Understanding Human Preferences: Towards More Personalized Video to Text Generation

Anonymous Author(s)*

ABSTRACT

While previous video to text models have achieved remarkable successes, they mostly focus on how to understand the video contents in a general sense, but fail to capture the human personalized preferences, which is highly demanded for an engaging multimodal chatbots. Different from user modeling in collaborative filtering, there is no other user behaviors in inference as a real-time video stream is coming. In this paper, we formally define the task of personalized video commenting task and design an end-to-end personalized framework for solving this task. In specific, we argue that the personalization for video comment generation can be reflected in two aspects, that is, (1) for the same video, different users may comment on different clips, and (2) for the same clip, different people may also express various opinions with diverse commentary styles. Motivated by these considerations, we design our framework based on two components. The first one is a clip selector, which is responsible for predicting the clips that the user may comment in the video. The second one is a text generator, which aims to produce the comment based on the above predicted clips and the user's preference. In our framework, these two components are optimized in an end-to-end manner to mutually enhance each other, where we design confidence-aware scheduled sampling and iterative inference strategies to solve the problem that the ground truth clips are absent in the inference phase. As the absence of personalized video to text dataset, we collect and release a new dataset for studying this problem. We conduct extensive experiments to demonstrate the effectiveness of our model.

KEYWORDS

video to text generation, user preference modeling, video comments dataset, personalized content generation, multimodal interaction

1 INTRODUCTION

With the advancement of Large Language Models (LLM), LLM based multi-modal models, such as Flamingo [1], BLIP-2 [15], and miniGPT-4 [42], have demonstrated excellent capability to describe an image precisely and with details. Such capabilities are fundamentally useful to test how well a bot can understand vision information; however, they cannot add much value when the bot is expected to accompany human users with shared vision content. Instead of answering some questions on what can be seen, a multimodal bot is more desired if it can select an appropriate moment based on vision and proactively give an interesting comment, as a family member or a friend can do.

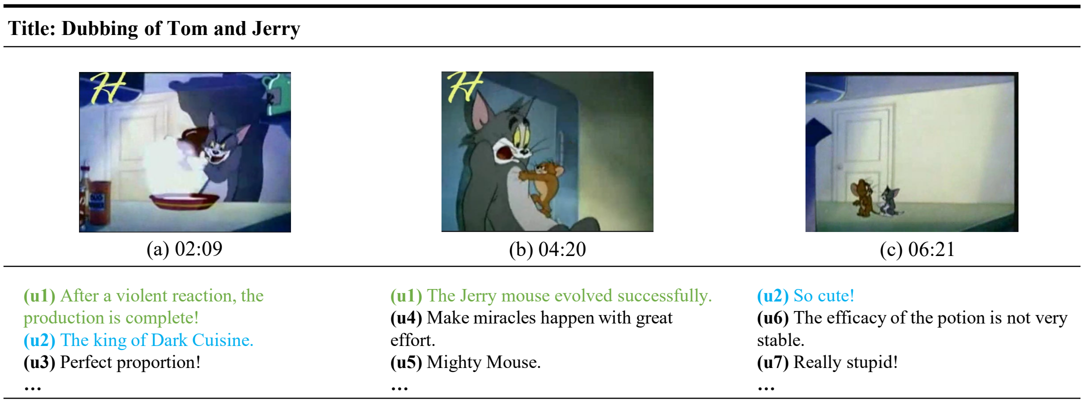
Video to text generation has attracted increasing attention from both academic and industry communities. Standard video to text generation tasks include video captioning [13, 18, 38], video question answering [10, 11, 35], and video comment generation [9, 21, 22, 31, 33]. They mostly focus on how to better represent the videos to generate more accurate texts while neglecting personalization in both selecting an appropriate visual moment and generating a styled text. Recently, some works [24, 28, 37, 41] generate engaging text with diverse style from image. They model user preference by personality traits, user identity, and text style. However, they do not model user preference in video scenarios, without considering user preference combining visual information. Different people may select different parts of a video for captioning and describe them in different ways, which have not been addressed yet.

Taking real-life communication among humans as an example, the human personalities play important roles in their interested clip and generated text. As shown in Fig. 1, different users may comment on various clips of the same video, *e.g.*, $u1$ comments on clip (a) and (b), while u2 comments on clip (a) and (c). Even for the same clip, different users may also have various styled comments. For example, $u1$ usually expresses her opinions by a long text with more details, while $u2$ prefers shorter and more succinct comments. Thus, $u1$ comments clip (a) and (b) by "After a violent reaction, the production is complete" and "The Jerry mouse evolved successfully", respectively, while, for clips (a) and (c), $u2$ only says "the king of dark cuisine" and "So cute". All the above examples suggest that the comments with obvious personal style makes communication more interesting. Personalized style is reflected not only in the textual style but also in the selection of video clips.

To enhance engaging multimodal chatbot performance, in this paper, we formally define the problem of personalized video commenting. While the task seems to be interesting, it is non-trivial due to the following challenges: to begin with, in real scenarios with multimodal bot companionship, visual information is input in the form of a stream rather than ready-made videos. This requires solutions solely make use of users' history and the vision content before this moment. Collaborative filtering methods are not applicable here because there is no comment from other users available in real-time applications. Second, there is few previous work on incorporating user preferences into the video to text generation process. Thus, incorporating what types of user information as user preference and how to incorporate such preference may challenge our model designs. Third, generating comments for a video includes two steps: (1) selecting the clips to comment and (2) generating comments based on the selected clips. Ideally, these two steps should be able to enhance each other. More accurately selected clips can lead to more consistent comments with the ground truth. If one can accurately predict the comments for a clip, then given

WWW '24, May 13 – 17, 2024, Singapore

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX



Title: Dubbing of Tom and Jerry

(a) 02:09

(b) 04:20

(c) 06:21

(u1) After a violent reaction, the production is complete!
(u2) The king of Dark Cuisine.
(u3) Perfect proportion!
...

(u1) The Jerry mouse evolved successfully.
(u4) Make miracles happen with great effort.
(u5) Mighty Mouse.
...

(u2) So cute!
(u6) The efficacy of the potion is not very stable.
(u7) Really stupid!
...

Figure 1: An example of comments from different users posted on clips of the video – " Dubbing of Tom and Jerry".

the ground truth, she can also identify which clips the comments should belong to. However, the clip selection process is a discrete operation, how to soften such operation to build an end-to-end framework to mutually enhance the above two steps needs our careful designs.

To overcome the above challenges, we design an end-to-end framework to incorporate the user personalized preferences into the video comments generation process (called **PVCG** for short). In specific, our framework is composed of two components. The first one is a clip selector, which aims to predict the clips that the user would like to comment on by three user preference modeling methods. We input the comment and all the clips of a video, and output the probabilities of all the comment-clip pairs, which are compared with the ground truth clip in the dataset for optimization. The second one is a comment generator, which is responsible for producing the comments according to the above predicted clips and the user preference. To bridge these components, we soften the clip selector by leveraging softmax to weighted average the visual features of all the clips in a video, and the obtained results are input into the comment generator. Since softmax is a differentiable operation, the above two components can be optimized in an end-to-end manner. Considering that the ground truth for the selected clips and user comments are not available in the testing phase, we design the strategies of **scheduled sampling** and **iterative inference** to bridge the gaps between the model training and testing processes.

The main contributions of this paper are as follows:

- We highlight the importance of personalization in the problem of video to text generation, and formally define the task of personalized video commenting in real-time accompanied scenarios.
- To solve the above problem, we propose a personalized framework, which contains a clip selector and a comment generator, and design strategies to enhance the quality of generated text.
- We release a new dataset for studying the problem of personalized video comments generation. We conduct extensive experiments on proposed datasets to demonstrate the effectiveness of our model.

2 RELATED WORK

We divide related works into three categories.

2.1 Video to Text Generation

Many works focus on exploring video to text generation tasks, including video captioning, video question answering, and video comments. Video captioning [13, 18, 38] takes video frames as input and generates a relevant description. Video question answering [10, 11, 35] asks machines to answer the question based on the video content. Video commenting generates human-like comments according to the video context. LiveBot [22] first proposes the video comment generation task. It constructs a large-scale video comment dataset and generate comments using surrounding comments from other users and visual content. Following LiveBot, Lv et al. [21] proposes an Embedding-based Generative Adversarial Framework to bridge the gap between visual and textual content. Also, some works [9, 31] introduce additional audio information, using a multi-modal and multi-task architecture to capture the relationships among comments, vision, and audio, generating comments with these information. Although Wu et al. [33] explores comments generation with few surrounding comments, it still ignores personal preference when selecting video clips and generating comments.

2.2 Personalization in Vision and Language Research

Despite many works focusing on standard video to text generation, some works explore personalization in vision and language research to meet users' real demand. AttendToYou [24] incorporate each user's active vocabularies into memory networks to capture their writing style. EICP [28] specifies 215 different personality traits as user preference. MHTH [37] learns user's short-term and long-term literal-preference by user identity and user recent comments respectively. SACO [41] proposes a style-aware triplet contrast learning method to generate text with desired style from image. These works consider the importance of user literal-preference in generating personalized text. However, they ignore the user preference in visual information selection. Some works [3, 32, 34] encode user preference based on both visual and textual information for recommending personalized key frames. But they do not further explore personalized text generation in video scenarios. Different from an image, a video always contains more diverse information. Thus, different people may select different parts of a video for captioning and describe them in different ways, which have not been addressed yet.

2.3 Explainable Recommendation Considering User Preference

Explainable recommendation refers to personalized recommendation algorithms that not only provide users with recommendation results, but also explain why such items are recommended. Similar with personalized video to text generation, it learns to perform rating prediction and user-dependent sentences related to this prediction. Explainable recommendations naturally involve humans in the loop. Similar to personalized video to text generation, explainable recommendation generates user-dependent explanation considering modeling users' behavior.

Based on the different forms of explanation, explainable recommendation systems can be categorized into many types. User/item-based explanations are usually provided based on users' feedback [4, 36]. Feature-based explanation using content-based recommendation methods to provide explanation [12, 39]. Besides, as users' reviews and social media posts contains their opinion, some works show that such information is quite beneficial in user preference modeling and recommendation [19, 27, 30]. Based on the development of natural language generation techniques and computer vision systems, many explainable recommendation systems generate explanation sentences using knowledge from other modality. They utilize item images or social information to generate explanation sentences [2, 20, 25]. User preference modeling is well-discussed in explainable recommendation systems, which can be applied to personalized video to text generation tasks. The difference is that for a new video in our scenarios, there is no other user behaviors available associated with the video, which is more challenging.

3 PERSONALIZED VIDEO COMMENTING TASK

In this section, we formally define the new problem of personalized video commenting and describe how we construct a dataset to support such kind of research.

3.1 Problem Formulation

We imagine that in the future a chatbot can accompany a lonely user watching films or walking together with camera opened. The chatbot can make funny comments whenever it finds some interesting points. Thus the user will never feel lonely. Moreover, in such a scenario, a chatbot can provide more emotion values if it can play as a character with a personalized taste of selecting an interesting clip and making styled comments.

Suppose we have a user set $\mathcal{U}$ and a video set $\mathcal{V}$. For each video $v \in \mathcal{V}$, we segment it into K clips, that is, $v = \{v_1, v_2, ..., v_K\}$. The user comments are collected in $D = \{(u_i, v_{ik}, t_i)\}_i^N$, where each element (u_i, v_{ik}, t_i) means that user u_i comments on the k-th clip v_k of video v with text t_i.

Given D, our task is to learn a model f, such that, for a new video v^*, it can accurately predict which clip of the video will be chosen and how it will be commented by each user corresponding to their personality. In order to accomplish this task, the model needs to firstly predict which clip the user may comment on, and then generate the comment contents.

It should be noted that for the new video there is no other users' behaviors available for collaborative filtering. Only the video content and the user's history in training data can be exploited. This is a big difference from user behavior based recommendation.

Table 1: Overall statistics of the Personalized VideoIC dataset.

Total number of users	7,065
Total number of videos	4,036
Total number of clips	94,831
Total number of comments	263,835
Ave. comments per video	65.4
Ave. clips per video	23.5
Ave. comments per user	37.3

3.2 Personalized Video Commenting Dataset

We build a new dataset to support research on our proposed new task. In specific, our dataset is built upon VideoIC [31], which is a time-synchronized comment (TSC) dataset, containing 4,951 videos and 5 million comments. However, the user IDs are not provided in VideoIC. Therefore, we crawl user IDs and corresponding comments from the website [1] where VideoIC is collected from and construct Personalized VideoIC dataset.

For each video, we segment it into clips every ten seconds, and each record in our dataset is a tuple (u, v, c, t), where u is a user, $v = \{v_1, ...v_K\}$ is a set of clips in a video, c is the index of the user selected clip and t is the comment posted by the user. If a user has multiple comments for a video, then we separate them into different records. Thus only one clip is commented in each record. We remove users with less than twenty comments and their corresponding records.

The final statistics of our dataset are summarized in Tab. 1. We split the data into 3636, 200, and 200 videos for training, validation, and testing respectively. The videos in the test set never appear in the training set or validation set, which ensure they are new. Unlike certain personalized image to text datasets [28], our Personalized VideoIC dataset does not rely on personality traits that are manually labeled by humans. A user's personality is implicitly represented by her selected clips and posted comments in the training set. If a method can better model the user, it will achieve better performance in predicting her selected clips and generated comments in the test set.

4 OUR PROPOSED PVCG MODEL

In this section, we introduce our model for solving the personalized video commenting problem in detail.

[1] https://www.bilibili.com/

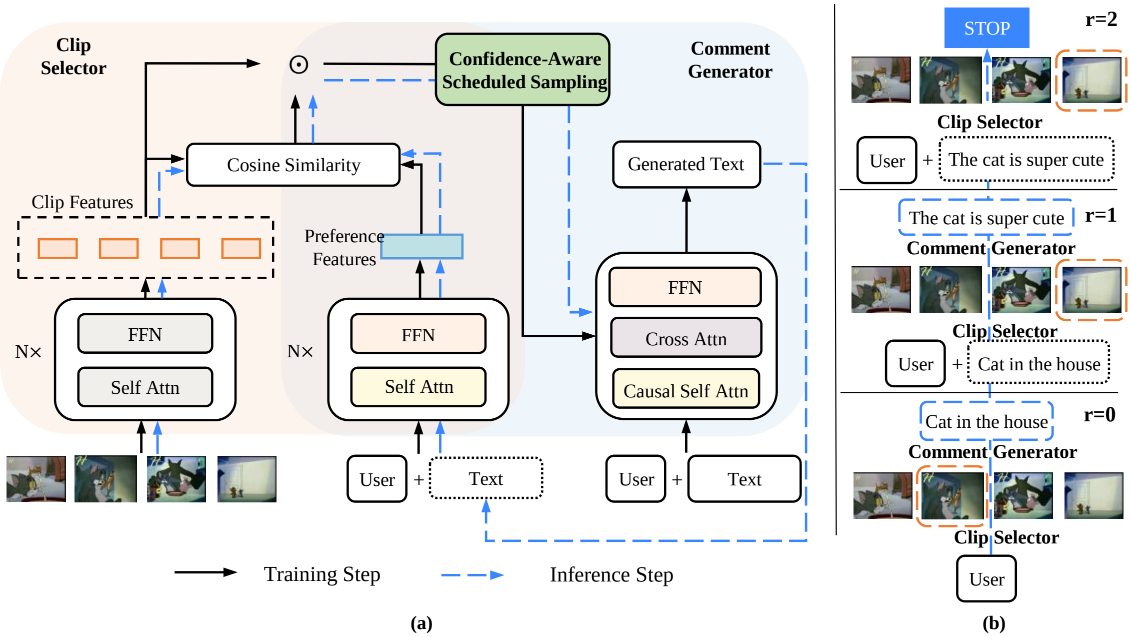


Figure 2: Overview of our proposed PVCG framework. (a) shows that the PVCG contains two modules: clip selector and comment generator. The black solid line represents the training stage, and the blue dashed line represents the inference stage. (b) illustrates iterative comment inference through an example. It starts with a user preference feature to select a clip, here is the second one, and then generates a comment "cat in the house" in the first round. Then concatenating the user ID with the generated comment can select the fourth clip and then generate another comment "the cat is super cute" at the second round. Finally concatenating the user with the new comment still selects the fourth clip at the top and thus stops the process of iterative comment inference at the third round.

4.1 Framework Overview

As shown in Fig. 2 (a), our proposed PVCG framework consists of two main components: the clip selector to predict the clip that the user may comment in a given video, and the comment generator to produce the comments on the predicted clip. The clip selector can match clip features with user preference features by optimizing a contrastive loss during pre-training and further select users' most interested clip in a given video by optimizing a cross-entropy loss. The comment generator is designed to estimate textual comment of a user on the selected video clip and user preference by optimizing a generation loss. In the example shown in Fig. 2 (b), the clip selector may select the second clip and the comment generator generate "Cat in the house" in the first round.

We argue that these two components can mutually enhance each other to predict the final user comments. On one hand, based on an accurate enough clip selector, the comment generator can find more accurate visual features of the clips, and thus can predict better comments. On the other hand, if the comment generator is sufficiently well, then the clip selector can be more accurately supervised, since a wrong clip can be more likely to lead to an inaccurate comment.

If the clip selector and comment generator are optimized separately, it fails to influence and mutually enhance each other. To build a bridge between them, we soften the prediction results of the clip selector, and leverage them as the input of the comment generator to derive an end-to-end framework and propose a confidence-aware strategy to schedule the process of training the model. In addition, we propose an iterative strategy to inference comment. Thus the generated comment in the first round can provide more information to user preference features and thus select a better matched clip in the second round. Later the clip can generate better comment then. For example, in Fig. 2 (b), the fourth clip, which is the ground truth clip, is selected in the second round. Thus the comment generator outputs a new comment "The Chat is super cute". Finally, the iterative comment inference stops in the third round when the selected clip is the same to that selected in the second round.

4.2 User Preference Modeling

User preference for a given video is related to both user's static interest and the content of current video clips. Thus, combining user's static preference and their dynamic preference for current video is crucial in modeling user preferences. Based on this, there are three methods to model user preferences.

(1) **User identity.** Users who post their comments online are typically associated with user IDs. We aim to leverage the IDs to learn static user preference. As shown in Fig 3 (a), we formulate the input format as "The comment of user <user id>" to indicate the user identity in both training and inference stage.

(2) **User's historical behavior.** Another way to model user's static preference is leveraging their historical behavior/interests. Here, we consider the video clips that users have commented on, along with their corresponding comments, as their historical interests. As shown in Fig 3 (b), we use an image encoder and a text

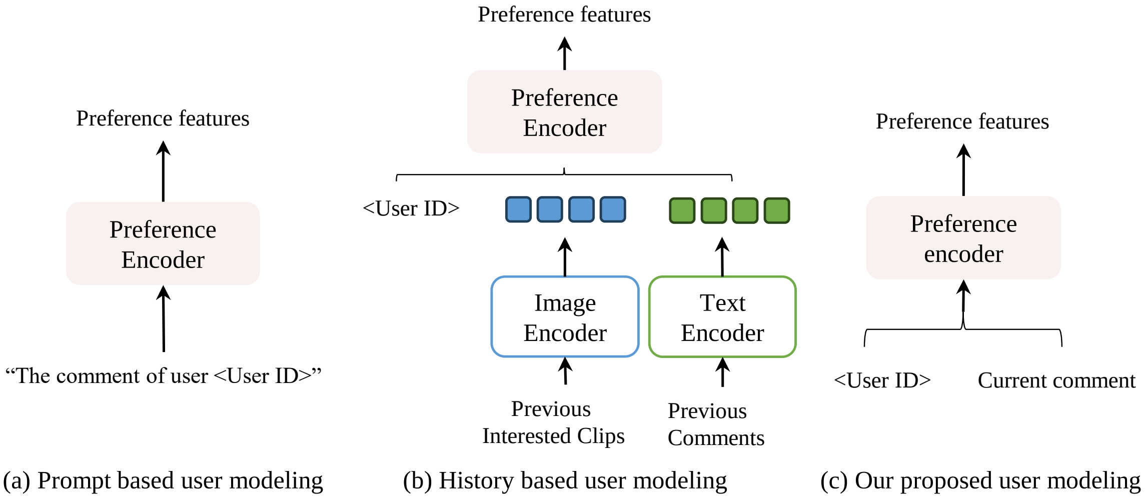


Figure 3: Illustration of different user preference modeling method.

encoder to encode these historical information respectively and concatenate them into a sequence to model users' preference.

(3) **User identity with current comment.** User's preference is not only related to her/his static preference, but also related to current videos. We map user identity to learnable user embedding, learning user's static preference. Also, we model user's dynamic preference from user's comments for their interested clips (As shown in Fig 3 (c)).

4.3 Clip Selector

The clip selector aims to predict where the user may comment on a video. The inputs include two parts: (1) the visual embeddings of all the clips of a video, and (2) the user preference features. The output is the distribution that reflects the selection preference of the user on the clips. Formally, for a video v with clips $\{v_1, v_2, ..., v_K\}$, we firstly use a visual transformer [8] to project each clip v_k into an embedding, that is:

$$e_{v_k} = \mathrm{ENC}_V(v_k), \tag{1}$$

where ENC_V is implemented based on the architecture introduced in [8]. $e_{v_k} \in \mathbb{R}^d$ is the obtained d-dimensional visual embedding for v_k. We use $E_v \in \mathbb{R}^{K\times d}$ to represent the embeddings of all the visual features of the clips in v.

For a user u, suppose she has posted a comment t for the k-th clip of video v, then we leverage the preference encoder to derive the user preference features. It learns both dynamic user preference which is related to current video and static user preference which is consistent among all videos. Here we leverage user ID to learn static preference and model dynamic user preference from comment t. The preference encoder is implemented based on a deep bidirectional Transformer [7], where we use a learnable look-up table mapping user ID to embedding, and append it before the comment t's word embedding sequence, that is:

$$e_u = \mathrm{ENC}_U(u, t), \tag{2}$$

where ENC_U is the encoder for generating the user preference features e_u.

The training of clip selector can be divided into two stages: In the first stage, we want to make sure the clip selector have the ability of understanding visual information and user preference representation. Therefore we use contrastive learning by encouraging positive clip-preference pairs to have similar representations in contrast to the negative pairs. We regard user preference features and clip features which the user is interested in as the positive pair, while the others in a mini batch are negative pairs. We sample negative clips across the entire dataset, rather than just sampling within the same video to avoid false negatives. In the second stage, we encourage clip selector to learn user's most interested clip in a given video. For the dataset D, we optimize the clip selector based on the following cross-entropy loss:

$$\mathcal{L}_{select} = -\sum_{(u,v_k,t)\in D} z_k \log g(e_{v_1}^T e_u, e_{v_2}^T e_u, ..., e_{v_K}^T e_u), \tag{3}$$

where z_k is a K-dimensional one-hot vector with the k-th element equal to 1. g is the softmax operator.

4.4 Comment Generator

The comment generator DEC_G aims to estimate the textual contents of the comment of a user on a video clip. Formally, suppose the visual embedding of the clip is $\hat{e}$, then the comment is generated based on a multi-modal text decoder with the user embedding as the initial input. We insert a cross-attention layer between the casual self-attention layer and the feed forward layer to integrate user and clip information $\hat{e}, e_u$ into the generated comment.

Suppose the words in comment t is $\{t_1, t_2, ..., t_l\}$, then the objective for optimizing the comment generator is:

$$\mathcal{L}_{gen} = -\sum_{(u,v_k,t)\in D} \sum_{i=1}^{l} \log p\ (t_i | t_{<i}, \hat{e}, e_u), \tag{4}$$

where p represents the probability of the next word given the previous ones based on BERT.

Instead of optimizing the two losses separately, we soften the prediction results of the clip selector, and concatenate user preference features as the input of the comment generator to derive an end-to-end framework. Formally, we let

$$\hat{e} = \sum_{k=1}^{K} [g(e_{v_1}^T e_u, e_{v_2}^T e_u, ..., e_{v_K}^T e_u)]_k e_{v_k}, \tag{5}$$

where $[g(e_{v_1}^T e_u, e_{v_2}^T e_u, ..., e_{v_K}^T e_u)]_k$ is the k-th element of the softmax output. This equation actually weighted averages all the embeddings of the video clips.

Since softmax is a differentiable operator, the final supervision signal (*i.e.*, the user comments) can be back propagated to influence the parameters of the clip selector via $\hat{e}$. Finally, the complete objective of our model is:

$$\mathcal{L} = \mathcal{L}_{select} + \mathcal{L}_{gen}. \tag{6}$$

4.5 Confidence-Aware Scheduled Sampling

While in the above formulation, the clip selector and comment generator can be optimized in an end-to-end manner, the weights $[g(e_{v_1}^T e_u, ..., e_{v_K}^T e_u)]_k$ leveraged to derive $\hat{e}$ can be inaccurate in the initial stage of model optimization, which may lower the final performance. To solve this problem, we propose a confidence-aware strategy to reschedule the model training process. Our general idea is that, when the confidence of the clip selector is not sufficiently high, we use the ground truth signal to guide the comment generator learning process. When the clip selector is more confident for its predicted results, then we use $\hat{e}$ as the input of the comment generator, which is more consistent with the model inference stage. Formally, we use the loss L_{select} to evaluate the confidence of the clip selector. Then, we have:

$$\hat{e} = \begin{cases} \sum_{k=1}^{K} z_k e_{v_k}, \text{ if } L_{select} > \theta, \\ \sum_{k=1}^{K} [g(e_{v_1}^T e_u, e_{v_2}^T e_u, ..., e_{v_K}^T e_u)]_k e_{v_k}, \text{ else,} \end{cases} \tag{7}$$

where θ is a threshold to measure whether the confidence of the clip selector is high enough.

The above strategy simultaneously considers the prediction error of the clip selector as well as the consistency between the model training and testing phases. If we always use the ground truth clip as the input of the comment generator, the learned model can be incompatible with the testing environment, where we do not know the clip ground truth. If we leverage $\hat{e}$ as the input of the comment generator in the complete optimization process, then for the cases where the clip selector is less confident, the comment generator may receive inaccurate input signals, which may impact the model optimization. Our designed strategy makes a trade-off between the above two settings, which is shown to be effective in the experiments.

4.6 Iterative Comment Inference

In Eq. 2, the preference feature is derived based on the ground truth comment, which is only available in the training phase. To make the inference, in the first round we only input u into ENC_U, and left t as an empty set. Then, we use the obtained embedding e_u to generate a comment t_0. In the next round, we input u and t_0 into ENC_U to generate a new comment t_1, which is then input into ENC_U again to generate t_2. After R round iteration, we use the output t_R as the final predicted user comment. The complete iterative comment inference process can be seen in Algorithm 1.

Algorithm 1 Iterative Comment Generation

Data : user u_i, a new video $v^* = \{v_1^*, ..., v_K^*\}$, and interation round r.
Result : clip v_p^* user may comment on, and corresponding text t_i.

```
r ← 0
t_i ← None
while not convergence do
    r ← r + 1
    $e_{v_k} = \text{ENC}_V(v_k)$
    $e_u = \text{ENC}_U(u, t_{i,p})$
    $\hat{e} = \sum_{k=1}^{K} [g(e_{v_1}^T e_u, e_{v_2}^T e_u, ..., e_{v_K}^T e_u)]_k e_{v_k}$
    $\hat{t}_i = \text{DEC}_G(\hat{e}, e_u)$
    $t_i \leftarrow \hat{t}_i$
end while
```

The strategy of iterative comment inference bridges the gap of the inconsistent inputs of ENC_U in the training and testing stages. Unlike previous methods [14, 28, 41], which only use an empty set to replace t, we further leverage the generated comments t_1, t_2, ... to enhance the inference stage. Ideally, as the number of iterations becomes larger, the inference stage is more consistent with the training phase, which facilitates more accurate user comment generation.

5 EXPERIMENTS

We focus on the following research questions:

RQ1: Whether our framework can achieve better performance for the task of clip selection?

RQ2: Given the user actually commented clips, whether our framework can achieve satisfactory performance for the task of comment generation?

RQ3: Without the ground truth clips, whether our method can achieve better performance in comment generation task?

5.1 Experiment Setup

5.1.1 Evaluation of Clip Selection. To evaluate how well a method can select a user's interested clip, we form it as a ranking problem and use Recall@k to measure. Recall@k means how many percentage of records a method can retrieve the ground truth clip at top k. We implement baselines with different user preference modeling:

- **MostPopular** This is a non-personalized method. We fine-tune BLIP [16] on the Personalized VideoIC dataset, without using user identity information. In the inference stage, it generates comments using most popular clip of the video. Although in our proposed task, there are no other users' behavior available for a new video, we still implement this method by leaking other users behavior as a reference.
- **BPR-MF** This is a personalized recommendation based method with implicit feedback. Here, we treat clips that users commented as positive items and those they didn't commented as negative items. The clips are ranked based on users' history behaviors,

Table 2: Comparison of clip selection accuracy. † and ∗ denote significant improvements over the Pers (history), PVCG (cascaded) methods results with p-value< 0.05 in t-test respectively. Here, "Pers" refers to "personalized."

Settings	Recall@1 ↑	Recall@5 ↑	Recall@10 ↑
MostPopular	3.45	10.17	22.99
BPR-MF	6.18	13.96	36.04
Pers (user id)	8.04	26.44	52.87
Pers (history)	18.06	**44.44**	59.72
PVCG (cascaded)	18.39	32.18	60.92
PVCG (r=0)	25.35	40.85	60.56
PVCG (r=4)	**27.54**$^{\dagger *}$	42.18*	**65.92**$^{\dagger *}$

without considering any comments information. The code is implemented through RecBole [40].

- **Personalized (user id).** It adopts a prompt-like mechanism to incorporate user preference information (as shown in Fig. 3 (a)). The structure of the user preference encoder is the same as PVCG, but uses prompts instead of user identity as the input. The prompt is set as "The comment of user ⟨user id⟩" in both training stage and inference stage.
- **Personalized (history).** It employs a user history encoder based on Transformer to model the user's historical information (as shown in Fig. 3 (b)). Specifically, we use CLIP [26] to map the video clip and user posted text pair to the same space and concatenate them together as the input sequence to the user preference encoder. Being the same as our PVCG, the encoded user preference feature is used to select video clip and generate comments.
- **PVCG (cascaded).** This is a cascaded version of our proposed PVCG model. The personalized clip selection module and text generation module are optimized separately.

5.1.2 *Evaluation of Comment Generation.*

To evaluate how well the generated comment is, we employ both reference-based metrics and ranking-based metrics to evaluate the generation performance. For reference-based metrics, we use BLEU [23], Meteor [6], CIDEr [29], and ROUGE [17] to evaluate the quality of generated text. For ranking-based metrics, we discriminate a good model according to its ability to rank the correct comments on the top of candidates. We construct a list of candidate comments for each user-clip pair in the test set. We pool all comments in the training set and sample negative comments in four ways:

- **Personalized comments:** The twenty comments from the same user in the training set, which can be considered as text with the same personal style but not necessarily relevant to the clip.
- **Plausible comments:** The twenty most similar comments in the training set to the video titles based on the cosine similarity of their TF-IDF vectors.
- **Popular comments:** the twenty comments randomly selected from the training set with the highest frequency.
- **Random comments:** the forty comments randomly selected from the training set.

Following VideoIC [31], we measure the ranking results with three types of metrics: (1) Recall@k: The proportion of ground truth comments in the top-k sorted text. (2) Mean Rank (MR): The mean rank of the ground truth text. (3) Mean Reciprocal Rank (MRR): The mean reciprocal rank of the ground truth text.

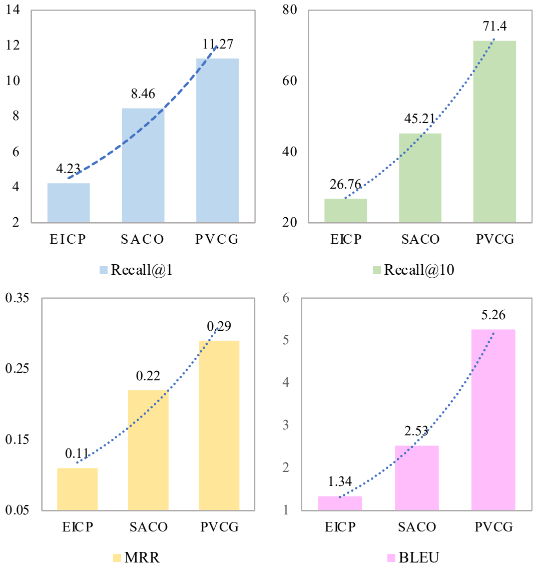


Figure 4: Comparing the quality of generated text generated by different methods given the golden clips.

We compare our model with baseline methods described in Section 5.1.1 and two personalized image caption baselines.

- **EICP [28].** EICP first proposes personalized image caption task. It combines one-hot encoding personality information with standard image caption model by adding it with the text decoder input at each step. To ensure fairness, we use a pre-trained BLIP as the standard image caption model here.
- **SACO [41].** SACO is one of the SOTA models on personalized image caption. It leverages a style-aware contrastive learning to integrate style and visual information into the caption.

As EICP and SACO cannot select personalized clip, we input user actually commented clips (golden clips).

5.1.3 *Model Configuration.*

The vision encoder is initialized from $ViT-B/16$ pre-trained on ImageNet [5], and the user preference encoder and text generator are initialized from $BERT_{base}$. The preference encoder and text generator share all parameters except self-attention layer and cross-attention layer. We load parameters from vision language pre-trained model – BLIP [16], and then fine-tune the model on Personalized VideoIC dataset for 20 epochs using a batch size of 256. We use AdamW optimizer with a weight decay of 0.05, and a cosine learning rate schedule. Please note that we pre-train model with contrastive loss before training to reduce the training difficulty. During pre-training, we use contrastive learning by encouraging positive clip-preference pairs to have similar representations in contrast to the negative pairs. We regard user preference features and clip features which the user is interested in as the positive pair, while the others are negative pairs. We sample negative clips across the entire dataset, rather than just sampling within the same video to avoid false negatives.

Table 3: Comparing the quality of generated text generated by different methods. † denotes significant improvements over the PVCG (cascaded) results with p-value< 0.05 in t-test.

Settings	Ranking-based metrics					Reference-based metrics			
	Recall@1 ↑	Recall@5 ↑	Recall@10 ↑	MR ↓	MRR ↑	BLEU ↑	METEOR ↑	ROUGE_L ↑	CIDEr ↑
MostPopular	1.57	5.63	8.45	40.51	0.06	2.03	2.23	3.16	2.03
Pers (history)	3.63	6.70	14.50	34.78	0.09	3.35	2.33	2.77	3.39
PVCG (cascaded)	4.23	18.31	30.99	28.00	0.13	3.30	2.66	3.87	4.27
PVCG(r=0)	8.45	**49.30**	70.42	11.03	0.27	4.54	3.29	4.77	**4.93**
PVCG(r=4)	**9.85**†	47.89†	**71.83**†	**10.87**	**0.28**	**4.67**†	**3.34**†	**4.84**†	3.51

Table 4: Ablation studies by measuring both clip selection and comment generation. Here, scheduled sampling refers to confidence-aware scheduled sampling described in Section 4.1.

Methods	Clip selection			Comment Generation			
	Recall@1 ↑	Recall@5 ↑	Recall@10 ↑	BLEU ↑	METEOR ↑	ROUGE_L↑	CIDEr ↑
w/o pre-train	4.64	29.01	63.02	3.99	3.64	3.73	2.85
w/o scheduled sampling	8.45	29.58	36.62	4.18	3.10	4.77	4.81
PVCG (r = 0)	25.35	40.85	60.56	4.54	3.29	4.77	4.93
PVCG (r = 2)	25.99	41.52	63.56	4.56	3.32	4.80	4.90
PVCG (r = 4)	**27.54**	**42.18**	**65.92**	**4.67**	**3.34**	**4.82**	4.87

5.2 Performance on Clip Selection (RQ1)

The selection results of different models are shown in Tab. 2. First, our proposed PVCG achieves the best performance on both Recall@1 and Recall@10. It verifies the effectiveness of user preference modeling method in PVCG. The improvement over PVCG (cascaded) indicates that joint learning can improve the clip selection accuracy significantly and mutual benefit between clip selection and text generation. Second, as expected, all personalized models perform much better than MostPopular baseline which does not consider personal preference in clip selection. Third, BPR-MF method performs not as good as other personalized models, it demonstrates that using users' interaction information is not enough in video clip selection task. Forth, when compared to personalized (user id), PVCG demonstrates a higher selection accuracy, highlighting that explicit modeling user identity performs much better than introducing user identity into plain text. Fifth, PVCG outperforms Personalized (history) in terms of Recall@1 and Recall@10. This could be due to the presence of noise in the user history information, which leads to a decrease in selection accuracy.

5.3 Performance on Comment Generation Only (RQ2)

To evaluate the comments generation performance in multi-modal scenarios, first we evaluate the text generation quality given the user actually commented clips. As shown in Tab. 4, we compare PVCG with two personalized image caption models – EICP [28] and SACO [41]. Compared with EICP, which represents personal preference by user identity only, SACO and PVCG both perform much better in ranking-based metrics and reference-based metrics. It proves that the importance of modeling user preference considering textual information. Besides, compared with SACO, our proposed PVCG model performs better considering both ranking-based metrics and reference-based metrics. Especially, in Recall@10 and BLEU score, we find that PVCG brings significant improvements over baseline methods. it shows the effectiveness of our proposed training strategies and and architecture, especially in complicated multi-modal scenario.

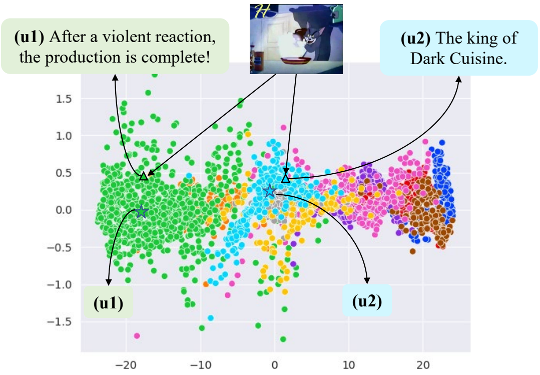


Figure 5: Visualization of the preference representation from different users with t-SNE.

5.4 Performance of Clip Selection and Comment Generation (RQ3)

We verify the performance without golden clips. As shown in Tab. 3, we have some observations. First, all personalized video to text generation methods work much better than the MostPopular baseline without personalization, which indicates that user generated text does have obvious individual preferences, thus understanding human preference is definitely helpful. Second, our proposed methods based on joint optimization of clip selection and comment generation are significantly better than the cascaded baseline. This is consistent with our observations on the Tab 2. Third, our full version PVCG method (i.e. PVCG (r=4)) performs better than the version without iterative inference in most metrics. This shows that our proposed iterative comment inference is effective.

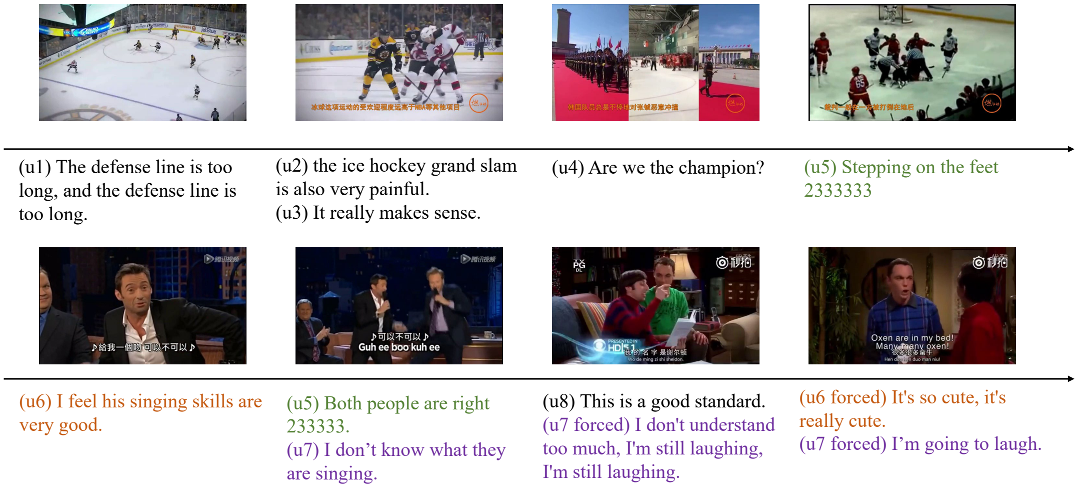

(u1) The defense line is too long, and the defense line is too long.
(u2) the ice hockey grand slam is also very painful.
(u3) It really makes sense.
(u4) Are we the champion?
(u5) Stepping on the feet 2333333
(u6) I feel his singing skills are very good.
(u5) Both people are right 233333.
(u7) I don't know what they are singing.
(u8) This is a good standard.
(u7 forced) I don't understand too much, I'm still laughing, I'm still laughing.
(u6 forced) It's so cute, it's really cute.
(u7 forced) I'm going to laugh.
Guh ee boo kuh ee
Oxen are in my bed!
Many many oxen!

Figure 6: Generated results for different users on the test data of Personalized VideoIC. For each video, PVCG can select clip and generate comments based on user preference with PVCG in an end-to-end manner. PVCG is also flexible to generate personalized comment by giving a specific video clip as input.

5.5 Ablation Study

We conduct ablation studies to investigate the impact of training strategy, validate the effectiveness of confidence-aware scheduled sampling and iterative comment generation strategy. Results are shown in Tab. 4. First, removing the pre-training stage leads to a decrease in selection accuracy and thus text generation quality. We attribute this to the fact that clips from the same video have higher similarity, making it difficult for the model to select the correct clip from those clips. The pre-training stage samples negative examples from the entire dataset, which makes it easier to select the ground truth clip. This method actually optimizes the model in an easy to hard manner, which smooths the training process, and facilitates better selection and generation performance. Second, removing confidence-aware scheduled sampling significantly lowers both the selection accuracy and the quality of the generated text. It also proves the importance of bridging the gap between the training and the inference stage in the personalized video to text generation task. Third, as the number of iterative rounds r increases, both the quality of the generated text and the selection accuracy improve. In the end-to-end framework, the higher the clip selection accuracy, the more accurate the visual information utilized during comment generation, resulting in comment content and style that are more similar to the target user. Then, using the generated text as additional input of user preference encoder can better model target user preference, in turn improving the clip selection accuracy.

5.6 Qualitative Analysis

To better illustrate how the proposed model captures user preference and behaves in clip selection and comment generation, we do some visualization and show some cases.

Fig. 5 visualizes the user preference features from ten different users who comments on video "Dubbing of Tom and Jerry". Different colors represent preference features of different users. We observe that different users can be effectively distinguished by our user preference encoder. We can see that the preference features from the same user are very closely clustered together, regardless of whether combining corresponding comments or not. Moreover, there is a large difference in preference between user $u1$ and user $u2$. Even for comments on the same video clip, their personal preference features differ greatly.

Fig. 6 shows the selected videos and generated comments for different users. The same color refers to the text generated given same user identity. Obviously, different video clips interest different users, resulting in diverse personalized comments. We also find that the generated texts for different videos given the same user identity have similar literal preference. For example, $u5$ prefers to use "2333..." (several 3's following a 2, which means "happiness" online). Furthermore, we force the model to generate text for users on clips they are not interested in. We observe that PVCG can generate personalized text related to the clip's visual information and the user's preference. Our framework is flexible and effective to select clips, or generate text for a given clip, or do both.

6 CONCLUSION

In this work, we highlight the importance of personalization in video to text generation for multimodal chatbots and define personalized video to text generation task formally. By taking video comments generation as a special application, we design a personalized video to text generation model PVCG to address the task in an end-to-end manner. PVCG consists a clip selector and a comment generator, with well-designed strategies to make this framework fully differentiable. We also construct Personalized VideoIC, a new dataset of personalized video comments generation, to support the research. Experimental results demonstrate the superiority of PVCG over all baselines and the positive contributions of our proposed ideas. For future work, we plan to explore personalized clip selection considering explicit temporal modeling. How to expand the model to a real-time streaming scenario is also interesting.

REFERENCES

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *NeurIPS*. http://papers.nips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html

[2] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized Fashion Recommendation with Visual Explanations based on Multimodal Attention Network: Towards Visually Explainable Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*. ACM, 765–774. https://doi.org/10.1145/3331184.3331254

[3] Xu Chen, Yongfeng Zhang, Qingyao Ai, Hongteng Xu, Junchi Yan, and Zheng Qin. 2017. Personalized Key Frame Recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 315–324. https://doi.org/10.1145/3077136.3080776

[4] Sergio Cleger-Tamayo, Juan M. Fernández-Luna, and Juan F. Huete. 2012. Explaining neighborhood-based recommendations. In *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012*. ACM, 1063–1064. https://doi.org/10.1145/2348283.2348470

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[6] Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*. 376–380.

[7] 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=YicbFdNTTy

[9] Chaoqun Duan, Lei Cui, Shuming Ma, Furu Wei, Conghui Zhu, and Tiejun Zhao. 2020. Multimodal Matching Transformer for Live Commenting. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020) (Frontiers in Artificial Intelligence and Applications, Vol. 325)*, Giuseppe De Giacomo, Alejandro Catalá, Bistra Dilkina, Michela Milano, Senén Barro, Alberto Bugarín, and Jérôme Lang (Eds.). IOS Press, 1998–2005. https://doi.org/10.3233/FAIA200320

[10] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. 2019. Heterogeneous Memory Enhanced Multimodal Attention Model for Video Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 1999–2007. https://doi.org/10.1109/CVPR.2019.00210

[11] Pranay Gupta and Manish Gupta. 2022. NewsKVQA: Knowledge-Aware News Video Question Answering. In *Advances in Knowledge Discovery and Data Mining - 26th Pacific-Asia Conference, PAKDD 2022, Chengdu, China, May 16-19, 2022, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 13282)*, João Gama, Tianrui Li, Yang Yu, Enhong Chen, Yu Zheng, and Fei Teng (Eds.). Springer, 3–15. https://doi.org/10.1007/978-3-031-05981-0_1

[12] Yunfeng Hou, Ning Yang, Yi Wu, and Philip S. Yu. 2019. Explainable recommendation with fusion of aspect information. *World Wide Web* 22, 1 (2019), 221–240. https://doi.org/10.1007/s11280-018-0558-1

[13] Vladimir Iashin and Esa Rahtu. 2020. Multi-modal Dense Video Captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*. Computer Vision Foundation / IEEE, 4117–4126. https://doi.org/10.1109/CVPRW50498.2020.00487

[14] Chengxi Li and Brent Harrison. 2021. 3M: Multi-style image caption generation using Multi-modality features under Multi-UPDOWN model. In *Proceedings of the Thirty-Fourth International Florida Artificial Intelligence Research Society Conference, North Miami Beach, Florida, USA, May 17-19, 2021*, Eric Bell and Fazel Keshtkar (Eds.). https://doi.org/10.32473/flairs.v34i1.128380

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *CoRR* abs/2301.12597 (2023). https://doi.org/10.48550/arXiv.2301.12597 arXiv:2301.12597

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 12888–12900. https://proceedings.mlr.press/v162/li22n.html

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

[18] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. SwinBERT: End-to-End Transformers with Sparse Attention for Video Captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 17928–17937. https://doi.org/10.1109/CVPR52688.2022.01742

[19] Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

[20] Yichao Lu, Ruihai Dong, and Barry Smyth. 2018. Why I like it: multi-task learning for recommendation and explanation. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*. ACM, 4–12. https://doi.org/10.1145/3240323.3240365

[21] Guangyi Lv, Tong Xu, Qi Liu, Enhong Chen, Weidong He, Mingxiao An, and Zhongming Chen. 2019. Gossiping the Videos: An Embedding-Based Generative Adversarial Framework for Time-Sync Comments Generation. In *Advances in Knowledge Discovery and Data Mining - 23rd Pacific-Asia Conference, PAKDD 2019, Macau, China, April 14-17, 2019, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 11441)*, Qiang Yang, Zhi-Hua Zhou, Zhiguo Gong, Min-Ling Zhang, and Sheng-Jun Huang (Eds.). Springer, 412–424. https://doi.org/10.1007/978-3-030-16142-2_32

[22] Shuming Ma, Lei Cui, Damai Dai, Furu Wei, and Xu Sun. 2019. LiveBot: Generating Live Video Comments Based on Visual and Textual Contexts. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 6810–6817. https://doi.org/10.1609/aaai.v33i01.33016810

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 311–318. https://doi.org/10.3115/1073083.1073135

[24] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to You: Personalized Image Captioning with Context Sequence Memory Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 6432–6440. https://doi.org/10.1109/CVPR.2017.681

[25] Haekyu Park, Hyunsik Jeon, Junghwan Kim, Beunguk Ahn, and U Kang. 2017. UniWalk: Explainable and Accurate Recommendation for Rating and Network Data. *CoRR* abs/1710.07134 (2017). arXiv:1710.07134 http://arxiv.org/abs/1710.07134

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 8748–8763. http://proceedings.mlr.press/v139/radford21a.html

[27] Zhaochun Ren, Shangsong Liang, Piji Li, Shuaiqiang Wang, and Maarten de Rijke. 2017. Social Collaborative Viewpoint Regression with Explainable Recommendations. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM 2017, Cambridge, United Kingdom, February 6-10, 2017*, Maarten de Rijke, Milad Shokouhi, Andrew Tomkins, and Min Zhang (Eds.). ACM, 485–494. https://doi.org/10.1145/3018661.3018686

[28] Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019. Engaging Image Captioning via Personality. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 12516–12526. https://doi.org/10.1109/CVPR.2019.01280

[29] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.

[30] Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. 2018. Explainable Recommendation via Multi-Task Learning in Opinionated Text Data. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, Kevyn Collins-Thompson,

Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz (Eds.). ACM, 165–174. https://doi.org/10.1145/3209978.3210010

[31] Weiying Wang, Jieting Chen, and Qin Jin. 2020. VideoIC: A Video Interactive Comments Dataset and Multimodal Multitask Learning for Comments Generation. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann (Eds.). ACM, 2599–2607. https://doi.org/10.1145/3394171.3413890

[32] Zhongke Wang. 2021. Analysis of User Personalized Retrieval of Multimedia Digital Archives Dependent on BP Neural Network Algorithm. *Advances in Multimedia* 2021 (2021), 1–7.

[33] Hao Wu, François Pitié, and Gareth Jones. 2021. Cold Start Problem For Automated Live Video Comments. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*. 54–62.

[34] Le Wu, Lei Chen, Yonghui Yang, Richang Hong, Yong Ge, Xing Xie, and Meng Wang. 2019. Personalized Multimedia Item and Key Frame Recommendation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus (Ed.). ijcai.org, 1431–1437. https://doi.org/10.24963/ijcai.2019/198

[35] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video Question Answering via Gradually Refined Attention over Appearance and Motion. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, Qiong Liu, Rainer Lienhart, Haohong Wang, Sheng-Wei "Kuan-Ta" Chen, Susanne Boll, Yi-Ping Phoebe Chen, Gerald Friedland, Jia Li, and Shuicheng Yan (Eds.). ACM, 1645–1653. https://doi.org/10.1145/3123266.3123427

[36] Markus Zanker and Daniel Ninaus. 2010. Knowledgeable Explanations for Recommender Systems. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010, Toronto, Canada, August 31 - September 3, 2010, Main Conference Proceedings*, Jimmy Xiangji Huang, Irwin King, Vijay V. Raghavan, and Stefan M. Rüger (Eds.). IEEE Computer Society, 657–660. https://doi.org/10.1109/WI-IAT.2010.131

[37] Wei Zhang, Yue Ying, Pan Lu, and Hongyuan Zha. 2020. Learning long-and short-term user literal-preference with multimodal hierarchical transformer network for personalized image caption. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9571–9578.

[38] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. 2020. Object Relational Graph With Teacher-Recommended Learning for Video Captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 13275–13285. https://doi.org/10.1109/CVPR42600.2020.01329

[39] Wayne Xin Zhao, Yanwei Guo, Yulan He, Han Jiang, Yuexin Wu, and Xiaoming Li. 2014. We know what you want to buy: a demographic-based system for product recommendation on microblogs. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*. ACM, 1935–1944. https://doi.org/10.1145/2623330.2623351

[40] Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, Yingqian Min, Zhichao Feng, Xinyan Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji-Rong Wen. 2021. RecBole: Towards a Unified, Comprehensive and Efficient Framework for Recommendation Algorithms. In *CIKM*. ACM, 4653–4664.

[41] Yucheng Zhou and Guodong Long. 2023. Style-Aware Contrastive Learning for Multi-Style Image Captioning. *CoRR* abs/2301.11367 (2023). https://doi.org/10.48550/arXiv.2301.11367 arXiv:2301.11367

[42] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592* (2023).