KVQ: Kwai Video Quality Assessment for Short-form Videos
Supplementary Material

Yiting Lu[1†], Xin Li[1†], Yajing Pei[1,2†], Kun Yuan[2(✉)],
Qizhi Xie[2,3], Yunpeng Qu[2,3], Ming Sun[2], Chao Zhou[2], Zhibo Chen[1(✉)]
[1]University of Science and Technology of China, [2]Kuaishou Technology, [3]Tsinghua University
{luyt31415,lixin666,peiyj}@mail.ustc.edu.cn, chenzhibo@ustc.edu.cn
{yuankun03,xieqizhi,quyunpeng,sunming03,zhouchao}@kuaishou.com

Section 1 clarifies the details of our practical and sophisticated video processing workflows. And Section 2 provides the analysis of the six feature distributions in our KVQ database. Section 3 includes the details about the test setup and data cleaning process. Section 4 compasses the details about region selection in the Quality-aware Region Selection (QRS) module and implementation details of our proposed KSVQE. In section 5, subsection 5.1, subsection 5.2 and subsection 5.3 provide more ablation studies for the QRS, Content-adaptive Modulation (CaM) and Distortion-aware modulation (DaM), respectively.

1. Details of Our Video Processing Workflows

In contrast to previous UGC databases [6, 19], which primarily focus on simulated compression artifacts, our proposed KVQ database is significantly different since its processing workflows are consistent with the practical applied workflows in the typical short-form video platform. Our processing workflow is composed of three cascaded parts, including video enhancement module $\phi_e(\cdot)$, pre-processing module $\phi_p(\cdot)$, and transcoding module $\phi_t(\cdot)$.

Video Enhancement Module $\phi_e(\cdot)$ is composed of three commonly used enhancement algorithms in short-form video platforms: De-Blur, De-Noise, and De-Artifact algorithm, where De-Blur aims to enhance the texture details of videos, and De-Noise is utilized to remove the structure/non-structure noises that are harmful to human perception. The De-Artifact algorithm is exploited to reduce other-form degradations, such as block artifacts.

Video Pre-processing Algorithms $\phi_p(\cdot)$ aims to reduce the high-frequency components that do not affect the human perception (*e.g.*, the non-ROI region) or the high-frequency distortions, such as noises. In this way, it can reduce the compression and transmission costs while preserving/improving the subjective quality of short-form videos. We select two pre-processing algorithms: global level pre-processing and region-of-interest (ROI) level pre-processing. The former aims to remove high-frequency information related to global-level impairment, while the latter focuses on eliminating high-frequency information associated with local-level impairment.

Video Transcoding Algorithms $\phi_t(\cdot)$ The quantization parameters (QP) are the crucial parameters used to adjust the compression ratios, where higher QP corresponds to a higher compression ratio and lower visual quality. However, it is costly and labor-intensive to compress each video traversing each QP value (*i.e.*, from 0-51). To mitigate this and ensure the diversity of QP values, we divided the commonly-used QP range (*i.e.*, 16-47) into six intervals, encompassing 16-23, 24-31, 32-35, 36-39, 40-43, and 44-47, and then randomly select one QP from each interval for the compression of each video.

To demonstrate the effects of different processing workflows, we provide some examples for our three typical processing workflows in Fig. 2 of our manuscript, *i.e.*, $\phi_t(\cdot)$, $\phi_t(\phi_e(\cdot))$ and $\phi_t(\phi_p(\phi_e(\cdot)))$, Concretely, the example for $\phi_t(\cdot)$ is shown in Fig. 1. The example for $\phi_t(\phi_e(\cdot))$ is shown in Fig. 2, and the example for $\phi_t(\phi_p(\phi_e(\cdot)))$ is shown in Fig. 3. The distorted patches are indicated by red boxes. Therefore, the KVQ dataset we established not only encompasses rich content within short video scenes but also spans more intricate video processing workflows, as illustrated in the comparisons across various UGC datasets in Table 1.

2. Feature Analysis

In summary, our KVQ database exhibits diverse feature characteristics across six video quality-related features, namely sharpness, blocky, blurriness, colorfulness, complexity, and noise. The distribution analysis, illustrated

† Equal contribution. ✉ Corresponding authors.

Table 1. Comparison of various dimensions among different UGC datasets.

UGC database	Video Sources	Num Ref/Dis	Distortion Type	Subjective Form
CVD2014 [9]	Captured	-/234	authentic	MOS
LIVE-VQC [11]	Captured	-/585	authentic	MOS
KoNViD-1k [3]	Flicker	-/1200	authentic (UGC)	MOS
YouTube-UGC [15]	YouTube	-/1380	authentic (UGC)	MOS
Youku-V1K [16]	Youku	-/1072	authentic (UGC)	MOS
LSVQ [17]	IA,Flicker	-/39075	authentic (UGC)	MOS
UGC-VIDEO [6]	TikTok	50/550	authentic+compression	MOS
TaoLive [19]	Taobao	418/3762	authentic+compression	MOS
KVQ	Short-form video platform	600/3600	authentic+enhancement+pre-processing+compression	MOS+Rank

Table 2. Annotation criteria for subjective labeling scores from 1 to 5.

Score	Annotation criteria
1 Bad	The Visual information within video content becomes challenging or impossible to distinguish.
2 Poor	The primary video content remains distinguishable but exhibits pronounced noise, block artifacts, and blurriness, along with substantial jitter and lag.
3 Fair	The primary video content is reasonably clear, but it includes noticeable distortions such as conspicuous noise, visual blurring, minor localized glare, or distinct edge sharpening. Additionally, the video exhibits a markedly blurry background texture.
4 Good	The videos feature a clear primary subject, free from substantial noise or visual blurring, and devoid of apparent distortions such as jitter or glare. However, they exhibit limited overall textural complexity.
5 Excellent	The primary video object is characterized by exceptional clarity, devoid of noise, block artifacts, blurriness, jitter, glare, or lag. It presents a high-quality spectacle distinguished by lucid textural elements.

Figure 1. Examples of transcoding.

in Fig. 4, highlights that the majority of features span a wide range, showcasing the feature diversity inherent in our database. Notably, blocky features and colorfulness features are more skewed towards the right, indicating a substantial presence of computer graphics, portraits, rich special effects, and common compression distortions, particularly on short-form video platforms. While complexity distribution and noise distribution skew towards lower values, the other features maintain closer adherence to middle values, with less pronounced spikes, providing an approximated overview of the distinctive feature characteristics on the typical short-form video platform.

3. More Detail About Human Study

3.1. Test Setup

The subjective experimentation involves a group of 15 observers, each tasked with assessing 4,200 videos. Uniform MAC devices are employed by the observers to ensure standardized screen brightness and video resolution support. During the scoring process, a consistent stimulus evaluation method is utilized, allowing for repeated viewing of the same video, ultimately leading to the evaluation of all videos. Continuous scales with intermediate numerical labels (ranging from 1 to 5 with a step size of 0.5) are employed for scoring.

Given the distinctive characteristics of the KVQ database, we establish the following instructions:

- Scoring of special effects is lenient, disregarding the impact of special effects on perceptual quality, such as variations in lighting and shadows.
- In cases where a video is composed of multiple segments of merged content and significant quality differences, the total video score is computed based on the proportional duration of each segment.
- For videos containing text, the evaluation takes into account both text and background distortions, simultaneously determining whether the primary focus of the video is on text or background to derive the final quality score.

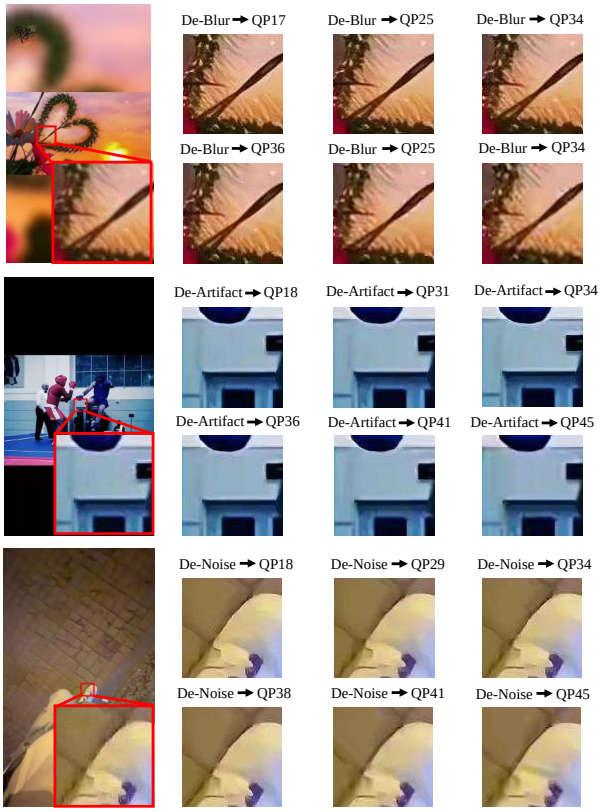


Figure 2. Examples of enhancement→transcode.

- For three-stage videos, only the content of the middle region is considered, with no regard for the influence of unrelated content in the upper and lower regions.

The guidelines for scoring, outlining the rules for assigning scores ranging from 1 to 5, are given in Table 2.

3.2. Data Clean

Fig. 5 depicts the pipeline of the data cleaning. First, it is necessary to ascertain the reliability of each observer's scores. After scoring overall videos in KVQ database, the correlation metrics (*i.e.*, SROCC, and PLCC) between an observer's scores and the average scores of other observers are computed. If the correlation falls below 0.7, retraining will be conducted for these observers.

Secondly, in accordance with ITU-R BT 500.13, further data processing is performed to screen observers for each video. Specifically, for each video, we compute the kurtosis of the scores to assess whether the ratings exhibit a normal distribution. Subsequently, based on the distribution of ratings, we calculate the quality score range for each video as 2 times the standard deviation from ratings or $\sqrt{20}$ times the standard deviation of the ratings. Based on this, We can determine the number of videos rated out of this range by each observer. For the i-th observer:

For all j in J:

If $u_{ij} \geq \overline{u}_j + \alpha * S_j$, then $P_i = P_i + 1$.

If $u_{ij} \leq \overline{u}_j + \alpha * S_j$, then $Q_i = Q_i + 1$.

If $\left|\frac{P_i - Q_i}{P_i + Q_i}\right| < 0.3$ and $\frac{P_i + Q_i}{J} > 0.05$, the annotation made by the i-th observer will be rejected.

Here, P_i denotes the number of videos that an observer has rated above the range, Q_i represents the number rated below the range, and J signifies the total number of videos rated by the observer, α can be 2 or $\sqrt{20}$, S represents the standard deviation of each video. Following this step, we ascertain that all observers are reliable.

Thirdly, for each video in the KVQ dataset, it is imperative to establish a corresponding confidence interval for all ratings. This interval relies on the standard deviation and mean quality score of each video. We opt for a 95% confidence interval, derived from:

$$S_j = \sqrt{\sum_{i=1}^{N} \frac{(u_{ij} - \overline{u}_j)^2}{(N-1)}} \tag{1}$$

which yields the standard deviation. Subsequently, we calculate the range of the 95% confidence interval as:

$$(\overline{u}_j - \delta_j, \overline{u}_j + \delta_j) \tag{2}$$

with $\delta_j = 1.96\frac{S_j}{\sqrt{N}}$, where $\overline{u}_j$ signifies the average rating for the j^{th} video, N is the number of observers that participate in the labeling of the j^{th} video, S_j represents the standard deviation of the j^{th} video. The ratings falling outside the 95% confidence interval range are then removed.

4. Details of Our Framework KSVQE

4.1. QRS Details

Spatial Region Selection The process involves selecting the most important fragments based on a quality-aware semantic importance score $I \in \mathbb{R}^N$, which contains two key points: i) how to make the selection operation differentiable, ii) how to preserve the spatial dependency within selected fragments. To preserve the problem of spatial dependency, following work [13], we select the most quality-aware fragments through an aggregate-then-select strategy to simulate the Top-k selection of fragments with a correct spatial dependency. The aggregation operation involves splitting the reshaped score map $I \in \mathbb{R}^{\sqrt{N} \times \sqrt{N}}$ into a list of non-overlapping smaller score maps. Then average pooling is applied to each of these smaller score maps and obtains the patch importance vector $\hat{I}_r$. After the importance score aggregation operation, we apply the TopK operator to obtain the most quality-aware patches $\hat{X}$. We denote the TopK operator [2] as :

$$inds = \mathrm{TopK}(\hat{I}_r) \tag{3}$$

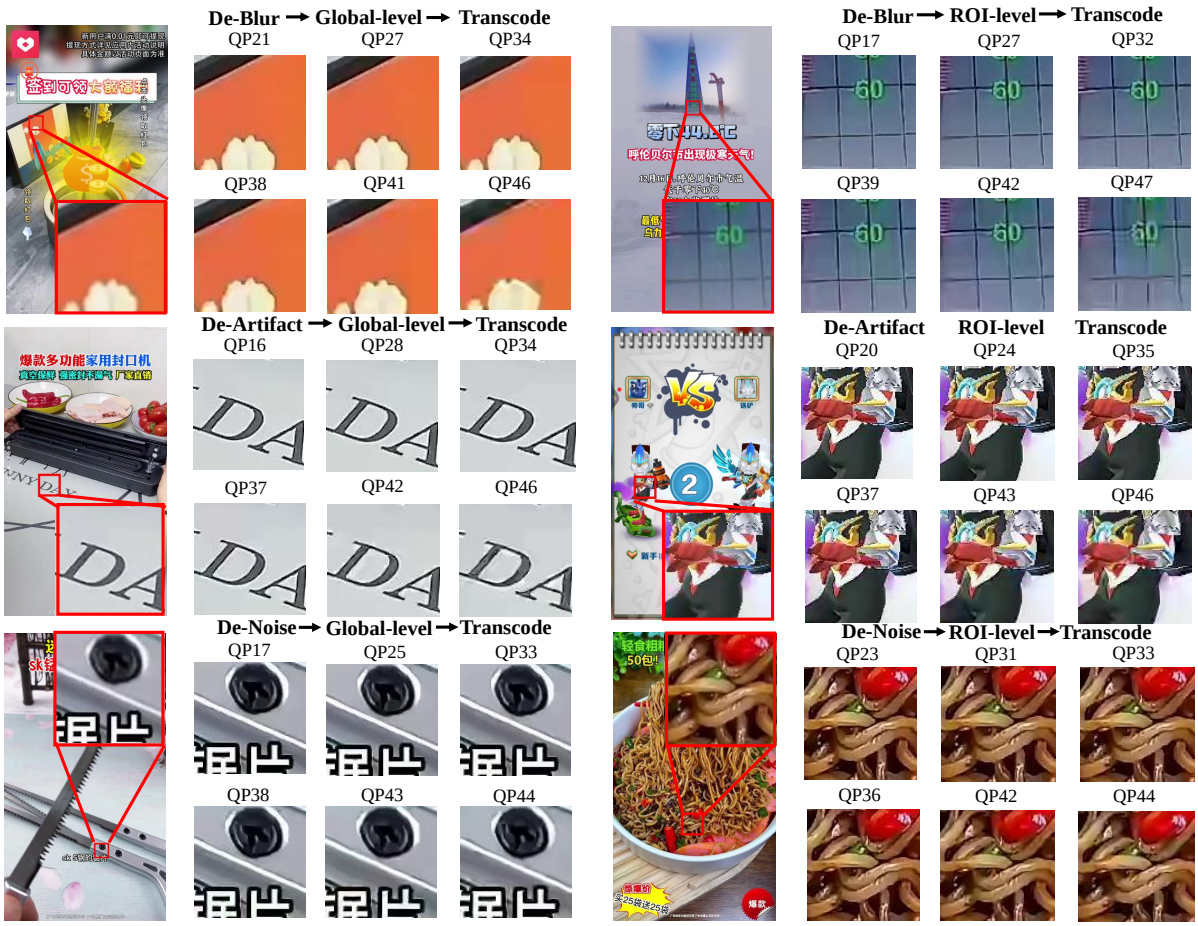


Figure 3. Examples of enhancement→pre-processing→transcode.

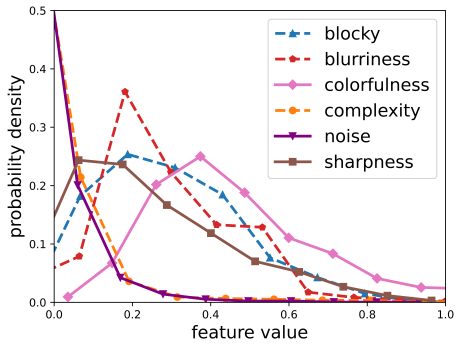


Figure 4. The feature distributions on the KVQ dataset.

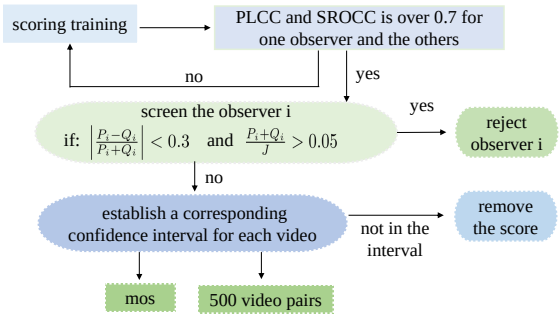


Figure 5. The overall data clean workflow. First, we ensure that the annotator achieves a correlation of 0.7, and then we screen the annotator via ITU-R BT 500.13 to confirm reliability. Finally, for each video, we set a corresponding confidence interval, scores that are outside this range will be removed.

However, the *inds* from the TopK operation are non-differentiable.

Inspired by the perturbed maximum method [1], the differentiable TopK can be realized through the solution for inputs with perturbation. The differentiable TopK operation shares a fundamental similarity with the Gumbel Softmax operation [4, 5, 8]. Specially, we sample uniform Gaussian noise Z and add it to input $\hat{I}_r$, then we can obtain the per-

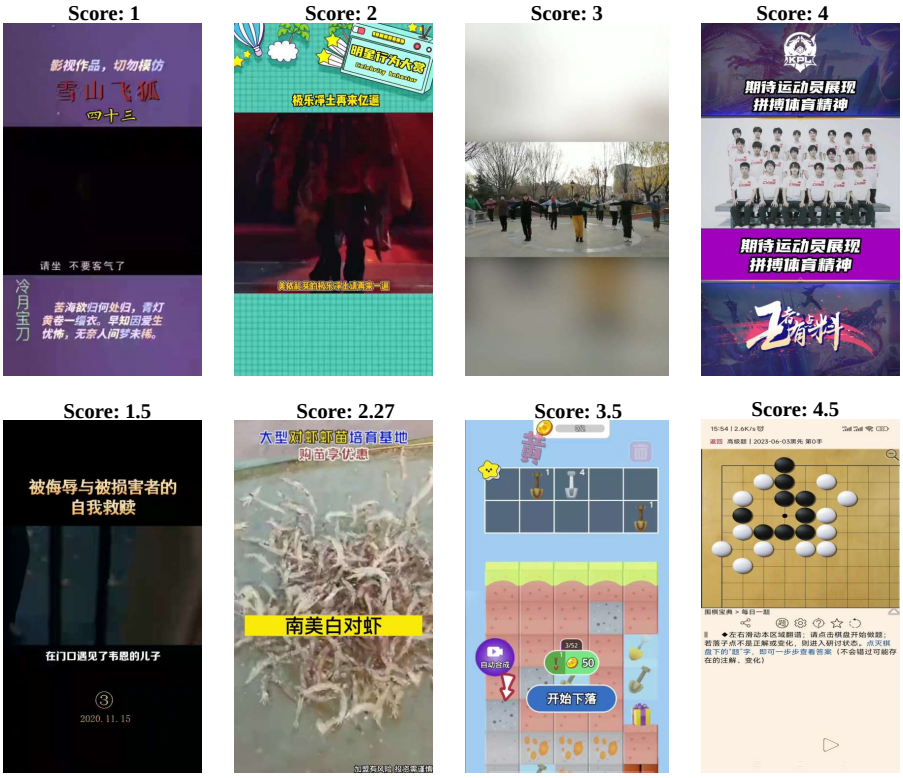


Figure 6. The video examples of quality score ranged from 1 to 5.

turbed maximizer:

$$Y_{inds,\sigma} = \underset{Y_{inds} \in \mathcal{C}}{\arg\max} \langle Y_{inds}, \hat{I}_r + \sigma Z \rangle \tag{4}$$

Where Y_{inds} is the one-hot vector of indices $inds$, σ is a hyper-parameter to control the level of added noise. And we fix the $\sigma = 0.5$ in our all experiments. And $\mathcal{C}$ is the convex polytope constrain set. For backward, the gradient can be passed from variable parameter Z to optimization variable $\hat{I}_r$.

4.2. Implementation Details

The semantic adapter f and the distortion adapter f_d each consist of several fully-connected (FC) layers with dimensional variations of '768-192-768' and '128-32-768', respectively. The Multi-Head Cross Attention for both semantic modulation and distortion modulation has a head number of 8 and a dimension of 768. The Multi-Head Self Attention in distortion modulation has a head number of 8 and a dimension of 768. The modulation parameter generator l_{ss} and l_{so} for semantic modulation are the convolutions with the kernel size of 1×1 and dimension variation of "768-1". And the modulation parameter generator l_{ds} and l_{do} for distortion modulation are the FC layer with the dimension variation of "768-768".

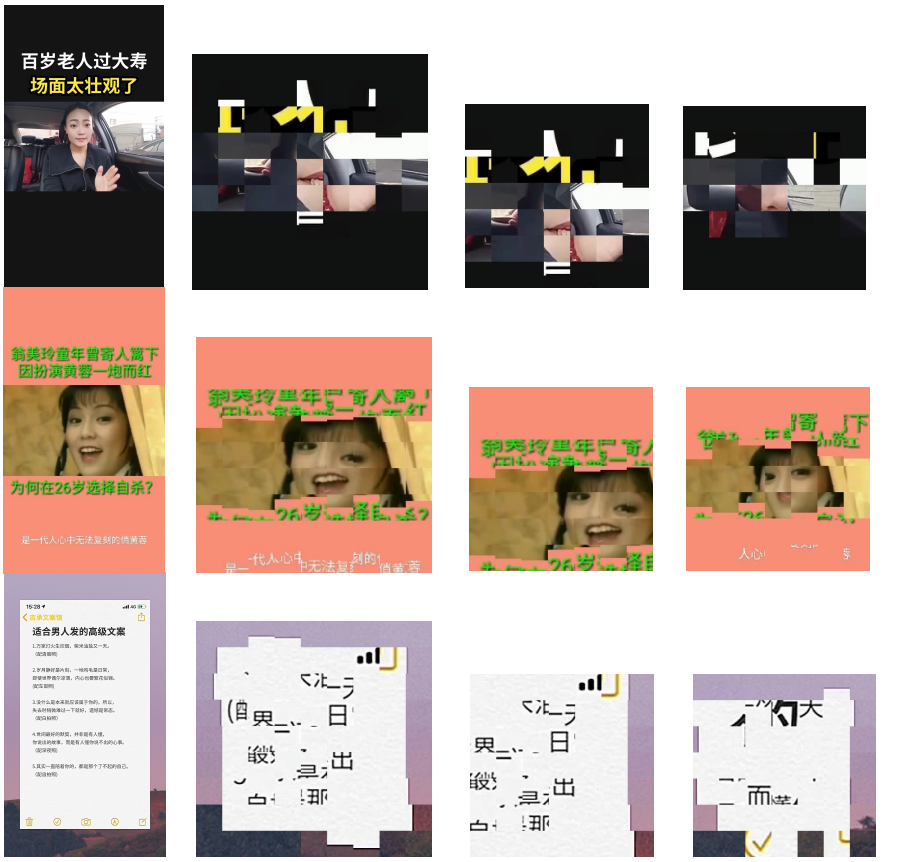
For N_t key frames in the input of the CLIP visual encoder, we partition the videos into segments and select a single frame from each segment to encapsulate the comprehensive semantic information of that segment. Subsequently, utilizing the acquired quality-aware importance vector and visual tokens for modulation guidance, we extend the temporal dimension of N_t to T.

In the training process, we utilize AdamW optimizer [7] with a learning rate of $3 \times e^{-5}$ and a weight decay of 0.05 for optimization. And batchsize set as 8.

5. More Experiment Results

5.1. QRS

More Variants About QRS In order to investigate the optimal original number of fragments for region selection in QRS, we compare multiple numbers of original fragments in Table 3. Notably, we observe that extracting 7×7 fragments from 9×9 input fragments for the 3D Swin Transformer yielded the most optimal performance. In the context of region selection, selecting an excessive number of original fragments results in the retention of excessive redundant information. Conversely, opting for too few numbers of original fragments leads to the absence of crucial,



(a) Original Frame **(b) Original 9×9 Fragments** **(c) Selected 7×7 Fragments** **(d) Original 7×7 Fragments**

Figure 7. The visualization of original fragments of 7×7, 9×9 and the selected fragments of 7×7.

Table 3. Ablation study for the number of fragments in QRS.

Region Selection	KVQ		KoNViD-1k		YouTube-UGC	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
$8 \times 8 \rightarrow 7 \times 7$	0.841	0.841	**0.918**	0.917	0.887	0.903
$9 \times 9 \rightarrow 7 \times 7$	**0.847**	**0.853**	0.917	**0.920**	**0.894**	**0.906**
$10 \times 10 \rightarrow 7 \times 7$	0.847	0.848	0.911	0.914	0.892	0.907

quality-aware visual information. The selection of 9×9 for original fragments strikes a balance, demonstrating superior performance by capturing essential visual features without succumbing to information redundancy or loss of significance.

Visualization About QRS Also, we visualize the selected fragments and original fragments in Fig. 7. And we can see that for the first row of three-stage video, our Quality Region Selection (QRS) method excels in extracting concentrated regions within the central regions. In videos characterized by extensive monochromatic backgrounds (*i.e.*, the second row), QRS is capable of capturing visually enriched regions for the face and text. Additionally, when dealing with videos incorporating much text (*i.e.*, the third row), QRS focuses mainly on the text area rather than the background, which is consistent with human attention.

5.2. CaM

The effectiveness of adapter-style training To verify the effectiveness of our adapter on cls token, we conduct an experiment by removing it and comparing the results with those obtained using the full modules for content understanding, as illustrated in Table 4. The results show that adding adapter-style training can bring a performance gain of 0.007/0.004 on SROCC and PLCC on the KVQ database. It illustrates that the feature adaptation to quality-aware space is necessary to incorporate content understanding and extract the quality-aware semantics to provide guidance.

5.3. DaM

More Variants About DaM Also, we analyze the effectiveness of our distortion adapter on DaM in Table 6. We can see that adapter-style training demonstrates an improve-

Table 4. Ablation study for adapter-style training in CaM.

Content Adapter	KVQ		KoNViD-1k		YouTube-UGC	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
QRS+CaM (w.o. adapter)	0.841	0.850	0.913	0.914	0.882	0.893
QRS+CaM	**0.848**	**0.854**	**0.918**	**0.922**	**0.895**	**0.901**

Table 5. Ablation study for multiple variants of selection in DaM.

DaM	KVQ		KoNViD-1k		YouTube-UGC	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
CA+CM	0.832	0.834	0.911	0.912	0.888	0.899
CASA+CM (DaM)	**0.839**	**0.843**	**0.915**	**0.914**	**0.893**	**0.910**

Table 6. Ablation study for adapter-style training in DaM.

Distortion Adapter	KVQ	
	SROCC	PLCC
DaM (w.o. adapter)	0.831	0.831
DaM	**0.839**	**0.843**

ment in the performance of 0.008/0.012 in terms of SROCC and PLCC on KVQ database. It reveals the significance of adapting knowledge from CONTRIQUE to distortion distribution in KVQ database.

For more variants for distortion modulation in DaM, we remove the multi-head self-attention as the variant "CA+CM" and compare it with our DaM (*i.e.*, CASA+CM) in Table 5. The results show that DaM benefits from the influence exerted by self-attention for temporal distortion extraction, resulting in a gain of 0.007/0.009 in terms of SROCC and PLCC on KVQ database.

5.4. The Combination of Content-Distortion Understanding

We also investigate another method to incorporate content prior and distortion prior into the original feature. We compared our proposed modulation method with the simplest fusion approach, concatenation, and the results are presented in Table 7. The results indicate that our modulation method is more effective in explicitly modeling the understanding of content and distortion.

5.5. The selection of content extractor and distortion extractor.

To verify the effectiveness of enhanced CLIP (with adapter-style training) for quality-ware content mining, we replace CLIP with CLIPIQA+ [14] and LIQE [18] in Table 8. From the results, our KSVQE with enhanced CLIP can obtain the optimal correlation performance on KVQ, which shows the ability of the enhanced CLIP to capture quality-aware content. The visualization results can be seen in Appendix. As for the effectiveness of CONTRIQUE for distortion identification in KVQ, we choose GraphIQA [12] or ReIQA [10] to substitute CONTRIQUE in KSVQE, which is shown in Table 9. From these results, we can see that CONTRIQUE with distortion-aware contrastive learning can be adapted well to distortion space in KVQ.

Table 7. Ablation study for multiple variants of combination.

DaM	KVQ		KoNViD-1k		YouTube-UGC	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
QRS+concat	0.853	0.856	0.912	0.914	0.891	0.894
QRS+CaM+DaM	**0.867**	**0.869**	**0.922**	**0.921**	**0.900**	**0.912**

Table 8. Different selection for content extractor in KSVQE, in which "XXX/XXX" represent "SROCC/PLCC".

Model	KVQ	KoNViD	YT-UGC
CLIPIQA+	0.862/0.856	0.916/0.915	0.888/0.899
LIQE	0.848/0.854	0.920/0.917	0.892/0.896
KSVQE	**0.867/0.869**	**0.922/0.921**	**0.900/0.912**

Table 9. Different selection for distortion extractor in KSVQE, in which "XXX/XXX" represent "SROCC/PLCC".

Model	KVQ	KoNViD	YT-UGC
ReIQA	0.858/0.851	0.921/0.921	0.892/0.891
GraphIQA	0.849/0.850	0.916/0.915	0.888/0.881
KSVQE	**0.867/.869**	**0.922/0.921**	**0.900/0.912**

References

[1] Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis R. Bach. Learning with differentiable pertubed optimizers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 4

[2] Jean-Baptiste Cordonnier, Aravindh Mahendran, Alexey Dosovitskiy, Dirk Weissenborn, Jakob Uszkoreit, and Thomas Unterthiner. Differentiable patch selection for image recognition. In *CVPR*, pages 2351–2360. Computer Vision Foundation / IEEE, 2021. 3

[3] 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 *QoMEX*, pages 1–6. IEEE, 2017. 2

[4] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 4

[5] Xin Li, Yiting Lu, and Zhibo Chen. Freqalign: Excavating perception-oriented transferability for blind image quality assessment from a frequency perspective. *IEEE Transactions on Multimedia*, 2023. 4

[6] Yang Li, Shengbin Meng, Xinfeng Zhang, Shiqi Wang, Yue Wang, and Siwei Ma. UGC-VIDEO: perceptual quality assessment of user-generated videos. In *3rd IEEE Conference on Multimedia Information Processing and Retrieval, MIPR 2020, Shenzhen, China, August 6-8, 2020*, pages 35–38. IEEE, 2020. 1, 2

[7] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 5

[8] Yiting Lu, Jun Fu, Xin Li, Wei Zhou, Sen Liu, Xinxin Zhang, Wei Wu, Congfu Jia, Ying Liu, and Zhibo Chen. Rtn: Reinforced transformer network for coronary ct angiography vessel-level image quality assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 644–653. Springer, 2022. 4

[9] 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 Trans. Image Process.*, 25(7):3073–3086, 2016. 2

[10] Avinab Saha, Sandeep Mishra, and Alan C. Bovik. Re-iqa: Unsupervised learning for image quality assessment in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 5846–5855. IEEE, 2023. 7

[11] Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Trans. Image Process.*, 28(2): 612–627, 2019. 2

[12] Simeng Sun, Tao Yu, Jiahua Xu, Wei Zhou, and Zhibo Chen. Graphiqa: Learning distortion graph representations for blind image quality assessment. *IEEE Transactions on Multimedia*, 2022. 7

[13] Junke Wang, Xitong Yang, Hengduo Li, Li Liu, Zuxuan Wu, and Yu-Gang Jiang. Efficient video transformers with spatial-temporal token selection. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXV*, pages 69–86. Springer, 2022. 3

[14] 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. 7

[15] Yilin Wang, Sasi Inguva, and Balu Adsumilli. Youtube UGC dataset for video compression research. In *MMSP*, pages 1–5. IEEE, 2019. 2

[16] Jiahua Xu, Jing Li, Xingguang Zhou, Wei Zhou, Baichao Wang, and Zhibo Chen. Perceptual quality assessment of internet videos. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 1248–1257. ACM, 2021. 2

[17] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan C. Bovik. Patch-vq: 'patching up' the video quality problem. In *CVPR*, pages 14019–14029. Computer Vision Foundation / IEEE, 2021. 2

[18] 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. 7

[19] 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 *CVPR*, pages 1746–1755. IEEE, 2023. 1, 2