This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

An Empirical Study for Efficient Video Quality Assessment

Wei Sun[1]*, Kang Fu[2]*, Linhan Cao[2], Dandan Zhu[1]†, Kaiwei Zhang[2],
Yucheng Zhu[2], Zicheng Zhang[2], Menghan Hu[1], Xiongkuo Min[2], Guangtao Zhai[2]†
[1]East China Normal University, [2]Shanghai Jiao Tong University

Abstract

Video quality assessment (VQA) plays a critical role in optimizing various video processing systems, yet achieving both high accuracy and computational efficiency remains a challenging task. Recent advances in deep neural network (DNN)-based VQA models have led to notable performance improvements, but often at the cost of high computational complexity and memory consumption, limiting their applicability in resource-constrained scenarios. In this paper, we empirically investigate a set of good practices for building efficient yet effective VQA models. Specifically, we decompose the VQA training pipeline into three components: video preprocessing, quality-aware feature extraction, and optimization techniques. For each component, we identify and validate effective practices using the KVQ dataset—a user-generated content (UGC) VQA dataset that includes both in-the-wild distortions and processing-induced artifacts such as compression and enhancement. Based on these findings, we propose E-VQA, an efficient VQA model that combines the best-performing practices. Experiments conducted on the KVQ dataset, as well as the large-scale UGC VQA dataset LSVQ, demonstrate that E-VQA achieves competitive performance while significantly reducing computational complexity. Furthermore, E-VQA ranked third in the NTIRE 2025 Short-form UGC Video Quality Assessment Challenge, highlighting its practical effectiveness. The code is available at https://github.com/sunwei925/E-VQA.

1. Introduction

The proliferation of user-generated content (UGC) video on the Internet and the increasing demand for high-quality video delivery have made video quality assessment (VQA) [27, 37] a critical component in modern multimedia systems. From video capture to processing, VQA serves as an essential tool for optimizing processing pipelines and ensuring satisfactory user experiences. Recently, with the rapid advancement of deep neural network (DNN) techniques [1, 5, 21], VQA research has made significant progress, demonstrating a strong correlation with human perception (*e.g.*, mean opinion scores). However, as shown in Figure 1, designing VQA models that simultaneously achieve high accuracy and computational efficiency remains a challenging task.

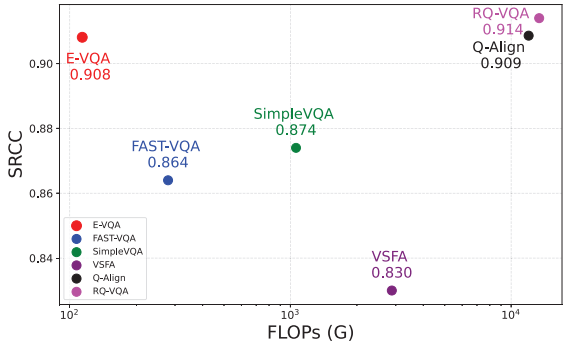


Figure 1. Performance comparison of different VQA methods in terms of FLOPs (G) vs. SRCC on the KVQ validation set. Each point represents a method, where higher SRCC and lower FLOPs indicate better performance-efficiency trade-off.

In general, the VQA models are categorized into three types: full-reference (FR) [4, 19, 30, 50, 57], reduced-reference (RR) [25], no-reference (NR) [12, 13, 15, 22, 23, 32, 36, 39, 51]. FR and RR VQA models require complete and partial pristine video references, whereas NR VQA, also referred to as blind VQA (BVQA), operates without any reference information, relying solely on distorted video content. Given that most UGC videos lack corresponding undistorted pristine counterparts, BVQA has gained broader practical applicability. Therefore, we focus on BVQA in this paper.

In previous studies, BVQA is commonly approached from two paradigms: handcrafted feature-based methods [13, 29, 32, 44, 45] and deep neural network (DNN)-based methods [14, 15, 18, 20, 42, 49, 54–56]. Handcrafted feature-based BVQA methods design quality-related features according to prior knowledge of human visual percep-

*These authors contributed equally to this work.
†Corresponding authors.

tion to extract features for quality regression. For instance, natural scene statistics (NSS) [29, 32] have been demonstrated to be highly effective to synthetic distortions (*e.g.,* noise, blur, compression, etc.) and are among the most widely used features in handcrafted featured-based BVQA approaches. Besides, some visual descriptors [13]—such as texture, noise, contrast, motion vectors, and optical flow—have also been explored for quality assessment. However, handcrafted feature-based methods are typically limited to synthetic distortions and fail to capture the influence of video content [38], leading to suboptimal performance on UGC videos. Moreover, the handcrafted feature-based methods are normally computationally expensive and unsuitable for real-time applications.

As for DNN-based BVQA methods, they automatically extract quality-aware features via a learning-driven paradigm. Current methods generally adopt three architectural strategies: (1) extracting frame-wise features via 2D networks followed by temporal aggregation using sequence models to derive video-level representations [15, 55]; (2) combining 2D networks for keyframe feature extraction with 3D networks capturing motion features from video segments, subsequently fused through concatenation or sequence modeling [38, 52]; and (3) directly extracting spatiotemporal features from video segments using 3D networks [51]. Leveraging the powerful feature representation capabilities of DNNs, these methods have demonstrated substantial progress in recent years, achieving state-of-the-art performance on UGC VQA benchmark datasets [10, 11, 31, 35, 48, 56]. However, their computational complexity—driven by the massive parameter count of DNNs and the need to process dense video frames—poses significant challenges for real-world deployment, particularly in large-scale video streaming applications.

To tackle this challenge, we propose **E-VQA**, a computationally efficient video quality assessment model that achieves state-of-the-art performance while requiring substantially fewer computational resources. Building upon recent exploration in efficient VQA frameworks, such as SimpleVQA [38], FAST-VQA [51], and MinimalisticVQA [40], we systematically decompose the VQA model training process into three key components: video preprocessing, quality-aware feature extraction, and optimization strategies. For each component, we conduct extensive ablation studies to strike an optimal balance between efficiency and effectiveness. Specifically, in video preprocessing, we explore spatial resolution choices and temporal sampling rates for key frames. For quality-aware feature extraction, we compare one single-branch architecture and two dual-branch architectures that capture complementary representations—motion features extracted by an extra action recognition network and distortion features using the same network extracted from image fragments sampled from key frames. In the optimization stage, we examine loss function design and pre-training strategies aimed at distilling knowledge from expert VQA models. We empirically validate these design decisions on the KVQ benchmark [24], establishing good practices for developing efficient VQA. Based on these insights, we propose E-VQA, which delivers performance close to the state-of-the-art method RQ-VQA [41] on the KVQ dataset while consuming substantially fewer computational resources than RQ-VQA. Furthermore, evaluations on the LSVQ dataset [56] further confirm the effectiveness of E-VQA.

Our key contributions are summarized as follows:

- We decompose the VQA training process into three essential components: video preprocessing, quality-aware feature extraction, and optimization strategies. We further identify good practices within each component to enable efficient VQA.
- Based on these insights, we propose an efficient and effective video quality assessment model named E-VQA that achieves high performance with low computational complexity.
- Extensive experiments on the KVQ and LSVQ datasets demonstrate the effectiveness and efficiency of the proposed E-VQA model.

2. Related Work

In this section, we briefly review the related work on VQA datasets and BVQA models, which have mutually promoted the development of BVQA studies in recent years.

2.1. VQA Datasets

Traditional VQA datasets primarily investigate synthetic distortions, particularly compression and transmission artifacts introduced by various codecs. Popular datasets include LIVE-VQA [33], EPEL-PoliMl [3], CSIQ-VQA [47], etc. These datasets typically comprise a small set of high-quality reference videos and their compressed counterparts generated using different coding standards. Their primary objective is to support the development of FR VQA models for evaluating and optimizing compression algorithms. However, the synthetic distortions in these datasets are overly simplistic and fail to capture the complex and diverse degradations commonly found in real-world scenarios, such as UGC on social media platforms. Hence, in-the-wild VQA datasets are needed to facilitate the development of general BVQA models capable of handling various kinds of authentic distortions.

To address these limitations, more recent efforts [10, 11, 31, 35, 48, 56] have shifted their focus on authentic, in-the-wild distortions. CVD2014 [31] contains 234 videos recorded across 5 different scenes using 78 cameras, including 3 DSLR, 4 digital video, 8 digital still, and 63 mobile cameras. Due to the diversity of capture devices and set-

tings, the video formats, resolutions, and frame rates are highly varied. LIVE-Qualcomm [10] comprises 208 videos captured by 8 smartphones across 54 diverse scenes, focusing on six dominant distortion types: artifacts, color, exposure, focus, sharpness, and stabilization. LIVE-VQC [35] contains 585 videos captured by 80 mobile cameras, each corresponding to a unique scene with diverse content and motion level under varying lighting conditions to reflect complex real-world distortions. KoNViD-1k consists of 1,200 videos sampled from YFCC100M [43] based on six attributes: blur, colorfulness, contrast, spatial information, temporal information, and video quality (V-NIQE [28]). YouTube-UGC [48] consists of 1,500 videos selected from 1.5 million YouTube videos across 15 categories using a distribution-matching strategy, covering various resolutions. LSVQ [56] is a large-scale VQA dataset comprising 38,811 Internet videos sampled based on six feature distributions—*i.e.,* absolute luminance, colorfulness, RME luminance contrast, number of detected faces, spatial Gaussian derivative filters, and temporal Gaussian derivatives. Notably, it also provides 116,433 quality-labeled space-time video patches to enable localized quality analysis.

In practical applications such as social media platforms (*e.g.*, Kwai and TikTok), UGC videos are often further compressed or enhanced after being uploaded. Therefore, the visual quality of these videos is influenced by both in-capture distortions and artifacts introduced by various video processing algorithms. To address this challenge, several VQA datasets have been developed to jointly consider in-the-wild distortions alongside compression and enhancement effects. For instance, Li *et al.* [18] select 50 UGC videos from TikTok and apply two encoders including H.264 and H.265 to compress each video at five quantization parameter (QP) levels. The TaoLive dataset [61] consists of 418 source videos from the TaoLive platform and 3,762 distorted ones compressed using the H.265 codec at 8 different constant rate factor (CRF) levels. The VDPVE dataset [8] includes 184 low-quality videos and 1,211 enhanced ones processed with various algorithms for light enhancement, contrast adjustment, color correction, deblurring, and stabilization, aiming to investigate the impact of enhancement techniques on perceived quality. The KVQ dataset [24] further explore the influence of complete video processing pipelines—including pre-processing, transcoding, and enhancement—on visual quality, comprising 600 user-uploaded videos from Kwai and 3,600 corresponding processed ones.

2.2. BVQA Models

Handcrafted Feature-based BVQA Methods. As stated in Section 1, NSS features are the most frequently used features for BVQA. For instance, Saad *et al.* [32] propose Video BLIINDS, a BVQA model that leverages spatio-temporal NSS and motion coherency features in the DCT domain to evaluate video quality without access to reference videos. Mittal *et al.* [29] propose VIIDEO, a training-free blind VQA model that evaluates video quality by modeling space-time statistical regularities in frame-difference signals, using NSS and temporal correlation features. TLVQM [13] is a two-level no-reference VQA framework, which first computes low-complexity motion and spatial features across the entire video , and then extracts high-complexity features from representative frames, followed by SVR or RFR regression to evaluate quality scores. VIDEVAL [44] is a fusion-based BVQA model that selects 60 statistically effective features from existing efficient BI/VQA models to achieve a balance between prediction performance and computational cost for UGC videos. RAPIQUE [45] combines low-level NSS with high-level convolutional neural network (CNN) features in a two-branch framework, incorporating efficient spatio-temporal sampling and bandpass-filtered temporal statistics to achieve competitive performance. Although diverse types of handcrafted features have been investigated for BVQA, these methods often suffer from high computational complexity and suboptimal performance on the UGC VQA datasets, thereby constraining their practical applicability.

DNN-based BVQA Methods. DNN-based BVQA models primarily utilize neural networks to learn quality-aware features directly from data. For example, VSFA [15] integrates content-aware features extracted from a pre-trained CNN and model temporal-memory effects using a GRU network [2] and a subjectively-inspired temporal pooling layer. Yi *et al.* [55] propose an end-to-end BVQA model that integrates non-local attention modules into VGG [34] for capturing spatial dependencies and combines GRU-based temporal modeling with hybrid temporal pooling to address the uneven distortion distribution in UGC videos. Ying *et al.* [56] introduce Patch-VQ, a local-to-global BVQAA model that extracts 2D spatial and 3D spatio-temporal features from full videos and their localized patches, and uses temporal regression via InceptionTime to obtain both global video quality and fine-grained space-time quality maps. Wang *et al.* [49] present CoINVQ, a multi-branch framework that decomposes UGC video quality into content, distortion, and compression features via three specialized subnetworks, and aggregates them to derive quality scores and interpretable indicators. Li *et al.* [14] propose a BVQA model that leverages quality-aware pre-training on IQA datasets [58] and motion features extracted from a 3D action recognition network, fused through a GRU network and optimized with a mixed list-wise ranking loss for in-the-wild video quality assessment. Sun *et al.* [38] develop SimpleVQA, which jointly learns end-to-end spatial quality-aware features and pretrained motion features, followed by MLP regression and multi-scale quality fusion

guided by human visual perception. Wu *et al.* [51] introduce FAST-VQA, which employ a fragment-based sampling strategy and a dedicated Fragment Attention Network (FANet) to enable efficient end-to-end VQA, significantly reducing computational cost while preserving both spatial and temporal quality cues for high-resolution videos. Wu *et al.* [53] further propose Q-Align, which teaches large multimodal models (LMMs) to perform visual quality assessment by emulating human rating behavior through instruction tuning with discrete text-defined levels. Sun *et al.* [40] propose a family of minimalistic BVQA models composed of four basic modules—video preprocessor, spatial quality analyzer, optional temporal quality analyzer, and quality regressor. Ge *et al.* [9] introduce LMM-VQA, which reformulates BVQA as an instruction-tuned question-answering task and introduces a spatiotemporal enhanced LMM that integrates 2D and 3D visual encoders for deriving quality scores and levels. Sun *et al.* [41] develop RQ-VQA, which enhances the SimpleVQA framework by integrating quality-aware features extracted from pre-trained BIQA and BVQA models (LIQE [59], Q-Align [53], and FAST-VQA [51]), and regress the concatenated features into quality scores via an MLP for robust assessment of UGC videos.

Although DNN-based BVQA models have achieved significant progress on popular UGC VQA benchmarks, their high computational complexity hinders deployment in large-scale real-world applications. To address this limitation, Zhang *et al.* propose Split-Conv [60], a resource-efficient structural pruning method for I/VQA models that splits convolutional layers into sub-kernels and decorators to evaluate filter importance, enabling effective model compression with minimal retraining on small I/VQA datasets. Mei *et al.* [26] develop a lightweight blind video quality assessment method GreenBVQA that combines unsupervised spatial, spatio-color, temporal, and spatio-temporal representations with supervised feature selection and ensemble-based regression, achieving high performance with low computational cost suitable for edge deployment.

3. Proposed Method

In this section, we aim to develop a VQA model that is both efficient and effective. To this end, we first constrain the computational complexity (measured in GFLOPs) of the VQA models to remain below a given threshold, and then explore the best practices that satisfy this constraint while achieving optimal performance.

3.1. Problem Formulation

Given a video $\boldsymbol{x} = \{\boldsymbol{x}_i\}_{i=0}^{N-1}$ with N frames at a frame rate of r, and a BVQA model $\mathcal{Q}$ that maps $\boldsymbol{x}$ to a scalar quality score q, our goal is to minimize its computational complexity $\mathcal{C}$ while maximizing its evaluation performance $\mathcal{P}$. To achieve this goal, we systematically decompose the VQA training pipeline into three key components: video preprocessing, quality-aware feature extraction, and optimization strategies. For each component, we ablate commonly adopted practices from prior VQA studies and identify the most effective configurations.

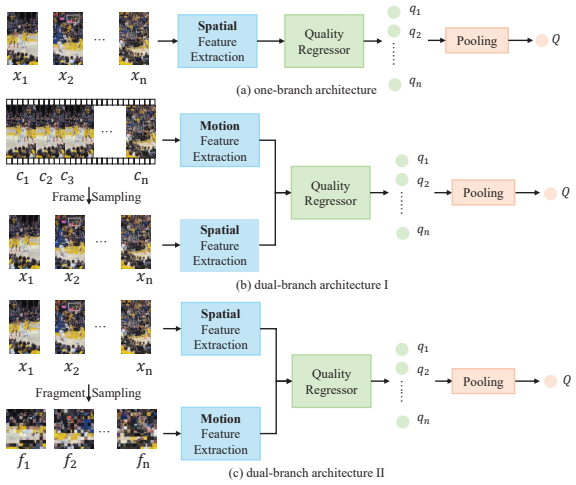


Figure 2. We evaluate three types of DNN architectures: one single-branch network (a) and two dual-branch variants (b) (c). In the dual-branch designs, one (b) incorporates a motion feature extraction branch, while the other (c) integrates a distortion-aware feature extraction branch.

3.2. Video Preprocessing

Videos inherently comprise dense frame sequences, where direct processing of individual frames entails substantial computational burdens. To alleviate complexity, effective preprocessing techniques are critical. In this paper, we adopt a similar video preprocessing approach as in MinimalisticVQA, where a sparse set of key frames is sampled for quality-aware feature extraction. Accordingly, we systematically investigate the impact of different preprocessing parameters, including five spatial resolution 224×224, 320×320, 384×384, 448×448, and 512×512, and temporal sampling rate for key frames, including four temporal sampling rates $1/8$, $1/4$, $1/2$, and 1 fps. It is noted that with the increasing the spatial resolution and temporal sampling rate of key frames, the computational complexity also increases. Notably, increasing either the spatial resolution or the temporal sampling rate leads to higher computational costs. Therefore, it is necessary to balance these parameters to achieve an optimal trade-off between model performance and computational efficiency.

3.3. Quality-aware Feature Extraction

Previous VQA models generally employ two distinct paradigms for quality-aware feature extraction. The first paradigm utilizes features derived from pre-trained com-

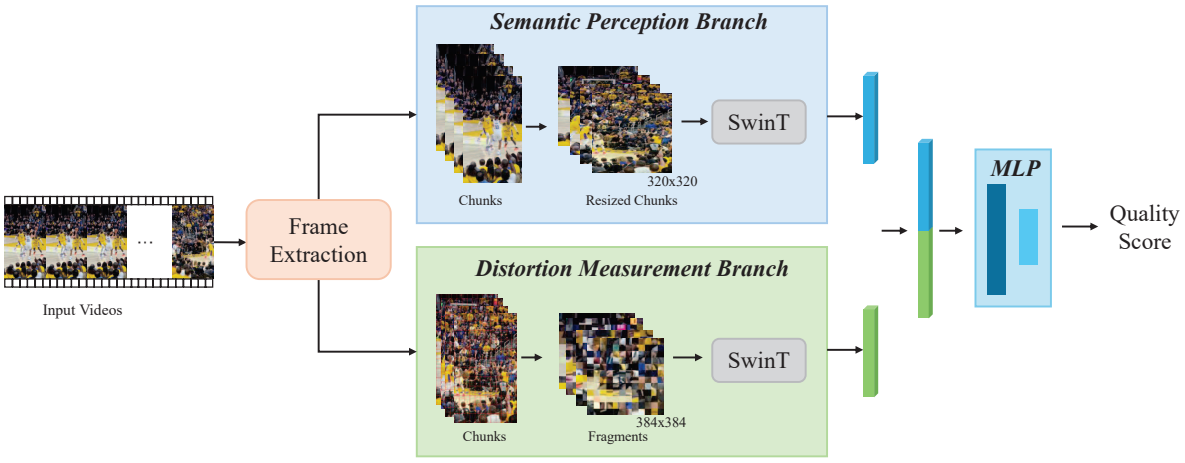


Figure 3. The architecture of E-VQA consists of two branches: one branch takes the resized full-frame image as input for semantic feature extraction, while the other processes a cropped fragment image to extract distortion-aware features.

puter vision models without fine-tuning on the training set [14, 15, 56]. The second paradigm adopts end-to-end training of the feature extraction network directly on the training set [38, 40, 51]. Pre-trained models often require large-scale architectures with substantial computational complexity, while end-to-end training enables learning of quality-aware features through relatively small networks. Motivated by these considerations, this work mainly focuses on the end-to-end fine-tuning paradigm.

As illustrated in Figure 2, we investigate two distinct DNN architectures: single-branch and dual-branch configurations. The single-branch architecture employs a standard backbone network (*i.e.,* Swin Transformer [21]) as the spatial feature extractor, followed by a multi-layer perceptron (MLP) regression head that maps features to quality scores. While this design is simple and computationally efficient, it primarily focuses on spatial information and lacks sensitivity to temporal characteristics. The dual-branch architecture extends this framework through parallel processing streams: one branch retains the spatial feature extraction pipeline, while the secondary branch incorporates complementary feature extraction mechanisms to enhance VQA performance. We explore two implementations for the secondary branch: (1) motion-aware variant, which integrates a pre-trained action recognition network (*i.e.,* X3D [6]) to capture temporal dynamics; (2) distortion-aware variant, which leverages a modified 2D FAST-VQA architecture to quantify local distortions using fragment images sampling from key frames [51].

3.4. Optimization Techniques

We investigate two optimization techniques for training efficient VQA models: loss function analysis and knowledge distillation. Our framework evaluates three fundamental loss function: (1) A ranking loss that enforces ordinal consistency between predicted and ground-truth quality scores, (2) L1/L2 regression losses minimizing pointwise distance between score distributions, and (3) Pearson Linear Correlation Coefficient (PLCC) loss that jointly optimizes both rank preservation and distributional alignment through statistical correlation maximization. For knowledge distillation, we adopt an offline strategy where a high-capacity expert VQA model serves as the teacher. The teacher model is used to pseudo-label a large-scale unlabeled video dataset, which is then used to pre-train a lightweight student model. The pre-trained student model is subsequently fine-tuned on the target dataset. This two-stage procedure effectively transfers knowledge from the expert model and accelerates convergence during optimization.

4. Experiments

4.1. Experimental Implementations

Test Datasets. We evaluate our model on two UGC VQA datasets: KVQ [24] and LSVQ [56]. The KVQ dataset consists of source UGC videos and their corresponding processed ones, including both compressed and enhanced variants. In contrast, LSVQ contains only in-the-wild UGC videos. For KVQ, we train our model on the publicly available training data from the NTIRE 2025 Short-form UGC Video Quality Assessment Challenge[1], and evaluate it on both the validation and test sets. Specifically, the dataset includes $2,926$ training videos, 420 validation videos, and 854 test videos. For LSVQ, we follow the standard training and testing split proposed in [56], consisting of $28,013$ training videos, $7,220$ test videos, and an additional $3,561$

[1] https://codalab.lisn.upsaclay.fr/competitions/21335

Table 1. Performance of the baseline model (single-branch network structure) with different spatial resolutions on the KVQ validation set

Backbone	Resolution	SRCC	PLCC	GFLOPs
SwinT	224	0.814	0.834	36.1
	320	0.859	0.869	79.5
	384	0.859	0.871	115
	448	0.864	0.874	144
	520	0.879	0.886	213
SwinB	224	0.848	0.856	124
	320	0.851	0.866	267
	384	0.859	0.872	401
	448	0.850	0.872	494
	520	0.875	0.882	724

high-resolution videos in the test-1080p subset.

Evaluation Criteria We adopt two standard metrics to evaluate VQA model performance: the Spearman's rank-order correlation coefficient (SRCC) and the Pearson linear correlation coefficient (PLCC). SRCC measures the monotonicity between predicted and ground-truth scores, while PLCC assesses both the linearity and accuracy of the predictions. An effective VQA model should achieve SRCC and PLCC values close to 1. Before calculating the PLCC, we follow the procedure described in [46] and apply a four-parameter logistic function to map the predicted objective scores to the subjective MOS scale, ensuring consistency in score alignment.

Competitors. We compare the proposed method against nine representative VQA methods, including four handcrafted feature-based methods: NIQE [28], TLVQM [13], VIDEVAL [44], and RAPIQUE [45], and five DNN methods: VSFA [15], SimpleVQA [38], FAST-VQA [52], Q-Align [53], and RQ-VQA [41]. For fair comparison, all models are trained under the same conditions.

Training Details. We train the proposed model using the Adam optimizer with an initial learning rate of 1×10^{-5} and a batch size of 6 on a server equipped with two NVIDIA RTX 3090 GPUs. For the KVQ dataset, the learning rate is decayed by a factor of 10 after 10 epochs, and training is conducted for a total of 30 epochs. For the LSVQ dataset, the learning rate is decayed by a factor of 10 after 2 epoch, with training proceeding for a total of 10 epochs.

4.2. Empirical Evaluations on KVQ

We evaluate the effectiveness of the proposed practices in Section 3 using the KVQ dataset [24] for both training and validation. To ensure model efficiency, we impose a computational complexity threshold of 120 GFLOPs, and restrict all evaluated models to remain within this limit.

Table 2. Performance of the baseline model (single-branch network structure) with different temporal sampling rates on the KVQ validation set

Backbone	Temporal Rate	SRCC	PLCC	GFLOPs
Input Resolution: 320				
SwinT	1	0.859	0.869	79.5
	1/2	0.859	0.868	39.8
	1/4	0.821	0.832	19.9
	1/8	0.754	0.761	9.94
SwinB	1	0.851	0.866	267
	1/2	0.843	0.858	134
	1/4	0.836	0.849	66.8
	1/8	0.755	0.763	33.4
Input Resolution: 384				
SwinT	1	0.859	0.871	114
	1/2	0.859	0.871	57.4
	1/4	0.830	0.844	28.7
	1/8	0.791	0.794	14.4
SwinB	1	0.859	0.872	401
	1/2	0.854	0.867	200
	1/4	0.846	0.858	100
	1/8	0.786	0.786	50.1

4.2.1. Video Preprocessing

Spatial Resolution for Key Frames. We adopt Swin Transformer Tiny (Swin-T) [21] and Swin Transformer Base (Swin-B) [21] as backbone networks due to their strong feature representation capabilities. To ensure consistency and simplicity, we employ a single-branch structure and fix the temporal sampling rate to 1 fps—one key frame is extracted per second (8 frames per video for KVQ). The models are optimized using the PLCC loss. We vary the spatial resolution while keeping all other training strategies unchanged. Table 1 presents the experimental results of the baseline model under different spatial resolutions. As shown in Table 1, increasing the spatial resolution generally improves VQA performance; however, this gain comes at the cost of higher computational complexity. For instance, a resolution of 520×520 for Swin-T and resolutions larger than 320×320 for Swin-B exceed the predefined GFLOPs constraint. Considering the GFLOPs constraint, we identify resolutions of 320×320 and 384×384 as effective trade-offs between accuracy and efficiency.

Temporal Sampling Rate for Key Frames. Following the spatial resolution experiments, we further fix the spatial res-

Table 3. Performance of dual-branch network structures on the KVQ validation set

Backbone	Temporal Rate	SRCC	PLCC	GFLOPs
Input Resolution: 320				
SwinT+2D FAST-VQA	1/2	0.870	0.874	79.5
	1/4	0.856	0.857	39.8
SwinT+X3D	1/2	0.848	0.864	67.7
	1/4	0.830	0.844	33.9
Input Resolution: 384				
SwinT+2D FAST-VQA	1/2	0.875	0.877	115
	1/4	0.848	0.855	57.4
SwinT+X3D	1/2	0.841	0.858	82.7
	1/4	0.794	0.813	42.7

Table 4. Performance of the dual-branch network structure (with the distortion-aware branch) with different loss functions on the KVQ validation set

Loss	Resolution: 320		Resolution: 384	
	SRCC	PLCC	SRCC	PLCC
PLCC	0.870	0.874	0.875	0.877
L1	0.861	0.871	0.875	0.876
L2	0.870	0.875	0.874	0.878
Rank	0.870	0.874	0.875	0.877

Table 5. Performance of the dual-branch network structure (with the distortion-aware branch) with and without the knowledge distillation on the KVQ validation set. The spatial resolution and the temporal sampling rate are set as 384×384 and $1/2$ fps respectively

Knowledge Distillation	SRCC	PLCC
✓	0.875	0.877
×	0.908	0.907

olution to 320×320 and 384×384, and evaluate the impact of different temporal sampling rates. Table 2 summarizes the results. It is observed that reducing the sampling rate from 1 fps to $1/2$ fps yields negligible performance degradation, indicating that $1/2$ fps is sufficient for capturing spatial features. However, further reducing the rate to $1/4$ fps and $1/8$ fps leads to a significant performance drop. Based on this observation, we select $1/2$ fps as the optimal temporal sampling rate for subsequent experiments.

4.2.2. Quality-aware Feature Extraction

Table 1 and Table 2 have presented the performance of the single-branch network architecture.As discussed in Section 3.3, we further introduce two dual-branch network architectures. In each case, one branch retains the structure of the single-branch baseline. For the additional branch in the first dual-branch architecture, we employ the pretrained X3D network [6] to extract motion features. Specifically, we adopt the video sampling strategy proposed in MinimalisticVQA [] to select video chunks around each key frame for motion feature extraction. In the second dual-branch architecture, we introduce a distortion-aware branch based on a 2D variant of FAST-VQA, where fragment images are sampled from each key frame to extract distortion features.

The performance of the dual-branch architectures is reported in Table 3. We observe that incorporating the motion feature extraction module does not lead to a noticeable performance improvement. This finding is inconsistent with conclusions drawn in previous studies. We hypothesize two possible reasons for this discrepancy. First, the KVQ dataset may contain limited temporal distortions, reducing the importance of motion-related features. Second, X3D [6]—a lightweight action recognition network used in our experiments—may have insufficient sensitivity to subtle temporal artifacts compared to more powerful alternatives such as SlowFast [7]. In contrast, the distortion-aware branch yields a significant performance gain compared to the single-branch baseline, highlighting the effectiveness of incorporating distortion-level features. Therefore, we adopt the dual-branch network architecture in the subsequent experiments.

4.2.3. Optimization Techniques

Loss Function. The performance of different loss functions is summarized in Table 4. As shown in Table 4, the choice of loss function does not have a significant impact on overall model performance. Among the regression losses, L1 outperforms L2 at a spatial resolution of 384×384, whereas L2 performs better at 320×320. The rank loss yields slightly worse results compared to the regression losses, which may be attributed to the use of an approximate ranking formulation, potentially introducing bias during training. The PLCC loss achieves slightly better performance than the regression losses, possibly due to its ability to capture both ranking consistency and distributional alignment. In summary, the effect of different loss functions is relatively minor, and we adopt the PLCC loss as the optimization objective in our subsequent experiments.

Knowledge Distillation. We employ RQ-VQA [] as the teacher model to generate annotations for a large-scale set of unlabeled web videos. Specifically, we collect $40,000$ source videos from YouTube, TikTok, Bilibili, and YouKu. Additionally, $2,000$ of these videos are compressed using the H.264 codec at six different compression levels, resulting in $12,000$ compressed versions. In total, $52,000$ videos are pseudo-labeled using the RQ-VQA model. These

Table 6. Performance of the compared methods and the proposed E-VQA on the KVQ validation set, KVQ test set, LSVQ test set, and LSVQ test-1080p set. The best three results in each column are highlighted

VQA Methods		KVQ Validation		KVQ Test		LSVQ Test		LSVQ 1080p		GFLOPs
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	
Handcrafted Feature-based Methods	NIQE	0.239	0.24	0.272	0.281	0.442	0.332	0.489	0.459	-
	BRISQUE	0.472	0.480	0.489	0.493	0.579	0.576	0.497	0.531	-
	TLVQM	0.490	0.509	0.511	0.524	0.772	0.774	0.589	0.616	-
	VIDEAL	0.369	0.639	0.425	0.652	0.794	0.783	0.545	0.554	-
	RAPIQUE	0.803	0.801	0.815	0.818	0.827	0.819	0.655	0.701	-
DNN-based Methods	VSFA	0.830	0.834	0.843	0.840	0.801	0.796	0.675	0.704	2880
	SimpleVQA	0.874	0.875	**0.881**	**0.877**	0.867	0.861	0.764	0.803	**1,060**
	FAST-VQA	0.864	0.865	0.871	0.870	**0.876**	**0.877**	**0.779**	**0.814**	**279**
	Q-Align	**0.909**	**0.907**	0.664	0.693	**0.883**	**0.882**	**0.797**	**0.830**	12,000
	RQ-VQA	**0.914**	**0.918**	**0.926**	**0.924**	**0.894**	**0.894**	**0.809**	**0.842**	13,400
	Proposed	**0.908**	**0.907**	**0.924**	**0.925**	0.862	0.863	0.771	0.806	**114**

pseudo-labels are then used to pre-train our dual-branch model with the distortion-aware branch. Following pre-training, we fine-tune the model on the KVQ training set. The experimental results, shown in Table 5, demonstrate that knowledge distillation significantly improves model performance, highlighting its effectiveness.

4.3. Performance Comparison with State-of-the-art

Based on the findings in Section 4.2, we develop E-VQA, a dual-branch network architecture for video quality assessment. As illustrated in Figure 3, one branch takes as input the resized full image frame, while the other processes a cropped fragment image. Both inputs are resized to a resolution of 384×384. The model is initialized using a pretrained version obtained via knowledge distillation from RQ-VQA. The performance of E-VQA, along with other competing methods, is reported on the KVQ validation set, KVQ test set, LSVQ test set, and LSVQ test-1080p set in Table 6.

On the KVQ dataset, E-VQA outperforms most competing methods—except for RQ-VQA—on both the validation and test sets, while maintaining significantly lower computational complexity. On the LSVQ dataset, E-VQA achieves performance comparable to SimpleVQA, but falls behind FAST-VQA, Q-Align, and RQ-VQA. This may be attributed to the fact that LSVQ contains a larger proportion of motion-related distortions, which our method does not explicitly model. Furthermore, E-VQA ranked third in the NTIRE 2025 Short-form UGC Video Quality Assessment Challenge [16, 17], highlighting its practical effectiveness. Nevertheless, E-VQA remains far more computationally efficient than SimpleVQA, further demonstrating the effectiveness and efficiency of the proposed approach.

5. Conclusion

In this paper, we explore the design of a VQA model that is both efficient and effective. To this end, we decompose the VQA training pipeline into three components: video preprocessing, quality feature extraction, and optimization strategies. For each component, we empirically validate best practices derived from previous VQA studies. Building on these findings, we propose E-VQA, which achieves competitive performance on both the KVQ and LSVQ datasets, while maintaining significantly lower computational complexity compared to existing VQA models.

6. Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grants 62301316, 62377011, the Fundamental Research Funds for the Central Universities, the National Key R&D Program of China (2021YFE0206700), the Science and Technology Commission of Shanghai Municipality (2021SHZDZX0102), and the Shanghai Committee of Science and Technology (22DZ2229005).

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1

[2] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and

Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 3

[3] Francesca De Simone, Marco Tagliasacchi, Matteo Naccari, Stefano Tubaro, and Touradj Ebrahimi. A h. 264/avc video database for the evaluation of quality metrics. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2430–2433. IEEE, 2010. 2

[4] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1

[5] 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. In *International Conference on Learning Representations*, 2020. 1

[6] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020. 5, 7

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 7

[8] Yixuan Gao, Yuqin Cao, Tengchuan Kou, Wei Sun, Yunlong Dong, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. Vdpve: Vqa dataset for perceptual video enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1474–1483, 2023. 3

[9] Qihang Ge, Wei Sun, Yu Zhang, Yunhao Li, Zhongpeng Ji, Fengyu Sun, Shangling Jui, Xiongkuo Min, and Guangtao Zhai. Lmm-vqa: Advancing video quality assessment with large multimodal models. *arXiv preprint arXiv:2408.14008*, 2024. 4

[10] Deepti Ghadiyaram, Janice Pan, Alan C Bovik, Anush Krishna Moorthy, Prasanjit Panda, and Kai-Chieh Yang. In-capture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9):2061–2077, 2017. 2, 3

[11] 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. 2

[12] Ziheng Jia, Zicheng Zhang, Jiaying Qian, Haoning Wu, Wei Sun, Chunyi Li, Xiaohong Liu, Weisi Lin, Guangtao Zhai, and Xiongkuo Min. Vqa ̃2: Visual question answering for video quality assessment. *arXiv preprint arXiv:2411.03795*, 2024. 1

[13] Jari Korhonen. Two-level approach for no-reference consumer video quality assessment. *IEEE Transactions on Image Processing*, 28(12):5923–5938, 2019. 1, 2, 3, 6

[14] Bowen Li, Weixia Zhang, Meng Tian, Guangtao Zhai, and Xianpei Wang. Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):5944–5958, 2022. 1, 3, 5

[15] 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. 1, 2, 3, 5, 6

[16] Xin Li, Xijun Wang, Bingchen Li, Kun Yuan, Yizhen Shao, Suhang Yao, Ming Sun, Chao Zhou, Radu Timofte, and Zhibo Chen. NTIRE 2025 challenge on short-form ugc video quality assessment and enhancement: Kwaisr dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 8

[17] Xin Li, Kun Yuan, Bingchen Li, Fengbin Guan, Yizhen Shao, Zihao Yu, Xijun Wang, Yiting Lu, Wei Luo, Suhang Yao, Ming Sun, Chao Zhou, Zhibo Chen, Radu Timofte, et al. NTIRE 2025 challenge on short-form ugc video quality assessment and enhancement: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025. 8

[18] Yang Li, Shengbin Meng, Xinfeng Zhang, Meng Wang, Shiqi Wang, Yue Wang, and Siwei Ma. User-generated video quality assessment: A subjective and objective study. *IEEE Transactions on Multimedia*, 25:154–166, 2021. 1, 3

[19] Zhi Li, Christos Bampis, Julie Novak, Anne Aaron, Kyle Swanson, Anush Moorthy, and JD Cock. Vmaf: The journey continues. *Netflix Technology Blog*, 25, 2018. 1

[20] Wentao Liu, Zhengfang Duanmu, and Zhou Wang. End-to-end blind quality assessment of compressed videos using deep neural networks. In *ACM Multimedia*, pages 546–554, 2018. 1

[21] 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. 1, 5, 6

[22] Wei Lu, Wei Sun, Xiongkuo Min, Wenhan Zhu, Quan Zhou, Jun He, Qiyuan Wang, Zicheng Zhang, Tao Wang, and Guangtao Zhai. Deep neural network for blind visual quality assessment of 4k content. *arXiv preprint arXiv:2206.04363*, 2022. 1

[23] Wei Lu, Wei Sun, Zicheng Zhang, Danyang Tu, Xiongkuo Min, and Guangtao Zhai. Bh-vqa: Blind high frame rate video quality assessment. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2501–2506. IEEE, 2023. 1

[24] Yiting Lu, Xin Li, Yajing Pei, Kun Yuan, Qizhi Xie, Yunpeng Qu, Ming Sun, Chao Zhou, and Zhibo Chen. Kvq: Kaleidoscope video quality assessment for short-form videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3, 5, 6

[25] 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. 1

[26] Zhanxuan Mei, Yun-Cheng Wang, and C-C Jay Kuo. Blind

video quality assessment at the edge. *IEEE Transactions on Multimedia*, 2024. 4

[27] Xiongkuo Min, Huiyu Duan, Wei Sun, Yucheng Zhu, and Guangtao Zhai. Perceptual video quality assessment: A survey. *arXiv preprint arXiv:2402.03413*, 2024. 1

[28] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 3, 6

[29] Anish Mittal, Michele A Saad, and Alan C Bovik. A completely blind video integrity oracle. *IEEE Transactions on Image Processing*, 25(1):289–300, 2015. 1, 2, 3

[30] Anush Krishna Moorthy and Alan Conrad Bovik. Efficient video quality assessment along temporal trajectories. *IEEE transactions on circuits and systems for video technology*, 20 (11):1653–1658, 2010. 1

[31] 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

[32] Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind prediction of natural video quality. *IEEE Transactions on Image Processing*, 23(3):1352–1365, 2014. 1, 2, 3

[33] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack. Study of subjective and objective quality assessment of video. *IEEE transactions on Image Processing*, 19(6):1427–1441, 2010. 2

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

[35] Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2):612–627, 2018. 2, 3

[36] Wei Sun, Xiongkuo Min, Guangtao Zhai, Ke Gu, Huiyu Duan, and Siwei Ma. Mc360iqa: A multi-channel cnn for blind 360-degree image quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 14(1):64–77, 2019. 1

[37] 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

[38] 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, 3, 5, 6

[39] Wei Sun, Xiongkuo Min, Danyang Tu, Siwei Ma, and Guangtao Zhai. Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training. *IEEE Journal of Selected Topics in Signal Processing*, 17(6):1178–1192, 2023. 1

[40] Wei Sun, Wen Wen, Xiongkuo Min, Long Lan, Guangtao Zhai, and Kede Ma. Analysis of video quality datasets via design of minimalistic video quality models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2, 4, 5

[41] Wei Sun, Haoning Wu, Zicheng Zhang, Jun Jia, Zhichao Zhang, Linhan Cao, Qiubo Chen, Xiongkuo Min, Weisi Lin, and Guangtao Zhai. Enhancing blind video quality assessment with rich quality-aware features. *arXiv preprint arXiv:2405.08745*, 2024. 2, 4, 6

[42] Wei Sun, Weixia Zhang, Yuqin Cao, Linhan Cao, Jun Jia, Zijian Chen, Zicheng Zhang, Xiongkuo Min, and Guangtao Zhai. Assessing uhd image quality from aesthetics, distortions, and saliency. *arXiv preprint arXiv:2409.00749*, 2024. 1

[43] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 3

[44] 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. 1, 3, 6

[45] 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. 1, 3, 6

[46] VQEG. Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, 2000. 6

[47] Phong V Vu and Damon M Chandler. Vis 3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. *Journal of Electronic Imaging*, 23(1): 013016–013016, 2014. 2

[48] 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. 2, 3

[49] 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*, pages 13435–13444, 2021. 1, 3

[50] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

[51] 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. 1, 2, 4, 5

[52] 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, 6

[53] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang,

Erli Zhang, Wenxiu Sun, Yan Qiong, Min Xiongkuo, Zhai Guangtao, and Lin Weisi. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 4, 6

[54] Jiahua Xu, Jing Li, Xingguang Zhou, Wei Zhou, Baichao Wang, and Zhibo Chen. Perceptual quality assessment of internet videos. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1248–1257, 2021. 1

[55] Fuwang Yi, Mianyi Chen, Wei Sun, Xiongkuo Min, Yuan Tian, and Guangtao Zhai. Attention based network for no-reference ugc video quality assessment. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1414–1418. IEEE, 2021. 2, 3

[56] 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. 1, 2, 3, 5

[57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1

[58] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, 2021. 3

[59] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14071–14081, 2023. 4

[60] Yiming Zhang, Xiongkuo Min, and Guangtao Zhai. Split-conv: a resource-efficient compression method for image quality assessment models. In *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE, 2023. 4

[61] 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*, pages 1746–1755, 2023. 3