP) to tackle this challenge. Specifically, we propose to decouple the generation process into headline generation and headline customization. The goal of headline generation is to produce headlines targeting a general

[1]Our source code is available at https://github.com/yunzhusong/TACL-GTP.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 1588–1607, 2023. https://doi.org/10.1162/tacl_a_00621
Action Editor: Xiaojun Wan. Submission batch: 12/2022; Revision batch: 8/2023; Published 12/2023.

audience, while headline customization aims to further customize them based on the *control code*[2] of a specific user. With the task decoupling, we could pre-train the headline generator using body-headline pairs from a diverse range of news articles, thus improving the quality of generation. Afterward, the control code and generated headlines are jointly utilized by the headline customizer.

However, constructing the control code poses several challenges. In previous work, the control code typically refers to specific and discrete attributes of the target headlines, such as topics, sentiments, keywords, or descriptive prompts (Keskar et al., 2019; Chan et al., 2021a; He et al., 2022; Carlsson et al., 2022). While for PHG task, the target attributes are the user preferences encapsulated in the click histories, which cannot be directly defined. As such, we follow a similar approach to Ao et al. (2021) and pre-train recommendation models on impression logs to extract the features of click histories as the control codes.

To train the headline customizer utilizing the control codes, we could form the training samples with news articles and corresponding user click histories. However, since a news article can attract multiple users, each with potentially distinct preferences, the pairing between news and user click history is not one-to-one. This limitation impedes the learning of preference control. Thus, we construct a new dataset to alleviate the issue, which assumes that users who click on the same news have similar preferences. Specifically, we integrate the click histories of these users to synthesize a pseudo click history, which helps build a new user profile with more specific interests in the target news.

Moreover, due to the distinct latent space of control codes compared to regular text features, models may inadvertently disregard the control codes during the learning process. Thus, we design two mechanisms, Masked User Modeling (MUM) and Information Self-Boosting (ISB), to alleviate these issues. MUM serves as a pre-training objective to make the control code recognizable to the model, while ISB leverages the generated headline to recall information from the article, reducing information loss in the two-stage generation process.

Finally, the last issue is the lack of evaluation metrics. Previous works only depend on lexical similarity for evaluation, constraining the models to generate one type of headline. However, given an article, a user could be attracted by various headlines beyond ground-truth ones. This argument is akin to the multi-target summarization problem (Cachola et al., 2020), suggesting multiple valid summaries could exist for an article. To benchmark the degree of personalization, we propose the Anomaly-based Personalization Evaluation (APE) metric, inspired by the anomaly detection task and the evaluation metrics used in the field of vision domain. To implement APE, we train auto-encoder models to assess whether an input headline adheres to the same distribution as user-written ones. The detection model is expected to learn text style to distinguish the inputs, enabling us to consider the hidden states as style features. We quantify the results by measuring the feature distance between generated and user-written headlines. However, relying solely on distance measurements may not provide a comprehensive quality assessment. To offer a more intuitive metric, we introduce the editor headlines as reference points for comparison and employ relative values to convey the results. Unlike ROUGE scores, which focus on lexical similarity, APE enables a distribution-wise evaluation, providing a more flexible reference for assessment.

The contributions are summarized as follows:

- We propose to decouple the personal headline generation task into generation and customization for incorporating the user preference in a late fusion style. We propose two mechanisms, MUM and ISB, to leverage user and content information better.

- We propose a novel formulation for constructing the control code with one-to-one mappings between click histories and news headlines for better modeling text styles.

- We introduce a new evaluation metric, APE, from the perspective of anomaly detection to provide a more flexible reference and validate the metric with human evaluation.

[2]The control code refers to the information for controlling the generation process toward target headlines (Keskar et al., 2019).

- Extensive experiments, analysis, and user study demonstrate that the proposed GTP outperforms state-of-the-art approaches significantly under both zero-shot and few-shot settings.

2 Related Work

2.1 Control over Text Generation

Style-controlled Text Generation. The advancements in pre-training techniques have enabled modern language models (Chowdhery et al., 2022; Brown et al., 2020) to generate text that is nearly indistinguishable from human-written text (Clark et al., 2021). This has sparked increased interest among researchers in modeling and controlling text attributes, giving rise to the field of *controllable text generation (CTG)* (Prabhumoye et al., 2020). Previous work has explored various attributes, such as keywords (He, 2021), specified entities (Dong et al., 2021), document diction (Dathathri et al., 2020), topics (Keskar et al., 2019), sentiments (Chan et al., 2021a), humor (Amin and Burghardt, 2020), authorship (Syed et al., 2020), and social bias (Barikeri et al., 2021). These examples show attributes can be approached from different perspectives, including grammatical, artistic, or cognitive aspects. Certain perspectives, notably sentiment and humor, are closely associated with general stylistic aspects, while others are more related to intrinsic text qualities, such as keywords and diction. In addition to controlling text by unconditional language models (Subramani et al., 2019), the techniques of CTG lead to the emergence of *controllable text summarization* and *controllable headline generation* tasks. For instance, He et al. (2022) prepend descriptive prompts to articles to enable controllability. Chan et al. (2021b) propose an RL framework based on a constrained Markov decision process. Yamada et al. (2021) propose a Transformer-based framework to generate summaries with specified phrases. Jin et al. (2020) apply multi-tasking to learn headline generation and denoising autoencoding for specific style corpora.

Style Transfer. Besides controlling text generation conditionally or unconditionally, another research line focuses on text attribute transfer (Hu and Li, 2021), aiming to edit the existing text to possess desired attributes without considering contextual information. Similar to the CTG, the attributes could be style-related or intrinsic text qualities. The approaches involve disentangling text into content and attributes in the latent space for manipulation (Yi et al., 2020; Li et al., 2020), editing based on sentence templates (Madaan et al., 2020; Li et al., 2018), and creating pseudo-parallel data (Jin et al., 2019; Nikolov and Hahnloser, 2019). These methods often focus on transferring attributes in short sentences, making them naturally suited for tasks like headline generation. However, many previous works consider well-defined properties. In this paper, we focus on injecting personal preferences into headlines, posing a great challenge since preferences can be vague and difficult to capture and utilize effectively.

2.2 Realization of Control Codes

The main component of text control is injecting target attribute information, i.e., *control codes* (Keskar et al., 2019), into models. For verbalizable control codes such as keywords, topics, or entities, an approach is to make the corresponding tokens as the *hard prompts* of inputs during inference (Keskar et al., 2019; Fan et al., 2018). Alternatively, some works learn continuous representations of target attributes, known as *soft prompts*, to enable the control over general attributes (Li and Liang, 2021; Yu et al., 2021). Another research line is to make the control codes as learning targets. Much research has attempted to train scorers for target attributes, and utilized them as reward functions within the RL framework (Song et al., 2020; Stiennon et al., 2020) or the sampling bias during the decoding process (Krause et al., 2021; Mireshghallah et al., 2022). However, for PHG, control codes are the user preferences encapsulated in click histories. Unlike categorical attributes such as topics or dictions, user preferences involve attributes from different perspectives, including grammatical, artistic, and cognitive. Previous work applies recommendation models to extract user representations and design a reward to match the representations of generated headlines (Ao et al., 2021). Although straightforward, such a scheme does not explore the benefit of news headlines. In this paper, we construct our control code in a fine-grained manner and

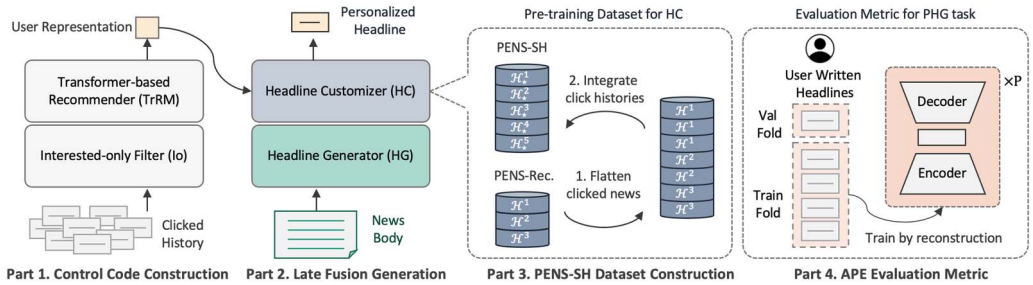

Figure 1: Overview of the proposed method. Parts 1 and 2 describe the frameworks for building and utilizing the user representation. Parts 3 and 4 present the training strategy and evaluation metric for the PHG task.

take click-through rate and news headlines into consideration to better capture user interests.

3 Problem Formulation

We denote the database of news articles as $\mathcal{D} = n_i = (x_i, y_i)_{i=1}^{|\mathcal{D}|}$, where x_i and y_i represent the body and headline of article n_i, respectively. The personalized headline generation task aims to generate user-specific headlines $Y^\tau = y_j^\tau$ for a given user τ, taking into account the user's preferences and the content of the news articles. To achieve this, the models need to understand the implicit user preference, which acts as the control code, from the user's click histories denoted as $\mathcal{H}^\tau = n_k$. Additionally, we include the user's impression logs, which contain information about the displayed news and click-through behaviors over a period of time, in the dataset to learn the user's preferences. The clicked news articles are considered positive samples $\mathcal{P}^\tau$, while the unclicked news articles are considered negative samples $\mathcal{N}^\tau$. For evaluation, we utilize the PENS dataset introduced by Ao et al. (2021), which provides click histories and a series of news articles with user-written headlines. Our work addresses both zero-shot and few-shot settings for personalized headline generation. In the zero-shot setting, the model learns to generate personalized headlines without ground-truth annotations. Furthermore, we explore the few-shot setting, where a limited number of user annotations are available during the learning phase. These two settings require different capabilities from the models (Yin et al., 2020), enabling us to investigate the proposed methods from diverse perspectives.

4 Methodology

In this section, we elaborate on how we tackle the PHG task as shown in Fig. 1. The following sections are organized as follows. Sec. 4.1 describes how to establish the control code from users' click histories. Sec. 4.2 introduces the generation framework to incorporate the control code effectively. Sec. 4.3 presents the training strategy and the process of pre-training dataset construction. Finally, Sec. 4.4 introduces an evaluation metric to quantify the degree of personalization.

4.1 Control Code Construction

The PHG task aims to generate user-tailored headlines that align with users' interests, but the lack of associated annotations necessitates extracting user preferences from historical click records. However, extracting relevant preference components from extensive click records is a significant challenge. To overcome this, a previous study (Ao et al., 2021) successfully trained a personalized news recommendation model to capture individual stylistic preferences. Following this approach, we also utilize a personalized news recommendation model to represent users' preferences derived from click records. Moreover, while prior work used different backbone models for recommendation and headline generation, we employ the same backbone and pre-trained models for both tasks to effectively integrate features from these distinct models. Furthermore, following the principles of content-based recommendation systems (Wu et al., 2019a,d; Li et al., 2022), we adopt a news encoder ϕ_{ne} and a user encoder ϕ_{ue} in the recommendation model. The textual information is then aggregated using *Attention Pooling* to construct the news representation. Similarly, the news representation is further aggregated by the user encoder to build the user representation $c^\tau \in \mathbb{R}^d$ as follows:

$$c^\tau = f(f(\mathcal{H}^\tau; \phi_{ne}); \phi_{ue}), \quad (1)$$

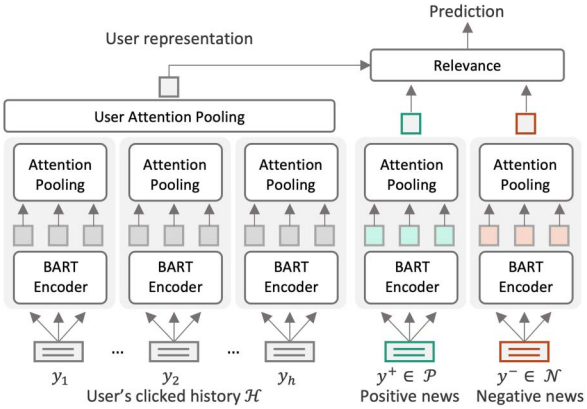

Figure 2: Framework of the personalized news recommendation model for extracting user representation.

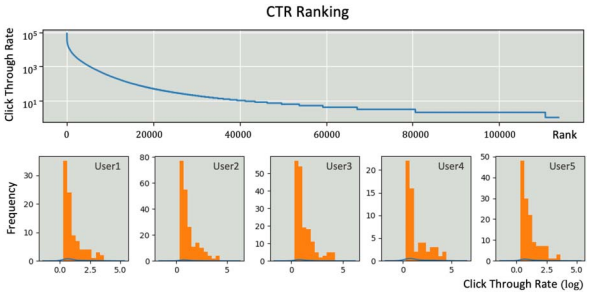

Figure 3: The CTR distribution of news (top) and the histograms of five users' click histories (bottom).

where $\mathcal{H}^\tau$ represents the click history of user τ. The overall framework is illustrated in Fig. 2. Unlike previous work, which solely learns textual representations from in-domain data (Wu et al., 2019b; An et al., 2019; Wang et al., 2020), recent research has started exploring pre-training techniques for news recommendation (Wu et al., 2021; Li et al., 2022). Our approach employs a pre-trained Transformer-based model as our Recommendation Model (TrRM) to enhance the textual representation, enabling us to achieve competitive performance with a simple configuration.

To train TrRM, we employ negative sampling techniques (Zheng et al., 2018; An et al., 2019). Specifically, we consider the news articles clicked by user τ as positive samples $\mathcal{P}^\tau$. We randomly sample M news articles from each user's negative sample set $\mathcal{N}^\tau$. A news representation v is obtained by encoding its headline: $v = f(y; \phi_{ne})$. The model jointly predicts the recommendation scores for the positive and negative samples by comparing them with user representation c^τ. Consequently, we formulate the training process as an M+1-way classification task as follows:

$$\mathcal{L}_{\text{rec}} = -\sum_{i \in \mathcal{P}^\tau} log(p_i),$$
$$p_i = \frac{exp(c^{\tau\top} v_i^+)}{exp(c^{\tau\top} v_i^+) + \sum_{j=1}^{M} exp(c^{\tau\top} v_{ij}^-)}, \quad (2)$$

where v_i^+ is the feature of the i-th positive sample and v_{ij}^- is the feature of the j-th negative sample of the i-th positive sample. This formulation allows the model to learn the relationships between the user representation and the positive and negative samples, facilitating effective recommendation for personalized headline generation.

Interested-News Only Filter. The concept of the long tail has been observed in various online businesses and utilized as a marketing strategy (Naik et al., 2022). To investigate this phenomenon in our task, we analyze the click-through rate (CTR) by examining the frequency of news articles in users' click histories and positive samples. Fig. 3 depicts the CTR rank distribution, which exhibits the characteristic of long-tail distribution. News articles with high CTR can be considered as high-impact news, while those with low CTR are regarded as low-impact news. Based on this observation, we make the assumption that a user's click history consists of both *popular news* and *interested news*. Popular news is widely circulated among users, and their clicks may be influenced more by general interest or current affairs rather than personal preferences. On the other hand, interested news is less widely disseminated and likely contains specific features that capture individual interests. Therefore, interested news is considered more indicative of user preferences. To construct the control code, we divide each user's click history into popular news and interested news using a quantile threshold based on the CTR. We focus on using only the interested news to build the control code. Consequently, the user representation c^τ is defined as:

$$c^\tau = f(f(\delta(\mathcal{H}^\tau|\gamma); \phi_{ne}); \phi_{ue}), \quad (3)$$

where $\delta(\cdot|\gamma)$ represents the interested-news filter, and γ is the quantile threshold used for identification. We refer to the TrRM model with the interested-news only filter as TrRMIo.

4.2 Late Fusion Generation Model

Late Fusion Framework. One approach to personalized headline generation (PHG) is to directly inject the user representation into a headline generator, creating early-fusion models. However,

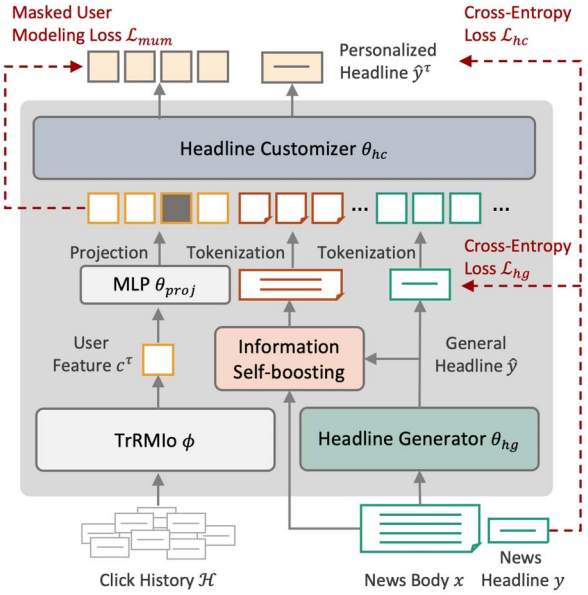

Figure 4: Framework of the proposed GTP, including the headline generator, headline customizer, and TrRMIo for generating personalized headlines.

this approach has limited effectiveness due to the absence of annotations. Moreover, early-fusion models often prioritize the headline generation task, potentially overshadowing the incorporation of user information, especially when the supervision for text style is weak. As a result, these models tend to disregard user information and function as regular headline generation models. Accordingly, we propose a two-stage generative model that exploits the news headlines while properly introducing user representation. While our goal is to generate personalized headlines, learning to produce a headline targeting a general audience first can be beneficial since personalized headlines and news headlines could share some commonality, such as the content and the grammar rules. Specifically, we decouple the generation process into headline generation (HG) θ_{hg} and headline customization (HC) θ_{hc} and apply them sequentially as depicted in Fig. 4. HG generates a general headline from the news body. HC further adjusts the generated headline using the user representation c^τ during the customization phase. The process of generating a personalized headline can be described as follows:

$$\widehat{y} = f(x; \theta_{hg}),\ \widehat{y}^\tau = f(\widehat{y}|c^\tau; \theta_{hc}), \qquad (4)$$

where $\widehat{y}$ and $\widehat{y}^\tau$ denote the generated and customized headline. By decoupling the process, HG can be trained on all news data without being constrained by user log impressions, leading to improved headline quality. Notably, adopting a late fusion style for incorporating the control code prevents its neglect, as the tasks of generation and personalization are treated separately. We provide further training details in Sec. 4.3.

Information Self-boosting. To address concerns about potential information loss arising from the two-stage generation process (Song et al., 2022), we propose a mechanism to incorporate supporting information from the news article into the generated headlines. Specifically, we extract information from the news body x based on the outputs $\widehat{y}$ generated by HG. We employ a greedy selection algorithm to retrieve relevant sentences, using lexical overlapping to measure text similarity. This extracted information denoted as s, is then concatenated with the outputs of HG and fed into HC. The customization process can be described as follows: $\widehat{y}^\tau = f([\widehat{y} \oplus s]|c^\tau; \theta_{hc})$. By incorporating relevant information, we enhance the customization stage and mitigate potential information loss during the two-stage generation.

Masked User Modeling. To enable the model to interpret the control code and avoid introducing strong bias (Carlsson et al., 2022), we incorporate the control code as part of the model inputs. Specifically, we concatenate the user representation with the token embeddings $\mathbf{e} = [e_1, \ldots, e_T]$ of the generated headline $\widehat{y}$ and the boosting information s. The user representation from the recommendation model, however, is not in the form of regular text, making it unrecognizable for pre-trained models. To address this, we propose Masked User Modeling (MUM) to enable the model to understand the heterogeneous input. In MUM, the user representation $c^\tau \in \mathbb{R}^d$ is first randomly projected by a fixed layer θ_{proj} to generate several user embeddings $\mathbf{u}^\tau = [u_1, u_2, \ldots, u_L] = f(c^\tau; \theta_{proj})$, each having the same dimension d as the token embeddings. The input sequence is then encoded using the headline customizer's encoder θ_{hce}, resulting in hidden states $\mathbf{h} = f([u_1, \ldots, u_L, e_1, \ldots, e_T]; \theta_{hce})$. During training, we randomly mask some user embeddings by replacing them with a special token [user]'s embedding $e_{[user]}$. The model is trained to reconstruct the masked user embeddings from the remaining inputs with Mean Square Error (MSE):

$$\mathcal{L}\text{mum} = \text{MSE}(\mathbf{h}[user], \mathbf{u}[user]), \qquad (5)$$

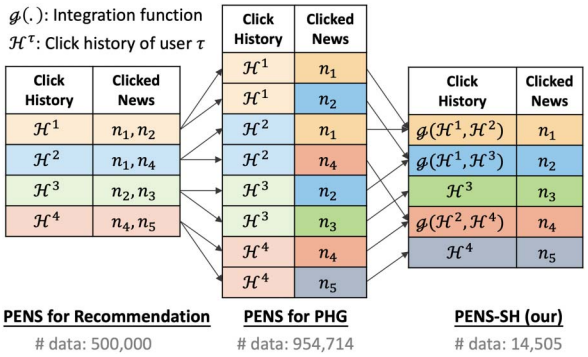

Figure 5: Illustration of the construction process for the PENS-SH dataset, which is applied for headline customizer pre-training.

where $\mathbf{h}[user]$ and $\mathbf{u}_{[user]}$ are the output hidden states and input user features of the masked tokens, respectively. This process aims to help the model understand the input structure and effectively incorporate user information.

4.3 Training Strategy

PENS-SH Dataset Construction. To train a personalized headline generation model without explicit annotations, previous research (Ao et al., 2021) encodes and aggregates the generated results into condensed features, subsequently matched with positive news items from the click history. However, we have identified significant potential for improving text quality over reinforcement learning-based methods. Specifically, we propose a two-stage approach to improve text quality while facilitating collaboration with preference information. In this methodology, editor headlines are treated as pseudo targets, and corresponding click histories are simulated based on users who have clicked on the same pseudo target. However, it is important to note that a single news article can attract multiple users, each with potentially distinct preferences, resulting in multiple click histories for a pseudo target. Learning from these potentially conflicting examples can profoundly impact the preference-aware generation process. Hence, we introduce PENS-SH to alleviate these issues as shown in Fig. 5, which assumes that users who click on the same news article have similar preferences. Specifically, we integrate the click histories of users who have clicked on a particular news article into a news pool. From this pool, we select news articles with higher occurrence frequencies to synthesize a pseudo click history. By leveraging shared information among these users, we construct a pseudo click history that helps establish a new user profile with clearer and more specific interests related to the target in a one-to-one manner.

Headline Generator Training. Through generation decoupling and late-fusion strategy, we separate the training process into two stages. The first stage of training is user-agnostic, allowing us to train the HG θ_{hg} on all news data pairs $\mathcal{D}$. We optimize HG by the cross entropy loss as follows:

$$\mathcal{L}_{hg} = \frac{-1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} logP(y|x;\theta_{hg}),$$
$$logP(y|x;\theta_{hg}) = \frac{1}{|y|} \sum_{t=1}^{|y|} logP(y_t|y_{<t}, x;\theta_{hg}). \tag{6}$$

Headline Customizer Pre-training. The goal of HC is to generate user-specific headlines by incorporating user representation. We adopt the proposed MUM loss to make the HC aware of the user representation. Besides, HC is trained on the PENS-SH database, denoted by $\mathcal{D}^\star$, which can stabilize the training since there is no duplicated news in the dataset. The overall training objective $\mathcal{L}_{hc}$ of HC θ_{hc} is defined as follows:

$$\mathcal{L}_{hc} = \mathcal{L}_{mum} + \mathcal{L}_{gen},$$
$$\mathcal{L}_{gen} = \frac{-1}{|\mathcal{D}^\star|} \sum_{(y,\mathcal{H}^\star)\in\mathcal{D}^\star} logP(y|\widehat{y}, s, c^\star;\theta_{hc}), \tag{7}$$

where $c^\star$ is the control code obtaining from the shared history $\mathcal{H}^\star$ by Eq.3, $\widehat{y}$ is the generated headline from HG, and s is the boosting information as mention in Sec. 4.2.

Headline Customizer Finetuning. We further consider the few-shot setting for PHG to explore model behaviors from different perspectives. Specifically, we finetune the headline customizer user-wisely with a few annotations. The HC is further finetuned by Eq. 7, where the generation target y is replaced by the user-written headline y^τ.

4.4 Anomaly-based Personalization Evaluation

In the PENS corpus, the degree of personalization is solely evaluated based on the ROUGE scores between the generated and user-written headlines.

Such a method poses strict requirements on models to generate one type of headline that is lexically similar to the ground-truths. However, given an article, users could be attracted by various headlines besides the ones written by themselves. This argument connects to the multi-target summarization problem (Cachola et al., 2020; Over, 2003), which suggests that multiple valid summaries could exist for a given document. Therefore, in addition to ROUGE scores, we propose quantifying the extent of personalization through the lens of *anomaly detection* (Chandola et al., 2009). The task of anomaly detection is to find patterns in data that do not conform to expected behavior. Leveraging this objective, we consider user-written headlines as *normal* data and evaluate whether the generated ones are *anomaly* accordingly. The anomaly detection model is expected to extract text style to better distinguish whether the input headlines are user-written, which enables us to consider the hidden states of the detection model as style features. Furthermore, inspired by the evaluation metrics for generative models in the computer vision domain, such as Fréchet Inception Distance (Heusel et al., 2017) and Fréchet Video Distance (Unterthiner et al., 2018), we quantify results by measuring the distance between the style features of generated data S_g and reference data S_r. The distance between them is defined by 2-Wasserstein distance as follows:

$$d(S_g, S_r) = |\mu_r - \mu_g|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}),$$

where μ_r and μ_g are the means, and Σ_r and Σ_g are the covariance matrices. A smaller $d(S_r, S_g)$ suggests the generated headlines are closer to the references distributionally. However, the distance without comparison is hard to tell how good is the generation results. Thus, we provide a more intuitive metric by normalizing the distance to the editor-written ones. Specially, the evaluation metric APE is calculated as follows:

$$\text{APE} = d(S_g, S_r)/d(S_e, S_r) \rightarrow [0, \infty),$$

where S_e is the style features of editor-written headlines. Hence, the APE score indicates the distance proportion between 1) generated and user-written headlines and 2) editor and user-written headlines. The minimal APE $= 0$ means that the generated and reference headlines have the same distribution. APE $= 1$ indicates that the generated results share a similar style similarity as the editor headlines, while APE < 1 implies that the generated results are better than the editor ones in terms of user style. The formulation enables us to gauge the magnitude of APE scores more effectively.

5 Experiments

5.1 Setups

Implementation Details. We conduct experiments on the PENS dataset (Ao et al., 2021), including 113,762 news articles and 500,000 impressions from online users. The testing set comprises data from 103 users. Each user's click history and 200 headlines written by the user are available. Thus, there are a total of 103 * 200 = 20,600 personalized headlines. The models will be evaluated on all 20,600 testing samples for the zero-shot setting. For few-shot learning, we divide the 200 headlines of each user into 80/20/100 splits for training, validation, and testing, respectively. Essentially, we train a model for a user with the 80 training examples, and evaluation is performed on the 100 testing examples. This setup can be understood as the *intra-user setting*, as the objective is to evaluate the model's generalization given a few examples specific to each user. Additionally, we also consider the *inter-user setting*. To achieve this, we utilize 40/13/50 users for training, validation, and testing. Specifically, the model is trained using 40 * 200 = 8000 examples. For testing, the split is chosen to encompass 50 users, resulting in 50 * 200 = 10,000 testing examples. This setting helps us evaluate the model's generalization across different users. The proposed PENS-SH dataset includes 14,505 pairs, and we take 12,505 and 2000 for training and validating the HC. Both HG and HC are initialized with BART (Lewis et al., 2020). For the APE metric, the fold number P is 2, and the fold size S is 100.

Baselines. Our baselines are PENS (Ao et al., 2021) and EUI-PENS (Zhang et al., 2022). EUI-PENS builds upon PENS using entity words from news and input-dependent user representations. Then, we compare with ChatGPT via the OpenAI API[3] to explore the benefits of using an

[3] https://openai.com/blog/openai-api.

	Methods	RM	Pre-train	ROUGE-1 / 2 / L ↑	BLEURT ↑	BARTScore ↑	APE ↓
Zero-Shot	Editor	–	–	47.81 / 26.67 / 36.74	51.47	3.71	0.71
	Pointer-Gen	–	–	19.86 / 7.76 / 18.83	–	–	–
	PG+RL-ROUGE	–	–	20.56 / 8.84 / 20.03	–	–	–
	PENS	EBRN	–	25.49 / 9.14 / 20.82	–	–	–
	PENS	DKN	–	27.48 / 10.07 / 21.81	–	–	–
	PENS	NPA	–	26.11 / 9.58 / 21.40	–	–	–
	PENS	NRMS	–	26.15 / 9.37 / 21.03	–	–	–
	PENS	LSTUR	–	24.10 / 8.82 / 20.73	–	–	–
	PENS	NAML	–	28.01 / 10.72 / 22.24	–	–	–
	EUI-PENS	Ent-CNN	–	32.34 / 13.93 / 26.90	–	–	–
	ChatGPT	–	–	29.80 / 11.04 / 24.15	40.97	2.32	13.04
	One-Stage	–	–	33.68 $_{0.007}$ / 14.09 $_{0.004}$ / 27.70 $_{0.010}$	42.22 $_{0.002}$	2.95 $_{0.001}$	1.59 $_{0.001}$
	One-Stage†	TrRMIo	PENS-SH	33.45 $_{0.008}$ / 13.97 $_{0.004}$ / 27.60 $_{0.009}$	41.77 $_{0.003}$	2.92 $_{0.001}$	3.69 $_{0.002}$
	GTP	TrRMIo	PENS	33.50 $_{0.008}$ / 14.03 $_{0.009}$ / 27.65 $_{0.002}$	41.85 $_{0.002}$	2.96 $_{0.001}$	1.92 $_{0.077}$
	GTP	TrRMIo	PENS-SH	**33.84**$^{*}_{0.002}$ / **14.23**$^{*}_{0.000}$ / **27.85**$^{*}_{0.001}$	**42.26**$^{*}_{0.002}$	**3.01**$^{*}_{0.001}$	**0.76**$^{*}_{0.003}$
Intra Few-Shot	One-Stage	–	–	33.87 $_{0.16}$ / 14.18 $_{0.11}$ / 27.83 $_{0.10}$	41.68 $_{0.002}$	2.92 $_{0.002}$	1.46 $_{0.002}$
	Two-Stage	–	✗	34.12 $_{0.17}$ / 14.46 $_{0.09}$ / 28.32 $_{0.06}$	41.76 $_{0.12}$	3.02 $_{0.004}$	1.20 $_{0.009}$
	Two-Stage‡	TrRMIo	✗	33.65 $_{0.11}$ / 14.26 $_{0.07}$ / 28.30 $_{0.06}$	41.39 $_{0.06}$	3.04 $_{0.001}$	2.24 $_{0.014}$
	GTP	TrRMIo	PENS-SH	**34.93**$^{*}_{0.16}$ / **15.23**$^{*}_{0.12}$ / **29.21**$^{*}_{0.08}$	**42.54**$^{*}_{0.06}$	**3.28**$^{*}_{0.002}$	**0.62**$^{*}_{0.004}$
Inter Few-Shot	One-Stage	–	–	34.10 $_{0.08}$ / 14.37 $_{0.06}$ / 28.05 $_{0.04}$	42.08 $_{0.281}$	3.01 $_{0.012}$	1.44 $_{0.005}$
	Two-Stage	–	✗	34.13 $_{0.37}$ / 14.55 $_{0.27}$ / 28.52 $_{0.08}$	**42.09** $_{0.282}$	2.92 $_{0.019}$	2.40 $_{0.386}$
	Two-Stage‡	TrRMIo	✗	33.48 $_{0.11}$ / 14.06 $_{0.07}$ / 28.19 $_{0.06}$	41.64 $_{0.083}$	3.05 $_{0.006}$	4.37 $_{0.941}$
	GTP	TrRMIo	PENS-SH	**34.61**$^{*}_{0.06}$ / **14.74** $_{0.03}$ / **28.55** $_{0.07}$	42.05 $_{0.192}$	**3.17*** $_{0.007}$	**1.12**$^{*}_{0.074}$

Table 1: The performance comparison of different baselines and GTP in zero-shot and few-shot finetuning scenarios. RM represents recommendation models that provide user representations. Our zero-shot results are averaged over three runs with different random seeds, while the few-shot results are averaged over three runs with different data splits. The subscripts denote the variances. † and ‡ indicate *Early* and *Late Fusion* settings, respectively. $*$ denotes GTP significantly improves over the strongest baseline (bootstrapping test, $p < 0.05$).

large language model on the personalized generation task. To facilitate the task with ChatGPT, we formulate the prompt as follows:

```
I want you to act as a personalized
headline writer. I will provide you
with some headlines clicked by a user
and a target document. You will answer
the personalized headline of the target
document for the user.

The clicked headlines are: <clicked>.
The target document is: <document>.
```

, where <clicked> comprises 50 concatenated headlines extracted from the user's click history.[4] As prior works primarily focus on the zero-shot setting, we adopt the ablations of GTP as our baselines for the few-shot setting.

Evaluation Metrics. *ROUGE-n* (Lin, 2004) evaluates the lexical similarity between the generation results and references. *BLEURT* (Sellam et al., 2020) involves pre-training BERT using millions of synthetic examples to enhance generalization and robustness in evaluation. *BARTScore* (Yuan et al., 2021), built upon BART, evaluates text by considering its probability of being generated from or generating other textual inputs and outputs. Both BLEURT and BARTScore utilize references, i.e., ground-truth, for evaluation. On the other hand, *G-Eval* (Liu et al., 2023) employs large language models with customized prompts to execute reference-free evaluation for different aspects of texts. The proposed *APE* metric evaluates the degree of personalization specifically.

5.2 Main Results

Table 1 summarizes the results obtained in zero- and few-shot settings. Notably, we observe that the APE metric exhibits higher sensitivity to out-of-distribution samples than ROUGE. This sensitivity can be attributed to the deliberate inclusion of user-written headlines in the APE learning process. APE models operate on the principle of anomaly detection, restricting the availability

[4]We use ''gpt-3.5-turbo'' for this experiment. To conform to the constraints of the employed OpenAI model, we truncate the prompt to 4000 tokens, and the output length is confined to 64 tokens.

of solely in-domain data (i.e., user-written headlines) during the learning phase. Consequently, when confronted with out-of-domain data during inference, the generated results could be subpar and sensitive. This situation resembles neural models' challenges in domain generation (Wang et al., 2022). We leverage this characteristic to emphasize the discrepancies between the generated results. Moreover, APE evaluates similarity based on the collective knowledge acquired by the model, which provides an evaluation from a different angle, compensating for the one-to-one comparison metrics.

Zero-shot Results. *One-Stage* approach is our first stage model, HG, which achieves exceptional performance, indicating the importance of leveraging general news headlines. The improvements can be attributed to the partial commonality between personalized and non-personalized headlines. In contrast to PENS, which employs reinforcement learning and utilizes the similarity with the news body as a learning reward, our decoupling scheme enables HG to exploit all news headlines and optimize specifically for headline generation. To generate personalized headlines, we incorporate user representations from a recommendation model. However, we encounter challenges in integrating user information without style annotations. Specifically, the *One-Stage (Early Fusion)*, which introduces the user representation alongside the news body, performs worse than the simple HG in terms of both lexical and style similarity metrics. Therefore, it is crucial to design pre-training objectives that facilitate model adaptation and the incorporation of control codes. By decoupling the generation process and introducing two proposed mechanisms for pre-training, namely ISB and MUM, *GTP* achieves significant improvements in all metrics. The decoupling and ISB mechanism allows the model to focus on transforming a general headline into a personalized one, while the MUM objective guides the model in utilizing user information effectively. Moreover, we validate the effectiveness of pre-training GTP with PENS-SH by replacing it with the original PENS dataset. The results show that directly pre-training with PENS yields inferior results compared to the tailored PENS-SH, showing the benefits of establishing a one-to-one mapping between the control code and the pseudo target. Finally, GTP significantly outperforms ChatGPT. We hypothesize that this discrepancy arises from the intricate and implicit nature of personal preferences, posing a challenge for effective utilization without appropriate design. Our methods leverage a recommender model to encapsulate the nuanced information. The information is subsequently employed with our specialized methodologies, making the generated headline more cognizant of the underlying preferences.

Intra Few-shot Results. We begin by presenting the performance of the first-stage outputs as the baseline in *One-Stage (w/o finetuning)*, as the testing data in the few-shot setting differs from the zero-shot setting. Subsequently, we investigate the benefits of finetuning using a small number of user-written samples within the two-stage framework, which is reflected in the results of *Two-Stage*. These results indicate that few-shot finetuning can enhance performance for both metrics, suggesting a distribution discrepancy between news and user-written headlines. This discrepancy further emphasizes the challenge of generating personalized headlines without any user annotations. Additionally, we explore the advantages of incorporating user representations extracted from the click history as shown in *Two-Stage (Late Fusion)*, where we employ the decoupling network and introduce the user representation in the second stage to facilitate few-shot finetuning. Similar to our observations in the zero-shot setting, directly adding user representation and finetuning the model lead to inferior performance. This finding underscores the difficulty of simultaneously utilizing out-of-distribution information while learning the user style from a limited number of samples. Consequently, we propose two mechanisms to pre-train the decoupled network, enabling better utilization of user representation. The results of *GTP* significantly outperform the baselines, underscoring the importance of making the control code recognizable to the model prior to few-shot finetuning.

Notably, Table 2 unveils that baseline models slightly outperform GTP in terms of the aspects such as coherence and consistency. To delve into this result, we also assess the G-Eval scores for editor-written and user-written headlines as shown in the first and second row. The results suggest that user-written headlines exhibit

	Methods	RM	Pre-train	Coherence (1∼5) ↑	Consistency (1∼5) ↑	Fluency (1∼3) ↑	Relevance (1∼5) ↑
Intra Few-Shot	Editor	–	–	3.89	4.21	2.73	4.16
	User	–	–	3.79	4.08	2.41	3.91
	ChatGPT	–	–	3.94	4.17	2.76	4.23
	One-Stage	–	–	3.88	4.23	2.74	4.18
	GTP	TrRMIo	PENS-SH	3.86	4.17	2.65	4.09

Table 2: The G-Eval performance comparison of different baselines and GTP in the intra few-shot finetuning scenarios. RM represents recommendation models that provide user representations.

Method	ROUGE-1 / ROUGE-2 / ROUGE-L ↑	BLEURT ↑	BARTScore ↑	APE ↓
(1) GTP	$34.93_{0.16}$ / $15.23_{0.12}$ / $29.21_{0.08}$	$42.54_{0.06}$	$3.28_{0.002}$	$0.62_{0.004}$
(2) w/o TrRMIo	$34.79_{0.12}$ / $15.21_{0.08}$ / $29.19_{0.07}$ (−0.06)	$42.41_{0.04}$ (−0.13)*	$3.18_{0.005}$ (−0.10)*	$0.85_{0.016}$ (↑ 0.23)*
(3) w/o MUM	$34.85_{0.18}$ / $15.18_{0.09}$ / $29.18_{0.13}$ (−0.05)	$42.26_{0.04}$ (−0.28)*	$3.26_{0.001}$ (−0.02)	$1.08_{0.028}$ (↑ 0.45)*
(4) w/o ISB	$34.26_{0.15}$ / $14.51^{*}_{0.12}$ / $28.51^{*}_{0.08}$ (−0.70)	$42.17_{0.05}$ (−0.37)*	$3.07_{0.001}$ (−0.21)*	$2.14_{0.001}$ (↑ 1.51)*
(5) w/o Pre-training	$33.65^{*}_{0.11}$ / $14.26^{*}_{0.07}$ / $28.30^{*}_{0.06}$ (−1.06)	$41.39_{0.06}$ (−1.15)*	$3.04_{0.001}$ (−0.24)*	$2.24_{0.014}$ (↑ 1.62)*
(6) w/o Late Fusion	$33.57^{*}_{0.14}$ / $14.08^{*}_{0.08}$ / $27.90^{*}_{0.07}$ (−1.27)	$41.27_{0.06}$ (−1.27)*	$2.90_{0.001}$ (−0.38)*	$3.13_{0.191}$ (↑ 2.51)*

Table 3: Ablation study for the intra few-shot finetuning. The performances are averaged over three different data splits with subscripts denoting the variances. The results indicate that removing either mechanism leads to a degradation in GTP's performance. Decoupling the generation framework and pre-training contribute the most to the overall performance. $*$ denotes the result is significantly different from GTP (bootstrapping test, $p < 0.05$).

relatively weaker performance in these aspects. This observation is expected as users tend to prioritize their preferences over exhibiting superior text quality compared to well-trained editors. As a result, we could note that the performance of GTP closely aligns with that of user-written headlines. Overall, the evaluation from various metrics demonstrates that GTP ensures not only effective personalization but also maintains the textual quality compared to the baselines.

Inter Few-shot Results. The results in Table 1 indicate that GTP can enhance performance even in inter-user scenarios for ROUGE, BARTScore, and APE metrics. Regarding BLEURT metrics, we observe significant variance, with similar performance levels across different methods. Overall, GTP could still offer advantages under inter-user settings, particularly benefiting applications where titles for new users are unavailable.

5.3 Ablation Study

Table 3 provides detailed ablation studies on GTP. The experiments are performed in the few-shot settings using three different data splits. Firstly, we replace the user encoder TrRMIo with a language model (Lewis et al., 2020) to evaluate the importance of using a recommendation model for obtaining the control code. From row 2, we identify that the APE score is degraded, indicating that the recommendation model is better equipped to capture user preferences beyond textual and content information. Additionally, row 3 presents the results without adopting the MUM pre-training objective, where the APE score drops more than the ROUGE scores. We consider the reason is that the MUM aims to assist the model in utilizing the text style encoded in the control code, which is better reflected in the APE metric compared to ROUGE. In another way, ISB aims to mitigate the information loss in the two-stage framework and can greatly contribute to both metrics, as shown in row 4, suggesting that ISB provides valuable information to enhance the customization process. Row 5 demonstrates the results without pre-training, where the model is instead initialized from a general language model. The results highlight the necessity of enabling the model to recognize the input formulation before few-shot learning. Lastly, row 6 presents the model's performance without late fusion, where personalized headlines are generated in one stage, and the control code is injected along with the input article. The results indicate that such a scheme fails to effectively leverage the control codes, leading to significant performance drops in both metrics. Overall, these ablation studies provide insightful analysis of the various components and mechanisms in GTP, highlighting contributions of the

Sampling Strategy	ROUGE-Avg	APE
Random	27.17 $_{0.11}$	0.64 $_{0.01}$
Diversity	27.05 $_{0.08}$	0.79 $_{0.02}$
Similarity	27.15 $_{0.09}$	0.87 $_{0.07}$

Table 4: Performance comparison of different sampling strategies for the intra few-shot finetuning suggests that the random sampling performs better than others. The ROUGE-Avg reports the average of ROUGE-1/2/L/S.

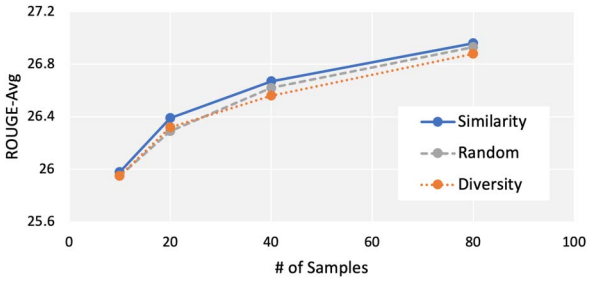

Figure 6: Performance comparison for different sample sizes (10/20/40/80) and sampling strategies (similarity/diversity/random) in the intra few-shot setting. The ROUGE-Avg reports the average of ROUGE-1/2/L/S.

proposed mechanisms, framework, and training scheme.

5.4 Analysis of Few-shot Sample

This section further analyzes the influence of sample selection for few-shot learning. We explore three strategies for selecting samples: 1) random sampling, where samples are randomly chosen from the news pool; 2) diversity sampling, which involves applying k-means clustering to identify distinctive data and selecting samples closest to the cluster centroids; and 3) similarity sampling, where samples with a higher similarity between the user-written and generated news headlines are chosen based on cosine similarity of sentence embeddings (Gao et al., 2021). The results presented in Table 4 reveal that random sampling achieves slightly better performance compared to the other two strategies, especially for the APE metric. As a result, we adopt random sampling as the default setting for few-shot finetuning. These findings also emphasize the challenges of effectively capturing the user style in personalized news headline generation. Furthermore, we analyze the influence of sample size on the performance. Fig. 6 provides an overview of the results, indicating that the performances improve as the sample size increases.

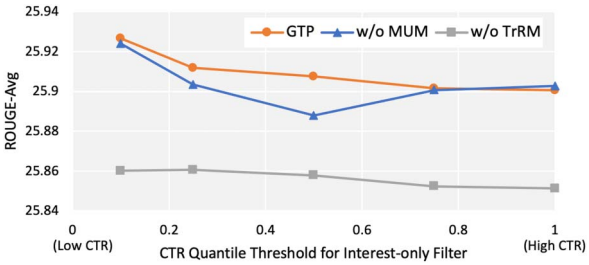

Figure 7: The comparison of Interested-only Filter's quantile threshold under intra 5-shot finetuning. The results show that forming user representations from the news with lower CTR could better indicate user preference.

This suggests that learning personal style from a limited number of samples is challenging.

5.5 Analysis of User Representation

This section analyzes the construction and the pre-training of user representations. Therefore, besides GTP, we consider the models without MUM (w/o MUM) and without training the user encoder with recommendation task (w/o TrRM). Fig. 7 shows the performances with different quantile thresholds of the Interested-only Filter under the three configurations. The results demonstrate that constructing user representations by the news with lower CTR performs better, especially when the user encoder is trained with the recommendation task. These observations also meet our hypothesis that a click history contains interested news and popular news, which could be identified by CTR. In addition, the results show that models achieve better performance with MUM and TrRM in various threshold settings.

5.6 Results of Personalized Recommendation

In the news recommendation literature, most methodologies rely on news titles to model news items due to their significant impact on users' clicking behaviors (Wu et al., 2023). Several studies have expanded their approach by incorporating supplementary features, such as keywords (Zhang et al., 2018a), entities (Qi et al., 2021), categories (Wu et al., 2019a), topics (Wu et al., 2019c), location (Xun et al., 2021), popularity (Tavakolifard et al., 2013), and others. Although integrating more textual features is feasible for the proposed TrRMIo, we have opted to exclusively employ news titles to ensure the generalizability of our method and concentrate on studying the

Model	AUC/MRR/NDCG@5/@10
EBRN (Okura et al., 2017)	63.97 / 22.52 / 26.45 / 32.81
DKN (Wang et al., 2018)	65.25 / 24.07 / 26.97 / 32.24
NPA (Wu et al., 2019b)	64.97 / 23.65 / 26.72 / 33.96
NRMS (Wu et al., 2019d)	64.27 / 23.28 / 26.60 / 33.58
LSTUR (An et al., 2019)	62.49 / 22.69 / 24.71 / 32.28
NAML (Wu et al., 2019a)	66.18 / 25.51 / 27.56 / 35.17
Entity-CNN (Zhang et al., 2022)	66.28 / 25.34 / 27.58 / 35.53
TrRMIo (title)	68.88 / 27.27 / 30.98 / 38.81
TrRMIo (title + keyword)	69.01 / 27.05 / 30.72 / 38.61
TrRMIo (keyword)	65.51 / 25.21 / 28.12 / 35.79

Table 5: The performance of baselines and TrRMIo on news recommendation task. The results underscore the advantage of using a pre-trained language model. Different input configurations of TrRMIo suggest that titles are necessary for the recommendation task, and incorporating auxiliary features could further enhance the performance.

proposed methodologies since the additional information may be unavailable. Nevertheless, to provide a more comprehensive discourse on TrRMIo, we have conducted additional experiments by incorporating title keywords as auxiliary inputs for learning. The results are presented in the last three rows of Table 5. These results indicate that the performance can be slightly enhanced by incorporating additional textual elements (*TrRMIo(title+keyword)*). However, the exclusion of titles, with only keywords under consideration (*TrRMIo(keyword)*), significantly affects the performance, underscoring the need to incorporate titles. It is imperative to emphasize that the TrRMIo is designed as a general-purpose model with the objective of serving the proposed GTP framework. Table 5 demonstrates the superiority of TrRMIo over previous approaches.

5.7 Human Evaluation

In addition to automated evaluation, the proposed methods are assessed by soliciting human judgments. Our human evaluation necessitates participants to answer a set of binary-choice questions. Each question comprises a target headline authored by a user from the PENS corpus and two test headlines generated by two distinct models. Participants are required to select the test headline that demonstrates a greater resemblance to the target headline in terms of text style, encompassing factors such as length, vocabulary, structure, tone, and other pertinent aspects. Participants are instructed not to base their choices on personal preferences but to assume that the target headline represents their preferred option for answering the questions. Furthermore, an additional ''tie'' option is provided if participants cannot decide after careful consideration. The evaluation is conducted separately for the zero-shot and few-shot settings. In the zero-shot setting, 26 randomly sampled questions are presented. The corresponding test headlines are generated using the *GTP* and the method of *One-Stage* model with early fusion. In the few-shot setting, a similar approach is adopted, with 40 questions provided, where 20 questions are the comparison between *GTP* and *One-Stage* and the remaining 20 are associated with *GTP* and *Editor* headlines. The order of the questions is randomly permuted to mitigate potential recognition of the underlying generation methods.[5] The win rates of GTP under various settings are presented in Table 6. The findings show that the headlines generated by GTP more closely resemble the desired headlines compared to various baselines, including editor-written ones, suggesting that GTP can better utilize the user information.

	GTP	Baseline	Tie
zero-shot			
GTP *vs* One-Stage	60.88%	22.30%	16.82%
few-shot			
GTP *vs* One-Stage	59.41%	26.92%	13.68%
GTP *vs* Editor	58.66%	29.97%	11.37%

Table 6: Human evaluation on style similarity. The results of win rates suggest that evaluators tend to consider headlines from GTP closer to reference headlines than those from baselines in terms of text style.

	Averaged Score		Win Rate	
zero-shot	chosen	unchosen	chosen	unchosen
GTP *vs* One-Stage	$0.827_{0.007}$	$0.881_{0.006}$	69.05%	30.95%
few-shot	chosen	unchosen	chosen	unchosen
GTP *vs* One-Stage	$1.094_{0.002}$	$1.152_{0.004}$	80.49%	19.51%
GTP *vs* Editor	$0.972_{0.006}$	$0.987_{0.008}$	61.90%	38.10%

Table 7: The APE metric validation by human evaluation. The subscripts denote the score variances between evaluators. The chosen groups consistently outperform the unchosen ones in APE by both the averaged score and win rate, matching the human judgments.

[5]The evaluation process requires approximately 20 minutes. We recruited 50 participants to evaluate both segments. Before engaging in the tasks, all participants provide informed consent and are duly compensated for their time, which is set at $5 per participant.

Furthermore, it is noteworthy that a certain proportion of tie options were selected, indicating that the manifestation of preference may be subtle or inconspicuous in some instances, which could be attributed to the intrinsic content and topic of the news. These observations necessitate future works to investigate the varying level of difficulty associated with customizing distinct headlines. It is important to note that no personally identifiable information was collected, and participants were not exposed to offensive content.

5.8 APE Validation

We leverage the human evaluation outcomes from Sec. 5.7 to validate the effectiveness of the APE metric in aligning with human judgments, as shown in Table 7. First, we separate the chosen and unchosen headlines into two groups for each participant. Next, we compute the APE scores for both groups, considering each participant individually, and then calculate the overall average scores across all participants. If the APE metric agrees with human judgments, we would expect to observe lower APE scores for the chosen headline group, indicating a reflection of human perspectives. In addition to the APE scores, we present the win rate for the chosen and unchosen groups. Specifically, we designate the group with a lower APE score as the winner for each participant and calculate the corresponding win rate. The APE metric provides a high-level view of the agreement between the metric and human tendencies. At the same time, the win rate offers a low-level perspective for the agreement of each participant. The results consistently show that the chosen groups outperform the unchosen ones in both the high- and low-level APE scores, thereby confirming the APE's alignment with human judgments.

6 Conclusion

In this paper, we propose a novel framework named General Then Personal (GTP) to tackle the challenges of constructing and incorporating control code for personalized headline generation. Specifically, we propose a late fusion model by decoupling the generation process. Two mechanisms are further introduced to facilitate the framework and enable text control. Additionally, we construct a pre-training dataset, PENS-SH, to build an effective control code, which enhances the connections between click history and target news. Moreover, we introduce a novel evaluation metric, APE, to quantify the degree of personalization. The extensive experiments and human evaluation demonstrate the necessity of all designs and show that GTP significantly outperforms state-of-the-art under both zero-shot and few-shot settings.

Acknowledgments

The authors would like to thank the anonymous reviewers and the action editor (Xiaojun Wan) for their valuable discussions and feedback. Lu Wang is supported by National Science Foundation through grant IIS-2046016. This work was supported in part by the National Science and Technology Council of Taiwan under Grants NSTC-109-2221-E-009-114-MY3, NSTC-112-2221-E-A49-059-MY3, NSTC-111-2221-E-001-021, and NSTC-112-2221-E-A49-094-MY3.

References

Miriam Amin and Manuel Burghardt. 2020. A survey on approaches to computational humor generation. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online. International Committee on Computational Linguistics. https://aclanthology.org/2020.latechclfl-1.4

Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural news recommendation with long- and short-term user representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 336–345, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1033

Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021. PENS: A dataset and generic framework for personalized news headline 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 82–92, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.7

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models. 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 1941–1955, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.151

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.428

Fredrik Carlsson, Joey Öhman, Fangyu Liu, Severine Verlinden, Joakim Nivre, and Magnus Sahlgren. 2022. Fine-grained controllable text generation using non-residual prompting. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6837–6857, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.471

Alvin Chan, Yew-Soon Ong, Bill Pung, Aston Zhang, and Jie Fu. 2021a. Cocon: A self-supervised approach for controlled text generation. In *International Conference on Learning Representations*.

Hou Pong Chan, Lu Wang, and Irwin King. 2021b. Controllable summarization with constrained Markov decision process. *Transactions of the Association for Computational Linguistics*, 9:1213–1232. https://doi.org/10.1162/tacl_a_00423

Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3). https://doi.org/10.1145/1541880.1541882

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*. Version 5. https://doi.org/10.48550/arXiv.2204.02311

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. 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 7282–7296, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.565

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Xiangyu Dong, Wenhao Yu, Chenguang Zhu, and Meng Jiang. 2021. Injecting entity types into entity-guided text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 734–741, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.56

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-2706

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910 Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.552

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRLsum: Towards generic controllable text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5879–5915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.396

Xingwei He. 2021. Parallel refinements for lexically constrained text generation with BART. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8653–8666, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.681

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Zhiting Hu and Li Erran Li. 2021. A causal lens for controllable text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 24941–24955. Curran Associates, Inc.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. 2020. Hooks in the headline: Learning to generate headlines with controlled styles. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5082–5093, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.456

Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. 2019. IMaT: Unsupervised text attribute transfer via iterative matching and translation. 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 3097–3109. https://doi.org/10.18653/v1/D19-1306

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A conditional transformer language model for controllable generation. *CoRR*, abs/1909.05858. Version 1.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.424

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. https://doi.org/10.18653/v1/2020.acl-main.703

Jian Li, Jieming Zhu, Qiwei Bi, Guohao Cai, Lifeng Shang, Zhenhua Dong, Xin Jiang, and Qun Liu. 2022. MINER: Multi-interest matching network for news recommendation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 343–352, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.29

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. 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 1865–1874.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.353

Yuan Li, Chunyuan Li, Yizhe Zhang, Xiujun Li, Guoqing Zheng, Lawrence Carin, and Jianfeng Gao. 2020. Complementary auxiliary classifiers for label-conditional text generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8303–8310. https://doi.org/10.1609/aaai.v34i05.6346

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. https://aclanthology.org/W04-1013

Dayiheng Liu, Yeyun Gong, Yu Yan, Jie Fu, Bo Shao, Daxin Jiang, Jiancheng Lv, and Nan Duan. 2020. Diverse, controllable, and keyphrase-aware: A corpus and method for news multi-headline generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6241–6250, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.505

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*. Version 3.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W. Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881. https://doi.org/10.18653/v1/2020.acl-main.169

Fatemehsadat Mireshghallah, Kartik Goyal, and Taylor Berg-Kirkpatrick. 2022. Mix and match: Learning-free controllable text generationusing energy language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 401–415, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.31

Aakanksha Naik, Jill Lehman, and Carolyn Rosé. 2022. Adapting to the long tail: A meta-analysis of transfer learning research for language understanding tasks. *Transactions of the Association for Computational Linguistics*, 10:956–980. https://doi.org/10.1162/tacl_a_00500, PubMed: 36303892

Nikola I. Nikolov and Richard Hahnloser. 2019. Large-scale hierarchical alignment for data-driven text rewriting. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 844–853. https://doi.org/10.26615/978-954-452-056-4_098

Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based news recommendation for millions of users. In

Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17, pages 1933–1942, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3097983.3098108

Paul Over. 2003. An introduction to duc 2003: Intrinsic evaluation of generic news text summarization systems. In *Proceedings of Document Understanding Conference 2003*.

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. https://doi.org/10.18653/v1/2020.coling-main.1

Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2021. PP-rec: News recommendation with personalized user interest and time-aware news popularity. 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 5457–5467, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.424

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.704

Kai Shu, Suhang Wang, Thai Le, Dongwon Lee, and Huan Liu. 2018. Deep headline generation for clickbait detection. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 467–476. https://doi.org/10.1109/ICDM.2018.00062

Yun-Zhu Song, Yi-Syuan Chen, and Hong-Han Shuai. 2022. Improving multi-document summarization through referenced flexible extraction with credit-awareness. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1667–1681, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.120

Yun-Zhu Song, Hong-Han Shuai, Sung-Lin Yeh, Yi-Lun Wu, Lun-Wei Ku, and Wen-Chih Peng. 2020. Attractive or faithful? Popularity-reinforced learning for inspired headline generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8910–8917. https://doi.org/10.1609/aaai.v34i05.6421

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Nishant Subramani, Samuel Bowman, and Kyunghyun Cho. 2019. Can unconditional language models recover arbitrary sentences? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Bakhtiyar Syed, Gaurav Verma, Balaji Vasan Srinivasan, Anandhavelu Natarajan, and Vasudeva Varma. 2020. Adapting language models for non-parallel author-stylized rewriting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9008–9015. https://doi.org/10.1609/aaai.v34i05.6433

Mozhgan Tavakolifard, Jon Atle Gulla, Kevin C. Almeroth, Jon Espen Ingvaldesn, Gaute Nygreen, and Erik Berg. 2013. Tailored news in the palm of your hand: A multi-perspective transparent approach to news recommendation. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, pages 305–308, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/2487788.2487930

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards accurate generative models of video: A new metric & challenges. *CoRR*, abs/1812.01717. Version 2.

Heyuan Wang, Fangzhao Wu, Zheng Liu, and Xing Xie. 2020. Fine-grained interest matching for neural news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 836–845, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.77

Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. Dkn: Deep knowledge-aware network for news recommendation. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 1835–1844, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3178876.3186175

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1. https://doi.org/10.1109/TKDE.2022.3178128

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019a. Neural news recommendation with attentive multi-view learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3863–3869. International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2019/536

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019b. Npa: Neural news recommendation with personalized attention. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '19, pages 2576–2584, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3292500.3330665

Chuhan Wu, Fangzhao Wu, Mingxiao An, Yongfeng Huang, and Xing Xie. 2019c. Neural news recommendation with topic-aware news representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1154–1159, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1110

Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019d. Neural news recommendation with multi-head self-attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6389–6394, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1671

Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2023. Personalized news recommendation: Methods and challenges. *ACM Transactions on Information Systems*, 41(1). https://doi.org/10.1145/3530257

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Empowering news recommendation with pre-trained language models. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11–15, 2021*, SIGIR '21, pages 1652–1656, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3404835.3463069

Jiahao Xun, Shengyu Zhang, Zhou Zhao, Jieming Zhu, Qi Zhang, Jingjie Li, Xiuqiang He, Xiaofei He, Tat-Seng Chua, and Fei Wu. 2021. Why do we click: Visual impression-aware news recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, pages 3881–3890, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3474085.3475514

Kosuke Yamada, Yuta Hitomi, Hideaki Tamori, Ryohei Sasano, Naoaki Okazaki, Kentaro Inui, and Koichi Takeda. 2021. Transformer-based lexically constrained headline generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4085–4090, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.335

Xiaoyuan Yi, Zhenghao Liu, Wenhao Li, and Maosong Sun. 2020. Text style transfer via learning style instance supported latent space. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3801–3807. International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2020/526

Mingzhang Yin, George Tucker, Mingyuan Zhou, Sergey Levine, and Chelsea Finn. 2020. Meta-learning without memorization. In *International Conference on Learning Representations*.

Dian Yu, Zhou Yu, and Kenji Sagae. 2021. Attribute alignment: Controlling text generation from pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2251–2268, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.194

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*.

Jiaao Zhan, Yang Gao, Yu Bai, and Qianhui Liu. 2022. Stage-wise stylistic headline generation: Style generation and summarized content insertion. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4489–4495. International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2022/623

Kui Zhang, Guangquan Lu, Guixian Zhang, Zhi Lei, and Lijuan Wu. 2022. Personalized headline generation with enhanced user interest perception. In *Artificial Neural Networks and Machine Learning – ICANN 2022*, pages 797–809, Cham. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-15931-2_65

Lemei Zhang, Peng Liu, and Jon Atle Gulla. 2018a. A deep joint network for session-based news recommendations with contextual augmentation. In *Proceedings of the 29th on Hypertext and Social Media*, HT '18, pages 201–209, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3209542.3209557

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, Huanhuan Cao, and Xueqi Cheng. 2018b. Question headline generation for news articles. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pages 617–626, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3269206.3271711

Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. 2018. Drn: A deep reinforcement learning framework for news recommendation. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 167–176, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3178876.3185994