Delving Deep into Engagement Prediction of Short Videos

Dasong Li[1,*] Wenjie Li[2] Baili Lu[2] Hongsheng Li[1,3] Sizhuo Ma[2]
Gurunandan Krishnan[2] Jian Wang[2,†]
[1]MMLab, CUHK [2]Snap Inc.
[3]Centre for Perceptual and Interactive Intelligence Limited
dasongli@link.cuhk.edu.hk, jwang4@snapchat.com

Abstract. Understanding and modeling the popularity of User Generated Content (UGC) short videos on social media platforms presents a critical challenge with broad implications for content creators and recommendation systems. This study delves deep into the intricacies of predicting engagement for newly published videos with limited user interactions. Surprisingly, our findings reveal that Mean Opinion Scores from previous video quality assessment datasets do not strongly correlate with video engagement levels. To address this, we introduce a substantial dataset comprising 90,000 real-world UGC short videos from Snapchat. Rather than relying on view count, average watch time, or rate of likes, we propose two metrics: normalized average watch percentage (NAWP) and engagement continuation rate (ECR) to describe the engagement levels of short videos. Comprehensive multi-modal features, including visual content, background music, and text data, are investigated to enhance engagement prediction. With the proposed dataset and two key metrics, our method demonstrates its ability to predict engagements of short videos purely from video content.

Keywords: Engagement Prediction · Short-form Videos

1 Introduction

With the rapid advancement of social media, an increasing number of content creators post short videos to document and share their daily lives on streaming media platforms such as TikTok, Instagram Reels, Youtube Shorts, and Snapchat Spotlight. Simultaneously, a substantial portion of users spend a significant amount of time in consuming short videos across these platforms.

Social media platforms receive a constant stream of newly published short videos. Therefore, it is important to determine to what extent each video should be recommended to users. Recommending high-quality User Generated Content (UGC) videos enhances viewer engagement and consequently encourages content creators, especially novice creators. The effective dissemination of newly

*First author. Main work was completed during an internship at Snap.
†Corresponding author

Method	Trained Dataset	Correlation of different durations			
		[19, 21)	[29, 31)	[39, 41)	[49, 51)
UVQ [40]	UGC [40]	0.084	0.156	0.290	0.289
DOVER [43]	LSVQ [49]	0.073	0.148	0.305	0.286

Table 1: Correlation between the predicted mean opinion score (MOS) scores and average watch time. The correlations are separately calculated for videos from 4 disjoint ranges of durations. "[19, 21)" refers to the videos of durations in the range of 19s to 21s, and similarly for "[29, 31)", "[39, 41)", and "[49, 51)". Small ranges are chosen to minimize the variation within each group.

published videos remains a core goal of social media platforms. However, owing to their limited user reactions, accurate recommendation of such *cold-start items* is usually a challenge. Typically, platforms would present each new video to a restricted number of users, *e.g.* one hundred. The latent popularity of each video is estimated based on the engagement metrics such as watch times from these initial users, serving as a basis for further recommendations. The cold start problem [19, 27, 39, 53] arises from the sampling bias in such limited initial interactions, resulting in noisy and inaccurate predictions of recommendation extents. This creates a negative feedback loop within the ecosystem, hindering the recommendation of high-quality videos to users. Content creators may also face delays in gauging their videos' popularity, slowing their adjustments based on viewer feedback and thus discouraging them from posting more quality content.

Previous video quality assessment (VQA) datasets [13, 34, 40, 48, 49] rely on subjective scores from relatively small groups of annotators (*e.g.* 40). These subjective scores often exhibit biases due to raters' diverse preferences and limited participation, which may not faithfully reflect a video's popularity among its true audience, gauged via metrics like average watch times. Our experiments in Table 1 reveal that VQA models [40, 43, 49] trained on these existing datasets yield very poor correlation with the popularity of short videos. While these VQA methods mainly focus on video visuals, short video engagement can be influenced by other factors like background music, content category, title, *etc.* Existing engagement prediction datasets such as Wu *et al.* [3,44] focus on limited categories of longer videos, which is not suitable for studying the engagement of short-form videos across diverse categories. Moreover, certain prerequisites [44] for historical creator information limits their applicability to videos from new creators.

To overcome the issues encountered in previous VQA datasets, we collect a large-scale UGC short video dataset named SnapUGC, which comprises publicly accessible short videos from Snapchat Spotlight. To mitigate potential biases arising from limited number of annotators, we propose to leverage engagement data from *real users*. For quantifying engagement levels of short videos, we propose to employ two key metrics: normalized average watch percentage (NAWP) and engagement continuation rate (ECR). NAWP provides an indication of the overall engagement level for videos with different durations. Meanwhile, ECR represents the probability of watch time exceeding 5 seconds, which assesses whether the video's outset is captivating enough to retain viewers' interest in

continuing to watch. It is worth noting that the two metrics are derived through aggregation from more than 2000 viewers and the dataset does not contain individual viewers' history or personal information, ensuring user privacy.

To predict engagement levels with limited user interactions, we formulate the challenge as extracting engagement solely from video content, independent of user, creator, or contextual cues. To enhance the modeling of engagement in short videos, we move beyond previous visual features [11,40,49,52,52]. Our methodology incorporates comprehensive multi-modal features such as video captioning, sound classification, titles, descriptions, and more to model the engagement levels of short videos. The seamless integration of these multi-modal features is achieved through the adoption of a cross-modal attention mechanism, enabling the harmonious fusion of visual and language-based attributes. In contrast to previous Video Quality Assessment (VQA) methods, our approach capitalizes on the incorporation of these comprehensive multi-modal features, resulting in superior performance in the engagement prediction for short videos.

The contributions of this study include: 1) We introduce a large-scale dataset to facilitate research in predicting engagement for real UGC short videos. 2) We employ two novel metrics, normalized average watch percentage and engagement continuation rate, to characterize engagement levels of short videos. 3) We investigate a diverse set of multi-modal features to strengthen the capacity of engagement prediction. 4) Using the proposed dataset and engagement metrics, our method demonstrates the ability to estimate short videos engagement in a cold start setup, highlighting its significance in the field.

2 Related Works

Video quality assessment methods Classical VQA methods [18, 21, 25, 32, 36, 37] utilize handcrafted features to evaluate video quality. Given the subjectivity and complexity of video quality, handcrafted features fall short in capturing the nuances of video quality assessment. Most previous deep VQA methods [4, 5, 9, 20, 23, 40, 49, 52] follow a two-step process: they begin by extracting deep features and subsequently train a temporal regression network using these fixed features. These deep features involves per-frame semantic features [40,49,52] from image classification networks [12,35] trained on ImageNet-1k [7], per-frame low-level distortion features [40] from low-level distortion recognition networks, and multi-frame semantic features [52] from action recognition networks [11] trained on Kinetics-400 [15]. Gated Recurrent unit (GRU) [6], InceptionTime [14] and simple average operations [52] are utilized for temporal regression of Mean Opinion Scores (MOS). Recent approaches [41–43] have emerged that opt for an end-to-end methodology, jointly optimizing feature extraction and final regression. However, these aforementioned VQA methods focus on exploring visual features while disregarding the potential contributions of additional information provided by content creators, such as background sound, title, descriptions, *etc*. The underexplored domain of vision-language correspondence [30, 51] in video quality assessment becomes apparent.

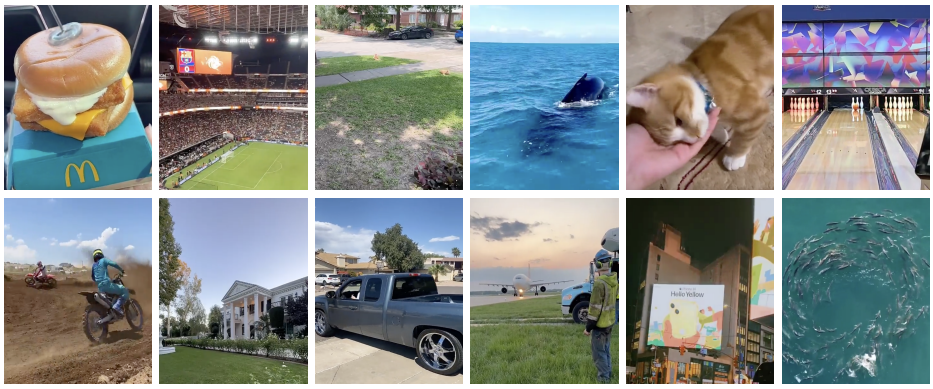


Fig. 1: Sample frames of the short videos in our dataset. The frame samples are cropped to exclude sensitive content such as human faces and watermarks for display.

Video quality datasets Early datasets [8, 26] are often designed with specialized distortions to facilitate the examination of low-level video quality. In contrast, more recent VQA datasets, such as KoNViD-1k [13], YouTube-UGC [48], LIVE-VQC [34], YT-UGC$^+$ [40], and LSVQ [49], are introduced with the aim of characterizing the subjective quality of videos. These datasets typically involve the labeling of Mean Opinion Scores (MOS) by a relatively small group of individuals. However, a notable domain gap exists between short videos and the videos in these VQA datasets. On social media platforms, users may swiftly skip uninteresting videos instead of watching the whole video, while annotators of VQA datasets tend to watch the entire video. This unique property of short videos introduces a discrepancy between engagement levels and previous MOS scores. **Engagement prediction** Previous datasets focus on analyzing engagement of video lectures [3] and YouTube videos [44]. Regrettably, there is a scarcity of publicly available datasets specifically tailored for predicting engagement for short videos. Commonly employed metrics for video engagement include view counts, average watch time, and average watch percentage. Video duration emerges as a critical covariate affecting both average watch time and average watch percentage, as illustrated in *et al*. [44,50]. Intuitively, longer videos are less likely to be watched in their entirety compared to shorter videos, a phenomenon attributed to the diminishing attention span of viewers. In response, Wu *et al*. [44] propose a relative engagement metric that accounts for varying video durations. However, the relative engagement metric takes into account the mutual connections and ranking orders among videos with similar durations. This approach may yield unstable results in the presence of sparse or uneven distributions of average watch times, as mentioned in Figure 2. Zhan *et al*. [50] propose to train the videos of different durations separately to remove the bias of video duration.

	Content			Metrics	
	Video	Audio	Text	Annotators number	Metric Sources
VQA datasets	✓	✗	✗	≤ 40	Labeling Scores
Our datasets	✓	✓	✓	≥ 2000	Real User Interactions

Table 2: We provide a detailed comparison with the VQA datasts. Our dataset contains multi-modal content to better measure the quality of videos. Moreover, our metrics are derived from thousands of real-world user interactions.

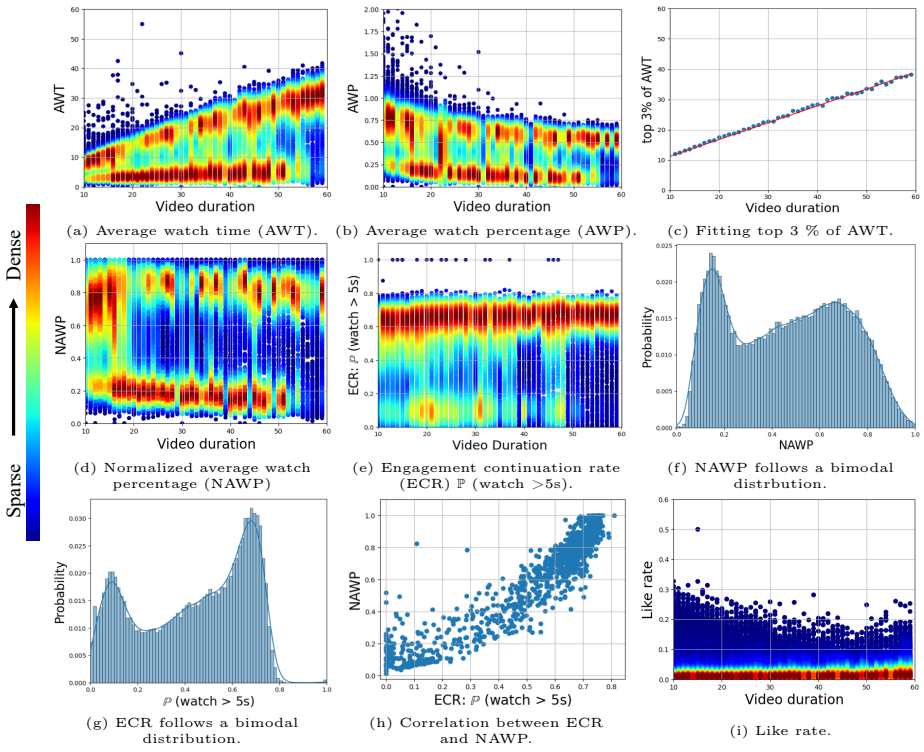
3 SnapUGC Engagement Dataset

3.1 Pilot Study

To model the engagement levels of the videos, we initially explore the use of mainstream video quality assessment (VQA) methods, commonly used for evaluating video quality. We conducted assessments using state-of-the-art video quality assessment methods [40,43] on a collection of real-world UGC short videos sourced from Snapchat Spotlight. These VQA methods were originally pre-trained on diverse VQA datasets [40,49]. As shown in Wu *et al.* [44], the average watch time can reflect the engagement levels of the videos with similar durations. Consequently, we conduct an evaluation aimed at evaluating the generalization capability of models trained on VQA datasets by calculating the correlation between Mean Opinion Score (MOS) and the engagement levels. To mitigate the potential influence [44] of video duration, we categorize the real short videos into distinct groups based on their respective durations. Within each group of similar durations, we assessed the correlation between the average watch time and the predicted MOS scores for videos. Our observation, as shown in Table 1, reveals a lack of correlation between the learned quality of pre-trained VQA methods and the engagement levels of the videos. This observation demonstrates that existing MOS scores provided by mainstream VQA datasets have difficulties in accurately reflecting the engagement levels.

3.2 Dataset Collection

While several previous datasets [10,29,45] are proposed for applications on short videos, they do not focus on video engagement analysis. To precisely model the engagement levels of real UGC short videos, we first collect a large-scale short video dataset, named SnapUGC. Our dataset comprises 90,000 short videos, all of which were published on Snapchat Spotlight. For each video, we have curated corresponding aggregated engagement data derived from viewing statistics. All short videos in our dataset have a duration ranging from 10 to 60 seconds. To mitigate sampling bias from small number of views, only short videos with view numbers exceeding 2000 are selected. The dataset is notably diverse, encompassing a wide range of video types, including Family, Food & Dining, Pets, Hobbies, Travel, Music Appreciation, Sports, etc. Several frames are shown in Figure 1. We provide a comprehensive comparison with traditional VQA datasets in Table 2.



(a) Average watch time (AWT). (b) Average watch percentage (AWP). (c) Fitting top 3 % of AWT.

(d) Normalized average watch percentage (NAWP)

(e) Engagement continuation rate (ECR) ℙ (watch >5s).

(f) NAWP follows a bimodal distrbution.

(g) ECR follows a bimodal distribution.

(h) Correlation between ECR and NAWP.

(i) Like rate.

Fig. 2: (a), (b), (e): The distributions of average watch time (AWT), average watch percentage (AWP) and engagement continuation rate (ECR), respectively. ECR, calculated as the probability of watch time exceeding 5 seconds: $\mathbb{P}$ (watch $>$ 5s), is more duration-independent. (c): We fit top 3% of average watch times to derive a universal metric for videos of different durations. (d): Further normalization of the average time is achieved by fitting a line, resulting in the normalized average watch percentage (NAWP). A color mapping is used to encode the distribution densities in (a), (b), (d) and (e). (f), (g): Distributions of NAWP and ECR. Both two metrics follow bimodal distribution, reflecting the unique property of user's swiftly skipping uninteresting videos or spend relative longer time on their interesting videos in short videos platforms. (h): The strong correlation between ECR and NAWP. (i): The distribution of like rate.

3.3 Engagement Metrics Analysis

For short videos, there are three straightforward metrics to measure viewer engagement: view numbers, like rates, and average watch time. However, each metric has its drawbacks. View numbers can be heavily influenced by recommendation systems, leading to potential bias. Short videos created by well-known content creators may receive significantly higher view numbers compared to those of new creators. Like rates, although reflective of viewer interest, often yield extremely small and indistinguishable values across different videos, pos-

ing challenges for effective learning. A detailed study on like rates is shown in Figure 2(i) and supplementary. Average watch time (AWT), while common, faces limitations when comparing videos of different durations. In this section, we first analyze the distribution and drawback of AWT, and then propose normalized average watch percentage (NAWP) as a novel engagement metric. Recognizing that users swiftly navigate through uninteresting content but persist in watching engaging videos, we introduce an additional metric: engagement continuation rate (ECR). Calculated for each video, this metric represents *the proportion of viewers who watched the video for at least 5 seconds*. It serves as an indicator of a video's ability to captivate viewers at the beginning. Unlike Kim *et al*. [16] measuring entire videos' dropout probability, ECR focuses on he contents of first several seconds, which determines whether the users would continue to watch and substantially affects watch times. The experiment in Table 5 also demonstrates the effectiveness of ECR on help learning NAWP.

Average watch time (AWT). We analyze average watch times (AWT) of various video durations d in Figure 2(a). A similar metric, average watch percentage (AWP), is calculated as AWT divided by d, and its distribution with video duration is shown in Figure 2(b). When the AWT of a video surpasses its duration, AWP exceeds 1, signifying that the video is popular to be watched repeatedly. Importantly, the distributions of AWT and AWP vary for different video durations, showing diverse user engagement patterns. Videos exhibit decreasing AWP as video duration increased, suggesting users' reduced likelihood of watching longer videos, potentially a result of declining attention spans. Due to this duration-dependent behavior, comparing the popularity of short videos with different durations using AWT or AWP is challenging. For instance, a 30-second video with an AWT of 30 seconds and a 60-second video with the same watch time tend to have different engagement levels. Similarly, a 10-second video with an AWP of 1.0 and a 30-second video with an AWP of 1.0 may differ in engagement levels, because a shorter video is easier to be fully watched.

Normalized average watch percentage (NAWP). We introduce a straightforward metric called normalized average watch percentage (NAWP) to provide a generalized measure for videos with different durations. It is observed in Figure 2(a) that the largest values under different durations align with a linear trend. Based on the observation, we make the assumption that videos with top 3% of highest AWT, regardless of their durations, are equally most popular, while videos with an average watch time of 0 seconds are deemed the least popular. For example, a 40-second video with an AWT of 30 seconds and a 60-second video with an AWT of 40 seconds are regarded as equally most popular. Similarly, a 40-second video with an AWT of 0 seconds and a 60-second video with an AWT of 0 seconds are considered equally least popular. The maximum average watch time $f_{\max}(d)$ for most popular videos and minimum average watch time $f_{\min}(d)$ for the least popular videos can be modeled by two linear functions:

$$f_{\max}(d) = 0.556 \times d + 5.64;\ \ f_{\min}(d) = 0. \tag{1}$$

$f_{\max}(d)$ is shown in Figure 2(c). The NAWP for any video of d seconds, with average watch time t is derived through normalization between $f_{\min}(d)$ and $f_{\max}(d)$:

$$\mathrm{NAWP}(\mathrm{AWT}, d) = \min\left(\frac{\mathrm{AWT} - f_{\min}(d)}{f_{\max}(d) - f_{\min}(d)}, 1\right). \quad (2)$$

The relationship between the video duration and NAWP is depicted in Figure 2(d). The NAWP falls within the range of [0, 1] and NAWP of videos with top 3% average watch time is set to be 1. The experiments in Table 4 shows that training with NAWP achieves much better performances than AWT or AWP.

Engagement continuation rate (ECR). As shown in Figure 2(e), engagement continuation rate (ECR), calculated as $\mathbb{P}$ (watch >5s), demonstrates stable behavior across different video durations. The majority of values fall within the range of [0, 0.8]. The observation aligns with the metric's focus on frames within first 5 seconds. Furthermore, we observe a robust correlation of 0.926 between ECR and NAWP, as shown in Figure 2(h). Videos with higher probabilities of watch time surpassing 5 seconds tend to exhibit longer average watch times, illustrating a strong correlation between these two metrics. This finding offers valuable insights for designing the network structure and joint training strategy, to be shown in Section 4.3 and Table 5.

Bimodal distributions. It is observed in Figure 2(f) and (g), that distributions of NAWP and ECR exhibit a bimodal pattern. Compared with the single peak distribution of MOS scores [1, 52], this bimodal distribution is **unique** to our dataset. This behavior exists due to the common UI designs that encourages "swiping" to skip boring videos in short video platforms. Users usually quickly skip through uninteresting videos, whereas they tend to dedicate relatively longer time to engaging with videos they find interesting. Consequently, it results in two separate peaks in the distributions.

Generalizability of NAWP. While NAWP is designed based on the linearity observation on our SnapUGC dataset, It is obversed in supplementary that *the linear approximation* can generalize to average watch time of *Kuaishou [50] and Youtube [44] datasets* for videos with short durations ($\leq$ 60s), which are exactly the domain of most short videos, explored in this paper.

4 Methods

In this section, we begin by presenting the natural bias of recommendation systems and formulate the engagement prediction. Then we conduct an in-depth exploration of the multi-modal features that aid engagement prediction in Section 4.2. In Section 4.3, we provide details about our network, and in Section 4.4, we outline the evaluation criteria.

4.1 Problem Formulation

Notably, the normalized average watch percentage (NAWP) and engagement continuation rate (ECR) are contingent upon the recommendation system, denoted as **R**. Recommendation systems often employ machine learning classifiers [10, 11] to categorize short videos and analyze user preferences based on

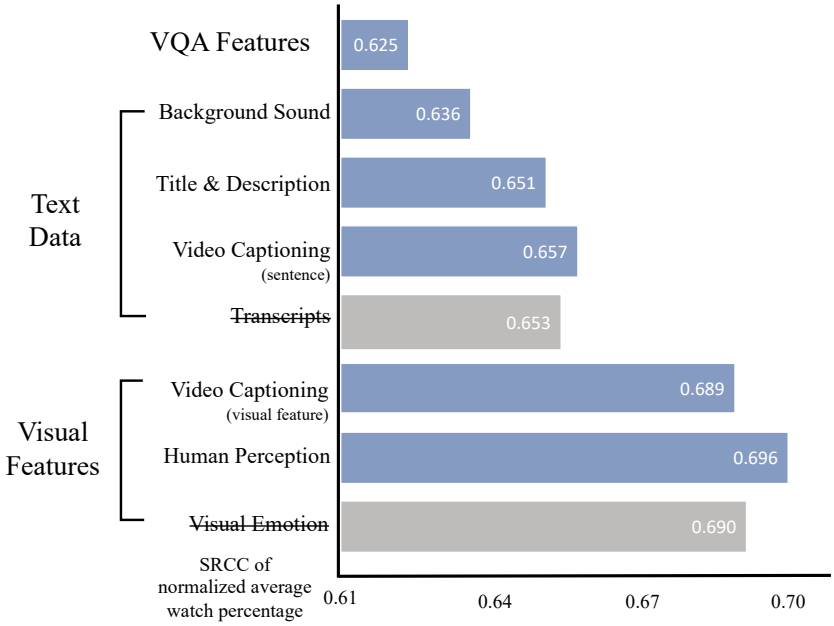


Fig. 3: The effectiveness of comprehensive multi-modal features to enhance engagement prediction. The blue bars represent incrementally incorporating new features to achieve improved SRCC, while a gray bar indicates that the modification was not adopted. These multi-modal features incorporated into our network leads to increasingly better performance than previous VQA features.

their historical engagements with various video types. These systems balance exploitation (recommending familiar contents and familiar creators) and exploration (introducing new contents and creators) to users. Consequently, the preference distribution for a given short video may vary depending on the exploitation strategy employed by different recommendation systems. The engagement metrics are biased due to the preference distribution provided by the recommendation system $\mathbf{R}$. Therefore, we formulate engagement prediction as a realistic conditional problem. For a given short video v and the recommendation system $\mathbf{R}$, our network G predicts the normalized average watch percentage $\widehat{\text{NAWP}}$ and the engagement continuation rate $\widehat{\text{ECR}}$ as follows:

$$(\widehat{\text{NAWP}},\ \widehat{\text{ECR}}) = G(v \mid \mathbf{R}). \tag{3}$$

We only focus on aggregated metric in this work as individual user's metric is subject to legal and privacy concerns.

4.2 Comprehensive Features for Engagement Prediction

To precisely model the engagement levels of short videos, we investigate a comprehensive set of multi-modal features. The evaluation of various features is con-

ducted using the Spearman Rank Correlation Coefficient (SRCC) of the normalized average watch percentage (NAWP). We utilize T5 [31] as the text encoder to encode the text data. In Figure 3, we show the procedure and the incremental performance achieved by gradually incorporating each feature. In particular, our exploration focuses on the following aspects:

- **VQA features.** Building on established video quality assessment methods UVQ [40] and MD-VQA [52], we extract per-frame semantic features [35] per-frame distortion features [40], and action recognition features [11] for video clips. These features collectively offer a fundamental assessment of both content and objective quality. This baseline gives a correlation of 0.625.
- **Background sound.** Creators usually incorporate background music in short videos to enhance the atmosphere and attract viewers. We employ YAMNet [2], a 521-class audio event classification model, to discern various types of background music. The top 5 classification results, presented as text, are then utilized as an additional network input to augment the modeling of video engagement. This improves the performance from 0.625 to 0.636.
- **Title and descriptions** are usually provided along with the short videos by the creators, which can emphasize key content and provide additional context information, enhancing the overall understanding of the videos. Incorporating the title and description leads to an increase from 0.636 to 0.651.
- **Video captioning.** Video captioning provides fine-grained understanding of the short videos. Leveraging mid-layer features and captions generated by mPLUG-2 [46] as complementary features enhances engagement predictions. The captions would also provide new insights for interpreting video popularity. The inclusion of captions increases the performance slightly to 0.657. Adding intermediate features as additional input visual features brings a significant improvements from 0.657 to 0.689.
- **Transcripts.** Ideally, transcripts would facilitate a better understanding of video content. However, our findings indicate that adding transcripts does not yield improvements. This observation can be attributed to the fact that only 30% of short videos include effective transcripts. Additionally, viewers often decide whether to continue watching based on the initial seconds, during which they only catch a small amount of the spoken content.
- **Human asethetic preference.** While the semantic [35] and action features [11] described above contain semantic information, they may not directly capture human reactions and feelings when watching videos. In response, Wu *et al.* [43] proposed a mean aesthetic option score to measure human quality opinions solely from an aesthetic perspective. Leveraging human aesthetic preferences may contribute to modeling the popularity of short videos. Therefore, we integrate the aesthetic features extracted from pretrained models in [43], resulting in an increase from 0.689 to 0.696.
- **Visual emotion.** Creators often convey emotions through short videos, and these emotions can be reflected in the visual sentiment captured in individual frames. To evaluate the potential benefits of emotion information, we employ WSCNet [33, 47], trained on the WEBEmo dataset [28] to obtain

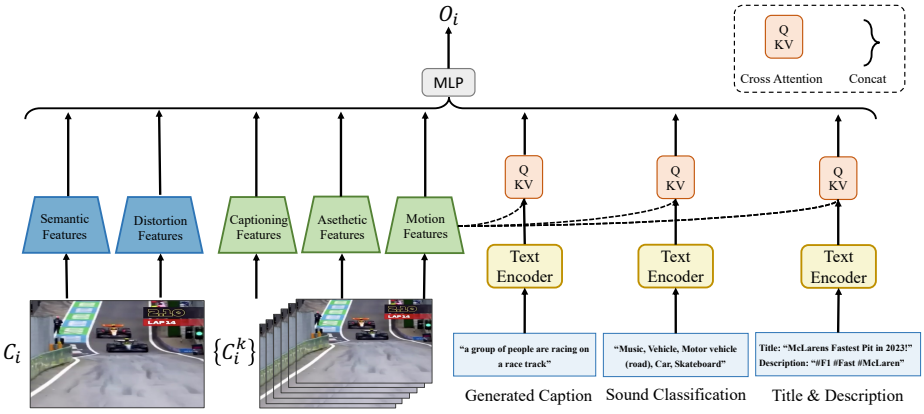


Fig. 4: The overview of multi-modal feature extractions. The learnable Multilayer Perceptron (MLP) to process extracted features is omitted for simplicity.

intermediate features. The observed change from 0.696 to 0.690 suggests a limited correlation between visual sentiment and engagement levels.

4.3 Network Details

Following MD-VQA [52], we split the video into several clips for efficient feature extraction. Given a video with frame count M and frame rate r, we create $\frac{M}{L}$ clips $\{C_i\}_{i=1}^{M/L}$ with each clip C_i containing L frames $\{C_i^k\}_{k=1}^{L}$. Our network takes visual features and text data as inputs. The feature extraction is shown in Figure 4. For each clip C_i, we extract semantic and distortion features for each of the L frames C_i^k, while the entire clip C_i is used for action feature extraction, asethetic feature extraction and video captioning feature extraction. The text data, which include background sound classification, title, descriptions, and generated captions, are shared among all the clips. We process the visual features with learnable Multi-Layer Perceptrons (MLP) and employ cross-attention to merge visual action features with text data. Then the multi-modal features are fused by 8 MLP layers to obtain the fused features $\{O_i\}_{i=1}^{M/L}$. Subsequently, we utilize a 8-layer self-attention architecture to combine the fused features $\{O_i\}_{i=1}^{M/L}$ of all the clips to obtain temporal aggregated features $\{H_i\}_{i=1}^{M/L}$. Finally, our network utilizes 2 MLP layers $F_{\text{out}}^1, F_{\text{out}}^2$ to jointly predict $\widehat{\text{NAWP}}$ and $\widehat{\text{ECR}}$:

$$\widehat{\text{NAWP}} = \frac{L}{M}\sum_{i=1}^{M/L} F_{\text{out}}^1(H_i); \widehat{\text{ECR}} = \frac{L}{5r}\sum_{i=1}^{5r/L} F_{\text{out}}^2(H_i), \tag{4}$$

where $\widehat{\text{ECR}}$ is derived from frames within first 5 seconds. The joint training loss L for $\widehat{\text{NAWP}}$ and $\widehat{\text{ECR}}$ is derived:

$$L = ||\text{NAWP} - \widehat{\text{NAWP}}||_2 + ||\text{ECR} - \widehat{\text{ECR}}||_2. \tag{5}$$

Method	NAWP			ECR			NAWP	ECR
	SRCC↑	PLCC↑	RMSE↓	SRCC↑	PLCC↑	RMSE↓	$\text{RMSE}_{\text{top 10\%}}$ ↓	$\text{RMSE}_{\text{top 10\%}}$ ↓
VSFA [20]	0.609	0.615	0.192	0.576	0.591	0.197	0.199	0.174
PVQ [49]	0.590	0.607	0.197	0.587	0.602	0.194	0.189	0.170
MD-VQA [52]	0.606	0.614	0.193	0.592	0.608	0.191	0.187	0.166
FastVQA [41]	0.587	0.590	0.218	0.581	0.585	0.223	0.232	0.201
DOVER [43]	0.635	0.636	0.206	0.619	0.622	0.203	0.216	0.189
Ours-VQA	0.625	0.632	0.188	0.605	0.620	0.189	0.191	0.171
Ours	**0.696**	**0.701**	**0.172**	**0.675**	**0.688**	**0.174**	**0.181**	**0.152**

Table 3: Experimental performances of NAWP and ECR on the proposed engagement prediction dataset. "Ours-VQA" denotes merely utilizing VQA features (per-frame semantic features, per-frame distortion features and per-clip action recognition features).

It is observed in our experiments (Table 5) that training these two highly correlated metrics jointly leads to enhanced overall performance.

4.4 Evaluation Criteria

We evaluate our method using common criteria in Video Quality Assessment (VQA) research, including Spearman Rank Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC) and Root Mean Square Error (RMSE) for both NAWP and ECR. Drawing insights from the observations in Figure 2(f) and (g), we empirically note that only 10% to 20% of uploaded short videos, centered around the second peak of the bimodal distributions, emerge as popular and are prioritized by the recommendation system. Therefore, we consider the top $K\%$ of N test videos with the highest NAWP. For these selected videos $\{v_j\}_{j=0}^{K\times N/100}$, we calculate RMSE of top 10% of NAWP as follows:

$$\text{RMSE}_{\text{top10\%}} = \sqrt{\frac{(\sum_{j=0}^{K\times N/100}(\widehat{\text{NAWP}}_j - \text{NAWP}_j)^2)}{K \times N/100}}, \tag{6}$$

which is similar for the RMSE of top $K\%$ of ECR. (K=10 in our evaluation.)

5 Experiments

5.1 Implementation Details

For the SnapUGC dataset, we adhere to the common practice and spilt the dataset with an 90%∼10% train-test ratio. Our network G takes extracted features as input and regresses $\widehat{\text{NAWP}}$ and $\widehat{\text{ECR}}$. All feature extraction networks are pre-trained separately. We follow UVQ [40] to train a distortion recognition network on KADIS-700K and KADID-10K [22]. The per-frame semantic features are extracted by EfficientNet [35], pre-trained on ImageNet [7]. The per-clip action recognition features are extracted by ResNet-3D [11], pre-trained on Kinetics-400 [15]. The video caption and mid-layer features are extracted from

Learning metrics	Duration as input	Average SRCC
AWP	✗	0.665
AWP	✓	0.681 (↑)
AWT	✗	0.668
AWT	✓	0.683 (↑)
NAWP	✗	**0.696**
NAWP	✓	0.689 (↓)

Table 4: We compare proposed NAWP with average watch percentage (AWP) and average watch time (AWT). "Duration as input" means adding the video duration as a network input. We divide the videos to different groups according to their video durations and average the SRCC of different groups to obtain "Average SRCC".

Training setting	NAWP	ECR
Separate training	0.662	0.681
Joint training	**0.675**	**0.696**

Table 5: Ablation of joint training normalized average watch percentage (NAWP) and engagement continuation rate (ECR).

the pre-trained video captioning model mPLUG-2 [46]. Human aesthetic features are extracted by the pre-trained model in DOVER [43]. We utilize T5 [31] as a text encoder to encode the text data, including generated captions, sound classification results, titles, and descriptions. The network is trained with a batch size of 8 for 70,000 iterations. We use the Adam optimizer [17] and the learning rate is decreased from 1×10^{-4} to 1×10^{-7} according to the cosine annealing strategy [24]. The parameter L is set to be 16. We optimize $\widehat{\text{NAWP}}$ and $\widehat{\text{ECR}}$ jointly following Eq (5). More details are provided in supplementary materials.

5.2 Engagement Results

To evaluate the performance of the proposed framework, we select popular quality assessment methods for comparisons, including VSFA [20], PVQ [49], MD-VQA [52], FastVQA [41], and DOVER [43]. *Our network and these VQA methods are trained* on the proposed engagement dataset to learn the normalized average watch percentage (NAWP) and engagement continuation rate (ECR) jointly. As conventional distortion features in MD-VQA [52] are not available, we substitute them with distortion networks from UVQ [40]. We enhance the models by adding an additional final layer of VSFA [20], PVQ [49], and MD-VQA [52] to make them adaptive for joint training with two metrics. Due to the frames sampling in FastVQA [41] and DOVER [43], we train two separate models to predict NAWP and ECR. The sampling range for NAWP is set to frames of whole videos, while the sampling range for ECR is set to frames within the first 5 seconds. All VQA models are trained with the default parameters defined by their respective authors.

The experimental performance on the proposed dataset is shown in Table 3. "Ours-VQA" denotes the model merely incorporating VQA features (per-frame

semantic features, per-frame distortion features and per-clip action recognition features). The difference between "Ours-VQA" and MD-VQA [52] lies in the utilization of self-attention layers [38], resulting in a 0.17 improvement in SRCC for NAWP. Although DOVER [43] outperforms 'Ours-VQA' on SRCC, it exhibits significantly poorer results on RMSE. "Ours-VQA" achieves balanced performance across SRCC, PLCC, and RMSE. Benefiting from the integration of complementary multi-modal features, our method outperforms state-of-the-art VQA models by a clear margin.

5.3 Ablation Study

Normalized average watch percentage. To evaluate the performance of normalized average watch percentage (NAWP), we train two models with average watch time (AWT) and average watch percentage (AWP). Since both AWT and AWP are duration-dependent metrics, calculating the correlation among videos of different durations is not feasible. Therefore, we categorize videos into groups based on their durations and compute SRCC for average watch percentage within each group. The average SRCC across these groups served as the evaluation metric in this ablation study. Table 4 illustrates that the proposed NAWP outperforms AWT and AWP by a significant margin. We also explore incorporating video duration as a network input, as Wu *et al*. [44] do. Although incorporating video duration as input leads to improvements in the learning performance associated with AWT and AWP, their performances are still worse than learning the proposed NAWP. Given that NAWP is more duration-independent, the incorporation of video duration as a network input cannot yield better improvements and can lead to potential confusion and overfitting. Furthermore, the models trained with AWT and AWP are unsuitable for comparing two videos with different durations, as detailed in supplementary materials.
Jointly training with two metrics. We conducted an experiment between joint training of two metrics and separate training of each metric. As illustrated in Table 5, joint training significantly enhances the performance of both metrics. The boost performance indicates that the strong correlation between the two metrics contributes to the ability of joint training to achieve higher performances.

6 Conclusion

In this paper, we first reveal the limitation of using mean opinion scores from previous video quality datasets to model popularity. To overcome this, we curate a large-scale dataset of real-world short videos and conduct a detailed analysis of engagement metrics and their correlations. We further investigate comprehensive multi-modal features to enhances the model's performance. The resultant model achieves state-of-the-art performance in predicting engagement for short videos.

References

1. Wang, Haiqiang and Li, Gary and Liu, Shan and Kuo, C.-C. Jay, "ICME 2021 UGC-VQA Challenge.", [Online] Available: http://ugcvqa.com/
2. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), `https://www.tensorflow.org/`, software available from tensorflow.org
3. Bulathwela, S., Perez-Ortiz, M., Yilmaz, E., Shawe-Taylor, J.: VLEngagement: A Dataset of Scientific Video Lectures for Evaluating Population-based Engagement. arXiv e-prints arXiv:2011.02273 (Nov 2020). `https://doi.org/10.48550/arXiv.2011.02273`
4. Chen, B., Zhu, L., Li, G., Lu, F., Fan, H., Wang, S.: Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment. IEEE Transactions on Circuits and Systems for Video Technology **32**(4), 1903–1916 (2022). `https://doi.org/10.1109/TCSVT.2021.3088505`
5. Chen, P., Li, L., Ma, L., Wu, J., Shi, G.: Rirnet: Recurrent-in-recurrent network for video quality assessment. In: Proceedings of the 28th ACM International Conference on Multimedia. p. 834–842. MM '20, Association for Computing Machinery, New York, NY, USA (2020). `https://doi.org/10.1145/3394171.3413717`, `https://doi.org/10.1145/3394171.3413717`
6. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1724–1734. ACL (2014)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255 (2009)
8. Ghadiyaram, D., Pan, J., Bovik, A.C., Moorthy, A.K., Panda, P., Yang, K.C.: In-capture mobile video distortions: A study of subjective behavior and objective algorithms. IEEE Transactions on Circuits and Systems for Video Technology **28**(9), 2061–2077 (2018)
9. Götz-Hahn, F., Hosu, V., Lin, H., Saupe, D.: Konvid-150k: A dataset for no-reference video quality assessment of videos in-the-wild. In: IEEE Access 9. pp. 72139–72160. IEEE (2021)
10. Gupta, V., Mittal, T., Mathur, P., Mishra, V., Maheshwari, M., Bera, A., Mukherjee, D., Manocha, D.: 3massiv: Multilingual, multimodal and multi-aspect dataset of social media short videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21064–21075 (June 2022)
11. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6546–6555 (2018)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)

13. Hosu, V., Hahn, F., Jenadeleh, M., Lin, H., Men, H., Szirányi, T., Li, S., Saupe, D.: The konstanz natural video database (konvid-1k). In: Ninth International Conference on Quality of Multimedia Experience (QoMEX). pp. 1–6 (2017)
14. Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D.F., Weber, J., Webb, G.I., Idoumghar, L., Muller, P.A., Petitjean, F.: Inceptiontime: Finding alexnet for time series classification. Data Mining and Knowledge Discovery (2020)
15. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, A., Suleyman, M., Zisserman, A.: The kinetics human action video dataset. ArXiv **abs/1705.06950** (2017)
16. Kim, J., Guo, P.J., Seaton, D.T., Mitros, P., Gajos, K.Z., Miller, R.C.: Understanding in-video dropouts and interaction peaks inonline lecture videos. In: Proceedings of the First ACM Conference on Learning @ Scale Conference. p. 31–40. L@S '14, Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2556325.2566237, https://doi.org/10.1145/2556325.2566237
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980
18. Korhonen, J.: Two-level approach for no-reference consumer video quality assessment. IEEE Transactions on Image Processing **28**(12), 5923–5938 (2019)
19. Lee, H., Im, J., Jang, S., Cho, H., Chung, S.: Melu: Meta-learned user preference estimator for cold-start recommendation. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 1073–1082. KDD '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3292500.3330859, https://doi.org/10.1145/3292500.3330859
20. Li, D., Jiang, T., Jiang, M.: Quality assessment of in-the-wild videos. In: Proceedings of the 27th ACM International Conference on Multimedia. p. 2351–2359. MM '19, Association for Computing Machinery, New York, NY, USA (2019)
21. Liao, L., Xu, K., Wu, H., Chen, C., Sun, W., Yan, Q., Lin, W.: Exploring the effectiveness of video perceptual representation in blind video quality assessment. In: Proceedings of the 30th ACM International Conference on Multimedia (ACM MM) (2022)
22. Lin, H., Hosu, V., Saupe, D.: Kadid-10k: A large-scale artificially distorted iqa database. In: 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX). pp. 1–3 (2019). https://doi.org/10.1109/QoMEX.2019.8743252
23. Liu, Y., Zhou, X., Yin, H., Wang, H., Yan, C.: Efficient video quality assessment with deeper spatiotemporal feature extraction and integration. Journal of Electronic Imaging **30**, 063034 (Nov 2021). https://doi.org/10.1117/1.JEI.30.6.063034
24. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net (2017), https://openreview.net/forum?id=Skq89Scxx
25. Mittal, A., Saad, M.A., Bovik, A.C.: A completely blind video integrity oracle. IEEE Transactions on Image Processing **25**(1), 289–300 (2016)
26. Nuutinen, M., Virtanen, T., Vaahteranoksa, M., Vuori, T., Oittinen, P., Häkkinen, J.: Cvd2014—a database for evaluating no-reference video quality assessment algorithms. IEEE Transactions on Image Processing **25**(7), 3073–3086 (2016)

27. Pan, F., Li, S., Ao, X., Tang, P., He, Q.: Warm up cold-start advertisements: Improving ctr predictions via learning to learn id embeddings. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 695–704. SIGIR'19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3331184.3331268, https://doi.org/10.1145/3331184.3331268
28. Panda, R., Zhang, J., Li, H., Lee, J.Y., Lu, X., Roy-Chowdhury, A.K.: Contemplating visual emotions: Understanding and overcoming dataset bias. In: European Conference on Computer Vision (2018)
29. Qing-Yuan, J., Yi, H., Gen, L., Jian, L., Lei, L., Wu-Jun, L.: SVD: A large-scale short video dataset for near-duplicate video retrieval. In: Proceedings of International Conference on Computer Vision (2019)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021), http://proceedings.mlr.press/v139/radford21a.html
31. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020), http://jmlr.org/papers/v21/20-074.html
32. Saad, M.A., Bovik, A.C., Charrier, C.: Blind image quality assessment: A natural scene statistics approach in the dct domain. IEEE Transactions on Image Processing **21**(8), 3339–3352 (2012)
33. She, D., Yang, J., Cheng, M.M., Lai, Y.K., Rosin, P.L., Wang, L.: Wscnet: Weakly supervised coupled networks for visual sentiment classification and detection. IEEE Transactions on Multimedia (2019)
34. Sinno, Z., Bovik, A.C.: Large-scale study of perceptual video quality. IEEE Transactions on Image Processing **28**(2), 612–627 (2019)
35. Tan, M., Le, Q.V.: Efficientnetv2: Smaller models and faster training. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 10096–10106. PMLR (2021), http://proceedings.mlr.press/v139/tan21a.html
36. Tu, Z., Chen, C.J., Wang, Y., Birkbeck, N., Adsumilli, B., Bovik, A.C.: Efficient user-generated video quality prediction. In: 2021 Picture Coding Symposium (PCS). pp. 1–5 (2021)
37. Tu, Z., Wang, Y., Birkbeck, N., Adsumilli, B., Bovik, A.C.: Ugc-vqa: Benchmarking blind video quality assessment for user generated content. IEEE Transactions on Image Processing **30**, 4449–4464 (2021)
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6000–6010. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
39. Volkovs, M., Yu, G., Poutanen, T.: Dropoutnet: Addressing cold start in recommender systems. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 4964–4973. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)

40. Wang, Y., Ke, J., Talebi, H., Yim, J.G., Birkbeck, N., Adsumilli, B., Milanfar, P., Yang, F.: Rich features for perceptual quality assessment of ugc videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13435–13444 (June 2021)
41. Wu, H., Chen, C., Hou, J., Liao, L., Wang, A., Sun, W., Yan, Q., Lin, W.: Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. Proceedings of European Conference of Computer Vision (ECCV) (2022)
42. Wu, H., Chen, C., Liao, L., Hou, J., Sun, W., Yan, Q., Gu, J., Lin, W.: Neighbourhood representative sampling for efficient end-to-end video quality assessment (2022)
43. Wu, H., Zhang, E., Liao, L., Chen, C., Hou, J., Wang, A., Sun, W., Yan, Q., Lin, W.: Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20144–20154 (October 2023)
44. Wu, S., Rizoiu, M.A., Xie, L.: Beyond views: Measuring and predicting engagement in online videos. Proceedings of the International AAAI Conference on Web and Social Media **12**(1) (Jun 2018). https://doi.org/10.1609/icwsm.v12i1.15031, https://ojs.aaai.org/index.php/ICWSM/article/view/15031
45. Wu, X., Hu, P., Wu, Y., Lyu, X., Cao, Y.P., Shan, Y., Yang, W., Sun, Z., Qi, X.: Speech2lip: High-fidelity speech to lip generation by learning from a short video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22168–22177 (October 2023)
46. Xu, H., Ye, Q., Yan, M., Shi, Y., Ye, J., Xu, Y., Li, C., Bi, B., Qian, Q., Wang, W., Xu, G., Zhang, J., Huang, S., Huang, F., Zhou, J.: mplug-2: A modularized multi-modal foundation model across text, image and video. ArXiv **abs/2302.00402** (2023)
47. Yang, J., She, D., Lai, Y.K., Rosin, P.L., Yang, M.H.: Weakly supervised coupled networks for visual sentiment analysis. In: The IEEE Conference on Computer Vision and Pattern Recognition (2018)
48. Yim, J.G., Wang, Y., Birkbeck, N., Adsumilli, B.: Subjective quality assessment for youtube ugc dataset. In: 2020 IEEE International Conference on Image Processing (ICIP). pp. 131–135 (2020)
49. Ying, Z., Mandal, M., Ghadiyaram, D., Bovik, A.: Patch-vq: 'patching up' the video quality problem. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)). pp. 14019–14029 (June 2021)
50. Zhan, R., Pei, C., Su, Q., Wen, J., Wang, X., Mu, G., Zheng, D., Jiang, P., Gai, K.: Deconfounding duration bias in watch-time prediction for video recommendation. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. p. 4472–4481. KDD '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3534678.3539092, https://doi.org/10.1145/3534678.3539092
51. Zhang, W., Zhai, G., Wei, Y., Yang, X., Ma, K.: Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14071–14081 (June 2023)
52. Zhang, Z., Wu, W., Sun, W., Tu, D., Lu, W., Min, X., Chen, Y., Zhai, G.: Md-vqa: Multi-dimensional quality assessment for ugc live videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1746–1755 (June 2023)

53. Zhu, Y., Xie, R., Zhuang, F., Ge, K., Sun, Y., Zhang, X., Lin, L., Cao, J.: Learning to warm up cold item embeddings for cold-start recommendation with meta scaling and shifting networks. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1167–1176. SIGIR '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3404835.3462843, https://doi.org/10.1145/3404835.3462843