

# Supplementary Material–Semantic-Aware and Quality-Aware Interaction Network for Blind Video Quality Assessment

Anonymous Author(s)

## ABSTRACT

This is the supplementary material for ACMMM 2024 submission, Semantic-Aware and Quality-Aware Interaction Network for Blind Video Quality Assessment, which provides additional experimental results and visual analysis.

## CCS CONCEPTS

• **Computing methodologies** → Modeling and simulation.

## KEYWORDS

Video quality assessment, semantic- and quality-aware, cross-aware guided interaction, cross-aware temporal modeling.

## 1 MORE DETAILED FOR TWO AWARE INFORMATION

In this subsection, we provide more sample videos to demonstrate the relevance of quality-aware and semantic-aware to subjective video quality. In Figure 1, we compare three high-quality and three low-quality videos, and sampled five frames with a sampling interval of around two seconds. Generally, high-quality videos have the characteristics of smooth scene transitions and high frame-level quality. Videos with fast scene switching or low frame-level quality are judged as low-quality. In addition, we further visualize the semantic-aware and quality-aware features from six sample videos.

Similarly, we choose ResNet-50 [3] pre-trained on the ImageNet [2] dataset and the KoNIQ-10k dataset [5] for semantic-aware and quality-aware feature extraction, respectively. The training details of quality-aware feature extraction can refer to [8]. In Figure 2, it can be observed that semantic-aware focus on representing the object features and are robust to spatial quality degradation. The responses of the quality-aware feature maps of three low-quality video frames are significantly stronger than that of three high quality video frames, indicating that the quality-aware features are sensitive to quality degradation. Meanwhile, two aware features have redundancy, such as quality degradation appearing at the edges and textures of the object, or redundant areas that neither perceptual features pay attention to.

To explore the relationship between temporal features and video quality, we further analyzed the temporal curves of two aware

features from videos. It can be found that videos with slower fluctuations in temporal content information and higher overall frame-level quality typically have better perceptual quality. Among them, the fluctuation of temporal content information is related to scene transitions or shot changes, and frame-level quality determines the spatial quality of videos. Overall, the spatial and temporal characteristics of two features are related to video perceptual quality. Based on the above findings, we design corresponding modules to enhance the representation for video quality.

## 2 MORE DETAILS ON TEMPORAL NETWORK AND TEMPORAL POOLING

Similar to previous work [6], we use gated recurrent units (GRUs) [1] as the temporal network to capture long-term dependencies and measure temporal distortion from distorted videos. To improve the learning efficiency of GRUs, the feature  $F^z$  is reduced by using fully connected (FC) layer and the feature  $\tilde{F}^z = \{f_t^z | t = 1, \dots, T\} \in \mathbb{R}^{128 \times T}$  is obtained, where  $F^z$  is the output features from temporal saliency quality perception and content variation perception blocks. Next, the  $\tilde{F}^z$  is input into GRUs  $M_{GRU}(\cdot)$ , the current state  $h_t$  of GRUs is used as the output feature, it is determined by the feature  $f_t^z$  and the previous state  $h_{t-1}$ , as follow:

$$h_t = M_{GRU}(f_t^z, h_{t-1}) \in \mathbb{R}^{32 \times 1} \quad (1)$$

Then a FC layer is used to regress  $\{h_t\}$  into frame-level scores  $q = \{q_t\}$ .

To predict video-level quality score  $Q_{pred}$ , a differentiable temporal hysteresis model [6, 7] as a temporal pooling manner to aggregate frame-level quality scores  $\{q_t\}$ . Specifically, the hysteresis effect is introduced to approximate the frame-level subjective quality scores  $\hat{q}_t$ . The hysteresis effect suggests that the quality of the  $t$ -th frame will be affected by the quality of the previous and following  $\beta$  frames, which can be formulated as a linear combination of memory quality  $q_t^m$  and current quality  $q_t^c$ , as follows:

$$\hat{q}_t = \alpha q_t^m + (1 - \alpha) q_t^c \quad (2)$$

where  $\alpha$  is a hyper-parameter to balance the contributions of different component and is empirically set to 0.5, the memory quality  $q_t^m$  is related to the worst quality score of the first  $\beta$  frames and is defined as:

$$q_t^m = \begin{cases} q_t, & \text{for } t = 1, \\ \min(q_i), & \text{for } t > 1. \end{cases} \quad (3)$$

where  $i \in \{max(1, t - \beta), \dots, t - 2, t - 1\}$ , while the current quality  $q_t^c$  is calculated by weighted combination of the next  $\beta$  frames [6], as follows:

$$q_t^c = \sum_k w_{t,k} q_k, \quad w_{t,k} = \frac{e^{-q_k}}{\sum_j e^{-q_j}} \quad (4)$$

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '24, October 28–November 1, 2024, Australia, Melbourne

© 2024 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00

<https://doi.org/XXXXXXX.XXXXXXX>



**Figure 1: More sample videos are selected from the KoNViD-1k dataset [4] to further demonstrate the importance of semantic-aware and quality-aware for VQA, including (a) three high-quality videos, HV1 (8751538577.mp4), HV2 (9571377943.mp4) and HV3 (7177696763.mp4), and (b) three low-quality videos, LV1 (10672253555.mp4), LV2 (9445782126.mp4) and LV3 (4744073127.mp4). We adopt the mean opinion score (MOS, subjective score) to measure the visual quality of each video. The larger the MOS value, the better the subjective quality. And five representative frames are selected to visualize the subjective content and quality of the video. The subjective content of high-quality videos changes smoothly and the quality of each frame is higher. In contrast, low-quality videos have greater content variation or lower frame-level quality.**

where  $j, k \in \{t, t+1, \dots, \min(t+\beta, T)\}$ , and  $\beta$  is set to 6 as described [6]. Finally, the video-level quality score  $Q^{pred}$  is predicted by aggregating the scores  $\hat{q}_t$ , as follows:

$$Q^{pred} = \frac{1}{T} \sum_{t=1}^T \hat{q}_t \quad (5)$$

### 3 MORE QUALITATIVE ANALYSIS

In this section, we present more quantitative results.

#### 3.1 Scatter Plot of Prediction Results

We present the results of the **proposed FR(S+S) and FR(S+M)** models on six video quality assessment (VQA) datasets, illustrating the correlation between predicted scores and the subjective scores (*i.e.*, MOS) in Figure 3. We observe that most of the scatter points cluster around the red line, indicating a consistent alignment between the predicted scores and subjective scores.

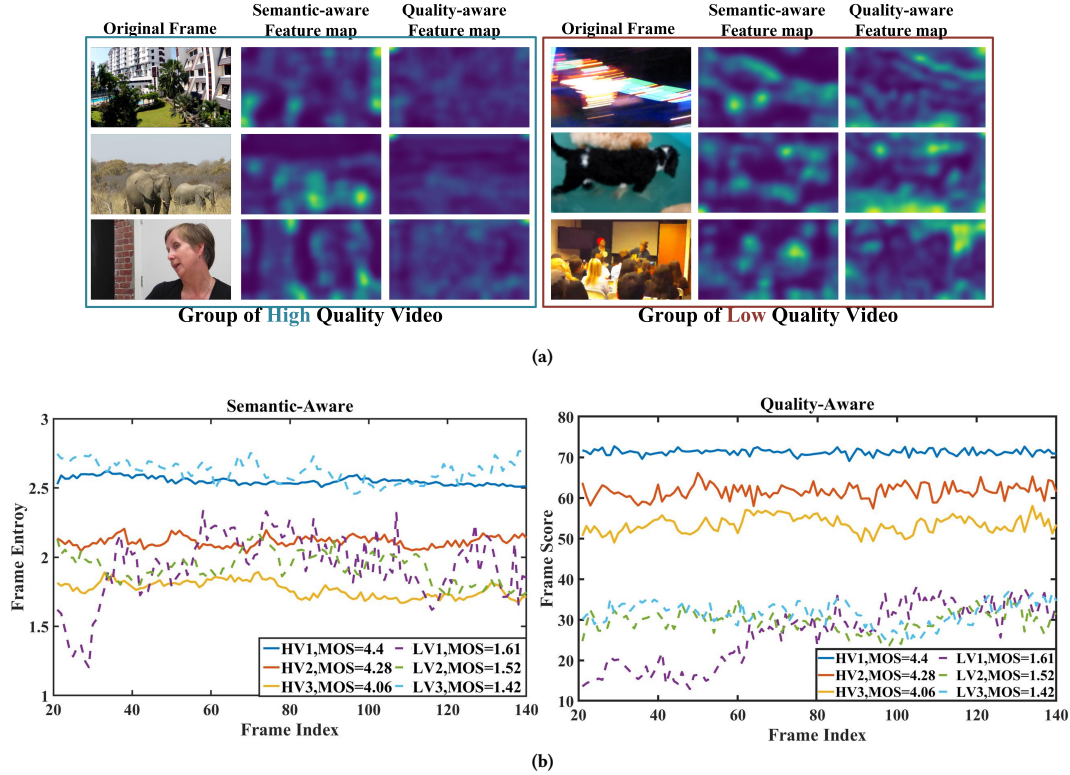


Figure 2: Visualize (a) the spatial feature maps of semantic-aware and quality-aware, as well as (b) the temporal distribution of two aware features. In the spatial domain, semantic-aware focuses on the object and are robust to spatial quality degradation, while the responses of quality-aware are stronger in low-quality frames. In (b), low quality videos usually exhibit large changes in information between frames and low frame-level quality, while the temporal distribution of content information in high-quality videos is smoother and the overall quality is higher.

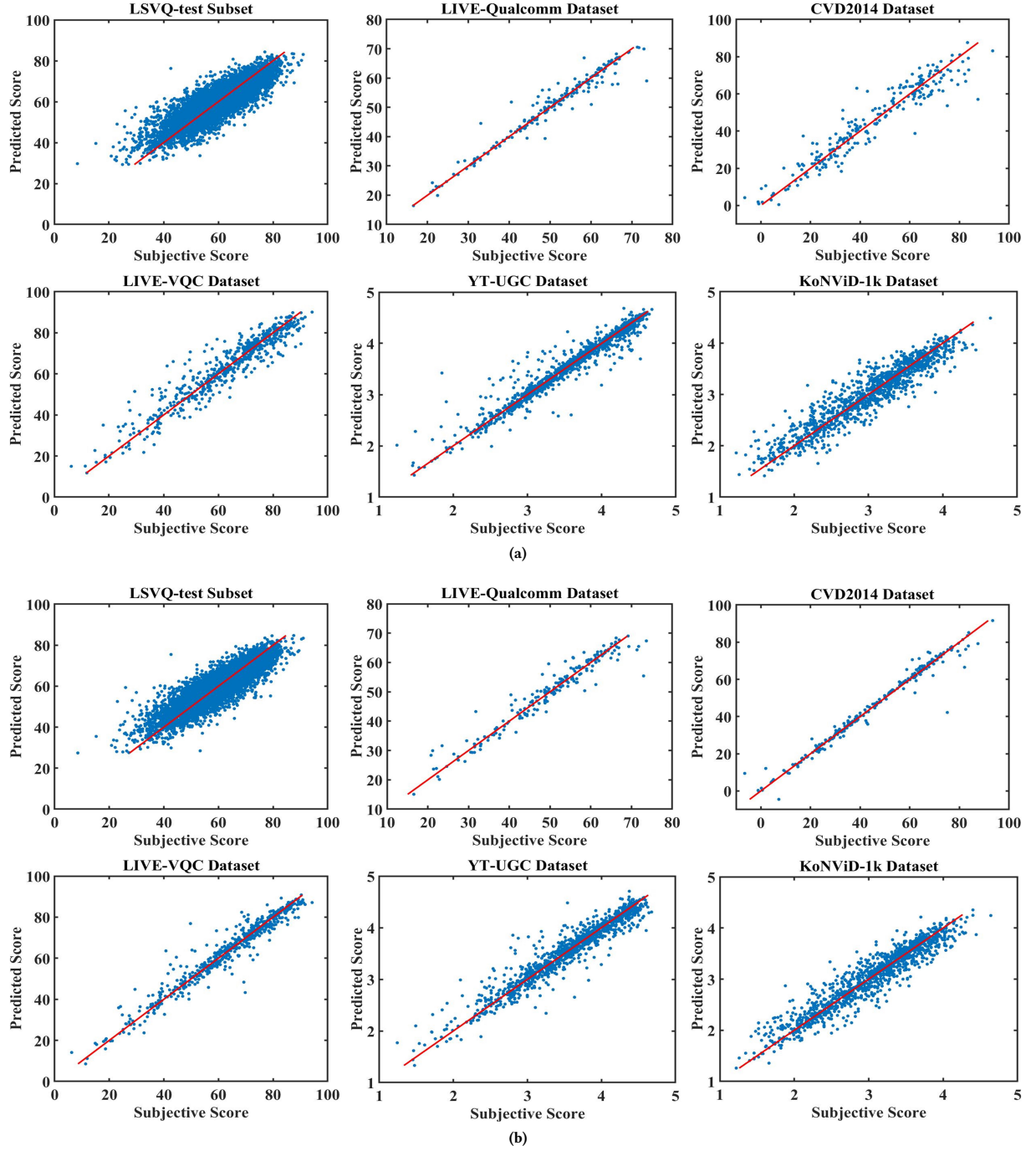
### 3.2 Visualization for Successful and Failure Cases

We showcase four successful and four failure video prediction cases of the proposed FR(S+M) model. Fig. 4a illustrate that the proposed FR(S+M) model accurately predicts most video scenes with obvious semantic information. Conversely, by Fig. 4b, the proposed (S+M) model has difficulty in predicting the quality of the video for which the semantic content is not obvious. One possible explanation is that the proposed model relies on consensus semantic-aware and quality-aware features to evaluate video quality. Consequently, the model cannot accurately predict video quality when the video semantics are incomplete. Through the analysis of the qualitative results of the model, the proposed model is still able to accurately assess the quality of most videos.

## REFERENCES

- [1] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).
- [2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 248–255.
- [3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 770–778.
- [4] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. 2017. The Konstanz natural video database (KoNViD-1k). In *IEEE International Conference on Quality of Multimedia Experience (QoMEX)*. 1–6.
- [5] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing* 29 (2020), 4041–4056.
- [6] Dingquan Li, Tingting Jiang, and Ming Jiang. 2019. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2351–2359.
- [7] Kalpana Seshadrinathan and Alan C Bovik. 2011. Temporal hysteresis model of time varying subjective video quality. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 1153–1156.
- [8] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinjia Sun, and Yanning Zhang. 2020. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3667–3676.





**Figure 3: Scatter plots of scores predicted by (a) the proposed FR(S+S) model and (b) the proposed FR(S+M) model versus subjective scores on six VQA datasets. The x-axis represents subjective scores, the y-axis is predicted scores. The red line is used as a reference line when the predicted score is the same as the subjective score. The closer the blue points are to the red line, the more relevant the predicted score is to the subjective score.**





Figure 4: (a) The four successful and (b) four failure prediction cases of the proposed FR(S+M) model, and five representative frames are presented. The successful cases (SV1, SV2, SV3 and SV4) have clear semantic information. For failure cases, the semantic information of FV1, FV3 and FV4 is difficult to describe directly, and the semantic information of FV2 is discontinuous. The predictive ability of model is insufficient for videos with ambiguous semantic information.