KVQ: Boosting Video Quality Assessment via Saliency-guided Local Perception —— Supplementary Material ——

Yunpeng Qu[1,2], Kun Yuan[2,†(✉)], Qizhi Xie[1,2], Ming Sun[2], Chao Zhou[2], Jian Wang[1,3(✉)]
[1] Tsinghua University, [2]Kuaishou Technology, [3] BNRist, Tsinghua University
{qyp21, xqz20}@mail.tsinghua.edu.cn, jian-wang@tsinghua.edu.cn
{yuankun03,sunming03,zhouchao}@kuaishou.com

Table 1. Content Scene diversity of LPVQ dataset.

Scene	Portrait	Landscape	Animation	Caption	Stage	Food	Crowd	Person
Num	5	10	5	3	10	3	4	6

Table 2. Qualtity factors diversity of LPVQ dataset.

Factor	Motion	Shake	Noise	Blur	Aircrafts	Compression	Bokeh	Exposure
Num	8	5	4	5	4	4	7	4

1. Details of KVQ

1.1. Architecture of the Designed Backbone

Utilizing the FWA as a component, we concatenate it with an FFN layer to obtain the fundamental block in the backbone, referred to as the Fusion Window Block. As illustrated in Fig. 3 of our paper, our backbone network follows the advanced vision transformers [6, 9] and employs a four-stage pyramid architecture. Since the preceding stages in the pyramid structure extract localized high-frequency features, we employ the basic modules of Swin-T to construct the first two stages. The latter two stages are constructed using Fusion Window Blocks to capture high-level semantic information and extract global visual saliency.

1.2. Implementation Details.

When applying the LPC, we divide the video into several spatiotemporal cubes of size of $[2, 28, 28]$. We feed these cubes separately into the network and calculate the loss function based on the results obtained from reassembling the cube outputs. k is set to 8 in the third stage and 4 in the fourth stage. In all our experiments, the hyper-parameters of λ_r and λ_p are set to 0.5 and 0.05. We use the Video Swin Transformer [7] pretrained on Kinetics-400 [3] as the backbone before training on VQA tasks. During training, we use an AdamW optimizer with a weight decay 5e-2 for 30 epochs. The learning rate is set to 5e-4. All experiments are conducted on 8 NVIDIA V100 GPUs.

2. Details of LPVQ Dataset

Why choose images for annotations? As stated in **Introduction** of our paper, it is a time-consuming and complicated task to annotate MOS for videos, as it requires a large number of participants to ensure reliability [1]. Indeed, annotating the local quality of spatiotemporal regions in videos incurs even more challenging costs, thereby escalating the annotation expenses by approximately $\mathcal{O}(N^3)$. This renders the acquisition of extensive local quality annotations nearly impracticable.

Considering the substantial cost of annotating videos, we build a dataset using images as static videos for annotating local quality. 2D images reflect spatial-level local perception ability. To maximize representativeness, all images in LPVQ are meticulously selected video screenshots to cover both temporal and spatial distortions. Tab. 1 and 2 show the categories of scenes and low-quality factors in LPVQ, along with the corresponding number of images. LPVQ covers a wide range of scenes and low-quality factors to ensure representativeness and the selected coverage of temporal distortions (e.g., motion blur, shake) making it suitable for evaluating videos.

Content Scope. LPVQ comprises a total of 50 images collected from a typical short-form video platform, where each image exhibits noticeable variations in quality across different regions. As shown in Fig. 1, images are meticulously selected to ensure diverse coverage of creation modes and content scenes, including scenery, stars, television dramas, games, and other scenes. The scope is broad with *34,000* annotations to ensure reliability.

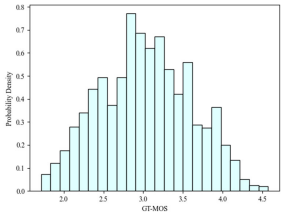
Annotation Process. We evenly divide each image into non-overlapping grid regions of 7×7. Following the standard subjective procedure in ITU-R BT 500.13 [2], we assign a subjective quality rating ranging from 1 to 5 points

[†] Project leader. [✉] Corresponding authors.

Table 3. Annotation criteria for subjective labeling scores from 1 to 5 [8].

Score	Annotation criteria
1 Bad	The visual information within the image content becomes challenging or impossible to distinguish.
2 Poor	The primary content remains distinguishable but exhibits pronounced noise, block artifacts, and blurriness, along with substantial jitter and lag.
3 Fair	The primary content is reasonably clear, but it includes noticeable distortions such as conspicuous noise, visual blurring, minor localized glare, or distinct edge sharpening. Additionally, the image exhibits a markedly blurry background texture.
4 Good	The images 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 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.

(a) Examples in the proposed LPVQ dataset



(b) The histogram of the MOS distribution

Figure 1. Examples and the overall MOS distribution in the proposed LPVQ dataset. Please zoom in for a better view. (Fig. 5 of our paper)

(interval of 0.5) to each patch, with the involvement of 14 professional visual researchers in the standard environment for annotation. Following the protocol of KVQ [8], Tab. 3 provides the rating guidelines, outlining the scoring rules.

After a glance at the entire image, all participants sequentially score each patch while other patches are occluded to avoid visual interference. Participants adhere to the same criteria and are specifically instructed to evaluate only the low-level quality perception aspects within each patch, such as distortion or sharpness, without considering the content or semantics of the patch. All annotations undergo a data-cleaning process after scoring for reliability. We calculate the mean of human opinions for each region as the final MOS. The MOS distribution in Fig. 1(b) demonstrates that it exhibits a normal distribution encompassing all quality levels. LPVQ will be open-sourced with detailed descriptions.

3. Additional Experimental Results

3.1. VQA Evaluation Results

Inference time. We deploy our KVQ and other baseline models on an NVIDIA V100, processing 8-second 1080p videos to compare the inference times required by different methods. As shown in Tab. 4, KVQ achieves breakneck speed comparable to Fast-VQA [10], far surpassing other baseline methods (*e.g.*, VSFA [5], PVQ [11], Li *et al.* [4]).

Table 4. Inference time (avg. of 20 runs).

Setting	VSFA	PVQ	Li *et al.*	Fast-VQA	**KVQ**
Time (/s)	11.14	13.79	27.6	0.045	0.056

Table 5. Ablation study on the saliency map.

Methods	$\text{LSVQ}_{\text{test}}$ SRCC/PLCC	LSVQ_{1080p} SRCC/PLCC	KoNViD-1k SRCC/PLCC	LIVE-VQC SRCC/PLCC
w/o multi-scale map	0.895/0.896	0.812/0.844	0.888/0.890	0.816/0.839
w multi-scale map	**0.896/0.897**	**0.814/0.846**	**0.890/0.892**	**0.820/0.843**

Table 6. Ablation study on the region selected input.

Input	$\text{LSVQ}_{\text{test}}$ SRCC/PLCC	LSVQ_{1080p} SRCC/PLCC	KoNViD-1k SRCC/PLCC	LIVE-VQC SRCC/PLCC
Random Crop	0.872/0.874	0.737/0.777	0.878/0.875	0.810/0.836
Center Crop	0.882/0.884	0.765/0.803	0.885/0.885	0.810/0.836
Fragments	0.891/0.892	0.808/0.841	0.887/0.8889	0.808/0.834
Ours	**0.896/0.897**	**0.814/0.846**	**0.890/0.892**	**0.820/0.843**

Effectiveness of the multi-scale saliency map. We ensemble the multi-scale saliency maps acquired from each block for the final saliency map. We compare the multi-scale approach with the method of solely using prediction heads, and the results are shown in Tab. 5. The performance improvement after incorporating multi-scale information validates the rationale behind this structural design, which integrates hierarchical features similar to the HVS.

Comparison with different region selected input. After resizing videos to a consistent size, KVQ inputs them into the model. This approach ensures that the input videos contain a complete spatial structure, allowing for better perception of high-level semantics for attention allocation. Our

Table 7. Number of correlated windows.

k	$LSVQ_{test}$ SRCC/PLCC	$LSVQ_{1080p}$ SRCC/PLCC	KoNViD-1k SRCC/PLCC	LIVE-VQC SRCC/PLCC
[4, 2]	0.894/0.894	0.810/0.841	0.887/0.888	0.815/0.840
[8, 4]	**0.896/0.897**	**0.814/0.846**	**0.890/0.892**	**0.819/0.842**

Table 8. Comparison under different resolution inputs.

Resolution	$LSVQ_{test}$ SRCC/PLCC	$LSVQ_{1080p}$ SRCC/PLCC	KoNViD-1k SRCC/PLCC	LIVE-VQC SRCC/PLCC
224×224	0.883/0.884	0.786/0.822	0.870/0.872	0.805/0.834
448×448	**0.896/0.897**	**0.814/0.846**	**0.890/0.892**	**0.819/0.842**

Table 9. Ablation of FWA on the proposed LPVQ dataset.

Method	Inter-sample		Intra-sample	
	SRCC	PLCC	SRCC	PLCC
IWA	0.572	0.553	0.570	0.579
CWA	0.565	0.557	0.549	0.553
FWA	**0.614**	**0.616**	**0.612**	**0.657**

model dynamically selects correlated regions and performs attention allocation to extract global saliency. We compare our approach with the other 3 methods of region selection types with defined static patterns for input: randomly cropping regions, center cropping regions, and the sampled fragments proposed in [10], which randomly samples mini-patches within the uniformly divided grids. As illustrated in Tab. 6, KVQ outperforms other types by preserving the complete spatial structure for global saliency extraction.

Number of correlated windows. In our paper, we employ the attention mechanism to compute the windows correlations. We select the top-k windows with the highest correlations. As shown in Tab.7, larger k in the third and fourth stages improves performance. A larger k implies that each window can establish long-distance connections with more regions, thereby enhancing attention allocation.

Comparison with Input Resolutions. Our KVQ calculates global quality by perceiving visual attention allocation and local texture. Therefore, we believe that preserving the detailed features is also crucial for evaluating video quality, especially for high-resolution videos where reducing the resolution can lead to a loss of local details. Taking this into consideration, we set the resolution of the input videos to 448×448. We compare videos adjusted to different resolutions, and the results in Tab. 8 demonstrate that higher resolutions correspond to better performance.

3.2. Local Perception on LPVQ Dataset

Effectiveness of the FWA module on LPVQ. We validate the assistance of the FWA module for local perception on the LPVQ dataset. All models are trained on VQA tasks and undergo zero-shot validation on the LPVQ dataset. In Tab. 9, we present the results of applying FWA compared to using IWA or CWA individually. The combined utilization of IWA and CWA outperforms using either module individually. We attribute this to the FWA's capability to allocate long-range attention and precisely assess global-wise saliency, allowing for better decoupling of visual correlations from local perception maps and resulting in more accurate predictions.

References

[1] 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. 1

[2] Telephone Installations and Local Line. Subjective video quality assessment methods for multimedia applications. *Networks*, 910(37):5, 1999. 1

[3] Will Kay, João 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. *CoRR*, abs/1705.06950, 2017. 1

[4] 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 Trans. Circuits Syst. Video Technol.*, 32(9):5944–5958, 2022. 2

[5] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *ACM Multimedia*, pages 2351–2359. ACM, 2019. 2

[6] 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 *ICCV*, pages 9992–10002. IEEE, 2021. 1

[7] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3192–3201. IEEE, 2022. 1

[8] Yiting Lu, Xin Li, Yajing Pei, Kun Yuan, Qizhi Xie, Yunpeng Qu, Ming Sun, Chao Zhou, and Zhibo Chen. Kvq: Kwai video quality assessment for short-form videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25963–25973, 2024. 2

[9] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan C. Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *ECCV (24)*, pages 459–479. Springer, 2022. 1

[10] 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 *ECCV (6)*, pages 538–554. Springer, 2022. 2, 3

[11] 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