COVER: A Comprehensive Video Quality Evaluator

Chenlong He[1], Qi Zheng[1], Ruoxi Zhu[1], Xiaoyang Zeng[1], Yibo Fan[1†], Zhengzhong Tu[2]
[1]Fudan University, [2]University of Texas at Austin
clhe22@m.fudan.edu.cn, fanyibo@fudan.edu.cn, zhengzhong.tu@utexas.edu

Abstract

Video quality assessment, especially for a massive scale of user-generated content, is an essential yet challenging computer vision and video analysis problem. Prior methods have been shown to be effective in mirroring subjective human opinion scores; however, they fail to capture the complicated, multi-dimensional aspects of factors that impact the overall perceptual quality. In this paper, we introduce COVER, a comprehensive video quality evaluator, a novel framework designed to evaluate video quality holistically — from a technical, aesthetic, and semantic perspective. Specifically, COVER leverages three parallel branches: (1) a Swin Transformer backbone implemented on spatially sampled crops to predict technical quality; (2) a ConvNet employed on subsampled frames to derive aesthetic quality; (3) a CLIP image encoder executed on resized frames to obtain semantic quality. We further propose a simplified cross-gating block to interact with the three branches before feeding into the predicting head. The final quality score is attained using a weighted sum of each subscore, making a multi-faceted metric. Our experimental results demonstrate that COVER exceeds the state-of-the-art models in multiple UGC video quality datasets. Moreover, COVER offers a diagnosable quality report to explain the quality score in multiple pillars, while it is capable of processing 1080p videos at 3x faster speed than the real-time requirement. To facilitate future research on efficient and explainable video quality research, the code is available at https://github.com/vztu/COVER.

1. Introduction

The widespread use of social media platforms like YouTube, Facebook, Instagram, and TikTok has called for a crucial research problem: developing effective and efficient tools to monitor, analyze, and process the perceptual quality of the vast amount of user-generated content (UGC) videos being uploaded, shared, and consumed globally at every moment. Unlike traditional video quality assessment (VQA) research, UGC videos often already suffer from various unpredictable distortions in their originals, ranging in multiple severities and types, significantly impacting users' quality of experience (QoE). This new form of prevailing content has facilitated recent advances in the so-called UGC video quality assessment (UGC-VQA) problem [32] that aims to build no-reference or blind video quality predictors to mimic the human's opinion scores on the perceptual quality of the presented videos.

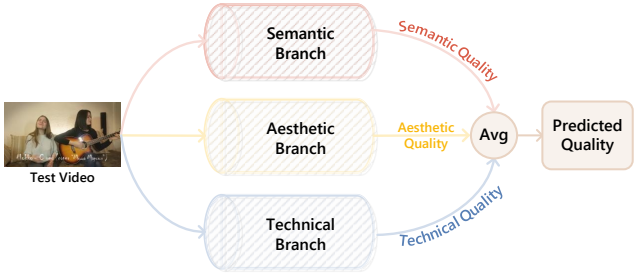


Figure 1. The flowchat of our proposed **CO**mprehensive **V**ideo quality **E**valuato**R** (**COVER**). COVER processes a video clip in three parallel branches: 1) a semantic branch, 2) an aesthetic branch, and 3) a technical branch. The final predicted quality is the average of three quality scores generated by these branches.

Perceptual video quality assessment of UGC can be approached via subjective and objective VQA studies. Subjective VQA studies [7, 25, 38, 50, 52] involve conducting a subjective experiment where human observers evaluate a diverse set of video sequences within a controlled environment or via crowdsourcing, then providing subjective opinion scores indicative of perceived quality. Despite accurately representing the 'actual' human judgments, subjective assessments are notoriously labor-intensive and time-consuming. Objective VQA methods, on the other hand, are studied to develop intelligent models that can automatically predict perceptual quality. In the context of

This work was supported in part by the National Key R&D Program of China (2023YFB4502802), in part by the National Natural Science Foundation of China (62031009), in part by the "Ling Yan" Program for Tackling Key Problems in Zhejiang Province (No.2022C01098), in part by Alibaba Innovative Research (AIR) Program, in part by Alibaba Research Fellow (ARF) Program, in part by the Fudan-ZTE Joint Lab, in part by CCF-Alibaba Innovative Research Fund For Young Scholars.
† Yibo Fan is the corresponding author.

UGC video quality analysis, only blind (no-reference or NR) VQA methods [12, 33, 41, 51] can be applied since access to the presumed pristine videos is unattainable.

However, holistic modeling of objective video quality assessment metrics tailored for UGC videos presents a significant challenge, given that multiple levels of characteristics intricately influence the perceptual quality of such videos. Firstly, the **low-level technical** quality degradations in UGC videos encompass authentic and commonly intermixed distortions that occurred in in-capture and/or post-capture processes [32, 55], including but not limited to flickering, judder, transmission distortions, and transcoding artifacts. Secondly, the **(mid-to-)high-level aesthetic** aspect encompasses abstract concepts such as emotional valence and the artistic quality within a scene [37, 44], which significantly influence human perceptions of the UGC video quality. Lastly, the **high-level semantic** aspect involves the understanding and recognizing semantic content and the consistency and logical coherence of semantic integrity [3, 12]. Factors such as video frame layout, object positioning and motion, and semantic distinguishability contribute to this aspect of quality perception. These multi-faceted perspectives collectively contribute to the perceptual quality assessment of UGC videos, making it a challenging visual task that involves multiple convoluted human perceptual factors.

Existing objective quality assessment models bespoke for UGC videos have not comprehensively addressed the aspects mentioned above. Traditional VQA models strive to quantify the quality degradation mainly from the technical aspects through designing handcrafted features and learning a mapping function from features to human opinions [4, 9, 17, 54], whereas neglecting the effects of higher-level quality factors results in their barely satisfactory performance. Deep learning-based VQA models [10, 12, 33, 42, 51] harness the capabilities of high-level semantic perception inherent in deep neural networks (DNN) that have been trained on semantic recognition tasks in computer vision. These pre-trained DNN models often manifest as the offline feature extraction or the initialization weights for the backbone. Among them, a few related works take both **technical and semantic effects** into consideration and develop dual-branch frameworks for quality prediction [10, 33, 56]. Moreover, quality perception [44] from the aesthetic aspect can be performed by leveraging the DNN models pre-trained on the aesthetic visual analysis task [19]. Recent VQA models leveraging aesthetic quality prediction [37, 43, 45, 46] propose to incorporate the merit of cross-modality learning, and apply the language-prompted approach to perform various levels of quality perception from low-level technical aspects to high-level aesthetic aspects. Among them, a few VQA models are developed to account for both **aesthetic and technical** perspectives [37, 44].

Nevertheless, there has been a limited effort to comprehensively model the perceptual quality assessment of UGC videos across the aforementioned three dimensions in a holistic approach. To fill this gap, we explore the possibility of a universal understanding of the video quality problem of UGC, and hereby propose an effective, efficient, and explainable VQA model, which we dub **COVER**, whose schematic overview is exemplified in Fig. 1. Unlike previous approaches, which inspect only one or at most two aspects of the perceptual quality, COVER presents a holistic strategy to model the video quality in three aspects: from technical, aesthetic, and semantic perspectives. We also introduce a novel cross-gating block to fuse the features coming from different branches to conduct feature interactions. Finally, these predicted scores for different axes are simply aggregated to arrive at the final quality prediction. Furthermore, we engineered efficient feature extraction designs that exploit the spatial and temporal frame redundancies, allowing COVER to run at a breakneck inference speed of 96 fps. Our key contributions are highlighted below:

- We present COVER, a comprehensive video quality evaluator that employs a parallel learning design to generate quality predictions from three different aspects: semantic, aesthetic, and technical dimensions.
- COVER utilizes a novel simplified cross-gating block to fuse features representing various aspects of the video perception to learn the interactions between different perceptual factors better.
- Extensive ablation studies have been conducted to demonstrate the effectiveness of each component design.
- Experiments show that COVER performs superior on the standard UGC-VQA databases, particularly excelling on the AIS 2024 UGC Video Quality Assessment Challenge.

2. Related work

2.1. Subjective studies on UGC-VQA

The earliest UGC-relevant VQA dataset is perhaps the Camera Video Database (CVD2014) [21] featuring authentic distortions from 78 video capture devices, followed by the LIVE-Qualcomm Mobile In-Capture Database [21]. These two databases primarily addressed in-camera distortions on limited video content. The KoNViD-1k video quality database [7] is a large-scale database comprised of 1,200 public-domain videos sampled from the YFCC100M dataset and annotated by 642 workers through online crowdsourcing. Subsequently, LIVE-VQC [25] emerged as another large-scale UGC video database with 585 videos containing human opinions from 4,776 crowdsourcing participants collected on Amazon Mechanical Turk. The authors of [38] collect 1,380 20-second video clips of diverse spatial resolutions (360p-4k) and frame rates (15-60 fps) from millions of YouTube videos to develop the YouTube-

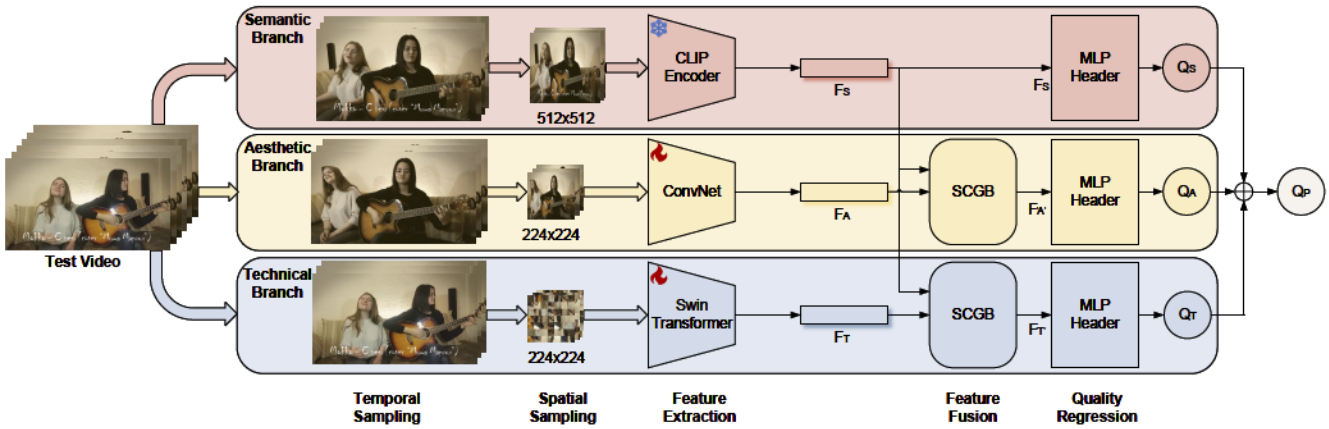


Figure 2. The architecture of our proposed **CO**mprehensive **V**ideo quality **E**valuato**R** (**COVER**). COVER processes a video clip in three parallel branches: 1) a semantic branch that extracts high-level object-semantics-related information using a pre-trained CLIP image Encoder; 2) an aesthetic branch that leverages a ConvNet run on subsampled image thumbnails to analyze their looking; 3) a technical branch utilizing Swin Transformer to execute on fragments. We also devise a simplified cross-gating block (SCGB) to fuse multi-branch features together, yielding the final quality score.

UGC dataset, with 8,000 human subjects included to rate these videos in an online subjective study. The LSVQ database [50] is by far the largest UGC dataset that contains 38,811 real-world distorted videos and 116,433 space-time localized video patches, whereby 5.5M human perceptual quality annotations are gathered to produce opinion scores. The most recently developed in-the-wild video database is DIVIDE-3k [44], consisting of 3,590 UGC videos and 450,000 human opinions collected from a multi-perspective subjective study, including aesthetic and technical rating.

2.2. Blind VQA models for UGC

Feature-based models. Previous conventional feature-based VQA models primarily account for technical quality degradations. The earliest algorithms are designed to quantify a single distortion type by measuring a small number of image/frame level features, such as blockiness, blur, ringing, banding, and compression [1, 5, 20, 31, 39, 40]. Afterward, a number of general-purpose BVQA models [4, 11, 17, 18, 32, 33, 47, 53–56] deliver impressive performance on UGC quality prediction, which may be attributed to the measurements of distortion-induced deviations of bandpass processed pictures/videos from perceptually relevant natural scene statistics (NSS) [23]. Diverse handcrafted quality-aware features are computed in [9] to account for spatial attributes, motion-induced statistics, and aesthetics features.

Deep learning-based models. The recent development in deep neural networks has considerably impacted UGC-VQA models. A plenty of deep learning-based VQA models [2, 12, 26, 33, 35, 36, 41, 42, 51] exploit off-the-shelf neural networks [6, 16, 24, 27, 28] pretrained on classic computer vision tasks to extract semantic-aware and technical-aware deep features. These pre-trained neural networks serve as either frozen feature extraction modules or initialization weights for the IQA/VQA model backbone. For example, VSFA [12] conducts subjectively-inspired temporal pooling on frame quality scores obtained from content-aware feature extraction followed by long-term memory modeling, wherein a frozen, pre-trained ResNet-50 [6] extracts feature maps from video frames. Patch-VQ [51] extracts 2D spatial features using PaQ-2-PiQ [49] and captures 3D spatio-temporal features using the 3D ResNet-18 [6]. FAST-VQA [41] and FasterVQA [42] employ the Swin-Transformer [16] pretrained on Kinetics-400 dataset [8] to initialize the backbone and introduced spatial and temporal grid mini-patch sampling to improve efficiency. DOVER [44] employs inflated-ConvNext [14] pretrained on AVA [19] (a database for aesthetic visual analysis) and Swin-Transformer [16] with GRPB [41] to initialize the backbones of aesthetic and technical branches, respectively. Recently, the rapid development of large multi-modality models [22, 29, 30, 57] has spurred a "paradigm shift" in visual quality prediction from the human-label regression manner to a language-prompted approach [37, 43, 45, 46]. For example, in CLIP-IQA [37], quality perception on both aesthetic and technical perspectives can be performed when feeding the multi-modality models with different quality-descriptive levels of prompts, such as 'a good/bad photo' and 'a sharp/fuzzy photo.'

3. Proposed Method

We present **COVER**, a **CO**mprehensive **V**ideo quality **E**valuato**R**, illustrated in Fig. 2. This network accepts videos and applies specific temporal-then-spatial sampling to obtain its input to the backbones. COVER employs three branches of backbones: a CLIP-based semantic branch, a

ConvNet-based aesthetic branch, and a Swin Transformer-based technical branch, each consisting of a feature extraction module and a quality regression module. Notably, aesthetic and technical branches additionally incorporate a feature fusion module to integrate features from the semantic branch via a channel cross-gating block. The input video is processed through these branches to generate three scores, reflecting the video's quality across the respective dimensions. The final score is the average of scores obtained from branches.

3.1. Temporal and Spatial Sampling

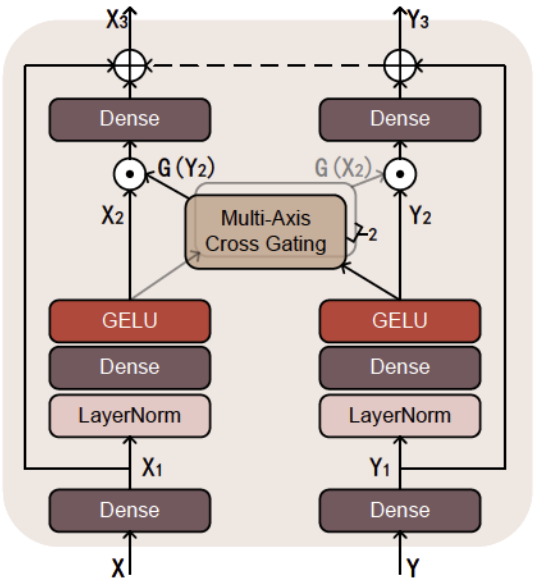
The input videos undergo bespoke temporal-spatial sampling before inputting into each branch's feature extraction module. As shown in Fig. 2, to improve the inference efficiency of the network, temporal sampling is designed to be very sparse, as we have observed that spatial perception is usually highly correlated and thus redundant in nearby frames. Specifically, the semantic branch samples one frame per thirty frames, while the aesthetic and technical branches sample two frames out of thirty. Since the temporal sampling processes of the three branches are independent, this approach ensures that, despite a reduced number of frames being processed, diverse information from the temporal domain is still adequately captured. We empirically observe that such a sparse sampling scheme is enough to secure a high performance with a small variance.

For spatial sampling, the semantic and aesthetic branches resize the video frames to 512x512 and 224x224, respectively, to align the input domain with the pre-trained setting. The technical branch, however, employs a fragment sampling operation introduced in [44], where a frame from the video is spatially divided into 7x7 sub-blocks. These sub-blocks are then randomly sampled and reassembled into a "scrambled" frame with a resolution of 224x224.

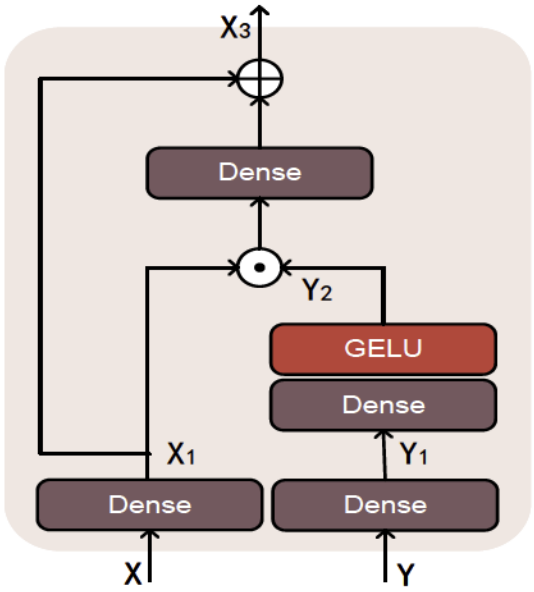
3.2. Feature Extraction and Fusion

Feature extraction backbones. Several studies have demonstrated the effectiveness of CLIP [22], a foundation model, in both IQA [37] and VQA [43] tasks. CLIP can accurately assess their subjective quality by extracting semantic information from images and videos. However, the above studies failed to address the challenging UGC-VQA task. This motivates us to employ the Image Encoder of CLIP as the backbone to extract semantics-aware features, and we then fine-tune a linear layer to map these features for quality prediction.

For the technical branch, the Swin Transformer [13] is utilized as the backbone of the feature extraction module. A CNN network, specifically the ConvNet [15], is used as the backbone of the feature extraction module for the aesthetic branch. These two branches are initialized with weights pretrained on the LSVQ [51] from DOVER [44], and it will be fine-tuned during subsequent training.



(a) Cross-Gating Block



(b) Simple Cross-Gating Block (SCGB)

Figure 3. Architecture of the original cross-gating block and its simplified version SCGB.

Feature fusion blocks. CLIP's image encoder enjoys robust capabilities in representing image semantics through the vision-language contrastive pre-training. Thus, the abundant information contained in CLIP's output features may inherently correlate with the features of the other branches. Here, we propose a feature fusion block to harness the representative features generated by the semantic branch and leverage them to modulate the other

branches that learn relatively lower-level quality aspects. More specifically, we modify the cross-gating block [34] to derive the novel channel-modulating block, which we dub Simple Cross-Gating Block (SCGB), for feature fusion between the semantic-aesthetic and semantic-technical feature pairs. Finally, The fused features from the aesthetic and technical branches, along with the features from the semantic branch, are then fed into their respective quality regression modules.

The detailed architecture of SCGB is depicted in Fig. 3(b). The input of the block is two tensors X and Y, where X is the feature from the technical or aesthetic branch, while Y is from the CLIP-based semantic branch. After input channel projections are applied, the projected CLIP features are fed to a gating pathway to yield the gating weights, which are then multiplied by the features from the other branch. Finally, a residual connection is adopted. Compared with the original cross-gating block in [34], SCGB has several novelties. Firstly, only one gating pathway is retained in SCGB, aiming to modulate the technical or aesthetic features using the semantic features. Secondly, considering that the spatial information is squeezed in the semantic branch, the operations concerning spatial interactions are eliminated in SCGB. In contrast, only channel-wise interactions are retained. Besides, the layer normalization is also removed because the embeddings have already been normalized in each branch. These simplifications effectively integrate semantic, aesthetic, and technical features by focusing on the representation dimensions. Compared with other feature fusion methods, such as cross-attention, the proposed SCGB enjoys reduced computational complexity while still capable of enhancing the network's performance in assessing video quality.

3.3. Quality Regression

The features from each branch are individually fed into a multi-layer perceptron (MLP) Header to predict quality scores, i.e., Q_S, Q_A, and Q_T, as shown in Fig. 2, and the final predicted quality, $Q_P = (Q_S + Q_A + Q_T)/3$. To enforce that each branch can independently capture the features of its focused dimension and accurately predict video quality, we adopted the limited view biased supervision scheme [44], which minimizes the relative loss between predictions in each branch with the overall opinion MOS, as formulated below:

$$\begin{aligned}\mathcal{L}_{all} =& \mathcal{L}_{rel}(Q_S, \text{MOS}) + \mathcal{L}_{rel}(Q_A, \text{MOS}) \\ &+ \mathcal{L}_{rel}(Q_T, \text{MOS})\end{aligned} \tag{1}$$

Our MLP Header comprises two Fully Connected (FC) layers, with the first FC layer followed by a GeLU activation layer. The second FC layer outputs a predicted score. To mitigate overfitting, dropout layers with a dropout ratio of 0.5 are inserted between the first FC layer and the input features, as well as between the GeLU activation layer and the second FC layer. This configuration not only enhances the model's ability to generalize by preventing it from relying too heavily on any single feature but also maintains a balance between its complexity and performance.

4. Experiments

4.1. Benchmarks

We evaluated our approach on three commonly used UGC databases with detailed metadata information shown in Table 1. The YouTube-UGC [38] is composed of UGC collected from YouTube, representing a diverse array of videos in terms of content, style, and quality. The KoNViD-1k [7] consists of 1,200 videos with a fixed resolution of 960x540. These videos are sourced from various open-source platforms and are representative of common quality distortions found in online video streaming. The LIVE-VQC [25] includes videos recorded using a variety of devices under different conditions. These datasets encompass a wide range of video content, quality levels, and resolutions, providing a comprehensive basis for testing and validating the effectiveness of our proposed COVER metric.

4.2. Implementation Details

The implementation details of COVER are explained below: i) the backbone of the feature extraction module for the semantic branch is the Image Encoder from CLIP [22] of type ViT-L/14; ii) the feature extraction backbone of the aesthetic branch is a ConvNet [15], structured into four stages. The configuration of each stage, defined by the number of blocks and feature dimensions, is (3, 96), (192, 3), (384, 9), and (768, 3); iii) the feature extraction backbone of the technical branch is a Swin Transformer [13], which also comprises four stages. Within each stage, the number of heads is set to 3, 6, 12, and 24, respectively, with the number of projection output channels being 96; iv) the SCGB module operates with input and output feature dimensions both set to 768, and its dropout layer has a drop ratio of 0.1; v) the input feature dimension for the MLP Header module is 768. It includes two dropout layers, both with a drop ratio of 0.5.

We train the COVER model in three consecutive stages:

1. **Initial training of technical and aesthetic branches:** We first train the entire network for both the technical and aesthetic branches. During this stage, the weights of both backbones and MLP Headers for all branches are fine-tuned.
2. **Integrating semantic branch:** Building on the best weights obtained from stage 1, we integrate the semantic branch into COVER. Then, MLP Headers of all branches, along with the backbones of both technical and aesthetic branches, are fine-tuned.

Table 1. Metadata of evaluated UGC databases, KoNViD-1k [7], LIVE-VQC [25], and YouTube-UGC [38].

Metadata	**KoNViD-1k [7]**	**LIVE-VQC [25]**	**YouTube-UGC [38]**
Publication year	2017	2018	2019
Source content	YFCC100m (Flickr)	Captured (mobile devices)	YouTube
Number of contents	1,200	585	1,380
Resolution	540p	1080p-240p	4k-360p
Framerate	24,25,30 fr/sec	20,24,25,30 fr/sec	15,20,24,25,30,50,60 fr/sec
Video duration	8 seconds	10 seconds	20 seconds
Experiment	Crowdsourcing (CrowdFlower)	Crowdsourcing (AMT)	Crowdsourcing (AMT)
Rating scale	Absolute category rating 1-5	Continuous rating 0-100	Continuous rating 1-5
Number of subjects	642	4776	>8,000
Number of ratings	136,800 (114 votes/video)	205,000 (240 votes/video)	170,159 (123 votes/video)

Table 2. Performance comparison of the VQA methods on YouTube-UGC [38] database, regarding SROCC, KROCC, PLCC, and RMSE. The validation set is specified by the AIS 2024 VQA challenge. The top performer is highlighted in boldface.

Method	SROCC ↑	KROCC ↑	PLCC ↑	RMSE ↓
BRISQUE [17]	0.4398	0.2934	0.4525	0.5608
GM-LOG [47]	0.3501	0.2336	0.3424	0.5904
VIDEVAL [32]	0.7946	0.5959	0.7691	0.4024
RAPIQUE [33]	0.7483	0.5556	0.7482	0.4177
FAVER [56]	0.7897	0.5832	0.7898	0.3861
NIQE [18]	0.2479	0.1689	0.3146	0.5976
HIGRADE [11]	0.7639	0.5524	0.7507	0.4156
FRIQUEE [4]	0.7182	0.5268	0.7091	0.4439
CORNIA [48]	0.5988	0.4113	0.5905	0.5064
TLVQM [9]	0.6690	0.4833	0.6412	0.4831
CLIP-IQA+ [37]	0.5374	0.3734	0.5801	0.5128
FasterVQA [42]	0.5345	0.3716	0.5438	0.5284
FAST-VQA [41]	0.6493	0.4676	0.6792	0.4621
DOVER [44]	0.7359	0.5391	0.7653	0.4053
FasterVQA*	0.6937	0.4965	0.6909	0.4552
FAST-VQA*	0.8617	0.6716	0.8669	0.3139
DOVER*	0.8761	0.6865	0.8753	0.3144
COVER	**0.9143**	**0.7413**	**0.9165**	**0.2519**

3. **Incorporation of SCGB:** Based on the optimal weights from stage 2, we add two SCGBs to the model. Subsequent fine-tuning of both SCGBs, along with all MLP Headers, is conducted.

Our multi-stage training approach uses the same data split across each step, allowing for incremental improvements in training effectiveness. Throughout different training stages, we use the ADAM optimizer with an initial learning rate of 1×10^{-3} and a cosine learning rate decay strategy with a decay weight of 0.05 over a total of 20 epochs. We adopt batch sizes across different stages, set to 10, 8, and 24, respectively. We implemented COVER in the Pytorch framework and trained it on an A6000 GPU card, which requires approximately one day to complete the entire training process.

4.3. Main Results

To validate the effectiveness of our proposed COVER method, we conducted evaluations on three widely recognized datasets in UGC-VQA research: YouTube-UGC [38], KoNViD-1k [7], and LIVE-VQC [25]. For a comprehensive assessment, we employed four key performance metrics: Spearman's Rank Order Correlation Coefficient (SROCC), Kendall's Rank Order Correlation Coefficient (KROCC), Pearson's Linear Correlation Coefficient (PLCC), and Root Mean Square Error (RMSE). Table 2 presents the comparison results of COVER and existing methods on the YouTube-UGC validation set specified by the AIS 2024 UGC video quality challenge. The notations FasterVQA*, FAST-VQA*, and DOVER* denote the results of the corresponding models after fine-tuning on the dataset. It may be seen that COVER exceeds all the other models in terms of all the metrics being evaluated: +4.3% and +4.7% improvement in SROCC and PLCC over the previous SoTA model DOVER, respectively.

Furthermore, we illustrate in Table 3 the comparison results on KoNViD-1k and LIVE-VQC. Evidently, on YouTube-UGC, our model outperforms all others across all metrics. On KoNViD-1k and LIVE-VQC, our COVER model always ranks within the top three in terms of both SROCC, PLCC, and RMSE metrics. These results further demonstrate the efficacy of our approach in a variety of UGC databases, particularly emphasizing its robustness across different content types and quality variations.

4.4. Ablations

Training strategy. The training strategy of a model is one of the key factors affecting its performance. As illustrated in Fig. 2, we do not freeze the backbones of aesthetic and technical branch of in COVER, as we found that fine-tuning these two models yielded better results. Specifically, we compared the impact of fine-tuning the backbones on the performance of COVER and COVER- on YouTube-UGC. Here, COVER- represents the initial version of COVER without the semantic branch and SCGB. The comparison

Table 3. Performance comparison of the VQA methods on two standard VQA databases, i.e., KoNViD-1k [7] and LIVE-VQC [25], in terms of SROCC, KROCC, PLCC, and RMSE. The top three performers are highlighted in boldface.

Method	KoNViD-1k [7]				LIVE-VQC [25]			
	SROCC ↑	KROCC↑	PLCC ↑	RMSE ↓	SROCC ↑	KROCC↑	PLCC ↑	RMSE ↓
BRISQUE [17]	0.6616	0.4784	0.6621	0.4799	0.5988	0.4248	0.6303	13.1100
GM-LOG [47]	0.6600	0.4778	0.6620	0.4801	0.5973	0.4243	0.6282	13.1318
VIDEVAL [32]	0.7857	0.5877	0.7813	0.3996	0.7138	0.5280	0.7260	11.6059
RAPIQUE [33]	0.7946	0.6004	0.8064	0.3809	0.7464	0.5601	0.7656	10.8694
FAVER [56]	0.7844	0.5893	0.7841	0.3967	0.7924	0.6041	0.7925	10.3067
NIQE [18]	0.5420	0.3794	0.5499	0.5351	0.5897	0.4185	0.6220	13.2455
HIGRADE [11]	0.7099	0.5222	0.7175	0.4458	0.6011	0.4287	0.6286	13.1543
FRIQUEE [4]	0.7457	0.5510	0.7483	0.4247	0.6653	0.4854	0.7049	11.9794
CORNIA [48]	0.7570	0.5570	0.7496	0.4265	0.6965	0.5094	0.7365	11.6949
TLVQM [9]	0.7688	0.5734	0.7657	0.4118	0.7965	0.6064	0.7991	10.1629
CLIP-IQA+ [37]	0.7813	0.5888	0.7817	0.3996	0.7276	0.5330	0.7789	10.6380
FasterVQA [42]	0.8272	0.6352	0.8289	0.3584	0.7728	0.5800	0.7906	10.4444
FAST-VQA [41]	**0.8543**	**0.6630**	**0.8508**	**0.3368**	**0.8211**	**0.6281**	**0.8359**	**9.3614**
DOVER [44]	**0.8752**	**0.6930**	**0.8816**	**0.3025**	**0.7989**	**0.6072**	**0.8348**	**9.3903**
COVER	**0.8933**	**0.7191**	**0.8947**	**0.2970**	**0.8093**	**0.6244**	**0.8478**	**9.3704**

Table 4. Comparison of performance on YouTube-UGC [38] database depending on whether finetuned the backbone of the model. The validation is specified by the AIS 2024 VQA challenge. The top one performing method is highlighted in boldface.

Diff. Finetune	YouTube-UGC [38] Validation			
	SROCC ↑	KROCC ↑	PLCC ↑	RMSE ↓
COVER- Frozen	0.8496	0.6543	0.8648	0.3162
COVER- Hot	0.8761	0.6865	0.8753	0.3144
COVER Frozen	0.8910	0.7106	0.8996	0.2821
COVER Hot	**0.9143**	**0.7413**	**0.9165**	**0.2519**

Table 5. Ablation studies of fine-tuning experiments on two standard VQA databases, i.e., KoNViD-1k [7] and LIVE-VQC [25], depending on which backbone of the model is used.

Diff. Backbone	KoNViD-1k [7]		LIVE-VQC [25]	
	SROCC ↑	PLCC ↑	SROCC ↑	PLCC ↑
COVER-org	**0.8933**	**0.8947**	0.8093	0.8478
COVER-ytb	0.8925	0.8938	**0.8220**	**0.8575**

results, as shown in Table 4, indicate that both COVER and COVER- achieve better performance when the backbones are fine-tuned.

Frozen vs. fine-tuning. To delve deeper into the impact of backbone fine-tuning, we conducted comparative experiments on KoNViD-1k and LIVE-VQC using two versions of COVER: one with the original, untrained backbone, denoted as COVER-org, and another with the backbone trained on YouTube-UGC, denoted as COVER-ytb. Notably, for both COVER-org and COVER-ytb, the backbones are kept frozen. As indicated in Table 5, the fine-tuned COVER-ytb shows slightly better performance on KoNViD-1k compared to COVER-org, whereas its performance on LIVE-VQC is worse than that of COVER-org. This suggests that fine-tuning backbones of COVER for a specific dataset can enhance its performance on that dataset while potentially diminishing the representativeness of features extracted by backbones for other datasets.

Table 6. Comparison of performance on YouTube-UGC [38] databasae accroding to the method of different temporal sampling. The validation is specified by the AIS 2024 VQA challenge. The top one performing method is highlighted in boldface.

Diff. Temp. Samp.	YouTube-UGC [38] Validation			
	SROCC ↑	KROCC ↑	PLCC ↑	RMSE ↓
Local Dense	0.8761	0.6865	0.8753	0.3144
Global Sparse	**0.8960**	**0.7180**	**0.8928**	**0.2916**

Temporal sampling. We conducted an additional experiment to compare the impact of different temporal sampling strategies on the model performance. For the technical branch, the temporal sampling method used in DOVER [44] involves randomly selecting a starting point in the video and consecutively sampling N frames, with an interval of T between two consecutive frames. This sampling method is referred to as Local Dense in Table 6. An alternative sampling method, termed Global Sparse, divides the video into M segments and randomly samples N/M frames from each segment, with an interval of T between two consecutive frames within the same segment. As shown in Table 6, for YouTube-UGC, the performance of global sparse sampling surpasses local dense. For a fair comparison, we

Table 7. Ablation studies of the component designs of COVER on the YouTube-UGC database [38].

No.	Branch			SCGB	YouTube-UGC [38] Validation			
	Technical	Aesthetics	Semantic		SROCC ↑	KROCC ↑	PLCC ↑	RMSE ↓
1	✔				0.8659	0.6759	0.8650	0.3159
2		✔			0.8234	0.6295	0.8439	0.3378
3			✔		0.8005	0.6096	0.8311	0.3502
4	✔	✔			0.8960	0.7180	0.8928	0.2916
5	✔		✔		0.8824	0.6997	0.8890	0.2883
6		✔	✔		0.8347	0.6455	0.8582	0.3232
7	✔	✔	✔		0.9006	0.7260	0.9052	0.2731
8	✔	✔	✔	✔	**0.9143**	**0.7413**	**0.9165**	**0.2519**

ensured that the total number of frames sampled by both methods was the same, and the model used for comparison was COVER-. This suggests that for YouTube-UGC, global technical distortion information reflects the objective quality of videos more effectively.

Component ablations. We conducted experiments on all the combinations of semantic, aesthetic, and technical branches on YouTube-UGC to demonstrate the importance of different dimensions of features and video quality. For models incorporating multiple branches, their final scores are calculated as the average of the scores from all included branches. The numerical results depicted in Table 7 show that when tested each branch independently (No. 1-3 in the table), technical branch has the best performance among the three dimensions. When testing various combinations of three branches (No. 4-6 in the table), we see significant performance improvements when aesthetic or semantic branch is combined with technical branch, possibly due to the consideration of both high-level and low-level features in videos. No. 7 in the table indicates that considering both aesthetic and semantic branches can further enhance the model performance, suggesting that although they both extract higher-level features, these two branches still focus on different and likely complementary aspects. Additionally, comparing No. 8 to No. 7, we show that adding the SCGB feature fusion block can further push foward the performance limit by around 1.5% SROCC.

4.5. Inference Time

VQA models are a highly practical tool that can be potentially deployed on large-scale video streaming platforms to monitor, analysis, and process millions of video streams every day. Therefore, the actual inference cost per video is highly significant to the system's total performance, which is directly tied to the annual revenue gain. To this point, the inference time of a VQA model is perhaps one of the most critical aspects that directly determine its larger impact, particularly in industry applications. Fortunately, we have imbued efficient modular design in every aspect of the COVER model, leading to high efficiency. We benchmarked the model inference time required by COVER on a video clip of 30 frames of 1080p resolution using a TITAN RTX graphic card. As shown in Table 8, COVER's semantic, aesthetic, and technical branch demands 191, 96, and 23 milliseconds, respectively, together adding up to a total inference time of 311 ms. In other words, this inference latency translates to a highly performant VQA metric with explainable properties and can inference at a speed of almost 3x faster than real-time processing speed.

Table 8. Inference time of COVER on a 30-frame chunk of a 1080p video on a TITAN RTX GPU card. The total 311 ms inference time translates to **96 fps**, 3x faster than real-time processing.

Branch	Semantic	Aesthetic	Technical	All
Time (ms)	191	96	23	311

5. Conclusion

In this paper, we present a novel blind video quality assessment model called COVER that can comprehensively predict multiple visual factors that conjointly impact the overall perceptual video quality. Specifically, COVER evaluates video quality from three distinct yet interconnected dimensions: from semantic, aesthetic, and technical perspectives. By incorporating the Image Encoder from CLIP to extract semantic features of videos and employing a cross-gating mechanism for feature fusion with the other two branches, COVER has achieved outstanding performance on three standard VQA databases, especially on the YouTube-UGC validation set, as specified by the AIS 2024 UGC Video Quality Assessment Challenge. Moreover, COVER has demonstrated superior efficiency thanks to our specific design of sparse spatial and temporal sampling strategies: It is capable of processing 1080p 30-frame clip with only 311 ms while exceeding the state-of-the-art to a large margin. COVER effectively extracts features from multiple dimensions of how humans perceive videos, and we expect that its efficacy can be extended to help researchers evaluate and improve AI-generated videos as well. We hope our work will facilitate future video-quality research on efficient and explainable perceptual modeling.

References

[1] David Boon Liang Bong and Bee Ee Khoo. Blind image blur assessment by using valid reblur range and histogram shape difference. *Signal Processing: Image Communication*, 29 (6):699–710, 2014. 3

[2] 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. 3

[3] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 2

[4] Deepti Ghadiyaram. Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of Vision*, 17(1)(32):1–25, 2017. 2, 3, 6, 7

[5] Rania Hassen, Zhou Wang, and Magdy M. A. Salama. Image sharpness assessment based on local phase coherence. *IEEE Trans. Image Process.*, 22(7):2798–2810, 2013. 3

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[7] 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 *2017 Ninth international conference on quality of multimedia experience (QoMEX)*, pages 1–6. IEEE, 2017. 1, 2, 5, 6, 7

[8] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. 3

[9] Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Trans. Image Process.*, 28(12):5923–5938, 2019. 2, 3, 6, 7

[10] Jari Korhonen, Yicheng Su, and Junyong You. Blind natural video quality prediction via statistical temporal features and deep spatial features. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3311–3319, 2020. 2

[11] Debarati Kundu, Deepti Ghadiyaram, Alan C Bovik, and Brian L Evans. No-reference quality assessment of tone-mapped HDR pictures. *IEEE Trans. Image Process.*, 26(6): 2957–2971, 2017. 3, 6, 7

[12] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2351–2359, 2019. 2, 3

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4, 5

[14] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022. 3

[15] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 4, 5

[16] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211, 2022. 3

[17] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.*, 21(12):4695–4708, 2012. 2, 3, 6, 7

[18] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 3, 6, 7

[19] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415, 2012. 2, 3

[20] Niranjan D. Narvekar and Lina J. Karam. A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection. In *2009 International Workshop on Quality of Multimedia Experience*, pages 87–91, 2009. 3

[21] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen. Cvd2014—a database for evaluating no-reference video quality assessment algorithms. *IEEE Transactions on Image Processing*, 25(7):3073–3086, 2016. 2

[22] 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. 3, 4, 5

[23] Daniel L Ruderman. The statistics of natural images. *Netw.: Comput. Neural Syst.*, 5(4):517–548, 1994. 3

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[25] Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2):612–627, 2018. 1, 2, 5, 6, 7

[26] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai. A deep learning based no-reference quality assessment model for ugc videos. *arXiv preprint arXiv:2204.14047*, 2022. 3

[27] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE con-*

ference on computer vision and pattern recognition, pages 2818–2826, 2016. 3

[28] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 3

[29] MosaicML NLP Team et al. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023. *URL www. mosaicml. com/blog/mpt-7b. Accessed*, pages 05–05, 2023. 3

[30] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[31] Zhengzhong Tu, Jessie Lin, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Bband index: a no-reference banding artifact predictor. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pages 2712–2716, 2020. 3

[32] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C. Bovik. UGC-VQA: Benchmarking blind video quality assessment for user generated content. *IEEE Trans. Image Process.*, 30:4449–4464, 2021. 1, 2, 3, 6, 7

[33] Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. RAPIQUE: Rapid and accurate video quality prediction of user generated content. *IEEE Open Journal of Signal Processing*, 2:425–440, 2021. 2, 3, 6, 7

[34] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5769–5780, 2022. 5

[35] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pages 459–479. Springer, 2022. 3

[36] Domonkos Varga. No-reference video quality assessment based on the temporal pooling of deep features. *Neural Processing Letters*, 50(3):2595–2608, 2019. 3

[37] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2555–2563, 2023. 2, 3, 4, 6, 7

[38] Yilin Wang, Sasi Inguva, and Balu Adsumilli. Youtube ugc dataset for video compression research. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE, 2019. 1, 2, 5, 6, 7, 8

[39] Zhou Wang, Alan C Bovik, and Brian L Evan. Blind measurement of blocking artifacts in images. In *Proc. IEEE Int. Conf. Image Process.*, pages 981–984, 2000. 3

[40] Zhou Wang, Hamid R Sheikh, and Alan C Bovik. No-reference perceptual quality assessment of JPEG compressed images. In *Proc. IEEE Int. Conf. Image Process.*, pages I–I, 2002. 3

[41] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling. In *European conference on computer vision*, pages 538–554. Springer, 2022. 2, 3, 6, 7

[42] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, Jinwei Gu, and Weisi Lin. Neighbourhood representative sampling for efficient end-to-end video quality assessment. *arXiv preprint arXiv:2210.05357*, 2022. 2, 3, 6, 7

[43] Haoning Wu, Liang Liao, Jingwen Hou, Chaofeng Chen, Erli Zhang, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring opinion-unaware video quality assessment with semantic affinity criterion. *arXiv preprint arXiv:2302.13269*, 2023. 2, 3, 4

[44] 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*, pages 20144–20154, 2023. 2, 3, 4, 5, 6, 7

[45] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Towards explainable in-the-wild video quality assessment: a database and a language-prompted approach. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1045–1054, 2023. 2, 3

[46] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al. Q-bench: A benchmark for general-purpose foundation models on low-level vision. *arXiv preprint arXiv:2309.14181*, 2023. 2, 3

[47] Wufeng Xue, Xuanqin Mou, Lei Zhang, Alan C Bovik, and Xiangchu Feng. Blind image quality assessment using joint statistics of gradient magnitude and laplacian features. *IEEE Trans. Image Process.*, 23(11):4850–4862, 2014. 3, 6, 7

[48] Peng Ye, Jayant Kumar, Le Kang, and David Doermann. Unsupervised feature learning framework for no-reference image quality assessment. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1098–1105, 2012. 6, 7

[49] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3575–3585, 2020. 3

[50] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-VQ: 'patching up' the video quality problem. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 14019–14029, 2021. 1, 3

[51] Zhenqiang Ying, Maniratnam Mandal, Deepti 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*, pages 14019–14029, 2021. 2, 3, 4

[52] Xiangxu Yu, Zhengzhong Tu, Zhenqiang Ying, Alan C Bovik, Neil Birkbeck, Yilin Wang, and Balu Adsumilli. Subjective quality assessment of user-generated content gaming videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 74–83, 2022. 1

[53] Qi Zheng, Zhengzhong Tu, Yibo Fan, Xiaoyang Zeng, and Alan C Bovik. No-reference quality assessment of variable frame-rate videos using temporal bandpass statistics. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1795–1799. IEEE, 2022. 3

[54] Qi Zheng, Zhengzhong Tu, Zhijian Hao, Xiaoyang Zeng, Alan C Bovik, and Yibo Fan. Blind video quality assessment via space-time slice statistics. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 451–455. IEEE, 2022. 2

[55] Qi Zheng, Zhengzhong Tu, Xiaoyang Zeng, Alan C Bovik, and Yibo Fan. A completely blind video quality evaluator. *IEEE Signal Processing Letters*, 29:2228–2232, 2022. 2

[56] Qi Zheng, Zhengzhong Tu, Pavan C Madhusudana, Xiaoyang Zeng, Alan C Bovik, and Yibo Fan. Faver: Blind quality prediction of variable frame rate videos. *Signal Processing: Image Communication*, 122:117101, 2024. 2, 3, 6, 7

[57] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3