

VQualA 2025 Challenge on Engagement Prediction for Short Videos: Methods and Results

Dasong Li* Sizhuo Ma* Hang Hua* Wenjie Li* Jian Wang* Chris Wei Zhou*
Fengbin Guan Xin Li Zihao Yu Yiting Lu Ru-Ling Liao Yan Ye Zhibo Chen
Wei Sun Linhan Cao Yuqin Cao Weixia Zhang Wen Wen Kaiwei Zhang
Zijian Chen Fangfang Lu Xiongkuo Min Guangtao Zhai Erjia Xiao
Lingfeng Zhang Zhenjie Su Hao Cheng Yu Liu Renjing Xu Long Chen
Xiaoshuai Hao Zhenpeng Zeng Jianqin Wu Xuxu Wang Qian Yu Bo Hu
Weiwei Wang Pinxin Liu Yunlong Tang Luchuan Song Jinxi He Jiaru Wu Hanjia Lyu

Abstract

This paper presents an overview of the VQualA 2025 Challenge on Engagement Prediction for Short Videos, held in conjunction with ICCV 2025. The challenge focuses on understanding and modeling the popularity of user-generated content (UGC) short videos on social media platforms. To support this goal, the challenge uses a new short-form UGC dataset featuring engagement metrics derived from real-world user interactions. This objective of the Challenge is to promote robust modeling strategies that capture the complex factors influencing user engagement. Participants explored a variety of multi-modal features, including visual content, audio, and metadata provided by creators. The challenge attracted 97 participants and received 15 valid test submissions, contributing significantly to progress in short-form UGC video engagement prediction.

1. Introduction

With the rapid rise of social media, a growing number of content creators are sharing short videos that capture their daily lives on platforms like TikTok, Instagram Reels, YouTube Shorts, and Snapchat Spotlight. At the same time, a large share of users are spending significant amounts of time watching this type of content across these platforms.

Social media platforms receive a constant stream of

newly published short videos. The effective dissemination of newly published videos remains a core objective for social media platforms. Recommending high-quality User Generated Content (UGC) videos enhances viewer engagement and consequently encourages content creators, especially novice creators. The effective dissemination of newly 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, such as 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 [23, 34, 44, 59] arises from the sampling bias in such limited initial interactions, resulting in noisy and inaccurate predictions of recommendation extents. Additionally, this conventional approach can result in time-sensitive short videos not being broadcast promptly, causing them to miss critical attention. Furthermore, emerging creators may struggle to gain sufficient visibility and recommendations, limiting their potential impact. 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. Consequently, an ineffective cold-start process may creates a negative feedback loop within the ecosystem, hindering the recommendation of high-quality videos to users, especially for some small-size or mid-size social media platforms.

A potential method for predicting engagement levels from video content is through user-generated content (UGC) video quality assessment (VQA). UGC VQA methods can be broadly classified into three categories based on the availability of reference information: full-reference [39,

*Dasong Li (dasongli@link.cuhk.edu.hk), Sizhuo Ma (sma@snap.com), Hang Hua (hhua2@cs.rochester.edu), Wenjie Li (wenjie.li@snap.com), Jian Wang (jwang4@snap.com), and Chris Wei Zhou (zhou26@cardiff.ac.uk) are the challenge organizers of this challenge.

The other authors are participants of the VQualA 2025 EVQA-SnapUGC: Engagement prediction for short videos Challenge.

The project page is https://github.com/dasongli1/SnapUGC_Engagement/tree/main/ECR_inference



Figure 1. Sample frames of the short videos in SnapUGC dataset [28].

	Multi-Modal Content			Annotators number	Metrics
	Video	Audio	Text		
VQA datasets [19, 37, 46, 52, 53]	✓	✗	✗	≤ 40	Labeling Scores
Our dataset [28]	✓	✓	✓	≥ 1000	Real User Interactions

Table 1. We provide a detailed comparison with the VQA datasets. 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.

[56], reduced-reference [32, 38], and no-reference approaches [16, 40, 43, 54]. The previous learning-based VQA methods [4, 5, 8, 15, 24, 30, 46, 53, 57] extract deep features via pre-trained models [12, 17, 18, 21, 42] and utilize these features to predict the MOS scores. With the emergence of large language models (LLMs) and large multimodal models (LMMs), recent studies [31, 48, 49] leverage their reasoning and interpretability capabilities to enhance the interactivity and explainability of VQA frameworks.

Despite the advancements in UGC VQA methods, Li et al. [28] demonstrated that VQA models [46, 47, 53] trained on existing VQA datasets [19, 37, 46, 52, 53] struggle to predict the popularity of short videos. This indicates that the mean opinion scores (MOS) annotated by small groups of human raters in video quality assessment datasets show a poor correlation with the popularity levels of these videos. This discrepancy may arise from the biases inherent in subjective MOS scores, which are influenced by the diverse preferences and limited participation of raters. As a result, these scores may not accurately reflect a video’s appeal to its broader audience, as assessed by metrics like average

watch time. Furthermore, while VQA methods primarily focus on video visuals, short video engagement can be affected by additional factors such as background music, content category, and titles. Therefore, engagement prediction and video quality assessment are distinct tasks due to the differing nature of their datasets.

To address these limitations, we introduce a large-scale SnapUGC dataset of publicly accessible short videos on Snapchat Spotlight directly model the engagement levels [3, 50, 55]. Unlike prior datasets, SnapUGC leverages real engagement data from over 2,000 users to mitigate the bias introduced by small-scale subjective annotation. We introduce two robust metrics to quantify engagement:

1. **Normalized Average Watch Percentage (NAWP):** Measures overall user engagement normalized across videos of varying lengths.
2. **Engagement Continuation Rate (ECR):** Represents the probability that a viewer watches beyond the initial 5 seconds, indicating the video’s ability to capture attention early on.

NAWP provides an indication of the overall engagement level for videos with different durations and ECR assesses

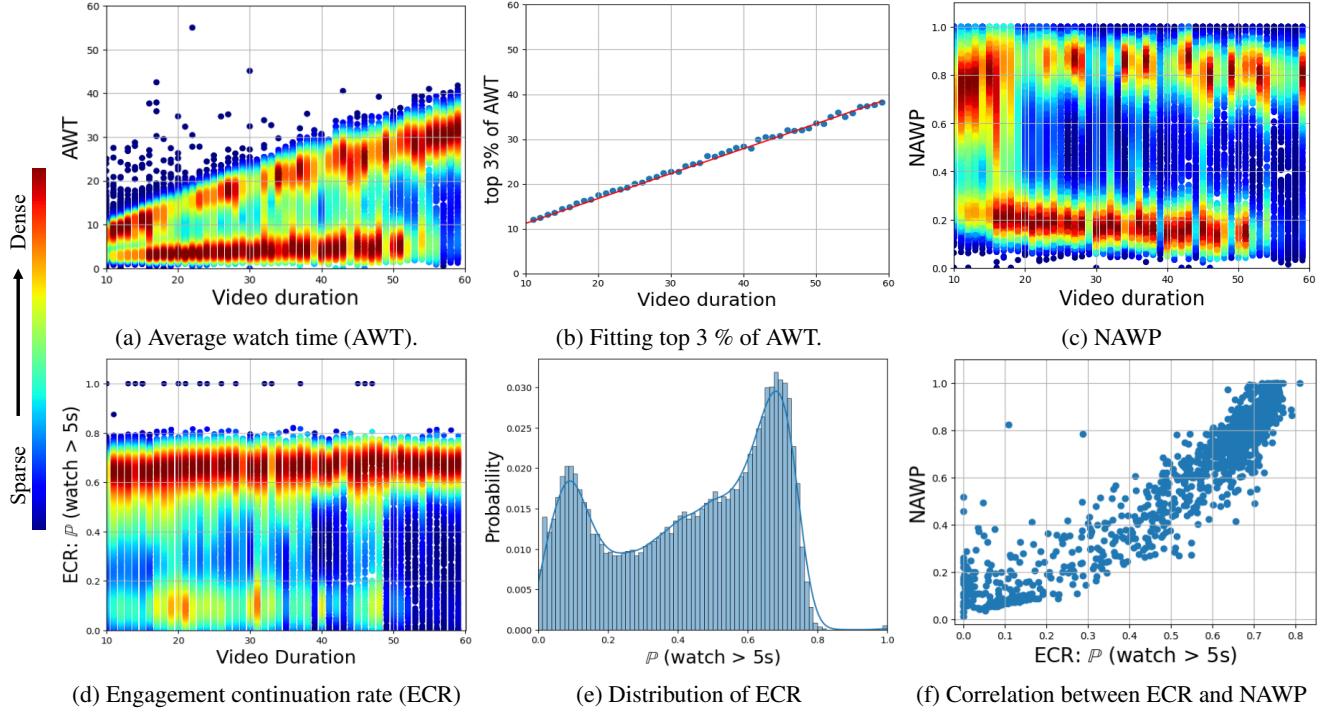


Figure 2. (a), (d): The distributions of average watch time (AWT) and engagement continuation rate (ECR), respectively. ECR, calculated as the probability of watch time exceeding 5 seconds: $\mathbb{P}(\text{watch} > 5\text{s})$, is more duration-independent. (b): We fit top 3 % of average watch times 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), (c), (d). (e): Distributions of ECR. ECR follows a 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. (f): The strong correlation between ECR and NAWP.

whether the video’s outset is captivating enough to retain viewers’ interest in continuing to watch. These metrics are computed in aggregate to ensure individual user privacy—no personal information or user histories are included in the dataset.

To further advance research on user engagement modeling for short-form videos, we are organizing the Engagement Prediction for Short Videos Challenge (EVQA) as part of the VQualA 2025 Workshop @ ICCV. This challenge aims to establish a practical and comprehensive benchmark for predicting viewer engagement, with a **specific focus on Engagement Continuation Rate (ECR) prediction** as the core task, selected for simplicity and clarity. We are grateful to participants from both academia and industry for contributing to this shared goal of advancing short-form video quality assessment and engagement prediction.

This Challenge is one of VQualA 2025 Workshop associated challenges on: ISRG-Q-image super-resolution generated content quality assessment [29], FIQA-face image quality assessment [33], Visual quality comparison for large multimodel models [58], GenAI-Bench AIGC video

quality assessment [9], and Document Image Quality Assessment [20].

2. Challenge Dataset

2.1. SnapUGC Dataset Collection

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 120,651 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 5 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 1. The dataset is shown in the following:

Rank	Team name	Team leader	Final Score	SROCC	PLCC	Features	Large Multi-modal Models
-	Baseline	-	0.660	0.657	0.665	Multi-Modal	-
1	ECNU-SJTU VQA	Wei Sun	0.710	0.707	0.714	Multi-Modal	Video-LLaMA (1.7B), Qwen2.5-VL (7B)
1*	IMCL-DAMO	Fengbin Guan	0.698	0.696	0.702	Multi-Modal	Qwen2.5-VL (7B)
3	HKUST-Cardiff-MI-BAAI	Xiaoshuai Hao	0.680	0.677	0.684	Visual Only	-
4	MCCE	Zhenpeng Zeng	0.667	0.666	0.668	Multi-Modal	-
4*	EasyVQA	Bo Hu	0.667	0.664	0.671	Multi-Modal	-
6	Rochester	Pinxin Liu	0.449	0.405	0.515	Multi-Modal	Skywork-VL-Reward (7B)
7	brucelyu	Hanjia Lyu	0.441	0.439	0.444	Textual Only	-

Table 2. Result of engagement prediction challenge.

1. **Train set:** 106,192 short-form videos. Each video is accompanied with title and descriptions provided by creators.
2. **Validation set:** 6000 short-form videos. Each video is accompanied with title and descriptions provided by creators.
3. **Test set:** 8,459 short-form videos. Each video is accompanied with title and descriptions provided by creators.

2.2. Engagement Metrics

Average watch time (AWT) is a naive and common metric to measure viewer engagement. However, AWT faces limitations when comparing videos of different durations. 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.* [22] measuring entire videos’ dropout probability, ECR focuses on the contents of first several seconds, which determines whether the users would continue to watch and substantially affects watch times.

Average watch time (AWT). We analyze average watch times (AWT) of various video durations d in Figure 2(a). Importantly, the distributions of AWT vary for different video durations, showing diverse user engagement patterns. Therefore comparing the popularity of short videos with different durations using AWT is challenging.

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. 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:

$$f_{\max}(d) = \alpha \times d + \beta; f_{\min}(d) = 0. \quad (1)$$

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

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

The relationship between the video duration and NAWP is depicted in Figure 2(c). The NAWP falls within the range of [0, 1] and NAWP of videos with top 3% average watch time is set to be 1.

Engagement continuation rate (ECR). As shown in Figure 2(e), engagement continuation rate (ECR), calculated as $\mathbb{P}(\text{watch} > 5\text{s})$, demonstrates stable behavior across different video durations. The majority of values fall within the range of [0, 0.8].

ECR for the Challenge. the α and β in NAWP may vary across different datasets or different platforms. The ECR and NAWP are observed to have a strong correlation of 0.928 in Figure 2(f). Therefore, we select ECR as the metrics in EVQA Challenge. To protect the private information of creators, the ECR used in this challenge is derived from normalizing the ranking of real ECR.

3. Challenge Results

The challenge results are summarized in Table 2, including the performance of all teams that submitted their fact sheets. As this is a novel task involving multi-modal features, we did not impose restrictions on model size in order to explore the upper bound of model capacity. We provide a baseline model achieving an SROCC of 0.660 and a PLCC of 0.657, based on the approach proposed by Li *et al.* [28].

The teams with top performances including ECNU-SJTU VQA, IMCL-DAMO, HKUST-Cardiff-MI-BAAI, MCCE(MCCE (Media Convergence and Communication Experimental)), EasyVQA achieved excellent results in both PLCC and SROCC, exceeding our baseline. Among

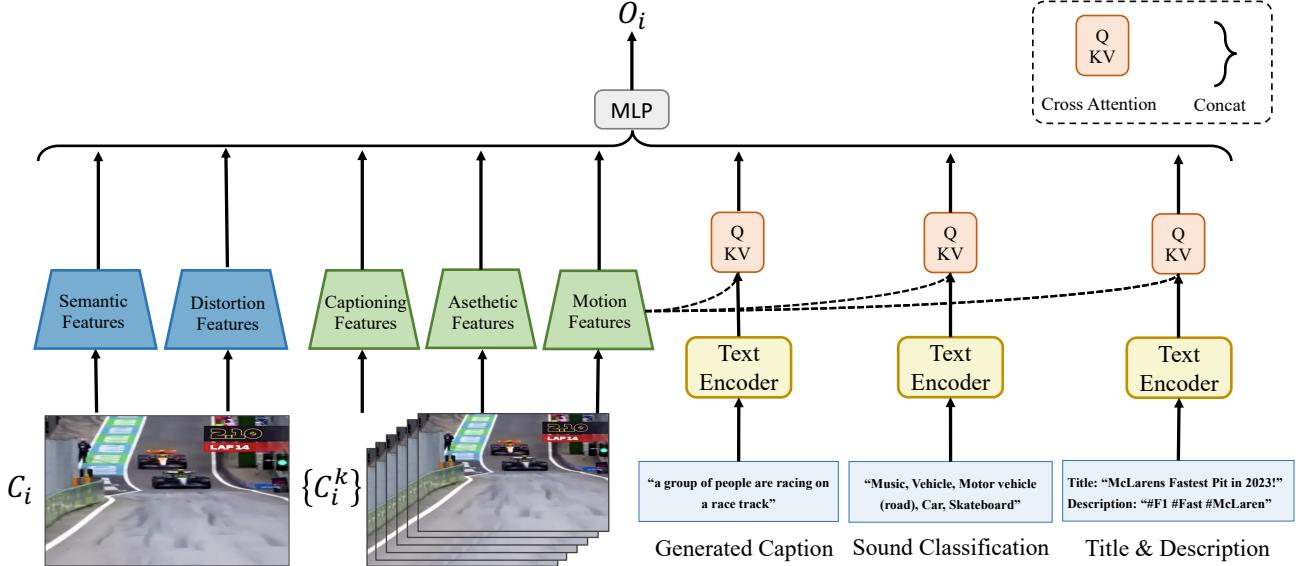


Figure 3. The overview framework of baseline.

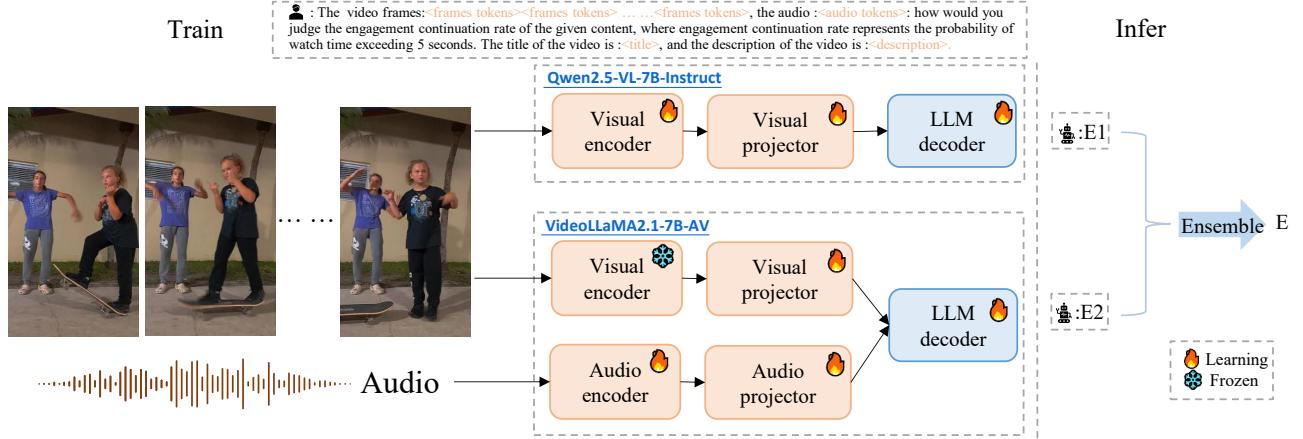


Figure 4. The overview framework provided by Team ECNU-SJTU VQA.

them, ECNU-SJTU VQA and ICML-DAMO demonstrated the most significant improvements over the baseline. Given their competitive performance, these two teams are recognized as co-first place. As shown in Table 2, large multi-modal models were widely adopted to boost performance. Interestingly, the team brucelyu achieved reasonable performance using only textual features (*e.g.*, title, description, and music classification), highlighting that non-visual information can also meaningfully contribute to predicting user engagement with short videos.

4. Teams and Methods

4.1. Baseline

The provided baseline, built on the Li *et al.* [28], are trained on the ECR prediction. A comprehensive set of multi-

modal features, including per-frame semantic features [42], per-frame pixel-level distortion features [46] from different degradations [25–27], sound classification [1], text descriptions from authors, video captioning [51], are used. The framework of baseline is shown in Figure 3.

4.2. ECNU-SJTU VQA Team

This approach utilizes an ensemble of Large Multimodal Models (LMMs) [41] for video quality score prediction. Specifically, they leverage two LMMs: Video-LLaMA2 [10], a high-performance model tailored for audio-visual language understanding, and Qwen2.5-VL [2], a powerful model focused on general vision-language tasks.

For Video-LLaMA2 [11], they provide the model with the first 8 (or 5) video frames, the audio track, and the

You are an expert in analyzing short-form video engagement. Your task is to predict the Engagement Continuation Rate (ECR) for the provided video.

****Task Definition:****

The ECR represents the probability, on a scale from 0.0 to 1.0, that a viewer will watch the video for more than 5 seconds. A high ECR indicates the video's beginning is highly captivating and effective at retaining viewer interest.

Please analyze the video's opening moments—considering visual elements, editing, pacing, and subject matter—to make your prediction.

****Required Output Format:****

Provide your final prediction in the following format and nothing else:

ECR_Score: [a single decimal value between 0.0 and 1.0]***

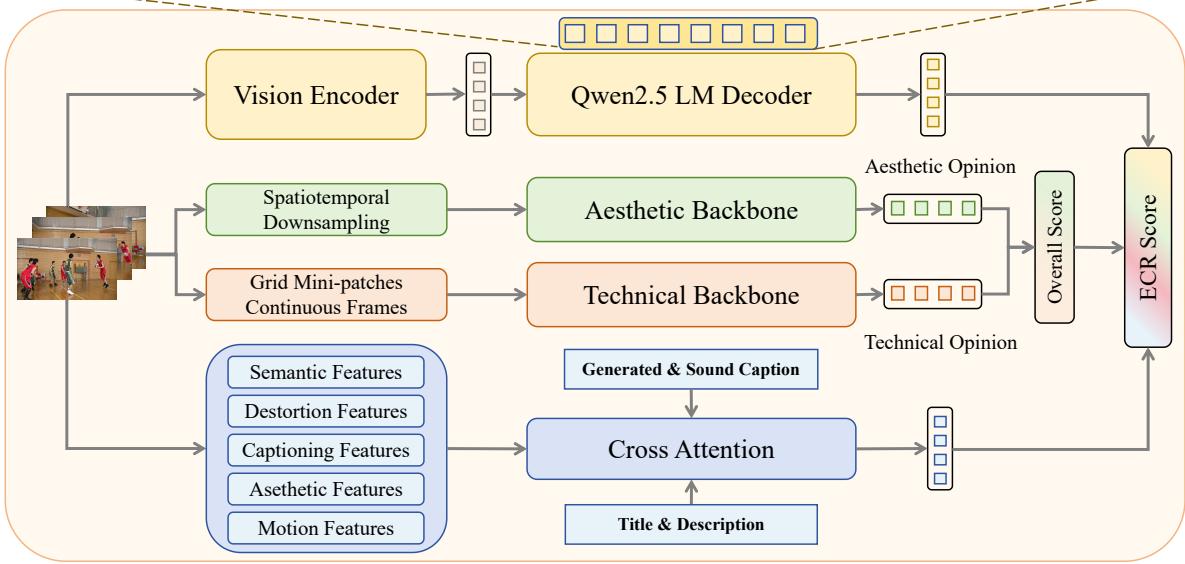


Figure 5. The overview framework provided by Team IMCL-DAMO.

associated video description text (including the video title and description; if unavailable, they use ‘None’ as the default input). To predict the Engagement Continuation Rate (ECR), they extract the hidden features from the last layer of Video-LLaMA2 and append a regression head. The regression head consists of a multilayer perceptron (MLP), which includes a dropout layer, a fully connected (FC) layer with 2048 neurons, a ReLU activation layer, and a final FC layer with a single neuron to predict the video engagement score.

For Qwen2.5-VL [2], they similarly provide the first 8 video frames and the corresponding video description text as input. For ECR prediction, they follow the original Qwen2.5-VL architecture and utilize the next-token output for regression. Finally, they ensemble the predictions from both models to obtain the final engagement score.

Training details. For VideoLLaMA2.1-7B-AV [11], each input image is first resized to a global resolution of 384×384 , then divided into multiple 384×384 patches using grid-based cropping and padding. These patches are jointly fed into the vision encoder. They finetune the model on 2 A800 GPUs with a batch size of 12 for one epoch. During training, the parameters of the vision encoder are frozen, while the remaining parameters are updated. The model is optimized using a learning rate of 5×10^{-5} . Then they train VideoLLaMA2.1-7B-AV with different random

seeds and number of input frames. For Qwen2.5-VL-7B-Instruct [2], they control the maximum number of image pixels to be image max pixels = $768 \times 28 \times 28$ to ensure efficient memory usage. The model is trained on 8 A800 GPUs with a batch size of 16 for one epoch. Similar to LLava, we update all parameters, applying a learning rate of 2×10^{-6} to the vision encoder and 1×10^{-5} to the rest of the model.

Testing details. They evaluate our framework on the EVQA dataset using three models. For VideoLLaMA2.1-7B-AV, each image is directly resized to 384×384 prior to inference. For the Qwen2.5-VL models, we apply the same preprocessing as used during training, constraining the maximum number of image pixels to $768 \times 28 \times 28$. The final ensemble prediction is obtained by combining the outputs of three VideoLLaMA2.1-7B-AV models and one Qwen2.5-VL-7B-Instruct model.

4.3. IMCL-DAMO Team

This team combines three main branches: (1) a baseline model that extracts diverse multi-modal features to enhance video-content relevance; (2) fragment-based sampling leveraging DOVER [47]’s technical and aesthetic branches to capture multiple quality perspectives; and (3) a Qwen2.5-VL-7B [2] branch utilizing full-parameter super-

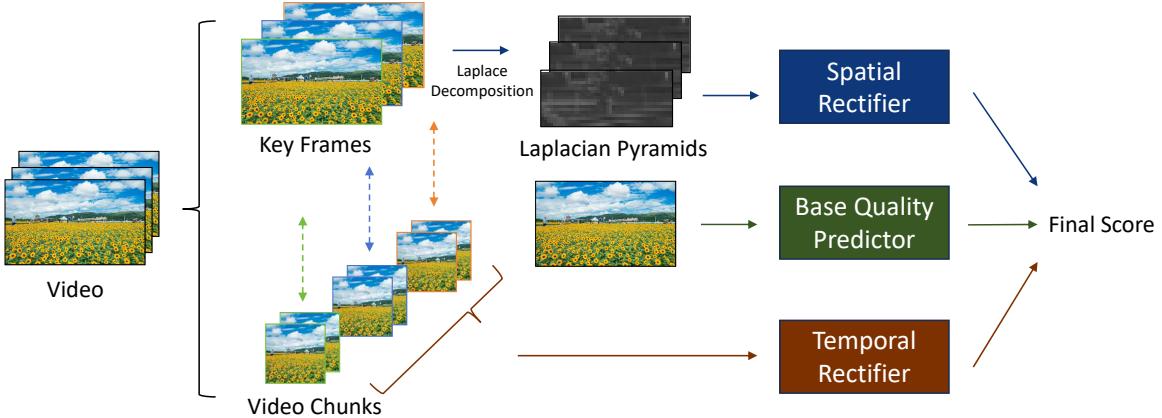


Figure 6. The overview framework provided by Team HKUST-Cardiff-MI-BAAI .

vised fine-tuning to incorporate strong multi-modal priors for ECR prediction.

Training details. They used PyTorch and the Transformers library to implement all models. The training involved only the official competition training dataset. Full parameter fine-tuning was conducted for both the DOVER branch and the Qwen2.5-VL-7B model. Experiments were run on 4 NVIDIA A100 GPUs, requiring approximately $48 \sim 72$ hours of training time. They employed mixed precision training and standard optimizer configurations to improve efficiency.

Testing details. Evaluation was performed on the official competition test set with a batch size of 1. No multi-scale or test-time augmentation strategies were applied. The baseline branch required over 12 hours of inference, the Qwen2.5-VL-7B branch about 6 hours, and the aesthetic/technical DOVER branches between 10–20 minutes.

4.4. HKUST-Cardiff-MI-BAAI Team

This team's approach integrates three main components:

1. **Base Quality Predictor:** This module takes a sparse set of spatially downsampled key frames as input and uses a pretrained Vision Transformer (ViT) [14] from CLIP [35] to generate a scalar quality estimate.
2. **Spatial Rectifier:** This component processes Laplacian pyramids of key frames at the original spatial resolution to compute scaling and shift parameters, which are used to refine the base quality score.
3. **Temporal Rectifier:** This module processes spatially downsampled video chunks centered around key frames at the original frame rate, producing another set of scaling and shift parameters to further adjust the quality estimate.

Training details. They used PyTorch to implement all models. The training involved only the official competition training dataset. They fine-tune the model with a batch

size of 32 for two epochs. Spatial and temporal rectifiers are randomly dropped out during training with probabilities of 0.1. The model is optimized by Adam optimizer using a learning rate of 5×10^{-5} .

4.5. MCCE (Media Convergence and Communication Experimental) Team

This team enhanced the provided baseline by incorporating multi-modal fusion techniques, novel perceptual features, and temporal aggregation of video features:

1. **Novel Image Features Fusion:** They generate attention weights from semantic features to dynamically adjust the fusion ratio of distortion features, placing greater emphasis on distortion features in key frames.
2. **Novel Perception Features:** A more advanced visual encoder (perceptual encoder) is introduced to extract richer video perception features, enabling more comprehensive video representation learning.
3. **Motion Features Temporal Aggregation:** They employ a global-local temporal aggregation mechanism to capture dynamic changes in short videos, such as action intensity and transition rhythm.

The team also achieves a $3 \sim 5\times$ speedup in training by leveraging advanced memory optimization, distributed computing, and intelligent resource management. Key techniques include LRU-based memory management and GPU memory-aware batch size adjustment.

4.6. EasyVQA Team

This team focuses on video content and leverages multi-modal information—visual, textual, and auditory—to enhance representation learning. They utilize CLIP [35] to extract features from video frames, BERT [13] for textual features from video titles, and BEATs [6] for audio features. As a preprocessing step, video data is processed based on the ECR metric: the first frame of each second within the

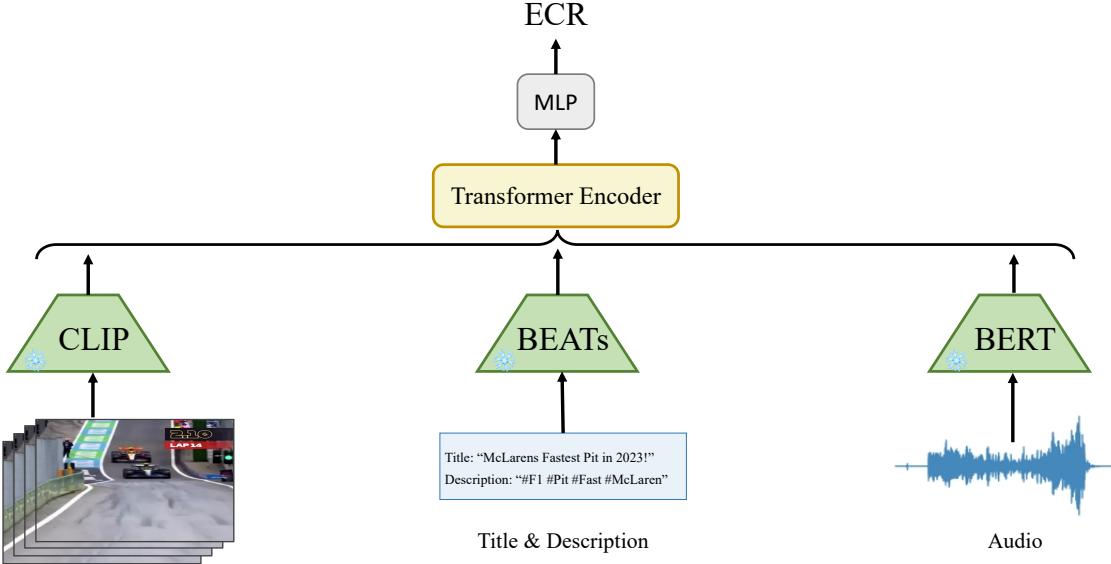


Figure 7. The overview framework provided by Team EasyVQA.

first 5 seconds of a video is extracted as a key frame. These frames are resized to 224×224 pixels and passed through CLIP’s visual encoder. Simultaneously, the audio and title text are processed to obtain audio tokens and text tokens, respectively. The model is trained to distinguish among the different modalities. To this end, modality-specific embeddings are added to the features of each modality, and positional encodings are applied to the video tokens. All tokens, including visual, textual, and auditory, are then concatenated and passed through a Transformer Encoder to generate a unified video feature representation. Finally, an MLP head predicts the ECR value.

4.7. Rochester Team

This team employs the pretrained Skywork-VL-Reward-7B vision-language model [45] to predict a continuous quality score in the range $[0, 1]$ for input videos. Sixteen frames are evenly sampled from each video and processed using the frozen Skywork-VL backbone, followed by a lightweight regression head that outputs the final score. Thanks to the strong multimodal foundation of the pretrained model and the simplicity of the regression head, the approach achieves high efficiency and strong alignment with human evaluations.

Training details. The model is implemented using PyTorch and trained using the AdamW optimizer with weight decay. The initial learning rate is set to 1×10^{-6} and decayed via cosine annealing across 5 epochs. Training was conducted using mixed-precision (FP 16) over 4 NVIDIA H100 GPUs for a total duration of 32 hours. They use the competition-provided dataset exclusively, without any external data. Frame sampling enables full-resolution input while maintaining computational efficiency. Gradient accu-

mulation was used to simulate larger batch sizes.

Testing details. During inference, 16 evenly sampled frames from the video are passed through the frozen Skywork-VL-Reward-7B [45] model. The features are then fed into a regression head to produce a single score between 0 and 1.

4.8. brucelyu17 Team

This team uses the XGBoost regression model [7], implemented via the xgboost Python package, to predict short video engagement levels using only textual features. The input comprises semantic representations derived from both textual and categorical data.

For textual features, the team applies the pre-trained Sentence-BERT model (all-MiniLM-L6-v2) [36] to convert video titles and captions into dense embeddings. These embeddings capture the contextual semantics of the text and provide compact, informative representations for modeling.

For categorical features, they incorporate background music category information. Initially encoded using multi-hot vectors, these features are further transformed into semantic embeddings to better reflect relationships among music categories.

To optimize model performance, they perform a grid search over key XGBoost hyperparameters and select the configuration that achieves the lowest Root Mean Squared Error (RMSE) on the validation set. The hyperparameter search space includes maximum tree depth (3, 6), learning rate (0.1, 0.05, 0.01), and subsample ratio (0.8, 1.0). Each parameter combination is evaluated using early stopping to avoid overfitting. The final model is trained using the best configuration and evaluated on a held-out test set.

References

- [1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. [5](#)
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. [5, 6](#)
- [3] Sahan Bulathwela, Maria Perez-Ortiz, Emine Yilmaz, and John Shawe-Taylor. VLEngagement: A Dataset of Scientific Video Lectures for Evaluating Population-based Engagement. *arXiv e-prints*, art. arXiv:2011.02273, 2020. [2](#)
- [4] Baoliang Chen, Lingyu Zhu, Guo Li, Fangbo Lu, Hongfei Fan, and Shiqi Wang. 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. [2](#)
- [5] Pengfei Chen, Leida Li, Lei Ma, Jinjian Wu, and Guangming Shi. Rirnet: Recurrent-in-recurrent network for video quality assessment. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 834–842, New York, NY, USA, 2020. Association for Computing Machinery. [2](#)
- [6] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022. [7](#)
- [7] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794, 2016. [8](#)
- [8] Wei-Ting Chen, Yu-Jiet Vong, Yi-Tsung Lee, Sy-Yen Kuo, Qiang Gao, Sizhuo Ma, and Jian Wang. Diffvqa: Video quality assessment using diffusion feature extractor. *arXiv preprint arXiv:2505.03261*, 2025. [2](#)
- [9] Ying Chen, Huasheng Wang, Pengxiang Xiao, Yukang Ding, Enpeng Liu, Wei Zhou, and et al. Vquala 2025 challenge on genai-bench aigc video quality assessment: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV) Workshops*, pages 1–11, 2025. [3](#)
- [10] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. [5](#)
- [11] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. [5, 6](#)
- [12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. [2](#)
- [13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. [7](#)
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. [7](#)
- [15] Franz Götz-Hahn, Vlad Hosu, Hanhe Lin, and Dietmar Saupe. Konvid-150k: A dataset for no-reference video quality assessment of videos in-the-wild. In *IEEE Access* 9, pages 72139–72160. IEEE, 2021. [2](#)
- [16] Fengbin Guan, Zihao Yu, Yiting Lu, Xin Li, and Zhibo Chen. Internvqa: Advancing compressed video quality assessment with distilling large foundation model. *arXiv preprint arXiv:2502.19026*, 2025. [2](#)
- [17] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 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)*, pages 6546–6555, 2018. [2](#)
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. [2](#)
- [19] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The konstanz natural video database (konvid-1k). In *Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2017. [2](#)
- [20] Fan Huang, Xiongkuo Min, Zhichao Ma, Xiaohong Liu, Chris Wei Zhou, Guangtao Zhai, and et al. Vquala 2025 document image quality assessment challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV) Workshops*, pages 1–8, 2025. [3](#)
- [21] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Apostol Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *ArXiv*, abs/1705.06950, 2017. [2](#)
- [22] Juho Kim, Philip J. Guo, Daniel T. Seaton, Piotr Mitros, Krzysztof Z. Gajos, and Robert C. Miller. Understanding in-video dropouts and interaction peaks in online lecture videos. In *Proceedings of the First ACM Conference on Learning @*

- Scale Conference*, page 31–40, New York, NY, USA, 2014. Association for Computing Machinery. 4
- [23] Hoyeop Lee, Jinbae Im, Seongwon Jang, Hyunsouk Cho, and Sehee Chung. Melu: Meta-learned user preference estimator for cold-start recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 1073–1082, New York, NY, USA, 2019. Association for Computing Machinery. 1
- [24] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, page 2351–2359, New York, NY, USA, 2019. Association for Computing Machinery. 2
- [25] Dasong Li, Yi Zhang, Ka Chun Cheung, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Learning degradation representations for image deblurring. In *Computer Vision – ECCV 2022*, pages 736–753, Cham, 2022. Springer Nature Switzerland. 5
- [26] Dasong Li, Yi Zhang, Ka Lung Law, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Efficient burst raw denoising with variance stabilization and multi-frequency denoising network, 2022.
- [27] Dasong Li, Xiaoyu Shi, Yi Zhang, Ka Chun Cheung, Simon See, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. A simple baseline for video restoration with grouped spatial-temporal shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9822–9832, 2023. 5
- [28] Dasong Li, Wenjie Li, Baili Lu, Hongsheng Li, Sizhuo Ma, Gurunandan Krishnan, and Jian Wang. Delving deep into engagement prediction of short videos. In *Computer Vision – ECCV 2024*, pages 289–306, Cham, 2025. Springer Nature Switzerland. 2, 4, 5
- [29] Yixiao Li, Xin Li, Wei Zhou, Shuo Xing, Hadi Amirpour, Xiaoshuai Hao, Guanghui Yue, Baoquan Zhao, Weide Liu, Xiaoyuan Yang, Zhengzhong Tu, and et al. Vquala 2025 challenge on image super-resolution generated content quality assessment: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV) Workshops*, pages 1–10, 2025. 3
- [30] Yinhao Liu, Xiaofei Zhou, Haibing Yin, Hongkui Wang, and Chenggang Yan. Efficient video quality assessment with deeper spatiotemporal feature extraction and integration. *Journal of Electronic Imaging*, 30:063034, 2021. 2
- [31] Yiting Lu, Xin Li, Haoning Wu, Bingchen Li, Weisi Lin, and Zhibo Chen. Q-adapt: Adapting lmm for visual quality assessment with progressive instruction tuning. *arXiv preprint arXiv:2504.01655*, 2025. 2
- [32] Lin Ma, Songnan Li, and King Ngi Ngan. Reduced-reference video quality assessment of compressed video sequences. *IEEE Transactions on circuits and systems for video technology*, 22(10):1441–1456, 2012. 2
- [33] Sizhuo Ma, Wei-Ting Chen, Qiang Gao, Jian Wang, Chris Wei Zhou, Wei Sun, Weixia Zhang, Linhan Cao, Jun Jia, Xiangyang Zhu, Dandan Zhu, Xiongkuo Min, Guangtao Zhai, Baoying Chen, Xiongwei Xiao, Jishen Zeng, Wei Wu, Tiexuan Lou, Yuchen Tan, Chunyi Song, Zhiwei Xu, MohammadAli Hamidi, Hadi Amirpour, Mingyin Bai, Jiawang Du, Zhenyu Jiang, Zilong Lu, Ziguang Cui, Zongliang Gan, Xinpeng Li, Shiqi Jiang, Chenhui Li, Changbo Wang, Weijun Yuan, Zhan Li, Yihang Chen, Yifan Deng, Ruting Deng, Zhanglu Chen, Boyang Yao, Shuling Zheng, Feng Zhang, Zhiheng Fu, Abhishek Joshi, Aman Agarwal, Rakhil Immidisetti, Ajay Narasimha Mopidevi, Vishwajeet Shukla, Hao Yang, Ruikun Zhang, Liyuan Pan, Kaixin Deng, Hang Ouyang, Fan Yang, Zhizun Luo, Zhuohang Shi, Songning Lai, Weilin Ruan, and Yutao Yue. Vquala 2025 challenge on face image quality assessment: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV) Workshops*, pages 1–10, 2025. 3
- [34] Feiyang Pan, Shuokai Li, Xiang Ao, Pingzhong Tang, and Qing He. 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*, page 695–704, New York, NY, USA, 2019. Association for Computing Machinery. 1
- [35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7
- [36] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 8
- [37] Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2):612–627, 2019. 2
- [38] Rajiv Soundararajan and Alan C Bovik. Video quality assessment by reduced reference spatio-temporal entropic differencing. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(4):684–694, 2012. 2
- [39] Wei Sun, Tao Wang, Xiongkuo Min, Fuwang Yi, and Guangtao Zhai. Deep learning based full-reference and no-reference quality assessment models for compressed ugc videos. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2021. 1
- [40] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. A deep learning based no-reference quality assessment model for ugc videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 856–865, 2022. 2
- [41] Wei Sun, Linhan Cao, Yuqin Cao, Weixia Zhang, Wen Wen, Kaiwei Zhang, Zijian Chen, Fangfang Lu, Xiongkuo Min, and Guangtao Zhai. Engagement prediction of short videos with large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV) Workshops*, 2025. 5
- [42] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 10096–10106. PMLR, 2021. 2, 5

- [43] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Ugc-vqa: Benchmarking blind video quality assessment for user generated content. *IEEE Transactions on Image Processing*, 30:4449–4464, 2021. 2
- [44] Maksims Volkovs, Guangwei Yu, and Tomi Poutanen. Dropoutnet: Addressing cold start in recommender systems. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 4964–4973, Red Hook, NY, USA, 2017. Curran Associates Inc. 1
- [45] Xiaokun Wang, Peiyu Wang, Jiangbo Pei, Wei Shen, Yi Peng, Yunzhuo Hao, Weijie Qiu, Ai Jian, Tianyidan Xie, Xuchen Song, Yang Liu, and Yahui Zhou. Skywork-vl reward: An effective reward model for multimodal understanding and reasoning, 2025. 8
- [46] Yilin Wang, Junjie Ke, Hossein Talebi, Joong Gon Yim, Neil Birkbeck, Balu Adsumilli, Peyman Milanfar, and Feng Yang. Rich features for perceptual quality assessment of ugc videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13435–13444, 2021. 2, 5
- [47] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. 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)*, pages 20144–20154, 2023. 2, 6
- [48] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching Imms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. Equal Contribution by Wu, Haoning and Zhang, Zicheng. Corresponding Authors: Zhai, Guangtao and Lin, Weisi. 2
- [49] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25490–25500, 2024. 2
- [50] Siqi Wu, Marian-Andrei Rizoiu, and Lexing Xie. Beyond views: Measuring and predicting engagement in online videos. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1), 2018. 2
- [51] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingen Zhou. mplug-2: A modularized multi-modal foundation model across text, image and video. *ArXiv*, abs/2302.00402, 2023. 5
- [52] Joong Gon Yim, Yilin Wang, Neil Birkbeck, and Balu Adsumilli. Subjective quality assessment for youtube ugc dataset. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 131–135, 2020. 2
- [53] Zhenqiang Ying, Maniratnam Mandal, Deepthi Ghadiyaram, and Alan Bovik. Patch-vq: ‘patching up’ the video quality problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14019–14029, 2021. 2
- [54] Zihao Yu, Fengbin Guan, Yiting Lu, Xin Li, and Zhibo Chen. Video quality assessment based on swin transformerv2 and coarse to fine strategy. *arXiv preprint arXiv:2401.08522*, 2024. 2
- [55] Ruohan Zhan, Changhua Pei, Qiang Su, Jianfeng Wen, Xueliang Wang, Guanyu Mu, Dong Zheng, Peng Jiang, and Kun Gai. Deconfounding duration bias in watch-time prediction for video recommendation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 4472–4481, New York, NY, USA, 2022. Association for Computing Machinery. 2
- [56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2
- [57] Zicheng Zhang, Wei Wu, Wei Sun, Danyang Tu, Wei Lu, Xiongkuo Min, Ying Chen, and Guangtao Zhai. Md-vqa: Multi-dimensional quality assessment for ugc live videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1746–1755, 2023. 2
- [58] Hanwei Zhu, Haoning Wu, Zicheng Zhang, Lingyu Zhu, Yixuan Li, Peilin Chen, Shiqi Wang, Wei Zhou, Linhan Cao, Wei Sun, Xiangyang Zhu, Weixia Zhang, Yucheng Zhu, Jing Liu, Dandan Zhu, Guantao Zhai, Xiongkuo Min, Zhichao Zhang, Xinyue Li, Shubo Xu, Anh Dao, Yifan Li, Hongyuan Yu, Jiaoqiao Yi, Yiding Tian, Yupeng Wu, Feiran Sun, Jiao Lijuan, and Song Jiang. Vquala 2025 challenge on visual quality comparison for large multimodal models: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision (ICCV) Workshops*, pages 1–11, 2025. 3
- [59] Yongchun Zhu, Ruobing Xie, Fuzhen Zhuang, Kaikai Ge, Ying Sun, Xu Zhang, Leyu Lin, and Juan Cao. 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*, page 1167–1176, New York, NY, USA, 2021. Association for Computing Machinery. 1

A. Teams and Affiliations

VQualA 2025 EVQA Track Organizers

Members: Dasong Li¹ (dasongli@link.cuhk.edu.hk), Sizhuo Ma² (sma@snap.com), Hang Hua³ (hhua2@cs.rochester.edu), Wenjie Li² (wjenjie.li@snap.com), Jian Wang² (jwang4@snap.com) and Chris Wei Zhou⁴ (zhouw26@cardiff.ac.uk)

Affiliations:

¹The Chinese University of Hong Kong.

²Snap Inc.

³University of Rochester.

⁴Cardiff University.

IMCL-DAMO

Members: Fengbi Guan^{1,2} (guanfengbin.gfb@alibaba-inc.com), Xin Li¹ (xin.li@ustc.edu.cn), Zihao Yu^{1,2} (zuhe.yzh@alibaba-inc.com), Yiting Lu^{1,2} (luyiting.lyt@alibaba-inc.com), Ru-Ling Liao² (ruling.lrl@alibaba-inc.com), Yan Ye² (yan.ye@alibaba-inc.com) and Zhibo Chen¹ ([chenzhibo@ustc.edu.cn\)](mailto:(chenzhibo@ustc.edu.cn)

Affiliations:

¹Intelligent Meida Computing Lab (IMCL).

²Alibaba.

ECNU-SJTU VQA Team

Members: Wei Sun¹ (sunguwei@gmail.com), Linhan Cao² (caolinhan@sjtu.edu.cn), Yuqin Cao² (caoyuqin@sjtu.edu.cn), Weixia Zhang² (zwx8981@sjtu.edu.cn), Wen Wen³ (wwen29-c@my.cityu.edu.hk), Kaiwei Zhang² (zhangkaiwei@sjtu.edu.cn), Zijian Chen² (zijian.chen@sjtu.edu.cn), Fangfang Lu⁴ (lufangfang@shiep.edu.cn), Xiongkuo Min² (minxiongkuo@sjtu.edu.cn) and Guangtao Zhai² (zhaiguangtao@sjtu.edu.cn)

Affiliations:

¹East China Normal University.

²Shanghai Jiao Tong University.

³City University of Hong Kong.

⁴Shanghai University of Electric Power.

HKUST-Cardiff-MI-BAAI

Members: Erjia Xiao¹ (exiao469@connect.hkust-gz.edu.cn), Lingfeng Zhang² (lfzhang715@gmail.com), Zhenjie Su³ (suzhen-jie2023@cuc.edu.cn), Hao Cheng¹ (hcheng046@connect.hkust-gz.edu.cn), Yu Liu⁴ (yuliu@hfut.edu.cn), Renjing Xu¹ (renjingxu@hkust-gz.edu.cn), Long Chen⁵ (longchen@xiaomi.com), Xiaoshuai Hao⁶ (xshao@baai.ac.cn)

Affiliations:

¹ The Hong Kong University of Science and Technology (Guangzhou).

²Tsinghua University.

³Communication University of China.

⁴Hefei University of Technology.

⁵Xiaomi EV.

⁶Beijing Academy of Artificial Intelligence.

Media Convergence and Communication Experimental Team

Members: Zhenpeng Zeng¹ (2473910949@qq.com), Jianqin Wu¹ (510483263@qq.com), Xuxu

Wang¹ (2330711901@qq.com) and Qian Yu¹ (1261591905@qq.com)

Affiliations: ¹Communication University of China.

EasyVQA

Members: HuBo¹ (hubo90@cqupt.edu.cn) and WangWeiwei¹ (s240231049@stu.cqupt.edu.cn)

Affiliations:

¹Chongqing University of Posts and Telecommunications, Chongqing.

Rochester

Members: Pinxin Liu¹ (pliu23@ur.rochester.edu), Yunlong Tang¹ (yunlong.tang@rochester.edu), Luchuan Song¹ (lsong11@rochester.edu), Jinxi He² (jhe44@u.rochester.edu) and Jiarui Wu² (jiaruiwu@andrew.cmu.edu)

Affiliations:

¹University of Rochester.

²Carnegie Mellon University.

brucelyu17

Members: Hanjia Lyu¹ (brucelyu17@gmail.com)

Affiliations:

¹University of Rochester.