Generating Hashtags for Short-form Videos with Guided Signals

Tiezheng Yu†*, Hanchao Yu‡, Davis Liang‡, Yuning Mao‡, Shaoliang Nie‡, Po-Yao Huang‡, Madian Khabsa‡, Pascale Fung†, Yi-Chia Wang‡

†Hong Kong University of Science and Technology ‡Meta AI

tyuah@connect.ust.hk, yichiaw@meta.com

Abstract

Short-form video hashtag recommendation (SVHR) aims to recommend hashtags to content creators from videos and corresponding descriptions. Most prior studies regard SVHR as a classification or ranking problem and select hashtags from a set of limited candidates. However, in reality, users can create new hashtags, and trending hashtags change rapidly over time on social media. Both of these properties cannot be easily modeled with classification approaches. To bridge this gap, we formulate SVHR as a generation task that better represents how hashtags are created naturally. Additionally, we propose the Guided Generative Model (GGM) where we augment the input features by retrieving relevant hashtags from a large-scale hashtag pool as extra guidance signals. Experimental results on two short-form video datasets show that our generative models outperform strong classification baselines, and the guidance signals further boost the performance by 8.11 and 2.17 absolute ROUGE-1 scores on average, respectively. We also perform extensive analyses including human evaluation, demonstrating that our generative model can create meaningful and relevant novel hashtags while achieving state-of-the-art performance on known hashtags [1].

Video Frames:

Video Description: First Snow Storm of 2008 in Toronto, Ontario, Canada.

Ground Truth Hashtags:
#canada #snowstorm #toronto #winter

V-MLM Classification Results:
#snow #winter #snowstorm #canada #2008

GGM Generation Results:
#snowstorm #winter **#cold-weather** #toronto

Figure 1: Video frames and the video description are the inputs. The Guided Generative Model (GGM) generates hashtags related to video frames (e.g., #winter and #cold-weather) as well as video description (e.g., #snowstorm and #toronto). The generated novel hashtag that never appears in the training set is highlighted. We use mock-up video frames in the paper.

1 Introduction

Short-form videos on social media are increasingly popular thanks to the proliferation of multimedia technologies and portable devices (Vandersmissen et al., 2014; Montag et al., 2021). To highlight the topics and salient information of the videos, hashtags – words or unspaced phrases prefixed with a "#" – have been widely used. Proper use of hashtags can also increase the probability of the videos being discovered (Cao et al., 2020). In light of this, short-form video hashtag recommendation (SVHR), which aims to suggest relevant and meaningful hashtags to content creators when they share videos, has received considerable attention from industry and academia (Li et al., 2019; Jain and Jindal, 2020; Mehta et al., 2021). However, most previous studies on SVHR consider it as a classification problem and rank the hashtags in a small and fixed-size set one by one (Li et al., 2019; Wei et al., 2019; Cao et al., 2020; Yang et al., 2020). These methods are time-consuming and far from the actual application, where users are free to create new hashtags, and trending hashtags change rapidly on social media platforms.

To fill this research gap, we formulate SVHR as a generation task that better represents the process through which hashtags are created by content creators. Figure 1 shows an example of the generation results that include a novel hashtag (#cold-weather). The generative model learns to generate hashtags related to video frames as well as video description.

[*] Work is done during an internship at Meta.

[1]The code is released at: https://github.com/facebookresearch/hashtag-generation

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, pages 9482–9495
July 9-14, 2023 ©2023 Association for Computational Linguistics

Additionally, we propose to retrieve hashtags from a large hashtag pool to augment input features and use the retrieved hashtags to guide hashtag generation. Inspired by the effectiveness of vision-language models (VLMs) (Radford et al., 2021; Jia et al., 2021; Fürst et al., 2022) in video-text retrieval tasks, we construct our hashtag retriever based on VLM. Then, we build a multimodal hashtag generator to generate hashtags from the retrieved hashtags, video frames, and user-written video descriptions. To leverage multimodal inputs, we introduce a cross-modal attention mechanism (CAM) to fuse information from different modalities. We name the whole architecture Guided Generative Model (GGM).

We conduct experiments to evaluate strong classification baselines, generative models and the proposed GGM on two well-known short-form video datasets: SFVD1 and SFVD2. For the classification models, we regard SVHR as a multi-label classification problem and compute probabilities over all hashtags that appear in training set by a softmax activation so the models can capture the interaction between the labels of each video. Experimental results demonstrate that the generative models outperform the classification models, and the guidance signals further boost the performance by 8.11 and 2.17 ROUGE-1 scores on average. We further create an unseen test set (Simig et al., 2022) for SFVD1 and analyze the models' performance on it. Results show that generative models are able to generate unseen hashtags that the classification models can never predict. In addition, we assess the generated hashtags with human evaluation since the automatic metrics might underestimate our models' ability to create novel hashtags. The results from our human evaluations show that GGM is able to create meaningful novel hashtags that are statistically comparable to the ground truth hashtags.

Our contributions are summarized as follows:

- We are the first to formulate SVHR as a generation task that better represents how hashtags are created naturally. We propose the Guided Generative Model (GGM), which leverages the retrieved hashtag to augment the input for hashtag generation.
- We present an extensive analysis of experimental results, including human evaluation, and demonstrate that GGM achieves state-of-the-art performance on two large-scale datasets (SFVD1 and SFVD2).
- Our work benchmarks classification and generative models on SVHR datasets and highlights the advantage of generative approaches, which we hope will catalyze research in this area.

2 Related Work

Short-form Video Hashtag Recommendation. Li et al. (2019) introduced the SVHR task and used graph convolutional network to deal with long-tail hashtags. Several works leveraged user information in SVHR (Wei et al., 2019; Liu et al., 2020). Yang et al. (2020) proposed incorporating sentiment features when recommending hashtags. Cao et al. (2020) focused on modeling the multimodal information of short-form videos. However, most of the previous works consider the SVHR as a binary classification problem and select hashtags from limited candidates by computing the recommended scores one by one (e.g., 101 candidates (Yang et al., 2020; Cao et al., 2020) and 1001 candidates (Wei et al., 2019)). Generally, these approaches are time-consuming and not practical for real-world applications. Therefore, we formulate SVHR as a generation task which better represents how hashtags are generated naturally by users.

Keyphrase and Microblog Hashtag Generation. Keyphrase generation (KPG) aims to generate phrases that highlight salient information for a piece of text. According to (Meng et al., 2021), existing KPG methods can be divided into `One2One` (Meng et al., 2017) which generates one keyphrase at a time and `One2Seq` (Yuan et al., 2020) which generates a sequence of keyphrases at once. In this work, we apply `One2Seq` framework and randomly shuffle the target hashtags in each batch to mitigate the effect of order when fine-tuning the model. Different from KPG, SVHR requires the model to process multimodal inputs. Recently, Wang et al. (2019b) introduced the microblog hashtag generation task, which can be viewed as a variation of the KPG for the social media domain. Additionally, topic-aware models (Wang et al., 2019a) and news articles (Zheng et al., 2021) are leveraged to improve hashtag generation. To our knowledge, we are the first to generate hashtags for short-form videos.

Retrieval-Augmented Generation. Retrieval-Augmented Generation (RAG) has been widely used in NLP tasks, such as neural machine transla-

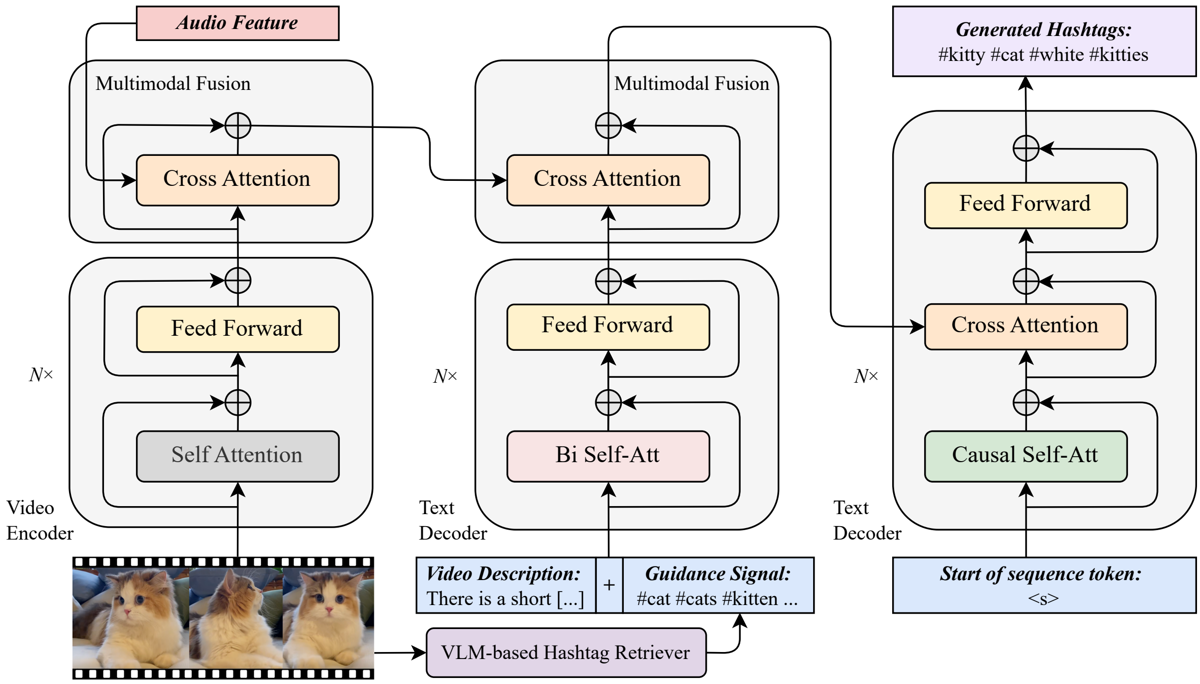

Figure 2: An overview of our proposed framework. The VLM-based Hashtag Retriever retrieves the relevant hashtags as extra features to guide the generative model. Information from different modalities is fused by cross-modal multi-head attention.

tion (Gu et al., 2018; Hossain et al., 2020), open-domain question answering (Lee et al., 2019; Guu et al., 2020; Lewis et al., 2020b) and knowledge-grounded dialogue generation (Lian et al., 2019). Recently, some works also utilized this framework in multimodal tasks. Zhang et al. (2021) proposed a Retrieve-Copy-Generate (RCG) model for open-book video captioning. To tackle the Outside-knowledge visual question answering task, (Gao et al., 2022) transformed the image into plain text, performed knowledge passage retrieval, and generated answers entirely in the natural language space. This work extends the RAG framework to multi-modal hashtag generation. Both hashtag retriever and generator can accept any advanced models as drop-in replacements.

3 Methodology

In this section, we first introduce our problem definition of the SVHR task. Then, we present our Guided Generative Model (GGM) which generates the hashtag from multimodal inputs.

3.1 Problem Definition

The main objective of SVHR is to generate recommended hashtags given a short-form video and its user-written textual description. To enrich the input signals, we construct a large-scale hashtag pool as the knowledge base. Note that recommended hashtags should not be limited to the hashtag pool since meaningful novel hashtags are also considered valuable. Formally, the visual information of the video is formulated as frames F, and A denotes the acoustic information of the video. The textual description is defined as a sequence of words $D = (d_1, ..., d_{|D|})$. K is the hashtag knowledge base, and all hashtags in the training set are included by default. We do not include hashtags from external sources since the hashtag styles can vary widely across datasets, as shown in Figure 4. Finally, the models need to recommend the optimal hashtags Y by finding:

$$\arg\max Prob(Y|F, A, D, K; \theta) \qquad (1)$$

where θ is the set of trainable parameters.

3.2 Guided Generative Model

Figure 2 depicts the architecture overview of GGM, which consists of a VLM-based hashtag retriever, a video encoder, a text encoder and a text decoder.

VLM-based Hashtag Retriever The goal of the hashtag retriever is to find the top-k most relevant hashtags from the hashtag knowledge base K given a video. Inspired by the recent improvements in video-to-text retrieval (Portillo-Quintero

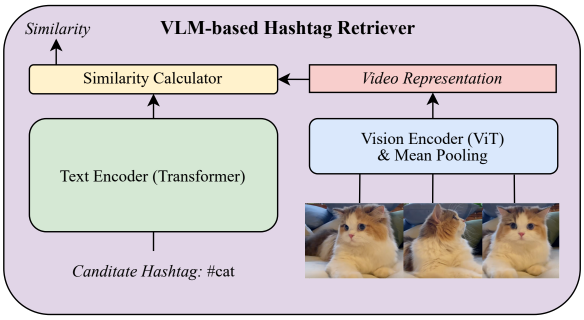

Figure 3: VLM-based hashtag retriever. Hashtags that have the highest similarity scores with video are retrieved.

et al., 2021; Luo et al., 2021; Fang et al., 2021), we built our hashtag retriever based on vision-language models. The hashtag retriever applies a Bi-encoders framework (Figure 3). The text encoder maps all hashtags in the pool to a list of hashtag representations $\mathbf{T} = (\mathbf{t}_1, ..., \mathbf{t}_j)$. The vision encoder calculates the frames' embedding $(\mathbf{w}_1, ..., \mathbf{w}_m)$ of each video, where each frame's embedding $\mathbf{w}_m$ comes from the representation (`[CLS]`) token of the vision encoder outputs. Similar to (Luo et al., 2021), the video representation is calculated by adopting a mean pooling mechanism to aggregate the embeddings of all frames, $\hat{\mathbf{w}} = \texttt{mean-pooling}(\mathbf{w}_1, ..., \mathbf{w}_m)$. Then, the similarity score sim between the video representation $\hat{\mathbf{w}}$ and each hashtag representation $\mathbf{t}_j$ is computed by:

$$sim = \frac{\mathbf{t}_j{}^T \hat{\mathbf{w}}}{\|\mathbf{t}_j\| \, \|\hat{\mathbf{w}}\|} \tag{2}$$

The hashtags with top-k similarity scores are selected as the guidance signal for our generative model's input.

To train the hashtag retriever, we adopt the pre-trained VLM (Radford et al., 2021) to initialize the model and follow the same contrastive learning loss. Each training sample consists of video frames and one hashtag corresponding to the video.

Audio-Grounded Video Encoder The audio-grounded video encoder creates the video representation given the audio and frames from the video. We employ a Transformer-based audio encoder (Wu et al., 2022) to encode the sound of the video. The acoustic information is mapped to an audio feature $\mathbf{a}$, and the audio feature will be used directly as the audio input to the audio-vision fusion mechanism. For the video frames, we use a N layers Vision Transformer (ViT) (Dosovitskiy et al., 2020) to encode them. Each frame is divided into patches and encoded as a frame embedding $\mathbf{w}_m \in \mathbb{R}^{n \times d}$, where n is the number of patches. All video frame embeddings are concatenated to build the visual representation $\mathbf{w} \in \mathbb{R}^{nl \times d}$.

After that, a cross-modal attention mechanism (CAM) is applied to get the audio-grounded video representation $\mathbf{v}$ (Eq. 3). CAM is based on the multi-head attention module in Transformer architecture (Vaswani et al., 2017). The query is linearly projected from the audio feature, and the key and value are linearly projected from the visual feature. In addition, we conduct residual connection (He et al., 2016) between the visual representation and the video representation.

$$\mathbf{v} = CAM(\mathbf{w}W_q, \mathbf{a}W_k, \mathbf{a}W_v) + \mathbf{w} \tag{3}$$

Video-Grounded Text Encoder The video-grounded text encoder takes the video representation and text (i.e., user-written description and guidance signal) as the input to produce the video-grounded text representation. We employ a N layers Transformer text encoder to get text representations. Each layer consists of a bi-directional self-attention sub-layer and a fully connected feed-forward network. In order to input both the description and the guidance signal to the text encoder, the input text is formatted into "`description [sep] guidance signal`", and the output is a text embedding $\mathbf{t} \in \mathbb{R}^{k \times d}$, where k denotes the number of input tokens. Similar to the audio-grounded video encoder, we use CAM to fuse the video and text representations (Eq. 4). Finally, the original text representation $\mathbf{t}$ is residual connected to the video-grounded text representation $\mathbf{t}'$.

$$\mathbf{t}' = CAM(\mathbf{t}W_q, \mathbf{v}W_k, \mathbf{v}W_v) + \mathbf{t} \tag{4}$$

Video-Grounded Text Decoder We construct an N-layers video-grounded text decoder following the standard Transformer text decoder. For each layer in the decoder, the bi-directional self-attention is replaced with causal self-attention. Meanwhile, an additional cross-attention sub-layer is inserted to perform multi-head attention over the video-grounded text representation. The decoder generates hashtags token by token. A beginning-of-sequence (BOS) token is used to indicate the start of decoding, and an end-of-sequence (EOS) token is used to signal its end. In addition, we use a separator token to separate the generated hashtags.

Dataset name	Avg len of videos [s]	Avg len of description	# of unique hashtags	# of hashtags per video
SFVD1	6.13	6.80	10,674	2.70
SFVD2	37.17	12.29	43,282	5.13

Table 1: Statistics of the datasets. We calculate the length of the video and description by the number of seconds and words, respectively.

4 Experimental Settings

4.1 Datasets

We evaluate our models on two well-known large-scale short-form video datasets, SFVD1 and SFVD2. After filtering out the videos that only have hashtags occurring lower than five times, we obtain 95,265 short-form videos for SFVD1 and 312,778 for SFVD2. The statistics of both datasets are shown in Table 1. Note that all videos in SFVD1 are shorter than seven seconds and videos in SFVD2 are shorter than 90 seconds. For SFVD2, we regard the user tags as the hashtags in our experiments. We randomly split the datasets into three disjoint subsets with 80%, 10% and 10% of the data for training, validation and test sets, respectively.

In order to simulate the real-world scenario where new and/or trending hashtags emerge, we construct an unseen test set for SFVD1 with two steps similar to (Simig et al., 2022): (1) Choose 500 hashtags that appear in the training split, and (2) Move all samples in the original training and test set that contain any of these 500 hashtags to the unseen test set[2]. Finally, the SFVD1 dataset covers 69,539 samples for training, 9,826 for validation, 8,690 for seen testing and 10,210 for unseen testing. Note that seen hashtags could also appear in the unseen test set samples because videos have multiple labels, and we only ensure that at least one unseen hashtag exists in each sample of the unseen test set. The numbers of seen and unseen hashtags are 10,104 and 570, respectively.

4.2 Implementation Details

For the video pre-processing, we extract frames with a frame rate of 2fps if the video duration is less than seven seconds and uniformly sample 15 frames if the video is longer than seven seconds. `ffmpeg` is used to extract the audio as a WAV format file from the video. The retrieval pool of hashtags contains all hashtags in the training set. We further evaluate the unseen hashtags to test the retriever's generalization ability as shown in Table 4. For the GGM, we initialize the video encoder with the ViT-base model and construct the text encoder-decoder based on the BART-base model. During decoding, we use beam search with a beam size of 5. The decoding process will not stop until the end-of-sequence token is emitted or the length of the generated hashtags reaches to maximum length. We set the maximum length as 32 for SFVD1 and 64 for SFVD2. See Appendix A for more implementation details.

4.3 Baselines

The following baselines are implemented for comparison: 1) **BERT** (Devlin et al., 2019) takes video description as input for classification. 2) **VLM** is the same as our VLM-based Hashtag Retriever. 3) **ViT** (Dosovitskiy et al., 2020) takes video as input for classification. 4) **ViT-BERT** concatenates the video and description embedding for classification. 5) **BART** (Lewis et al., 2020a) generates hashtags from the video description. 6) **Trocr-fid** (Li et al., 2021) generates hashtags based on the video. we apply the fusion-in-decoder (Izacard and Grave, 2021) strategy to let the model accept multi-frame input. 7) **VG-BART** (Yu et al., 2021) takes video and description to generate hashtags. We replace the visual feature extractor from 3D ResNet (Hara et al., 2017) to ViT for a fair comparison with GGM. Appendix A describes the details of the baseline implementation.

Similar to (Gong and Zhang, 2016; Mahajan et al., 2018), we build multi-label classifiers by minimizing the cross-entropy between the predicted softmax distribution and the target distribution. The reason we do not include previous binary classification approaches on SVHR (Yang et al., 2020; Cao et al., 2020; Wei et al., 2019) is three-fold: (1) They select the hashtags one by one from a pre-defined relatively small number of candidates (e.g., 101). In contrast, we only give the hashtags in training set as prior information. (2) For each test sample, the models need to compute a score for each video-hashtag pair across all hashtag candidates, which is time-consuming and far from the actual application. (3) Multi-label classifiers can encode interrelation among one video's hashtags, which binary classification approaches cannot.

[2]Due to strong correlations between labels, there are additional 70 hashtags removed together with the 500 hashtags from the training set and added to the unseen hashtag set.

Method	SFVD1				SFVD2			
	ROUGE-1	ROUGE-2	F1	BertScore	ROUGE-1	ROUGE-2	F1	BertScore
Classificaiton / Retrieval Methods								
BERT_d	14.84	7.69	13.43	59.85	32.92	7.70	37.40	60.11
VLM_v	12.06	5.96	9.72	56.72	10.48	2.54	9.59	52.09
ViT_v	19.18	9.88	17.88	61.39	28.29	6.75	32.20	58.71
$\text{ViT-BERT}_{d,v}$	20.51	10.74	19.33	61.95	36.33	8.24	41.13	61.46
Generation Methods								
BART_d	20.72	14.21	18.18	65.01	41.82	10.76	42.59	63.67
Trocr-fid_v	18.51	11.73	16.31	62.71	23.95	5.42	24.76	59.30
$\text{VG-BART}_{d,v}$	24.66	15.84	21.92	66.34	48.02	11.92	48.86	66.28
VLM Guided Methods								
$\text{GGM}_{d,v,g}$	28.71	18.24	24.95	67.77	48.68	**12.16**	49.72	66.54
$\text{GGM + Audio}_{d,v,g,a}$	**29.04**	**18.46**	**25.29**	**67.93**	**48.92**	12.05	**49.86**	**66.62**

Table 2: Main results of baselines and our proposed models on the seen test sets of SFVD1 and SFVD2 . For classification / retrieval methods, we choose the top five hashtags. We denote user-written description, video, guidance signal and audio inputs to *d*, *v*, *g* and *a* respectively.

4.4 Evaluation Metrics

We adopt ROUGE metrics (Lin, 2004) that were initially used for summarization evaluation since we consider SVHR as a generation task. ROUGE-1 and ROUGE-2 F1 are used to measure the n-gram overlaps. We include ROUGE-2 because some hashtags are combinations of words. We do not employ ROUGE-L because the order of hashtags should not affect the evaluation. The F1 score is used to calculate the exact match of the hashtags between the labels and predictions. In addition, we include the BERTScore (Zhang et al., 2019) to compute the semantic similarity of the hashtags. More details of the evaluation metrics are in Appendix B.1.

Methods	ROUGE-1	ROUGE-2	F1	BertScore
ViT-BERT n=1	23.21	14.57	22.38	65.48
ViT-BERT n=3	23.63	13.11	22.66	64.41
ViT-BERT n=5	20.51	10.74	19.33	61.95
ViT-BERT n=10	20.47	10.78	19.38	61.89
VG-BART	24.66	15.84	21.92	66.34
GGM_{ViT} k=50	27.01	16.64	23.93	67.52
GGM k=0	25.04	16.27	22.16	66.56
GGM k=25	28.33	17.85	24.56	67.65
GGM k=50	**28.71**	**18.24**	**24.95**	**67.77**
GGM k=100	27.16	17.23	23.35	67.11

Table 3: Results on the SFVD1 seen test set. n in ViT-BERT denotes the number of hashtags that were chosen as final results while k in GGM stands for the number of retrieved hashtags used as guidance signal. We also replace the VLM in GGM with ViT and use the top 50 hashtags selected from ViT to build the ViT-Guided Generative Model (GGM_{ViT}). We can see generative models overall outperform the classification models and GGM performs better than GGM_{ViT}.

5 Results and Analysis

5.1 Main Results

Effectiveness of Generative Models Table 2 shows the performance of the classification and generation models on the SVHR task. When we only use descriptions as input, BART achieves higher scores than BERT across all metrics. This could be because the BART decoder can better capture the textual information from the input description compared to the classification layer in BERT. On the contrary, ViT surpasses Trocr-fid when only taking videos as input. We speculate that the text decoder of Trocr-fid may lose some visual information through the cross-modal transformation from vision to text while ViT directly maps the video to hashtag labels. When taking both video and description as input, VG-BART achieves better performance than ViT-BERT. Most importantly, GGM performs significantly better than VG-BART (the SOTA multimodal summarization model) with the help of the guidance signals. Surprisingly, the audio information does not improve the performance much, especially on the SFVD2. It could be because lots of the audio is just random background music and more than 5% of the videos in SFVD2 do not have audio. Additionally, to verify that the reason of the low performance of the classification models is not because we choose top five hashtags, we evaluate ViT-BERT with different numbers of hashtags as final results. Table 3 shows that VG-

Test set	Model	k=1	k=5	k=10	k=50	k=100
Seen	ViT	**13.77**	**28.03**	**33.95**	**47.41**	**53.56**
	VLM	5.74	16.61	23.34	42.14	50.81
Unseen	ViT	**3.65**	**10.96**	14.54	23.51	27.60
	VLM	2.64	9.20	**14.29**	**29.72**	**37.81**

Table 4: Recall of ViT and VLM on SFVD1 test set, showing that VLM can really retrieve unseen hastags.

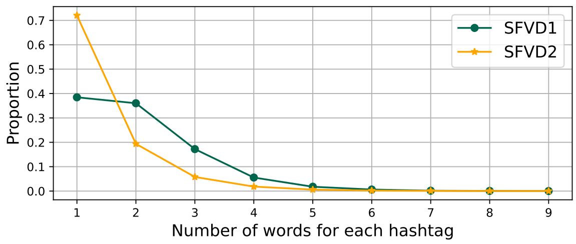

Figure 4: Distribution of the number of words that constitute each hashtag. SFVD1 has more multi-words hashtags compared to SFVD2.

BART outperforms ViT-BERT with all different settings, indicating the superiority of using generative models for SVHR.

SFVD1 versus SFVD2 We find that the models generally perform better on SFVD2 than SFVD1 except for ROUGE-2 scores (Table 2). There are two possible explanations. First, as shown in Figure 4, most of the hashtags in SFVD2 are made up of one word (e.g., cat, youth, family), while the hashtags in SFVD1 usually contain multiple words (e.g., #furrypride, #TheRealFamily, #dontjudgeme). It's easier for models to generate single-word hashtags and thus perform better on SFVD2. And due to the small proportion of multi-word hashtags in SFVD2, ROUGE-2 drops because it calculates the overlap of multi-word hashtags. Second, there are more hashtags (20.94% versus 4.48%) that appear in the corresponding video description of SFVD2 than SFVD1. Thus, the better model performance on SFVD2 could be partially attributed to the fact that it is easier for the models to extract words from the description as hashtags.

VLM Retrieval versus ViT Classification It's surprising that the performance of VLM is significantly lower than ViT on both datasets. To explore this phenomenon more deeply, we test how ViT and VLM perform discrepantly among varying top-k. Table 4 illustrates the recall of the models on the SFVD1 test set with different k. We can see that the performance gap between VLM and ViT

Methods	ROUGE-1	ROUGE-2	F1	BertScore
VG-BART	14.66	6.11	9.24	61.65
GGM	18.46	7.92	12.03	63.14
GGM$_{unseen}$	**18.72**	**8.15**	**12.48**	**63.26**

Table 5: Generative models' results on the SFVD1 unseen test set. Hashtag retriever only selects hashtags from seen hashtags for GGM and selects from both seen and unseen hashtags for GGM$_{unseen}$. The retrieved unseen hashtags help improve the performance.

decreases as k increases. We speculate it is because ViT is optimized by cross-entropy loss and therefore learns to predict a sharp distribution, while VLM applies a contrastive loss and outputs a more flattened distribution. The sharp distribution lets ViT perform better when k is small, and the flattened distribution makes VLM more stable when k grows. We further discuss the effectiveness of VLM retrieved hashtags in Section 5.2.

5.2 Effectiveness of Guidance Signal

To explore how different guidance signals can affect generation performance, we first use hashtags selected by ViT to create the ViT-Guided Generative Model (GGM$_{ViT}$) as a comparison of the GGM. As we can see from Table 3, the performance of GGM$_{ViT}$ is lower than the GGM when using the top-50 hashtags as the guidance, which contradicts the recall of ViT and VLM, as shown in Table 4. We conjecture that VLM-based Hashtag Retriever is trained on a contrastive learning loss which makes it provides more robust features for the next step generative model. Additionally, Table 3 depicts the performance over different numbers of hashtags in the guidance signal. Note that the result of GGM k=0 can be regarded as an ablation study of our approach without the guidance signal. We find that GGM guided with top-50 hashtags outperforms others. We speculate that fewer hashtags reduce the information in the guidance signal, while numerous hashtags will introduce extra noise to the model.

5.3 Performance on Unseen Test Set

To simulate the new trending hashtags, we construct an unseen test set for SFVD1 and investigate the models' performance on this test set. Since classification models will never predict unseen hashtags, we only evaluate the generative models for comparable performance. Table 2 and 5 show that both VG-BART and GGM perform much worse

Figure 5: Examples of the generated hashtags on SFVD2 test set with novel hashtags highlighted.

Hashtags	understandability	relevancy
Ground truth hashtags	4.59	3.13
Generated novel hashtags	4.35	3.58

Table 6: Human evaluation of the novel hashtags generated by GGM on the SFVD1 seen test set.

in the unseen test set than in seen test set. When we add the unseen hashtags in the hashtag retrieval pool for creating the guidance signal, GGM_{unseen} achieves better scores in all metrics. As discussed in Section 4.1, seen hashtags could also appear in the unseen test set due to the strong correlations between labels. To explicitly investigate the model's performance on those unseen hashtags, we calculate the number of unseen hashtags recalled at least once. Results show that GGM has recalled 13 unseen hashtags at least once. When we include the unseen hashtags in the hashtag retrieval pool, GGM_{unseen} recalls 51 unseen hashtags at least once. There are still more than 90% of the unseen hashtags which have never been recalled, indicating that generating unseen hashtags is challenging even with the guidance signal.

5.4 Novel Hashtag Analysis

One advantage of the generative models is that they can create novel hashtags that never appeared in the training set. These novel hashtags are valuable because they increase the diversity of the hashtag recommendation and could become new trends in the social media platform. However, our quantitative results focus on word overlap, which might underestimate the effectiveness of the novel hashtags. Thus, we conduct a case study and human evaluation on the generated novel hashtags.

Case Study We present a case study of the hashtags generated by GGM, shown in Figure 5. It is clear that the model can capture the video and its description to generate novel and meaningful hashtags. For instance, GGM generates #willowtree based on the video description in the first case and creates #cold-weather relevant to the video frames in the second case. Neither of the novel hashtags was in the training set, but they are still meaningful and relevant to the video and description.

Human Evaluation Novel hashtags should be understandable and relevant to the video and description so that users can easily access the correct information. Similar to (Simig et al., 2022), we randomly sample 100 novel hashtags created by GGM and manually evaluate their understandability and relevancy. For a fair comparison, we include and mix 100 ground truth hashtags from the same videos in the human evaluation. Assessments are scored on a scale of one to five, with higher scores being better. Each sample is evaluated by three people, and we average the three scores as the final result. As illustrated in Table 6, both generated novel hashtags and ground truth hashtags achieve high understandability scores (larger than four), indicating that hashtags created by our model are meaningful. Interestingly, our generated hashtags are significantly more relevant to the corresponding video and its description with p-value$<$0.05. Further analysis shows that ground truth hashtags consist of many generic hashtags such as "#funnny", "#follow" and "#remake", which are not closely related to the specific video content. In contrast, the generated novel hashtags can capture more details of the video, better representing the salient information in the video. Hence, our GGM is able to generate novel and meaningful hashtags to improve the diversity of recommended hashtags.

6 Conclusion and Future Work

In this paper, we formulate the short-form video hashtag recommendation (SVHR) as a generation task, and we propose a Guided Generative Model (GGM) that generates hashtags from multimodal inputs and guided signals from a VLM-based Hashtag Retriever. Our work benchmarks classification and generative models on SVHR datasets and highlights the advantage of using generative models. Our GGM achieves state-of-the-art performance, and the human evaluation results show that GGM is able to generate meaningful novel hashtags comparable to ground-truth hashtags. We hope our work can catalyze research on using generative models for SVHR.

For future work, since the hashtag recommendation is a highly concurrent task in the real-world application, we believe improving computational efficiency to strike a balance between accuracy and efficiency is one of the important directions. Besides, the popular trend of short-form videos changes rapidly on the internet, so it's important for systems to accurately generate hashtags for new trending videos.

7 Limitations

Our methods are currently trained and tested on two SVHR datasets. Gender biases and unethical hashtags could exist in the datasets, which may cause the model trained on these datasets to generate these biases. Besides, although our methods are not language-specific, we only choose the English dataset due to its rich resource. Furthermore, we regard the user tags as the hashtags for SFVD2 in our experiments and there are small differences between the user tags and hashtags. Experiments on more diverse languages and datasets are needed in the future.

As an initial work for SVHR, in our task formulation, our model only take the video and its description as input to predict hashtags and ignore user preference in hashtag recommendations. Exploring user preference is also a promising direction for future work.

We used up to eight A100 GPUs per experiment and it took more than one day to run experiments on SFVD2. More efficient models are needed for real-world applications. Besides, we hope our experimental results can be used as benchmarks for comparison in future work to avoid repeating training.

8 Ethics Statement

Although the generative models can create novel hashtags, which is beneficial for increasing hashtag diversity on social platforms, generative models also have the potential to generate toxic or offensive hashtags. Since we finetune the Pre-trained language models (PLMs) on existing SVHR datasets, undesirable hashtags could come from biases that are encoded in the PLMs (Blodgett et al., 2020) or the undesirable hashtags in the training set. We recommend that when generative models are deployed in real-world applications, additional post-processing should be carried out to remove undesirable and harmful hashtags. In addition, our hashtag generator will only recommend hashtags to human content creators who ultimately have the responsibility to decide which hashtags should be used.

References

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Da Cao, Lianhai Miao, Huigui Rong, Zheng Qin, and Liqiang Nie. 2020. Hashtag our stories: Hashtag recommendation for micro-videos via harnessing multiple modalities. *Knowledge-Based Systems*, 203:106114.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. 2021. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*.

Andreas Fürst, Elisabeth Rumetshofer, Johannes Lehner, Viet T Tran, Fei Tang, Hubert Ramsauer, David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto, et al. 2022. Cloob: Modern hopfield networks with infoloob outperform clip. *Advances in neural information processing systems*, 35:20450–20468.

Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. 2022. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5067–5077.

Yuyun Gong and Qi Zhang. 2016. Hashtag recommendation using attention-based convolutional neural network. In *IJCAI*, pages 2782–2788.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2018. Search engine guided neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, pages 3929–3938. PMLR.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2017. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3154–3160.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Nabil Hossain, Marjan Ghazvininejad, and Luke Zettlemoyer. 2020. Simple and effective retrieve-edit-rerank text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2532–2538.

Gautier Izacard and Édouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880.

Kirti Jain and Rajni Jindal. 2020. A survey on hashtag recommendations. In *Conference of Open Innovations Association, FRUCT*, 27, pages 323–327. FRUCT Oy.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. 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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Mengmeng Li, Tian Gan, Meng Liu, Zhiyong Cheng, Jianhua Yin, and Liqiang Nie. 2019. Long-tail hashtag recommendation for micro-videos with graph convolutional network. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 509–518.

Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. Trocr: Transformer-based optical character recognition with pre-trained models. *arXiv preprint arXiv:2109.10282*.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. *arXiv preprint arXiv:1902.04911*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Shang Liu, Jiayi Xie, Cong Zou, and Zhenzhong Chen. 2020. User conditional hashtag recommendation for micro-videos. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2021. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*.

Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. 2018. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196.

Swapneel Mehta, Somdeb Sarkhel, Xiang Chen, Saayan Mitra, Viswanathan Swaminathan, Ryan Rossi, Ali Aminian, Han Guo, and Kshitiz Garg. 2021. Open-domain trending hashtag recommendation for videos. In *2021 IEEE International Symposium on Multimedia (ISM)*, pages 174–181. IEEE.

Rui Meng, Xingdi Yuan, Tong Wang, Sanqiang Zhao, Adam Trischler, and Daqing He. 2021. An empirical study on neural keyphrase generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4985–5007.

Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. 2017. Deep keyphrase generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 582–592.

Christian Montag, Haibo Yang, and Jon D Elhai. 2021. On the psychology of tiktok use: A first glimpse from empirical findings. *Frontiers in public health*, 9:641673.

Jesús Andrés Portillo-Quintero, José Carlos Ortiz-Bayliss, and Hugo Terashima-Marín. 2021. A straightforward framework for video retrieval using clip. In *Mexican Conference on Pattern Recognition*, pages 3–12. Springer.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Daniel Simig, Fabio Petroni, Pouya Yanki, Kashyap Popat, Christina Du, Sebastian Riedel, and Majid Yazdani. 2022. Open vocabulary extreme classification using generative models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1561–1583.

Baptist Vandersmissen, Fréderic Godin, Abhineshwar Tomar, Wesley De Neve, and Rik Van de Walle. 2014. The rise of mobile and social short-form video: an in-depth measurement study of vine. In *Workshop on social multimedia and storytelling*, volume 1198, pages 1–10. Citeseer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yue Wang, Jing Li, Hou Pong Chan, Irwin King, Michael R Lyu, and Shuming Shi. 2019a. Topic-aware neural keyphrase generation for social media language. *arXiv preprint arXiv:1906.03889*.

Yue Wang, Jing Li, Irwin King, Michael R Lyu, and Shuming Shi. 2019b. Microblog hashtag generation via encoding conversation contexts. 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 1624–1633.

Yinwei Wei, Zhiyong Cheng, Xuzheng Yu, Zhou Zhao, Lei Zhu, and Liqiang Nie. 2019. Personalized hashtag recommendation for micro-videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1446–1454.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. 2022. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE.

Chao Yang, Xiaochan Wang, and Bin Jiang. 2020. Sentiment enhanced multi-modal hashtag recommendation for micro-videos. *IEEE Access*, 8:78252–78264.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3995–4007.

Xingdi Yuan, Tong Wang, Rui Meng, Khushboo Thaker, Peter Brusilovsky, Daqing He, and Adam Trischler. 2020. One size does not fit all: Generating and evaluating variable number of keyphrases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7961–7975.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Ying Shan, Bing Li, Ying Deng, and Weiming Hu. 2021. Open-book video captioning with retrieve-copy-generate network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9837–9846.

Xiuwen Zheng, Dheeraj Mekala, Amarnath Gupta, and Jingbo Shang. 2021. News meets microblog: Hashtag annotation via retriever-generator. *arXiv preprint arXiv:2104.08723*.

A Implementation Details

We initialize the models from Hugging Face Models. The name of the initial checkpoint and the number of trainable parameters for each model is shown in Table 7. We use learning rates $6e^{-5}$ following (Lewis et al., 2020a) and Adam optimizer (Kingma and Ba, 2014) to fine-tune the GGM. For all of our experiments, we use a batch size of 32. The final results are the average test set performance on the best three checkpoints in the validation set.

For classification baselines, we implement the multi-label classification models from the pre-trained vision and language models (e.g., ViT, BERT) because each short-form video could have multiple ground-truth hashtags. Firstly, we create a hashtag set that contains all hashtags in the training set (i.e., 10,674 hashtags for SFVD1 and 43,282 hashtags for SFVD2) as the candidates for the classification outputs. Then, similar to (Mahajan et al., 2018), we compute probabilities over the hashtag set using a softmax activation, and the models are trained to minimize the cross-entropy loss between

Model	Checkpoint	# of parameters
ViT	google/vit-base-patch16-224	94 M
BERT	bert-base-uncased	109 M
BART	facebook/bart-base	134 M

Table 7: Initialization checkpoint and number of trainable parameters for each model.

the predicted softmax distribution and the target distribution of each short-form video. The target distribution is a vector that only has k non-zero entries, each set to $1k$ corresponding to the ground-truth hashtags for the video. We also implement the multi-label classification models with per-hashtag sigmoid outputs and minimize each hashtag's average binary cross-entropy loss. However, the results are significantly worse; actually, we find the models only predict high-frequency hashtags for every test set sample.

For generative models, we follow the standard sequence generation models that generate the hashtags token by token. Decoding automatically stops when the end-of-sequence token is predicted. For Trocr-fid, we use the pre-trained ViT-base model to initialize the vision encoder and use the decoder of the BART-based model to initialize the text decoder.

B Implementation Details for Evaluation

B.1 Automatic Evaluation

For ROUGE and BERTScore, we randomly shuffle the predicted hashtags to remove the effect of hashtag order. Moreover, we split the multi-word hashtags into single words before calculating the ROUGE. We use `rouge` [3] to compute ROUGE scores and use `microsoft/deberta-xlarge-mnl` model to compute BERTScores as suggested [4]. `wordninja` [5] is used for separating words from the hashtags.

B.2 Human Evaluation

In Table 6, we conduct a human evaluation of the understandability and relevancy of the generated hashtags from the SFVD1 dataset. In detail, we randomly sample 100 novel hashtags from GGM and the ground truth hashtags from the same videos for comparison. Assessments are scored on a scale of one to five, with higher scores being better. Understandability means whether the hashtag is understandable given the context of the video and the corresponding description. Relevancy means whether the hashtag is relevant to the video or the corresponding description. Note that the annotators can search online for more information about the hashtags if they don't know it. We assign each hashtag to three annotators and take the average score as the final result. In total, we used six annotators from the US, and all annotators voluntarily participated in the human evaluation. All annotators agree to have their evaluation results included as part of this paper's results.

Here is an extra note for the annotators when they do the human evaluation. For evaluating relevancy, sometimes the hashtags can be very generic and the annotators should give lower scores for them. For example, given a family comedy show, the relevancy score of the hashtags will be "#familycomedy" > "#comedy" > "#room".

[3] https://huggingface.co/spaces/evaluate-metric/rouge

[4] https://github.com/Tiiiger/bert_score

[5] https://github.com/keredson/wordninja

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7 "Limitation"

☑ A2. Did you discuss any potential risks of your work?
Both Section 7 "Limitation" and Section 8 "Ethics Statement"

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Yes, we summarize the paper's main claims in "Abstract" and Section 1 "Introduction"

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

We create NLP models in Section 3 "Methodology"

☑ B1. Did you cite the creators of artifacts you used?
Section 2 and Section 3 and Section 4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In Appendix C, we discuss the license of the dataset that we use.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 3, Section 4 and Section 5

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Appendix C

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 1, Section 2 and Section3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
In Section 4.1 "Datasets" we discuss the data

C ☑ Did you run computational experiments?

Sections 4 and 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4 and Appendix A

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4 and Appendix B1

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 5

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix B2

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix B2

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix B2