Supplementary Materials – 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

1 Analysis of Different Categories

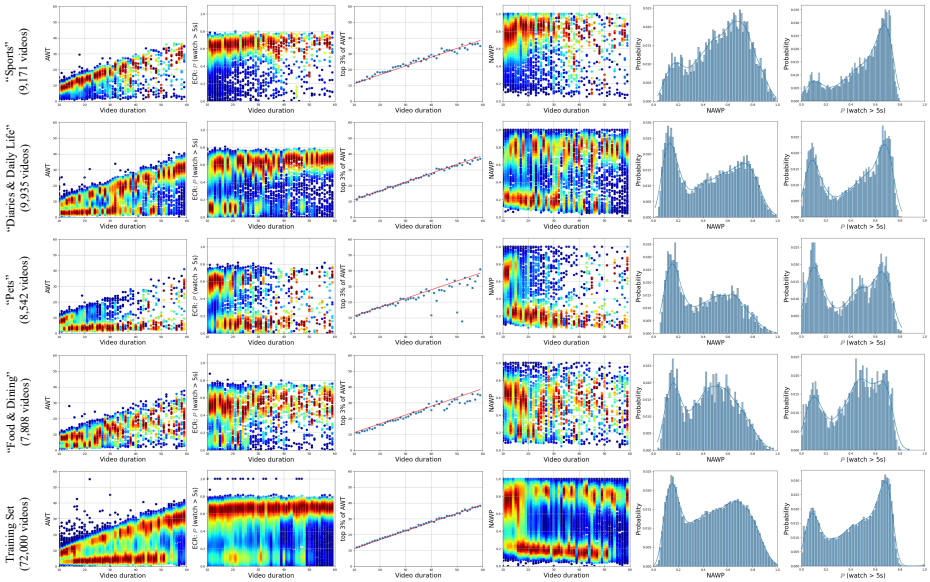
The short videos are systematically categorized by machine learning classifiers. Figure 1 illustrates the visualization of key metrics, including average watch time (AWT), engagement continuation rate (ECR), and Normalized Average Watch Percentage (NAWP). Notably, distinct distributions are observed across various video categories for AWT, ECR, and NAWP. For instance, the "Sports" category consistently demonstrates higher AWT and ECR values, whereas the "Pets" category exhibits comparatively lower values. The "Diaries & Daily Life" and "Food & Dining" categories mirror the overall dataset's distribution. In Figure 1(c), the red line, representing the top 3% of AWT, is observed to be applicable across different categories, indicating the effectiveness and generalizability of NAWP. In contrast, as demonstrated in Wu et al. [5], relative engagement relies on ranking and may prove less effective when confronted with the diverse distributions present in various video categories.

Adding category into network. Given the distinct distributions of Normalized Average Watch Percentage (NAWP) and Engagement Continuation Rate (ECR) across different categories of short videos, there arises a potential benefit in incorporating category classification information into the network to enhance the learning of engagement levels. However, our experimental findings reveal that the addition of category information does not improve the performance of our method. This observation may be attributed to the inherent inclusion of latent category information within the features derived from motion recognition [1] and video captioning [6] incorporated into the network.

2 Engagement Continuation Rate

The threshold of 5-second. The Engagement Continuation Rate (ECR) is a metric that captures the probability $\mathbb{P}$ (watch > 5), signifying the likelihood of viewers extending their watch time beyond 5 seconds. This metric serves to depict the video popularity within the initial 5 seconds of playback. The choice

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



(a) AWT (b) ECR (c) Fitting 3% of AWT (d) NAWP (e) Distribution of NAWP (f) Distribution of ECR

Fig. 1: Analysis of different video categories. (a): The relation between video duration and average watch time (AWT). (b): The relation between video duration and engagement continuation rate (ECR). (c) shows that the fitting line of top 3% average watch time (AWT) can fit all the videos categories. (d): The relation between video duration and normalized average watch percentage. (e) and (f): Probability distributions of NAWP and ECR. (a), (d), (e): The "Sports" category has more videos with higher AWT and NAWP, while the "Food & Dining" category has more videos with lower AWT and NAWP. The "Diaries & Daily Life" category has the similar distributions of AWT and ECR as the whole training set. (e) and (f) show that all the videos categories follow bimodal distributions.

of a 5-second threshold is motivated by following considerations: 1) Intutively, the unique user behavior of swiftly skipping through videos necessitates a relatively short threshold. This behavior is encapsulated by the first peak in the distributions of ECR and Average Watch Time (AWT). 2) As evident from the distribution of average watch time (AWT) in Figure 2(a) of the main paper, the first peak of AWT occurs before 5 seconds. Consequently, the threshold of 5-second is selected for meaningful analysis.

Value range. Observations from Figure 2(e) and 2(g) in the main paper highlight a predominant concentration of Engagement Continuation Rate (ECR) values within the [0, 0.82] range. The bimodal distribution features prominent peaks at approximately 0.1 and 0.7. The prevalence of values in the [0, 0.82] range implies that approximately 18% of users tend to exhibit a browsing behavior characterized by frames lasting less than 5 seconds per short video. This subset of users appears to have a relatively higher threshold for extending their

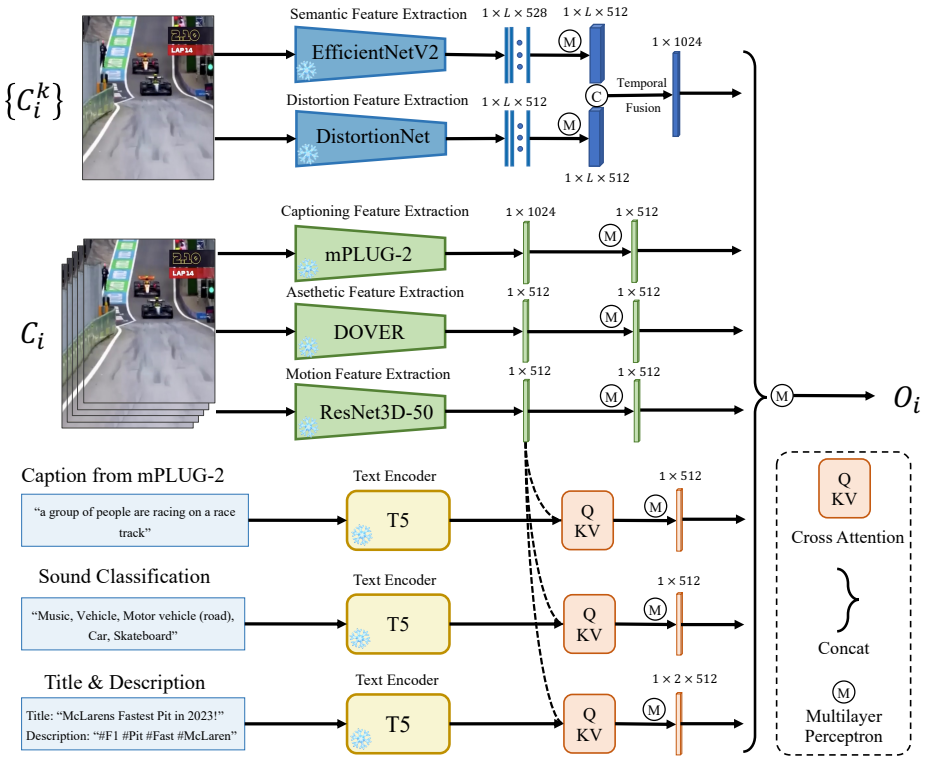


Fig. 2: The details of multi-modal feature extractions. Snowflakes refer to frozen parameters.

viewing beyond the initial frames. Additionally, Figure 1 demonstrates that ECR values within different categories share similar ranges, indicating similar browsing behaviors across categories.

Outliers. In Figure 2(e) of the main paper, several outliers with values exceeding 0.85 are apparent. Upon closer examination, it is discovered that the ECR values for 29 videos were recorded as 1.0, while one video had an ECR value of 0.875. Intriguingly, all 30 of these videos registered a like rate of 0.0. This anomaly suggests an incorrect storage of ECR outliers, possibly triggered by a storage or query bug associated with the like rate being equal to 0.0. Given the extremely small proportion of these outliers within the entire dataset and considering Normalized Average Watch Percentage (NAWP) as the primary metric, we opted to cease the gradient descent for learning ECR while continuing it for learning NAWP for these data points.

Method	Additional Text			Additional Visual		NAWP			ECR		
	Caption	Sound	T&D	Caption	Asethetic	SRCC↑	PLCC↑	RMSE↓	SRCC↑	PLCC↑	RMSE↓
Ours-VQA	✗	✗	✗	✗	✗	0.625	0.632	0.188	0.605	0.620	0.189
Ablations	✗	✓	✓	✓	✓	0.689	0.697	0.173	0.668	0.683	0.174
	✓	✗	✓	✓	✓	0.681	0.685	0.176	0.659	0.678	0.180
	✓	✓	✗	✓	✓	0.685	0.690	0.176	0.664	0.680	0.176
	✓	✓	✓	✗	✓	0.667	0.673	0.182	0.638	0.651	0.183
	✓	✓	✓	✓	✗	0.689	0.695	0.174	0.667	0.681	0.176
Ours	✓	✓	✓	✓	✓	**0.696**	**0.701**	**0.172**	**0.675**	**0.688**	**0.174**

Table 1: Ablation of comprehensive multi-modal features on SnapUGC dataset. "Additional Text" denotes text features obtained by feeding captions, sound classifications, and Title & Descriptions (T&D) through the text encoder. "Additional Visual" denotes visual features obtained from intermediate layers of the captioning model and the aesthetics model.

Method	Per-frame Semantic	Per-frame Distortion	Per-clip Motion	NAWP			ECR		
				SRCC↑	PLCC↑	RMSE↓	SRCC↑	PLCC↑	RMSE↓
Ablations	✓	✗	✗	0.587	0.593	0.201	0.590	0.604	0.195
	✓	✓	✗	0.590	0.597	0.200	0.594	0.601	0.194
	✓	✗	✓	0.618	0.627	0.189	0.600	0.617	0.193
Ours-VQA	✓	✓	✓	0.625	0.632	0.188	0.605	0.620	0.189

Table 2: Ablation of VQA features on SnapUGC dataset.

3 Network Details

We present the network details for feature extraction in Figure 2, specifying the channel numbers for each feature. Consistent with MD-VQA [8], we downsample the visual features to $C \times 1 \times 1$ to optimize computational costs and save the storage. For per-frame feature extraction, the semantic features S_i and distortion features D_i have dimensions $\mathbb{R}^{1\times L\times 512}$. The temporal fusion (TF) module reshapes the per-frame features S_i and D_i into the $S_i^r \in \mathbb{R}^{1\times(512\times L)}$ and $D_i^r \in \mathbb{R}^{1\times(512\times L)}$. Subsequently, a two-layer MLP merges these two features, expressed as:

$$M_i = \omega(S_i^r,\ D_i^r), \tag{1}$$

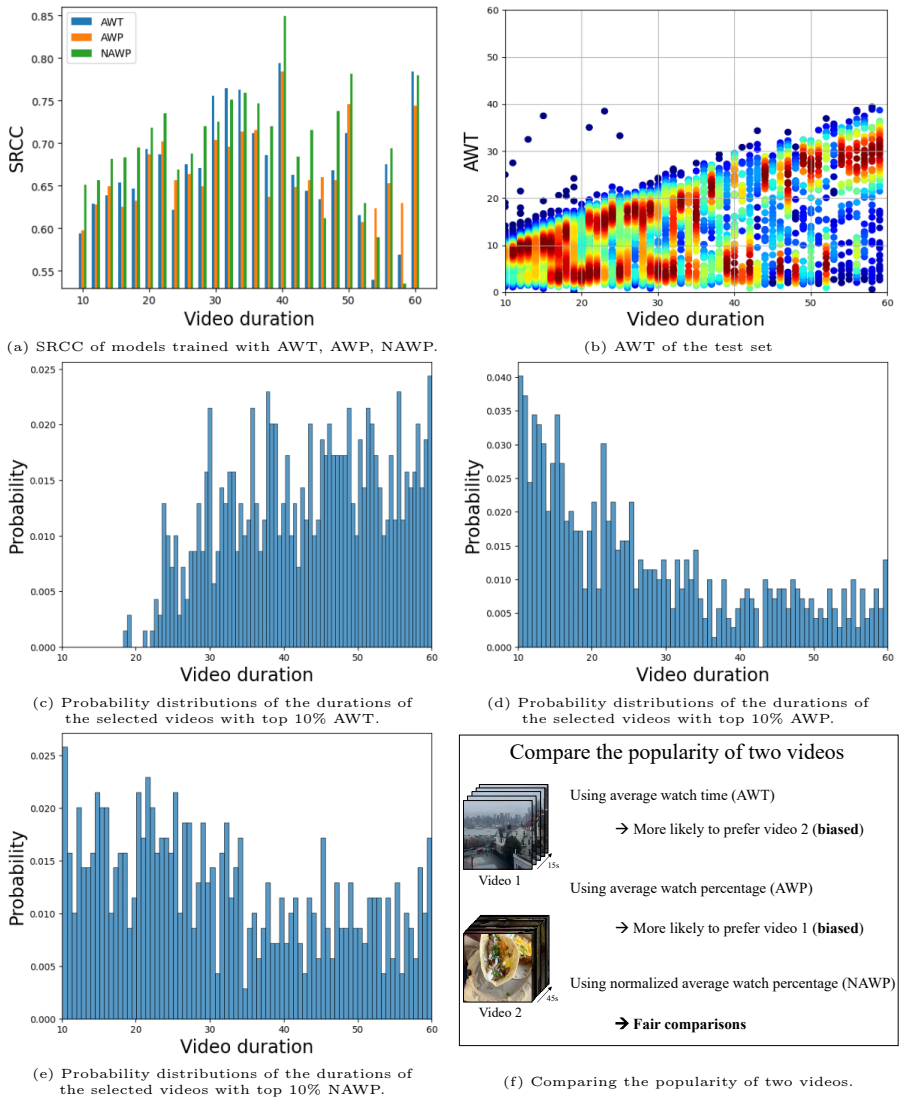
where ω denotes the two layer MLP and $M_i \in \mathbb{R}^{1\times 1024}$ represents the temporally merged feature.

For the processing of title & description, we employ the text encoder T5 [2] along with cross-attention to handle titles and descriptions separately. It is important to note that some videos may have either empty titles, empty descriptions, or both. In our approach, we consider empty text inputs as valid input for the text encoder.

4 More Ablation Study

4.1 Normalized Average Watch Percentage

We provide the details to illustrate the advantages of NAWP in Figure 3.



(a) SRCC of models trained with AWT, AWP, NAWP.

(b) AWT of the test set

(c) Probability distributions of the durations of the selected videos with top 10% AWT.

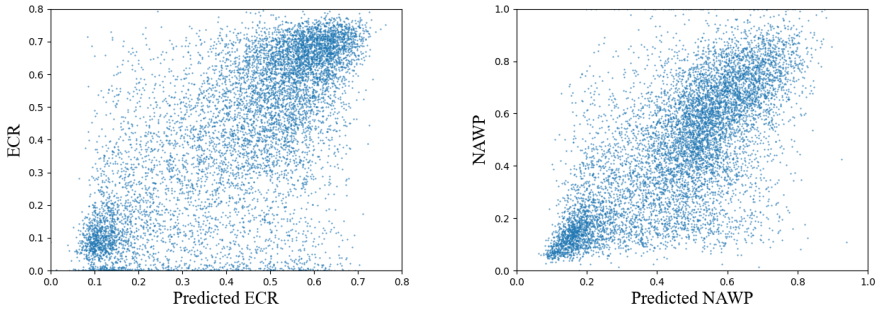
(d) Probability distributions of the durations of the selected videos with top 10% AWP.

(e) Probability distributions of the durations of the selected videos with top 10% NAWP.

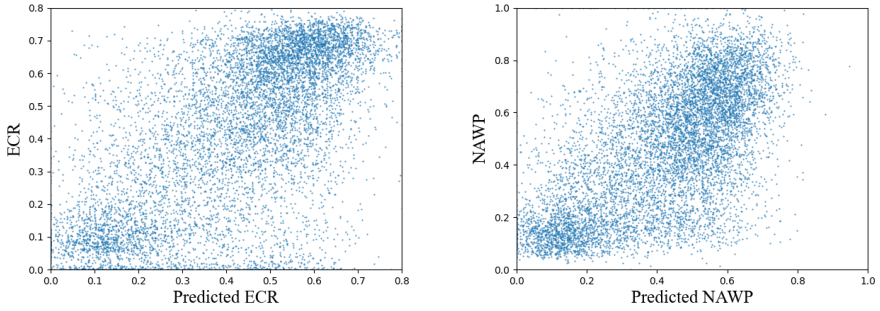
(f) Comparing the popularity of two videos.

Fig. 3: Ablation of normalized average watch percentage (NAWP). (a): comparison of SRCC on different durations. (b): The AWT distribution of test set. (c), (d), (e): We select the top 10% short videos from the test set (18,000 videos) based on our predicted AWT, AWP, and NAWP, respectively. NAWP can help select the top 10% videos evenly from different video durations. (f): Based on the results of (c), (d), and (e), using AWT and AWP would produce biased comparisons of two videos with different durations. NAWP enables a fair comparison.

Distribution of the test set. In Figure 3(a), the relation between video duration and average watch time (AWT) in the test set is presented, which is similar as the training set's shown in Figure 1(a).



(a) ECR of "Ours" (b) NAWP of "Ours"



(c) ECR of "Ours-VQA" (d) NAWP of "Ours-VQA"

Fig. 4: Our comparison with "Ours-VQA". The relationship between predicted ECR and ground truth ECR is shown via scatter points in (a), (c). The relationship between predicted NAWP and ground truth NAWP are shown via scatter points in (b), (d). The difference illustrates that "Ours" performs much better than "Ours-VQA", especially for relatively short and long videos.

Detailed SRCC comparisons. In the main paper, we compare Normalized Average Watch Percentage (NAWP) with Average Watch Time (AWT) or Average Watch Percentage (AWP) using average Spearman's rank correlation coefficient (SRCC). Figure 3(b) provides a detailed SRCC comparison across different durations. To minimize variation, the test set is partitioned into distinct groups, each spanning a small range (2 seconds). The SRCC is then separately calculated for each group. The results indicate that NAWP outperforms AWP and AWT on the most video durations.

Probability distributions of selected 10% videos. We further employ the AWT, AWP, and NAWP as metrics to select the top 10% videos. Figures 3(c), (d), and (e) depict the probability distributions of the durations of the selected videos with top 10% AWT, AWP, NAWP, respectively. Notably, the top 10% results by AWT tend to concentrate on videos with longer durations, while those by AWP concentrate on videos with shorter durations. NAWP, on the other

	Baseline	4 categories	Shorter ([10s, 25s))	Longer ([30s, 60s])
RMSE↓	0.197	0.203	0.196	0.197

Table 3: Ablation of dataset diversity. We sample 4 new training sets for fair comparisons. Each training set contains 34,000 videos.

Cross-attn with	Image Semantic	Motion	Video Captioning
SRCC↑ / RMSE↓	0.695 / 0.172	0.696 / 0.172	0.682 / 0.177

Table 4: Cross-attn with semantic, motion and caption features.

Metrics	ECR	NAWP	Like rate
SRCC	0.675	0.696	0.526

Table 5: Evaluation of like rate.

hand, exhibits the ability to evenly select the top 10% of videos across different durations.

Advantages of NAWP. The advantages of NAWP can be summarized in three key aspects: 1) NAWP effectively measures the engagement levels of videos across varying durations. Utilizing NAWP as the training metric enhances performance for videos of each duration, consequently improving overall model performance. 2) NAWP enables the fair comparison of any two videos, irrespective of their durations, as shown in Figure 3(f). Moreover, the real NAWP, calculated from the real AWT, can serve as an accurate indicator for generalized ranking within a recommendation system. 3) When selecting the top 10% of videos based on NAWP, the resulting distribution of durations is even. In contrast, AWT and AWP may yield biased results, especially when selecting videos of varying durations. While opting to select the top 10% of AWP or AWT for different duration groups separately is a viable alternative, it's crucial to acknowledge that without NAWP, ranking all selected top 10% videos of different durations in a recommendation system remains a challenging task.

4.2 Comprehensive Features

In Table 1, we systematically exclude each additional text or visual feature to assess the individual effectiveness of each component. Notably, adapting middle features from the video captioning model [6] results in a significant improvement in engagement prediction. The adaption of sound classification and Title & Description also demonstrates substantial enhancements. It is shown in Figure 4 that "Ours" with the multi-modal features achieves much better performances than "Ours-VQA" with only traditional video quality features.

4.3 VQA Features

As the baseline, "Ours-VQA" incorporates the per-frame semantic feature [3], per-frame distortion feature [4], and per-clip action recognition feature [1]. In Table 2, we perform ablation studies to assess the contributions of each feature. The results indicate that the addition of per-clip action recognition features leads

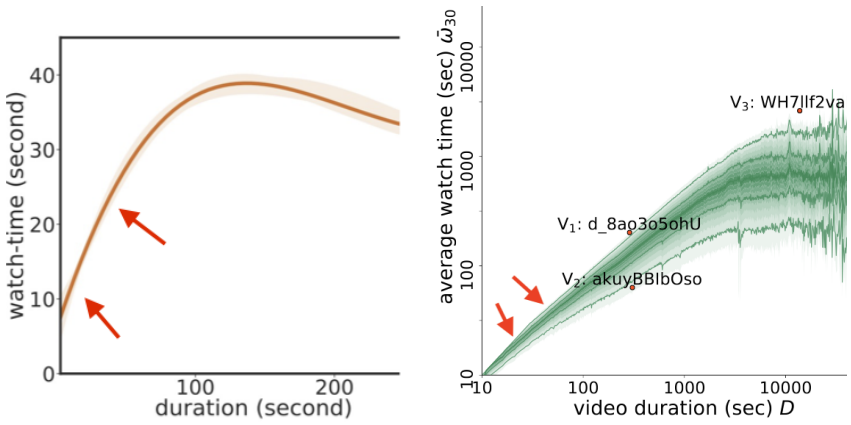


Fig. 5: Generalizability of linear coefficients in NAWP. Linear coefficients can generalize in smaller durations ($\leq$ 60s) in Kuaishou (*left*) [7] and Youtube (*right*) [5]. Figures are copied from [5, 7].

to a substantial improvement, whereas the inclusion of distortion features brings about a modest enhancement.

4.4 Dataset diversity

We construct 4 new training sets based on random sampling (Baseline), sampling within 4 categories, sampling shorter videos, sampling longer videos. Table 3 shows less diversity on categories leads to a drop. Less diversity on durations shows similar performances as the durations are normalized in the metrics.

4.5 Cross Attention

We provide ablation of different features connected to cross-attn. Table 4 shows image semantic and motion features achieve similar performances, while captioning feature causes a big drop.

4.6 Like rate

We perform an ablation study on like rate for SnapUGC. Training the model to learn the like rate, as shown in Table 5, reveals challenges in capturing this metric effectively.This challenge may arise from users' preference for swiftly consuming multiple videos rather than actively engaging with the like button, rendering the like rate indistinguishable for many videos.

5 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 [7] and Youtube [5] datasets* for videos with short durations ($\leq$ 60s), which are exactly the domain of most short videos, explored in this paper.

References

1. 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) 1, 7
2. 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 4
3. 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 7
4. 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) 7
5. 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 1, 8, 9
6. 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) 1, 7
7. 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 8, 9
8. 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) 4