Personalized Video Comment Generation

Xudong Lin[1], **Ali Zare**[1], **Shiyuan Huang**[1]
Ming-Hsuan Yang[2], **Shih-Fu Chang**[1], **Li Zhang**[2]
[1]Columbia University [2]Google Research
xudong.lin@columbia.edu

Abstract

Generating personalized responses, particularly in the context of video, poses a unique challenge for language models. This paper introduces the novel task of **Personalized Video Comment Generation** (PVCG), aiming to predict user comments tailored to both the input video and the user's comment history, where the user is unseen during the model training process. Unlike existing video captioning tasks that ignores the personalization in the text generation process, we introduce PerVidCom, a new dataset specifically collected for this novel task with diverse personalized comments from YouTube. Recognizing the limitations of existing captioning metrics for evaluating this task, we propose a new automatic metric based on Large Language Models (LLMs) with few-shot in-context learning, named FICL-Score, specifically measuring quality from the aspects of emotion, language style and content relevance. We verify the proposed metric with human evaluations. We establish baselines using prominent Multimodal LLMs (MLLMs), analyze their performance discrepancies through extensive evaluation, and identifies directions for future improvement on this important task. Our research opens up a new direction of personalizing MLLMs and paves the way for future research.

Figure 1: A novel task: **Personalized Video Comment Generation**, which is to generate a comment in a user's style given the user's comment history and a video as input. Besides the difficulty of detailed video understanding, PVCG is unique and challenging because models are required to understand the user's language style, subjects of interest, and possible emotion responses from a limited comment history from the user that is unseen during the model training process. For privacy, we mask faces and use pseudo user IDs.

1 Introduction

What if your AI assistant could not only help you with tasks but also share your unique sense of humor or express your excitement about your favorite sports team? While recent advancements in AI have led to impressive progress in foundational models (Brown et al., 2020; Reid et al., 2024) and task-oriented assistants (Wang et al., 2019a; Lin et al., 2022; Wang et al., 2023a; Lin et al., 2023c; Yildirim et al., 2024; Niu et al., 2024; Kazemitabaar et al., 2024), one crucial aspect of human interaction remains largely unexplored: **personalized** responses in the realm of entertainment, particularly online video. The widespread adoption of

Findings of the Association for Computational Linguistics: EACL 2024, pages 16806–16820

online video platforms has fostered a vibrant culture of user-creator interaction, with commenting being a prominent form of engagement. While significant effort has been made in general text generation with video input (Xu et al., 2016; Wang et al., 2019b; Ayyubi et al., 2023), the aspect of personalization, crucial for replicating natural user reaction, remains unexplored.

As shown in Figure 1, we make the first attempt by defining a novel task, **Personalized Video Comment Generation**, which goes beyond traditional video captioning and comment generation and requires strong generalization ability to adapt to unseen users. Unlike existing tasks that focus on generating generic descriptions or generating comments for seen users (Wu et al., 2024), PVCG requires models to generate comments tailored to both the input video content and the specific user's past commenting behavior, where the user is unseen during the training process of the model. This crucial distinction elevates the complexity of the task, as models must not only understand the video but also infer the user's *emotional response tendencies, individual preferences, and language style*.

To establish a base for this research problem, we curate a new benchmark dataset, **PerVidCom**, specifically designed for PVCG. We design a careful data collection process to ensure high-quality data with personalized comments from different users. To achieve this goal, we **leverage pretrained language models to determine whether a video has diverse comments**. The resulting dataset is verified to be high-quality with both models and manual assessment (Section 2.2).

Recognizing the limitations of existing automatic metrics like BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004) in capturing the nuances and variety of personalized generation, we propose a novel automatic metric based on LLMs, specifically measuring quality from the aspects of emotion, language style and content relevance. We further verify the proposed metric with human evaluations. This new metric, called **FICL-Score**, utilizes a few-shot in-context learning approach to assess the similarity of generated comments to target ones, considering emotion, language style, and content relevance, with a strong alignment with human evaluation scores (Section 5.1).

We establish baselines using prominent Multimodal LLMs (MLLMs) such as Gemini (Reid et al., 2024) and open-sourced models (Lin et al., 2023a). We also propose PV-LLM that is specifically fine-tuned for generating personalized comments given video and user comment history as input. Upon analysis of their performance discrepancies through extensive evaluation, we identify clear performance headroom and several possibilities for future research on this novel task. Our research opens up a new direction of personalizing MLLMs and paves the way for future research.

Our key contributions can be summarized as follows:

- We introduce the novel task of Personalized Video Comment Generation (PVCG), and present a new benchmark dataset, PerVidCom, for this task.
- We propose a new automatic evaluation metric, FICL-Score, leveraging few-shot in-context learning with LLMs to measure the quality of personalized video comments, which is strongly aligned with human judgement in terms of emotion, language style and content relevance.
- We establish strong baselines using prominent MLLMs, propose PV-LLM that is specifically fine-tuned for PVCG, and analyze their performance and identify interesting future directions.

This work paves the way for future research on personalized content generation. We will make our code and dataset publicly available to further accelerate progress in this exciting direction.

2 Dataset Collection and Statistics

This section details the process of constructing our personalized video comment dataset and the statistics of the resulting PerVidCom dataset.

2.1 Data Source and Filtering

To construct our dataset, we utilize the YouTube Data API[1] to collect videos and their associated comments from YouTube Shorts. We specifically focus on Shorts as they tend to attract a higher volume of comments, providing richer data for personalization. To ensure data quality, we apply the following filtering criteria:

- **User Comment History:** We only consider users with a minimum of 11 of comments

[1] https://developers.google.com/youtube/v3

to ensure sufficient comment history for personalization. Using extremely few historical comments might result in an ill-defined task, making it difficult to learn meaningful personalized generation patterns.

- **Video Selection:** We focus on videos that show strong personalization of comments. To measure this automatically during data collection process, we proposed a Personalization Score as defined below. We only keep videos with a Personalization Score ψ larger than 90%. For video quality and appropriateness, we rely on the algorithms of YouTube itself and only keep videos with more than 400 likes.

Motivated by observations of videos[2] where comments are almost identical, we proposed the Personalization Score to measure whether a video has good personalization effect or not. We first calculate the difference of two similarity scores as follows

$$d_{ij} = \max(S_{ij} - S_{it}, S_{ij} - S_{jt}), \quad (1)$$

where S_{ij} is the text similarity between the i-th comment and the j-th comment in this video, and S_{it} the text similarity between the i-th comment and the title of the video. We leverage a popular sentence embedding model all-MiniLM-L6-v2 (Reimers and Gurevych, 2019; Wang et al., 2020b)[3] that is known for efficiency and accuracy, to calculate textual similarities. Then for any pair of the comments of a video, we apply the following function,

$$p_{ij} = \begin{cases} 1, & \text{if } d_{ij} < 0 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Here we consider a pair to be positive if their similarity is smaller than the similarity to the title because personalization requires comments to be sufficiently different. Now we calculate the Personalization Score, defined as

$$\psi = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p_{ij}}{N(N-1)/2}. \quad (3)$$

With the proposed Penalization Score ψ, we find that a 90% threshold can help always rejecting videos with near identical comments in our pilot study of 100 videos. Since our goal is to achieve high precision in this filtering strategy to ensure the kept videos all have strongly personalized comments, we didn't apply further tuning of the threshold or filtering techniques.

[2] https://www.youtube.com/watch?v=ZW6MDdFzSi0

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

	Train	Validation	Test	Total
# of Users	1,0021	2,505	4,176	16,702
# of Videos	9,674	8,011	8,832	9,839
# of Comments	212,276	53,898	86,195	344,441
Avg. # of Words	10.95	10.55	10.79	10.85

Table 1: Dataset Statistics of PerVidCom.

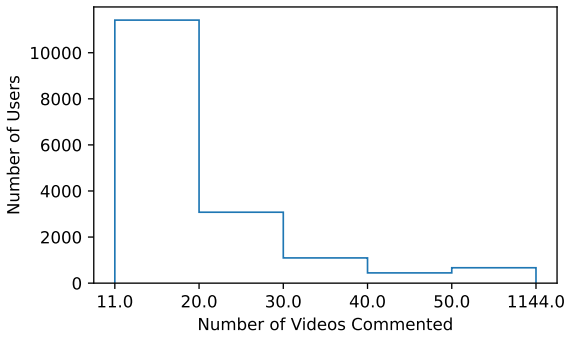

Figure 2: Histogram of the number of users commenting a certain number of videos.

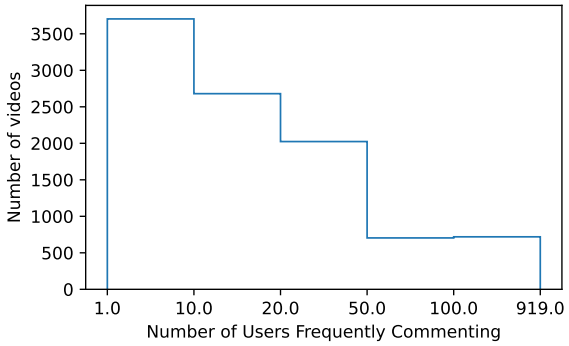

Figure 3: Histogram of the number of videos commented by a certain number of frequent users.

2.2 Dataset Statistics and Quality

The final dataset consists of 9,839 videos and 344,441 user comments from 16,702 frequent commenting users. In the final dataset, more than 96.3% of the videos have a Personalization Score larger than 95%, which indicating a strong personalization of comments in the collected videos. Figure 4 visualizes 4 randomly sampled users' comments, which clearly present diverse language styles and subject of interest in comments. Upon manual check of 200 randomly selected video, we did not

(a) (b) (c) (d)

Figure 4: Word cloud visualization of four randomly sampled users from the dataset. It is clear that each user has a unique language style and subject of interest in comments.

observe any inappropriate videos and we find the comments to be indeed diverse in each video.

To focus on personalization for users, we split the users into train, validation, and test sets with a ratio of 60%, 15%, and 25%. Detailed statistics as shown in Table 1. PerVidCom is unique as this is the first video dataset designed for user-specific comments, to the best of our knowledge. We also notice that compared to conventional video comment datasets (Ma et al., 2019; Wang et al., 2020a), our PerVidCom significantly contains more words per comment (11 versus 5), which makes it more possible to mine the personal preferences in comments. As shown in Figures 2 and 3, we successfully collect videos with sufficient common users that frequently (at least in 11 videos) make comments. Due to lack of access to larger-scale computational resources, we leave further scaling up the dataset as future work.

3 Methods

In this section, we first introduce all the baseline methods and then illustrate the proposed model.

3.1 Baseline Models

Generating personalized video comments is a brand-new task and there is no existing work, to the best of our knowledge. Therefore, we probe the following existing models that have multimodal video understanding abilities.

- Video-ChatGPT(-7B) (Maaz et al., 2023) is a recent open-sourced MLLM specifically tuned for chatting about videos.

- Video-LLaVA(-7B) (Lin et al., 2023a) is a recent open-sourced MLLM for joint image and video understanding tasks.

- Gemini-1.5-Flash (Reid et al., 2024) is a popular lightweight yet powerful MLLM, which has strong long-context and video capabilities, and its model details are not disclosed.

- Gemini-1.5-Pro (Reid et al., 2024) is one of the most powerful MLLM, which achieves state-of-the-art performance on many long-context/video capabilities and benchmarks, and its model details are not disclosed.

For simplicity, we use the same prompt format for all the models, as depicted in Figure 5. By default, we use 5 randomly sampled history comments from the user as input to the model (different models use the same sampled comments history), when not specified. For all the baseline models, we use their default sampling parameters, respectively.

3.2 Proposed Approach

To address the unique challenge of mimicking the user's emotion response, language style and subjects from the limited history comments, we propose a Personalized Video LLM (PV-LLM) that benefits from supervised fine-tuning. As shown in Figure 5, PV-LLM leverages a pretrained image encoder to extract features from the input frames that are sparsely sampled from the video. Then the extracted features are fed into the Projection Layer to be transformed into the input embedding space of the LLM pretrained for language tasks. The input text is fed into the LLM together with the projected features of video frames. Within LLM, the input text is transformed into a sequence of word embeddings in the LLM's word embedding layer, and then it is concatenated with the projected video features and fed to the rest of the layers in the LLM. The word embedding layer of LLM and the vision encoder are frozen during the supervised finetuning

```
From 0 (not similar or relevant at all), 0.25 (slightly similar), 0.5 (similar in some
aspects), 0.75 (similar in many aspects) to 1 (almost identical), rate the similarity between
generated comment and ground-truth in terms of emotion, language style, and content relevance.
Generate your final output in form of "Emotion: x, Style: y, Relevance: ", where x, y and z
are the scores you pick for each dimension.

Example 0
Generated comment: ...
Ground-truth: ...
Emotion similarity: ...
Language Style similarity: ...
Content Relevance similarity: ... ......

Now lets get started!
Generated comment: ...
Ground-truth: ...
```

Table 2: The text prompt template for obtaining FICL-Score.

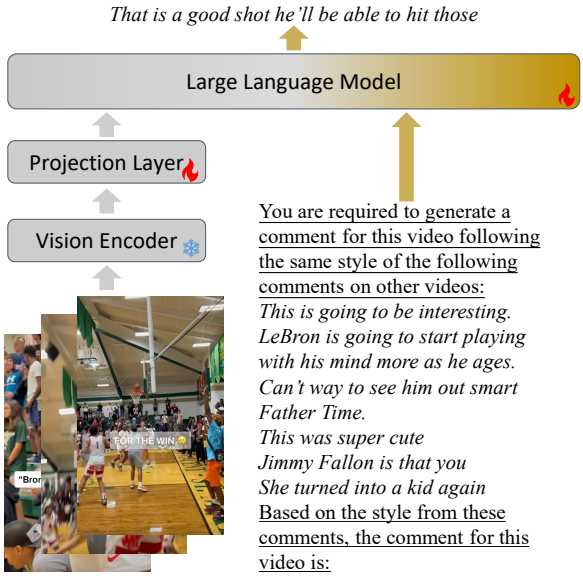

Figure 5: Diagram of the proposed PV-LLM model. We finetune the projection layer and the large language model towards generating comments personalized for a user, which is specified by the comment history in the input text.

process, following common practice (Maaz et al., 2023; Lin et al., 2023a).

During training, for each user, we randomly pick one video as the target video to generate comment and randomly sample from the rest of the videos to gather at most 15 comments as the comment history to increase data diversity. The number of comments history is randomly picked from 5 to 15. This strategy yields a training dataset of 10,0210 data samples. We initialize the model from Video-LLaVA-7B (Lin et al., 2023a) and fine-tune the model with standard cross-entropy loss. Training is performed on 8 NVIDIA A100 GPUs with a total batch size of 8 and accumulate gradients every 8 iterations. We follow the learning rate configurations of Video-LLaVA (Lin et al., 2023a) and observe that the training loss converges in the end of second training epoch. During inference, we use the default sampling parameter as Video-LLaVA (Lin et al., 2023a), without quantization of model parameters. Five history comments are used by default, same as the baselines.

4 Evaluation

Properly evaluating personalized video comments is challenging because of the rich variety in user's attention and emotion. Conventional metrics are used for widely used for evaluating text generation and captioning tasks such as BLEU-4 (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), METEOR (Banerjee and Lavie, 2005), and ROUGE-L (Lin, 2004) are not really suitable for evaluating the quality of generated comments because their string-matching design. A more suitable automatic evaluation metric is imperative for the PVCG task. To understand how to design a more suitable automatic metric, we first perform human evaluation.

4.1 Human Evaluation

Upon manual assessment of generated comments, we reach to a consensus that emotion and language style of the generated comments affect the satisfaction of it most. In the meantime, although one could possibly leave two completely irrelevant comments when the same video is watch at different circumstances, we still want to understand whether the generated comment is relevant to the video

content or not. Based on these observation and motivation, we design the following dimensions for human evaluation:

- **Emotion:** How well the generated comment aligns with the emotion reflected in the user's comment.
- **Language Style:** The extent to which the generated comment aligns with the user's language style.
- **Content Relevance:** The relevance between the generated comment and the target comment.

Specifically, we conduct human evaluation to assess the quality of the generated comments with the following rubric: 0, not similar or relevant at all; 0.25, slightly similar; 0.5, similar in some aspects; 0.75, similar in many aspects; 1, almost identical. To ease the costly human annotation process, we evaluate the generated comments of three models (PV-LLM, Gemini-1.5-Flash, and Gemini-1.5-Pro) on 100 random users (50 from the validation set and 50 from the test set). There are three expert annotators and we take the average score from them for each sample.

4.2 Few-shot In-Context Learning Score

Existing work (Cui et al., 2018) leverages supervised training to learn an auto-rater to evaluate image captioning tasks. However, such supervised approaches with small-scale model cannot ensure the resulting auto-rater possessing rich world knowledge to generalize well. Inspired by the recent progress on in-context learning (Wang et al., 2022b) for multimodal tasks, we proposed to leverage few-shot in-context learning (FICL) with LLMs to eliminate the costly process of curating human annotations of predictions of various models.

Specifically, we employ a prompt as shown in Table 2. We sample K samples as few-shot in-context learning examples, where each sample is associated with the results from three models and thus the effective number of shots is $K \times 3$. We observed that Gemini's output follows the desired format well, and we can easily use regular expressions to obtain the score for each dimension. The success rate is higher than 99.2%. In rare failure cases of regular-expressions-based-parsing, we further pass the prediction to Gemini again to help extract the exact scores. Through empirical analysis in Section 5.1, we use $K = 5$ and Gemini-1.5-Flash as the auto-rater by default.

Metric	Emotion	Style	Relevance
BLEU-4	0.68	0.68	0.65
METEOR	0.39	0.43	0.45
ROUGE-L	0.34	0.37	0.35
CIDEr	0.31	0.34	0.34
Gemini-1.5-Flash			
FICL-Score (0)	0.80	0.83	0.74
FICL-Score (5×3)	**0.94**	**0.97**	**0.94**
FICL-Score (10×3)	0.92	**0.97**	0.92
FICL-Score (15×3)	0.90	0.95	0.89
Gemini-1.5-Pro			
FICL-Score (0)	0.77	0.83	0.43
FICL-Score (5×3)	0.92	0.93	0.92
FICL-Score (10×3)	0.91	0.96	0.91
FICL-Score (15×3)	0.90	0.93	0.92
Gemini-1.5-Flash + Comment History			
FICL-Score (0)	0.79	0.87	0.74
FICL-Score (5×3)	0.89	0.89	0.90
FICL-Score (10×3)	0.92	0.93	0.91
FICL-Score (15×3)	0.88	0.91	0.91

Table 3: Average Normalized Discounted Cumulative Gain (NDCG) between automatic metrics and human evaluations. Bold indicates the best results.

5 Experiment Results

In this section, we first report and analyze the comparison between the proposed FICL-Score and conventional automatic evalution metrics. Then we provide benchmark results of all the methods considered for this task. Finally, we ablate and discuss important questions and design choices.

5.1 Is FICL-Score Better Aligned with Human Evaluation?

To understand whether the our proposed FICL-Score is more suitable for PVCG, we calculate the Normalized Discounted Cumulative Gain[4] (Wang et al., 2013) between the automatic metrics and the human evaluation score . We choose NDCG as the measure because each user could generate diverse comments and it is not feasible to assume that there is a unified absolute rubric reflecting how similar the generated comment is, compared to the user's ground-truth comment. What truly matters for future model development is whether the automatic score reflects the same rank of the evaluated models, against the rank from human annotators. Specifically, we calculate the NDCG score between the list of scores from automatic metric and that of the human evaluation for each sample. Then

[4] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.ndcg_score.html

Method	BLEU-4	METEOR	ROUGE-L	CIDEr	FICL-Score			Human Score		
					E	S	R	E	S	R
Validation Set										
Video-ChatGPT	0.25	0.42	0.98	3.32	24.1	49.0	30.9	-	-	-
Video-LLaVA	1.08	1.28	1.77	3.71	29.8	50.0	39.1	-	-	-
PV-LLM	1.42	2.63	4.18	5.43	30.3	49.9	43.7	36.5	51.0	43.6
Gemini-1.5-Flash	4.36	4.20	6.23	8.60	35.8	51.9	48.5	45.5	60.1	60.0
Gemini-1.5-Pro	2.76	3.47	6.12	10.53	38.4	51.5	48.3	45.0	58.5	55.3
Test Set										
Video-ChatGPT	0.34	0.52	1.07	3.39	24.3	49.4	31.2	-	-	-
Video-LLaVA	0.84	1.07	1.66	4.02	29.8	50.0	39.1	-	-	-
PV-LLM	1.31	2.69	4.19	4.76	30.1	49.7	43.5	30.0	50.1	35.0
Gemini-1.5-Flash	4.01	4.19	6.45	8.82	35.5	51.8	48.6	39.6	59.6	51.1
Gemini-1.5-Pro	1.54	2.70	5.03	6.50	38.1	51.4	48.4	42.0	56.3	47.5

Table 4: Results on the validation set and the test set of our PerVidCom dataset. All the scores are scaled by 100 following common practice. Human Score and FICL-Score are the main metrics. The other conventional metrics are reported for reference. E, S and R are short for emotion, language style, content relevance, respectively. Gemini models are grayed out since their parameter and training data scale are likely much larger.

we average scores over the 50 validation samples. Note that the few-shot examples are not from them.

As shown in Table 3, the answer is clearly **Yes**. The proposed FICL score is significantly more aligned with human evaluation by up to %55 absolute NDCG score, which validates the effectiveness of our propose new automatic metric specifically for the PVCG task. We also observe that Gemin-1.5-Flash overall achieves the best results with 5×3 few-shot in-context examples. Using larger-scale Gemini-1.5-pro does not help to further improve the scores and thus we decide to use Gemini-1.5-Flash since it is cheaper and more accessible. We also try to further augment the input prompt with each user's 5 history comments but we do not observe further improvement either.

5.2 Main Results on PerVidCom

We evaluate all the five models on both the validation set and the test set on our PerVidCom dataset. As shown in Figure 4, we observe that with the proposed supervised fine-tuning process, PV-LLM achieves better performance than existing open-sourced model without this task-specific training process, especially on the relevance between generated comment and ground-truth comments. This possibly indicates that it is easier to gain the ability to understand the user's subjects of interests from the proposed supervised fine-tuning process. However, the overall gain from the fine-tuning process is not as significant as directly using much larger and capable model such as Gemini, especially on emotion and language style. Interestingly, we observe that larger model achieves the better results in terms of emotion between the generated comments and the ground-truth comments, which indicates a similar trend to observations from emotion interpretation tasks (Huang et al., 2023).

We also find the FICL-Score well aligned with the human evaluation scores along all the evaluation dimensions, which again validates the effectiveness of the proposed FICL-Score.

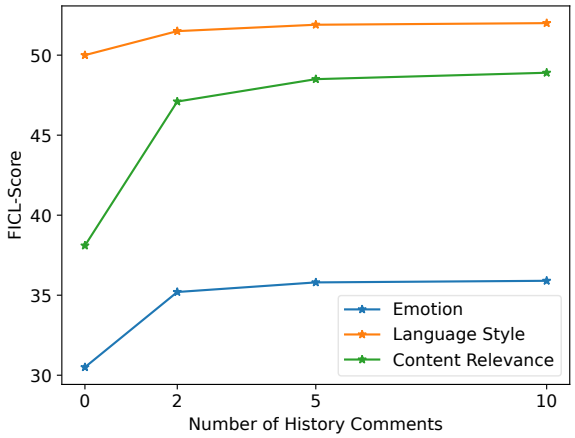

Figure 6: FICL-Scores of Gemini-1.5-Flash when varying the number of history comments.

Setting	BLEU-4	METEOR	ROUGE-L	CIDEr
No History	0.30	2.48	3.09	0.51
Others' History	0.21	2.10	3.24	0.52
Default	1.42	2.63	4.18	5.43

Table 5: Results of PV-LLM with different hisotry comments on the validation set.

5.3 Effect of Number of History Comments

As shown in Figure 6, using no history comments at all produces significantly lower results, which is expected and verifies that PVCG indeed requires user-specific information conveyed by the com-

Method	FICL-Score		
	Emotion	Style	Relevance
Video + Text	30.0	49.8	43.5
Video (Default)	30.3	49.9	43.7

Table 6: Results of PV-LLM variants on the validation set.

ment history. The improvement then starts to reach a plateau when 5 history comments are used, which may indicate either many more samples are needed to obtain further large improvement or models that are better at personalization through context are needed.

To comprehensively understand the effect of history comments and verify that this task indeed requires user history comments to generate user-specific comments, we further compare the results of the default setting, using no history and using others' history in Table 5.

5.4 Effect of Joint Training with Text Data

Existing MLLMs (Lin et al., 2023a) reported better video task performance when jointly trained video and text instruction tuning data. We compare these two variants in Table 6. We observe that the additional text data does not improve the FICL-Score on PVCG, which implies that the dialogue-style text data (Liu et al., 2023) is not beneficial for training the model to personalize its response in context.

6 Related Work

6.1 Video Captioning and Comment Generation

Video captioning aims to generate descriptions for videos. Research efforts usually focus on how to make the descriptions more comprehensive and precise. Earlier literature such as (Sun et al., 2019; Lei et al., 2020; Lin et al., 2021) focus on designing specific model architectures to better exploit video signals that transfer to texts. With the emergence of large-scale video-language datasets such as HowTo100M (Miech et al., 2019) and InternVid (Wang et al., 2023b), recent efforts have been made in developing strong video foundation models (Wang et al., 2022a; Li et al., 2023) that can be quickly fine-tuned for video captioning task. However, while personalized image captioning (Chunseong Park et al., 2017) has seen some exploration in tailoring descriptions based on user inputs, personalization in video captioning remains untapped.

On the other hand, video comment generation was initially introduced by (Ma et al., 2019) to address the challenge of generating user opinions from videos rather than generic descriptions. This task is inherently more demanding and closely aligns with real user-creator interactions since it requires an understanding of video content and user preferences. Subsequent efforts, including (Wang et al., 2020a; Meng et al., 2024) focus on data collections from different video domains, coupled with different modeling strategies.

The only highly relevant work (Wu et al., 2024) about personalized video comment generation is contemporaneous to our work and tackles a significantly different setting from ours: training and testing on the same set of users. Our work is the first to specifically focus on generalization to new users during test time. This requires models to adapt to unseen user preferences, a more challenging and practically relevant scenario than the contemporaneous work. Our PerVidCom contains over twice as many unique users and videos as the dataset in the contemporaneous work. Moreover, our dataset is newly collected from YouTube, providing a fresh and diverse data source for the community. We also introduce FICL-Score, the first automatic metric specifically designed for PVCG. We provide extensive human evaluation results validating its superiority over existing metrics like BLEU, METEOR, ROUGE, and CIDEr. The contemporaneous work does not propose such a dedicated evaluation metric. These key differences make our work a distinct and valuable contribution.

6.2 Personalizing/Customizing Language Models

Recent advancements in Large Language Models (LLMs) have opened new avenues for tailoring their strong capabilities to individual needs, preferences, and behaviors. The problem of personalized comment generation is closely related to personalized language generation, aiming to produce relevant text for users based on context and individual needs. Previous research has primarily focused on conditioning personalized agents with user-relevant text content. For instance, (Zhang et al., 2018) enhanced user dialogues by conditioning dialogue agents on user profile information, leading to more engaging conversations. (Dudy et al., 2021) emphasized the importance of leveraging additional context in natural language generation (NLG) for better personalization. (Salemi et al., 2024) addressed this issue using retrieval-augmented models for user-

specific information retrieval from user profiles. Other works (Jang et al., 2023; Li et al., 2024) have explored personalization through reinforcement learning. With the introduction of powerful LLMs like GPT-3.5 (Brown et al., 2020) and Gemini (Reid et al., 2024), there has been a new trend on personalization/customization through zero-shot reasoning (Lin et al., 2023b; Woźniak et al., 2024). In this work, we specifically concentrate on personalized comment generation, which aims to capture and replicate realistic user-specific interactions with video content, which necessitates personalization using multi-modal information, including video-specific content and user-specific profiles (i.e., comment interaction history).

7 Data Release Plan

To ensure responsible data sharing and comply with copyright regulations, we will not directly release the video data. Instead, we will provide a comprehensive code repository so that the audience can download the videos using the provided list of video IDs. Similarly, to protect user privacy, we will only provide the list of private user IDs that are are not visible publicly except to the users themselves. Our code repository will also include the code we used to call YouTube Data API to download comments with comment text and private user IDs. We encourage the audience to carefully read the Fair Use Notice in the appendix if there are any questions about the usage of the dataset.

8 Conclusion

We introduce a novel task of Personalized Video Comment Generation, along with PerVidCom, the first benchmark dataset for generalization to unseen users. We propose FICL-Score, a new automatic evaluation metric leveraging few-shot in-context learning with LLMs to measure the quality of generated comments, demonstrating strong alignment with human judgment. Our experiments with prominent MLLMs and a fine-tuned PV-LLM model highlight both the potential and limitations of current models. This work paves the way for future research on multimodal personalized content generation.

9 Acknowledgement

The research was supported in part by a Google gift award. This work gets support from the Google Cloud Platform on the computational resources for conducting our experiments. The research also gets support from the YouTube Researcher Program on accessing YouTube Data API. We would like to also thank all the other colleagues and anonymous reviewers for their valuable help.

10 Limitations

While we address the critical need for personalized video comment generation, our work has limitations. Firstly, our dataset is relatively small and are more sports-related due to the inherent difficulty of collecting high-quality personalized comments and accessing large-scale computational resources. We also acknowledge the possibility of biased and toxic content inherent in internet-sourced data, although we have tried our best to filtered out possible problematic data. Secondly, we primarily focus on textual personalization through comment history, but the current comment history is randomly sampled. It would be more realistic if users' recent watching/commenting history could be collected, which better reflects their current attention and emotion.

The automatic metric is not perfect. It's important to note that among all the 18 pairs of comparisons between any models on either validation or test set along one of the dimensions, only one pair (Gemini-1.5-Flash v.s. Gemini-1.5-Pro on validation set along Emotion) is unaligned. The successful rate is already 94.4% in this case. Another limitation is the limited variance in Style scores might make it harder to appreciate the magnitude of future model improvements. We plan to further explore possible improvements in the future to enlarge difference and increase alignment for a better automatic metric.

In this work, we focus on users with at least 11 comments mainly to ensure fair comparisons when varying the number of historical comments. Using extremely few historical comments might result in an ill-defined task, making it difficult to learn meaningful personalized generation patterns. We would like to note that evaluating on a wider range of users, especially those who comment less, is an important direction for future work. Finally, our human evaluations were conducted by annotators judging the predictions for other users. More accurate human evaluation could be done by letting users write comments for the video and then judge how much the annotator's comment aligns with the model's prediction.

References

Hammad A Ayyubi, Tianqi Liu, Arsha Nagrani, Xudong Lin, Mingda Zhang, Anurag Arnab, Feng Han, Yukun Zhu, Jialu Liu, and Shih-Fu Chang. 2023. Video summarization: Towards entity-aware captions. *arXiv preprint arXiv:2312.02188*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 895–903.

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. 2018. Learning to evaluate image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5804–5812.

Shiran Dudy, Steven Bedrick, and Bonnie Webber. 2021. Refocusing on relevance: Personalization in NLG. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5190–5202, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023. Emotionally numb or empathetic? evaluating how llms feel using emotionbench. *arXiv preprint arXiv:2308.03656*.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *Preprint*, arXiv:2310.11564.

Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L Berg, and Mohit Bansal. 2020. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. *arXiv preprint arXiv:2005.05402*.

Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. 2023. Lavender: Unifying video-language understanding as masked language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23119–23129.

Xinyu Li, Zachary C. Lipton, and Liu Leqi. 2024. Personalized language modeling from personalized human feedback. *Preprint*, arXiv:2402.05133.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023a. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

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

Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. 2021. Vx2text: End-to-end learning of video-based text generation from multimodal inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7005–7015.

Xudong Lin, Manling Li, Richard Zemel, Heng Ji, and Shih-Fu Chang. 2023b. Training-free deep concept injection enables language models for crossmodal tasks.

Xudong Lin, Fabio Petroni, Gedas Bertasius, Marcus Rohrbach, Shih-Fu Chang, and Lorenzo Torresani. 2022. Learning to recognize procedural activities with distant supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13853–13863.

Xudong Lin, Simran Tiwari, Shiyuan Huang, Manling Li, Mike Zheng Shou, Heng Ji, and Shih-Fu Chang. 2023c. Towards fast adaptation of pretrained contrastive models for multi-channel video-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14846–14855.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Shuming Ma, Lei Cui, Damai Dai, Furu Wei, and Xu Sun. 2019. Livebot: Generating live video comments based on visual and textual contexts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6810–6817.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Zixiang Meng, Qiang Gao, Di Guo, Yunlong Li, Bobo Li, Hao Fei, Shengqiong Wu, Fei Li, Chong Teng, and Donghong Ji. 2024. Mmlscu: A dataset for

multi-modal multi-domain live streaming comment understanding. In *Proceedings of the ACM on Web Conference 2024*, pages 4395–4406.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640.

Yulei Niu, Wenliang Guo, Long Chen, Xudong Lin, and Shih-Fu Chang. 2024. Schema: State changes matter for procedure planning in instructional videos. *arXiv preprint arXiv:2403.01599*.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization. *Preprint*, arXiv:2304.11406.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473.

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*, pages 4566–4575.

Jianbo Wang, Kai Qiu, Houwen Peng, Jianlong Fu, and Jianke Zhu. 2019a. Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance. In *Proceedings of the 27th ACM international conference on multimedia*, pages 374–382.

Weiying Wang, Jieting Chen, and Qin Jin. 2020a. Videoic: A video interactive comments dataset and multimodal multitask learning for comments generation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2599–2607.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020b. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. 2023a. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20270–20281.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019b. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. 2023b. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. 2022a. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of ndcg type ranking measures. In *Conference on learning theory*, pages 25–54. PMLR.

Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. 2022b. Language models with image descriptors are strong few-shot video-language learners. *Proc. The Thirty-Sixth Annual Conference on Neural Information Processing Systems (NeurIPS2022)*.

Stanisław Woźniak, Bartłomiej Koptyra, Arkadiusz Janz, Przemysław Kazienko, and Jan Kocoń. 2024. Personalized large language models. *Preprint*, arXiv:2402.09269.

Yihan Wu, Ruihua Song, Xu Chen, Hao Jiang, Zhao Cao, and Jin Yu. 2024. Understanding human preferences: Towards more personalized video to text generation. In *Proceedings of the ACM on Web Conference 2024*, pages 3952–3963.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Nur Yildirim, Hannah Richardson, Maria Teodora Wetscherek, Junaid Bajwa, Joseph Jacob, Mark Ames Pinnock, Stephen Harris, Daniel Coelho De Castro,

Shruthi Bannur, Stephanie Hyland, et al. 2024. Multimodal healthcare ai: identifying and designing clinically relevant vision-language applications for radiology. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–22.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

A Effect of Different History Comments Used

To understand the effect of different comment history, we sample another two sets of comment history and use them as input to Gemini-1.5-Flash. The resulted FICL-Scores are 35.7 ± 0.3, 51.8 ± 0.2, and 48.6 ± 0.2 for Emotion, Style and Relevance,, respectively. Overall, the variance is negligible.

B Qualitative Results

We visualize the predictions on two samples from PV-LLM, Gemini-1.5-Flash and Gemini-1.5-Pro in Figures 7 and 8. We first manually check that the predicted FICL-Score produces the same rank as human assessment. We then notice that the smaller PV-LLM may suffer from lack of detailed video understanding as depicted in Figure 7. The model probably overlooks the Frisbee and only sees the ocean in the background.

C Additional Details of Videos

We search for videos with hashtag of "#Shorts" and the resulted video have 883,893 views on average and the average duration of the videos is 21.6 seconds. Due to lack of information/API from YouTube Data API, we cannot check fine-grained distribution of categories in the collected video. We manually observe that many video are funny videos, amazing videos or related to certain sports.

D Human Evaluation Details

Annotators first iterate over manual inspect of models' predictions and then reach to a consensus of the human annotation protocol as illustrated in 4.1. Then during the annotation process, annotators choose from one of the score choices for each model's prediction along each evaluation dimension, using the interface as depicted in Figure 9.

E Implementation Details

To download comments of videos, we mainly utilize the YouTube Data API [5]. Since we focus on personalizing the comments, we only collect comments to videos in this dataset without considering the comments written to other comments.

To train our PV-LLM model, we modify the code repository from Video-LLaVA[6]. We follow the same hyper-parameter setting of Video-LLaVA's second training stage, and each video is subsampled to 8 frames. Training approximately takes 1 day in our configuration.

To access Gemini-1.5-Flash and Gemini-1.5-Pro, we utilize the Vertex AI APIs on Google Cloud[7]. Specifically, all the results we obtained from Gemini are based on their May 2024 versions. We use the default sampling parameter during comment generation process. However, for FICL-Score, we empirically find that using a temperature of 0 (instead of 1) helps to significantly reduce the relative standard deviation of scores from 5% to less than 1%.

F Additional Word Cloud Visualization

Figure 10 additionally visualizes the word cloud of 4 other randomly sampled users' comments, which again present diverse language styles and subject of interest in comments.

[5] https://developers.google.com/youtube/v3/docs/comments/list

[6] https://github.com/PKU-YuanGroup/Video-LLaVA

[7] https://cloud.google.com/vertex-ai?hl=en

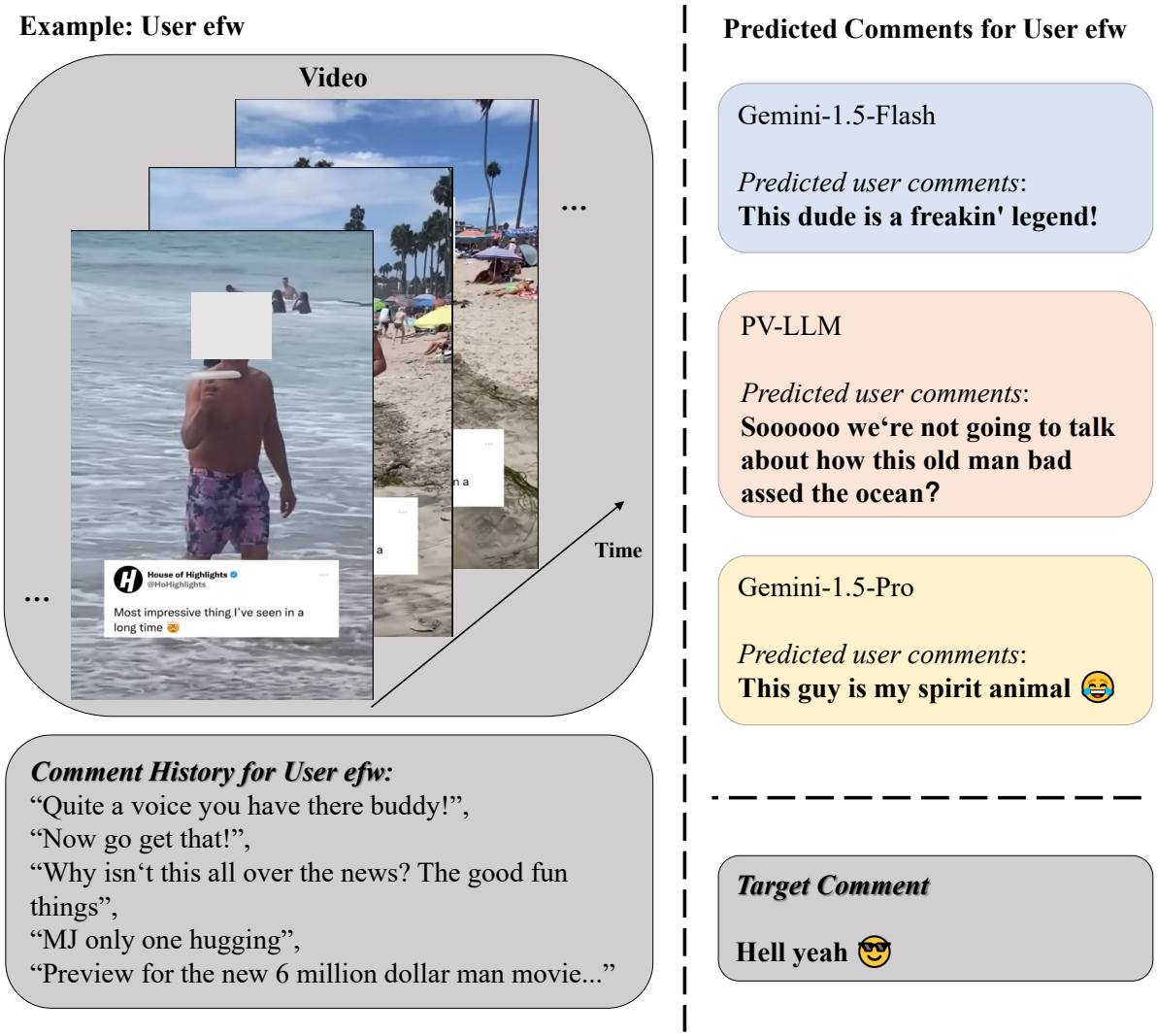

Figure 7: Visualization of the generated comments from three models.

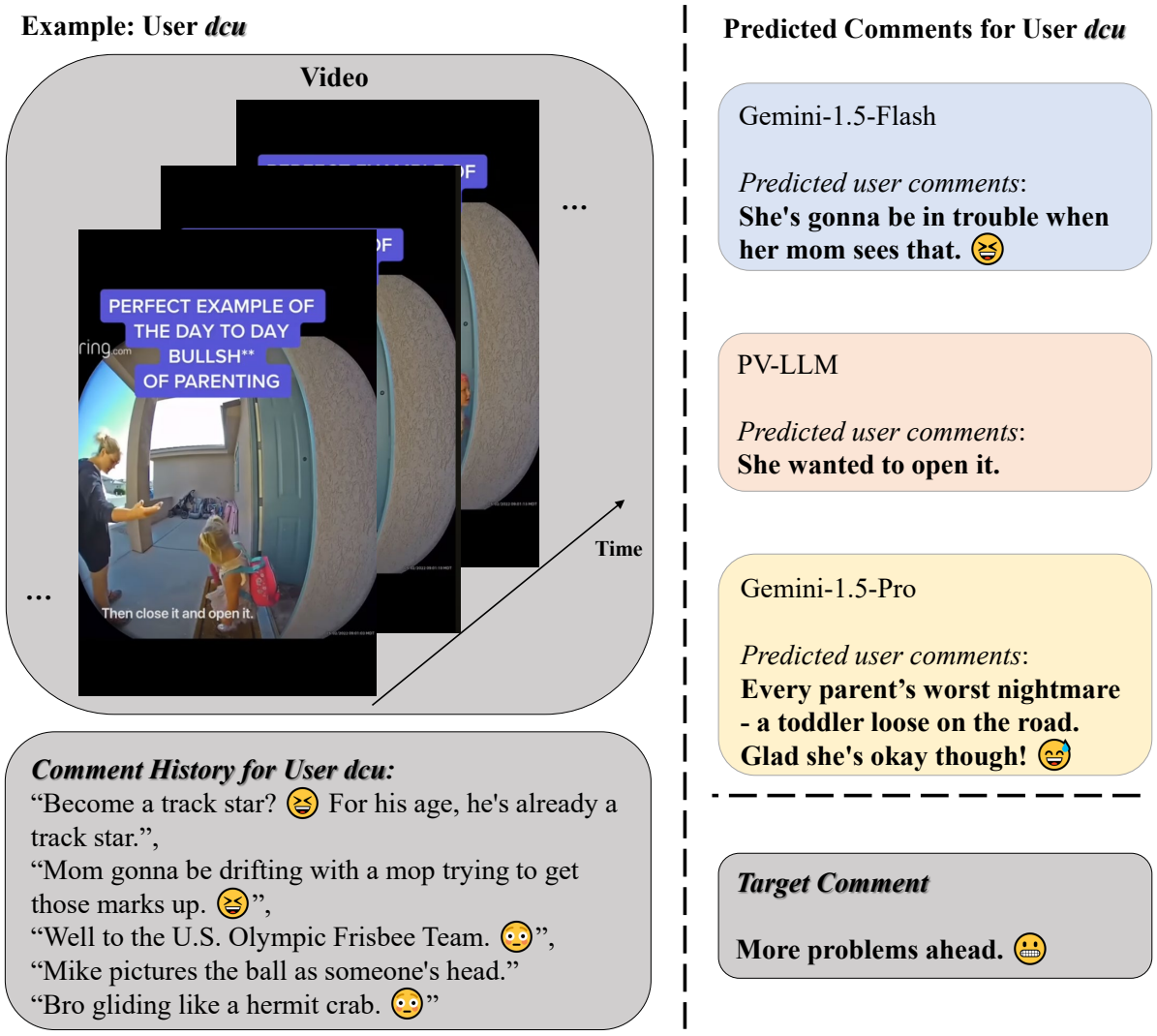

Figure 8: Visualization of the generated comments from three models.

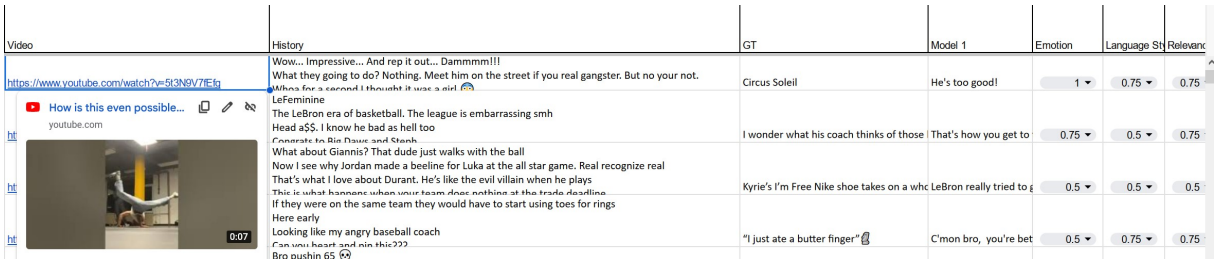

Figure 9: Human annotation interface.

(a) (b) (c) (d)

Figure 10: Word cloud visualization of four randomly sampled users from the dataset. It is clear that each user has a unique language style and subject of interest in comments.